JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 11, Numbers 2–3, 2004
© Mary Ann Liebert, Inc.
Pp. 507–518

Haplotype Reconstruction from SNP Alignment

LEI M. LI,[1] JONG HYUN KIM,[1] and MICHAEL S. WATERMAN[1,2]

ABSTRACT

In this paper, we describe a method for statistical reconstruction of haplotypes from a set of aligned SNP fragments. We consider the case of a pair of homologous human chromosomes, one from the mother and the other from the father. After fragment assembly, we wish to reconstruct the two haplotypes of the parents. Given a set of potential SNP sites inferred from the assembly alignment, we wish to divide the fragment set into two subsets, each of which represents one chromosome. Our method is based on a statistical model of sequencing errors, compositional information, and haplotype memberships. We calculate probabilities of different haplotypes conditional on the alignment. Due to computational complexity, we first determine phases for neighboring SNPs. Then we connect them and construct haplotype segments. Also, we compute the accuracy or confidence of the reconstructed haplotypes. We discuss other issues, such as alternative methods, parameter estimation, computational efficiency, and relaxation of assumptions.

Key words: haplotype, genotype, confidence, SNP, shotgun sequencing, alignment.

1. BACKGROUND AND INTRODUCTION

Churchill and Waterman (1992) described a method for the statistical reconstruction of a long DNA sequence from a set of assembled sequence fragments. In this work we adopt the same rationale, which is briefly summarized as follows. We consider DNA sequence accuracy as a function of the redundancy of coverage and the rates of sequencing errors in the fragment sequences. We assume that the fragments have been assembled and address the problem of determining the degree to which the reconstructed sequence is free from errors, i.e., its accuracy.

To facilitate the statistical analysis, we start off with a number of simplifying assumptions: A1—the DNA preparation and cloning stage are free of errors; A2—the fragment assembly is correct; A3—all positions within a fragment are equally reliable; A4—all fragment sequences are equally reliable; A5—sequencing errors are independent of their local context; A6—the sequencing error rates are constant across the entire sequence; A7—the composition of the sequence is independent, both of adjacent bases and over large regions. In the discussion, we consider relaxing some of these assumptions.

Next, a set of fragment sequences indexed by $i = 1, \dots, m$ are aligned by some procedure. The result of the alignment procedure is a matrix with m rows. Each row contains the ordered sequence of nucleotide

[1] Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089.

[2] Informatics Research, Celera Genomics, 45 West Gude Drive, Rockville, MD 20850.

bases in a particular fragment written in either direct or reverse complemented orientation. Gaps may be inserted internally, and each fragment is offset to produce a column-by-column correspondence among the fragments. The column index $j = 1, \dots, n$ runs from the leftmost base in the assembly to the rightmost. Let s_j denote the true state of the DNA sequence corresponding to column j in the fragment assembly. The true state may take any value in the set $\mathcal{A} = \{A, C, G, T, -\}$. The symbol "$-$" is included to allow for extra columns in the fragment assembly that do not correspond to any base in the true DNA sequence. The alphabet of redundant bases is defined to be the set of all nonempty subsets of $\mathcal{A}$. A partial notation for this set is given by the standard IUPAC DNA alphabet; see Cornish-Bowden (1984). Redundant bases are useful in the definition of consensus sequences and in the reduction of computational complexity. The elements of the fragment assembly matrix, denoted by x_{ij}, may take values in the set $\mathcal{B} = \{A, C, G, T, -, N, \phi\}$, where "$-$" denotes an internal gap, N denotes any ambiguous determination of a base, and null symbol ϕ is for nonaligned positions beyond the ends of a fragment. The consensus distribution is a probability distribution over the set $\mathcal{A}$ defined by $\pi_j(a) = Pr(s_j = a | x_{ij},\ i = 1, \dots, m)$. These probabilities can be computed using Bayes's rule based upon the rates of sequencing errors in the individual fragments. The consensus distribution allows us to reconstruct the consensus sequence. Churchill and Waterman proposed three different approaches to the definition of a consensus sequence. In *Procedure A*, the most likely value of s_j is the base that maximizes $\pi_j(a)$ over a. Denote this value by s_j. Then $s_j = \arg\max_{a \in \mathcal{A}} \pi_j(a)$. In *Procedure B*, redundant bases are allowed into the consensus. The goal is to find the character with minimum redundancy that has probability in excess of a specified level $1 - \alpha$. That is, the choice of consensus sequence is made so that at each position the confidence probability exceeds a specified level,

$$\sum_{a \in c_j} \pi_j(a) \geq 1 - \alpha.$$

To find the best redundant set, we begin by choosing the most likely base at position j; call it a_j. If $\pi_j(a_j) > 1 - \alpha$, we stop. Otherwise, we add to the redundant set the next most likely base. The process continues until the level exceeds $1 - \alpha$. *Procedure C* extends this idea of minimal redundant consensus to the entire sequence or regions of interests.

The parameters in this model include the composition probabilities and sequencing error rates. In practice, we need to estimate them from the data—the fragment assembly. Churchill and Waterman derived a likelihood-based procedure for the estimation of the parameters, which utilizes an E-M algorithm. Prior knowledge of the error rates is easily incorporated into the estimation procedure. They also derived formulas that relate depth of coverage to sequence accuracy.

In this paper, we describe a method for statistical reconstruction of haplotypes from a set of aligned SNP fragments. This problem is motivated by the sequencing projects of multiple chromosomes; see Venter *et al.* (2001). We consider the simple case of a pair of human chromosomes, one from the mother and the other from the father. After fragment assembly and SNP detection, we wish to reconstruct the two haplotypes of the parents. Given a set of SNP sites inferred from the assembly alignment, we wish to divide the fragment set into two subsets, each of which represents one chromosome. Lancia *et al.* (2001) proposed a deterministic method to solve the problem. Our method is based on a statistical model of haplotype composition, sequencing errors, and haplotype memberships. We calculate joint probabilities of different haplotype configurations for more than one SNP site at a time. Then we connect them and extend haplotype segments. Also, we propose strategies to determine the accuracy or confidence of the reconstructed haplotypes. In the literature, Irizarry *et al.* (2000) reported a SNP detection method based on statistical analysis of human expressed sequence tags (ESTs).

This paper is organized as follows. In Section 2, we describe our model and methodology. In Section 3, we show some simulation results. In Section 4, we discuss various issues, such as alternative methods, parameter estimation, computational efficiency, and relaxation of assumptions.

2. PROBABILISTIC MODEL AND METHODS

We now extend the work by Churchill and Waterman to statistical reconstruction of haplotypes from a set of aligned SNP fragments. To illustrate the idea, we first study a simple case in which we assume that

A1–A7 mentioned in Section 1 hold. Moreover, in this work we assume the two homologous chromosomes are from one person. The data is a set of fragment sequences that have been aligned by some procedure. We focus on potential SNP sites, ignoring all non-SNP positions, and this gives us the alignment of SNP fragments. We first omit the issue of orientation for the sake of simplicity and come back to it in the discussion section.

Our goal is to reconstruct the two haplotypes of the parents. That is, given a set of SNP sites inferred from the assembly alignment, we wish to divide the fragment set into two subsets, each of which represents one chromosome. Also, we wish to determine the accuracy of the reconstructed haplotypes. We keep the notation and definitions that are consistent with those in Section 1.

The dataset is an assembly matrix x_{ij} with m rows, where m is the total number of SNP fragments. Each row contains the ordered sequence of bases and possible gaps in a particular fragment. We denote the haplotypes by $\sigma_k = s_{k,1}s_{k,2}\dots s_{k,n}, k = 1, 2$, where $s_{k,j}$ represents the true state of the haplotype k corresponding to column j in the fragment assembly. The true state takes value in the set $\mathcal{A}$. Let f_i denote the membership of the i-th fragment, $i = 1, \dots, m$. That is, $f_i = 1$ if it is from haplotype 1, and $f_i = 2$ if it is from haplotype 2.

We denote the true base of fragment i at position j by y_{ij}. Then $y_{ij} = s_{f_i,j}$, $i = 1, 2$. In the following, we use capital letters to represent the random version of above variables. That is, we denote the assembly matrix by $\mathbf{X} = \{X_{ij},\ i = 1, \dots, m,\ j = 1, \dots, n\}$, their underlying true bases by $\mathbf{Y} = \{Y_{ij},\ i = 1, \dots, m,\ j = 1, \dots, n\}$, the haplotype composition variables by $\mathbf{S} = \{S_{i,j},\ i = 1, 2,\ j = 1, \dots, n\}$, and the haplotype memberships by $\mathbf{F} = \{F_i,\ i = 1, \dots, m\}$. Each fragment in the assembly and each haplotype are denoted by $\mathbf{X}_{i\cdot} = \{X_{ij},\ j = 1, \dots, n\}$ and $\mathbf{S}_{i\cdot} = \{S_{i,j},\ j = 1, \dots, n\}$, respectively. The data in column j is denoted by $\mathbf{X}_{\cdot j} = \{X_{ij},\ i = 1, \dots, m\}$.

An error occurs when the true base $Y_{ij} = a$ is misread to yield $X_{ij} = b, b \neq a$. Under our assumptions A5 and A6, we denote the sequencing error probabilities by

$$p(b|a) = \Pr(X_{ij} = b|Y_{ij} = a),\ a \in \mathcal{A},\ b \in \mathcal{B}.$$

We have assumed that bases occur independently with identical distribution across the assembly. The composition probabilities of the two haplotypes are denoted by

$$p_k(a) = \Pr(S_{k,j} = a),\ a \in \mathcal{A}.$$

Note that our definition of sequence composition includes the gap frequency. Namely, by including "-" in $\mathcal{A}$, we are also considering one-nucleotide-indel polymorphisms under the same machinery. We do not exclude homozygotes for the sake of model flexibility and simplicity. The rate at which the haplotype k emits a fragment is given by $p(\sigma_k) = \Pr(F_i = k) = \frac{1}{2}$, for $k = 1, 2$. Note that $\mathbf{X}$ is observed while $\mathbf{S}$ and $\mathbf{F}$ are not.

Our specific goal is to evaluate the conditional distribution of the haplotype composition given data: $\Pr(\mathbf{S}|\mathbf{X})$, over the product set $\mathcal{A}^{2n}$. The joint conditional distribution $\Pr(\mathbf{S}, \mathbf{F}|\mathbf{X})$ provides information regarding both haplotypes and memberships. First we consider the joint distribution:

$$\Pr(\mathbf{X}, \mathbf{F}, \mathbf{S}) = \Pr(\mathbf{X}|\mathbf{F}, \mathbf{S}) \Pr(\mathbf{F}) \Pr(\mathbf{S}),$$

$$\Pr(\mathbf{S}) = \Pr(\mathbf{S}_{1\cdot}) \Pr(\mathbf{S}_{2\cdot}) = \prod_{j=1}^{n} \Pr(\mathbf{S}_{1j}) \prod_{j=1}^{n} \Pr(\mathbf{S}_{2j}), \quad \Pr(\mathbf{F}) = \prod_{i=1}^{m} \Pr(\mathbf{F}_i),$$

$$\Pr(\mathbf{X}|\mathbf{F}, \mathbf{S}) = \prod_{i=1}^{m} \Pr(\mathbf{X}_{i\cdot}|\mathbf{F}_i, \mathbf{S}_{F_i\cdot}) = \prod_{i=1}^{m} \Pr(\mathbf{X}_{i\cdot}|\mathbf{Y}_{i\cdot} = \mathbf{S}_{F_i\cdot}).$$

These probabilities are, respectively, the composition probabilities, haplotype frequencies, and sequencing error rates. Consequently, Bayes's rule gives us the conditional distribution:

$$\Pr(\mathbf{S}, \mathbf{F}|\mathbf{X}) = \frac{\Pr(\mathbf{X}, \mathbf{F}, \mathbf{S})}{\Pr(\mathbf{X})}.$$

The formula to compute $\Pr(\mathbf{X}, \mathbf{S})$ is given by

$$
\begin{aligned}
\Pr(\mathbf{X}, \mathbf{S}) &= \Pr(\mathbf{X}|\mathbf{S}) \Pr(\mathbf{S}) = \Pr(\mathbf{S}) \prod_{i=1}^{m} \Pr(\mathbf{X}_{i\cdot}|\mathbf{S}) = \Pr(\mathbf{S}) \prod_{i=1}^{m} [\Pr(\mathbf{X}_{i\cdot}, F_i = 1|\mathbf{S}) + \Pr(\mathbf{X}_{i\cdot}, F_i = 2|\mathbf{S})] \\
&= \Pr(\mathbf{S}) \prod_{i=1}^{m} [\Pr(\mathbf{X}_{i\cdot}|\mathbf{S}, F_i = 1) \Pr(F_i = 1) + \Pr(\mathbf{X}_{i\cdot}|\mathbf{S}, F_i = 2) \Pr(F_i = 2)] \\
&= \left(\frac{1}{2}\right)^m \Pr(\mathbf{S}) \prod_{i=1}^{m} [\Pr(\mathbf{X}_{i\cdot}|\mathbf{Y}_{i\cdot} = \mathbf{S}_{1\cdot}) + \Pr(\mathbf{X}_{i\cdot}|\mathbf{Y}_{i\cdot} = \mathbf{S}_{2\cdot})] \\
&= \left(\frac{1}{2}\right)^m \Pr(\mathbf{S}) \prod_{i=1}^{m} \left[\prod_{j=1}^{n} \Pr(\mathbf{X}_{ij}|\mathbf{S}_{1j}) + \prod_{j=1}^{n} \Pr(\mathbf{X}_{ij}|\mathbf{S}_{2j}) \right]. \qquad (1)
\end{aligned}
$$

Next we look at a simple case.

2.1. One SNP locus

If we look only at j-th site, then (1) becomes

$$
\begin{aligned}
&\Pr(\mathbf{X}_{\cdot j} = x_{\cdot j}; S_{1,j} = s_{1,j}, S_{2,j} = s_{2,j}) \\
&= \left(\frac{1}{2}\right)^m \Pr(S_{1,j} = s_{1,j}, S_{2,j} = s_{2,j}) \prod_{i=1}^{m} [p(x_{ij}|s_{1,j}) + p(x_{ij}|s_{2,j})]. \qquad (2)
\end{aligned}
$$

Consequently, we obtain

$$
\Pr(S_{1,j} = s_{1,j}, S_{2,j} = s_{2,j}|\mathbf{X}_{\cdot j} = x_{\cdot j}) = \frac{\Pr(\mathbf{X}_{\cdot j} = x_{\cdot j}; S_{1,j} = s_{1,j}, S_{2,j} = s_{2,j})}{\sum \Pr(\mathbf{X}_{\cdot j} = x_{\cdot j}; S_{1,j} = s_{1,j}, S_{2,j} = s_{2,j})}.
$$

Based on the above calculations, we can find the most likely genotype at the j-th locus. We note that 20 out of the 25 cases of $(s_{1,j}, s_{2,j})$ correspond to heterozygous types. Because of the unknown parental origins, we cannot distinguish between heterozygotes such as (A, G) and (G, A). Consequently, the complete list of candidates in the one-locus case includes 15 different genotypes: 5 homozygotes and 10 heterozygotes as shown in Table 1. If we adopt the convention of 15 genotypes in the summation in the above formula, we need to add a multiplying factor of 2 for heterozygotes. Next, we rank these fifteen probabilities and report the most likely genotype that maximizes $\Pr(\mathbf{X}_{\cdot j} = x_{\cdot j}; S_{1,j} = s_{1,j}, S_{2,j} = s_{2,j})$. As far as the most likely solution is concerned, it is sufficient to consider a shorter list of plausible candidates rather than the complete list of 15 genotypes. For example, if we observe only one letter at one column other than "N"—the symbol representing the missing information—then the solution must be the homozygote of this letter. If we observe only two letters at one column other than "N," then we need to consider only two homozygotes and one heterozygote, etc. Although this improvement of computation is slight in the one-locus case, it is more significant in multiple-SNP cases. It is possible that two genotypes tie for the most likely solution. To achieve confidence of some predetermined levels, we can apply the idea of redundant sets, which in this case are genotypes; see Section 1.

For a specific SNP, we ignore those fragments outside the range under consideration. As a result, the size of m in (2) is approximately the coverage of the j-th site.

TABLE 1. THE GENOTYPES AT ONE SNP LOCUS

Homozygote	*Heterozygote*
AA, GG, CC, TT, --	AG, AC, AT, A-, GC, GT, G-, CT, C-, T-

TABLE 2. THE NUMBER OF HAPLOTYPES AT TWO SNPS

1st SNP	*Homozygote*	*Homozygote*	*Heterozygote*	*Heterozygote*
2nd SNP	*Homozygote*	*Heterozygote*	*Homozygote*	*Heterozygote*
# haplotypes	5 × 5	5 × 10	10 × 5	10 × 10 × 2

2.2. *Two SNP loci*

The haplotype information includes both polymorphisms at each locus and phases among loci. The consideration of SNPs one at a time provides no information regarding phases. In order to reconstruct the haplotypes, we have to consider more than one SNP at a time. Theoretically, we can compute $\Pr(\mathbf{S}|\mathbf{X})$ for any number of SNPs. However, computationally, we encounter a problem due to the large size of $\mathbf{S}$, which is of the order $|\mathcal{A}|^{2n}$. One strategy is to consider one pair of SNP sites at a time. That is, for $j_1 \neq j_2$ we calculate

$$\begin{aligned}\Pr(S_{k,j} &= s_{k,j}, k = 1, 2, j = j_1, j_2 | \mathbf{X}_{ij} = x_{ij},\ i = 1, \dots, m, j = j_1, j_2) \\ &= \frac{\Pr(\mathbf{X}_{ij} = x_{ij}, i = 1, \dots, m, j = j_1, j_2; S_{k,j} = s_{k,j}, k = 1, 2, j = j_1, j_2)}{\sum \Pr(\mathbf{X}_{ij} = x_{ij}, i = 1, \dots, m, j = j_1, j_2; S_{k,j} = s_{k,j}, k = 1, 2, j = j_1, j_2)}.\end{aligned} \tag{3}$$

Again, we need the formula (1), which now becomes

$$\begin{aligned}\Pr(\mathbf{X}_{ij} &= x_{ij},\ i = 1, \dots, m, j = j_1, j_2; S_{k,j} = s_{k,j}, k = 1, 2, j = j_1, j_2) \\ &= \left(\frac{1}{2}\right)^m \Pr(S_{k,j} = s_{k,j}, k = 1, 2, j = j_1, j_2) \prod_{i=1}^{m} \left[\sum_{k=1}^{2} p(x_{i,j_1}|s_{k,j_1}) p(x_{i,j_2}|s_{k,j_2})\right].\end{aligned} \tag{4}$$

Next, we combine the results from pairwise calculations to determine consensus haplotypes.

As in the one-SNP case, we can find the most likely haplotype after the pairwise computation. In Table 2, we count the number of haplotypes and classify them into four categories. The total is 325 different haplotypes instead of 225 = 15 × 15—there are 15 different genotypes at one locus; see Table 1. The extra 100 states are results from the phase complexity in the case of heterozygote by heterozygote. Note that each haplotype of a homozygote by a heterozygote, or of a heterozygote by a heterozygote, represents two cases in the 625 states of $(s_{1,j_1}, s_{2,j_1}, s_{1,j_2}, s_{2,j_2})$. If we adopt the convention of 325 haplotypes, we need to include a multiplying factor of 2 in the summation in (3) except for a homozygote-by-homozygote haplotype. As we argued earlier, we need to compute only probabilities in (4) for a subset of the 325 states depending on the observations. For example, if we observe only A and G in the first position and only T and G in the second position, then we consider the two haplotypes: AT with GG; AG with GT. For a heterozygote-by-heterozygote pair, it is possible that two haplotypes tie for the most likely solution. This means that we do not have enough information to infer the haplotype phase. To answer the question of whether to report a haplotype or not, we need a decision rule. One method is to check the odds ratio of the two most likely states. We report a haplotype only when the ratio is larger than a threshold; otherwise, we report that the phase between the two loci cannot be determined from the data. Alternatively, we can check the ratio

$$\frac{\Pr(\text{most likely haplotype})}{1 - \Pr(\text{most likely haplotype})}.$$

Another method is to choose the most likely haplotype phase for one adjacent pair if its conditional probability score as defined in (3) is above some threshold, say 25%. Although we can consider any pair of SNPs, it seems reasonable to restrict our focus to adjacent pairs of SNPs.

2.3. *Reconstructing haplotypes and confidence levels*

Next, we combine the phase information between adjacent pairs to construct haplotype segments. This is essentially a process of bookkeeping and linking. Let us illustrate the idea using the example shown in

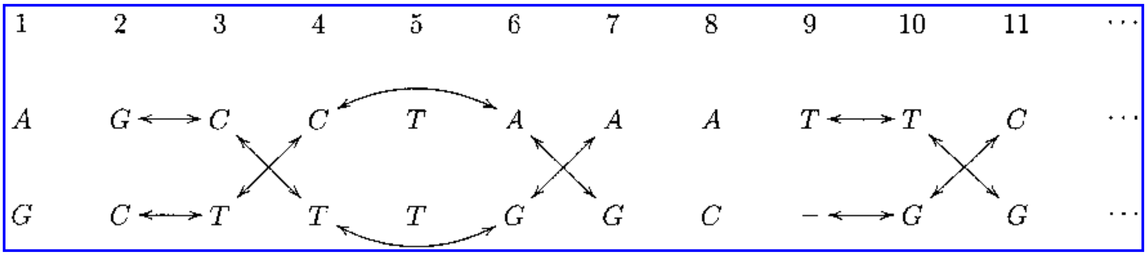

FIG. 1. Reconstructing haplotypes from pairwise calculations.

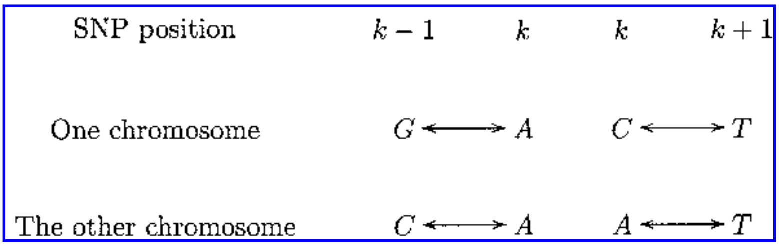

FIG. 2. Inconsistency between adjacent pairs.

Fig. 1. We index 11 SNPs according to their positions along the chromosomes. For any adjacent pair of SNPs whose phase pattern can be identified, we use two connectors to link letters on the same chromosome. Consequently, those nucleotide bases that are linked with one another through these connectors form one haplotype segment. For example, from position 2 to 7, we have two haplotype segments: GCTTGA and CTCTAG; from position 9 to 11, we have two haplotype segments: TTG and -GC. Please notice that no phase information is available from a homozygous locus such as site 5. In this case, we skip homozygous sites and compare the two nearest heterozygous SNPs.

The pairwise method cannot guarantee that the genotype at the k-th SNP obtained from the left-hand-side pair is always the same as that obtained from the right-hand-side pair. Figure 2 shows such an example. In this case, at least one of the pair configurations is incorrect. It might be helpful to jointly consider more than two SNP sites for these cases. The formula is similar to (3). We have done so for all neighboring triple SNPs in our simulation.

For the constructed haplotype segments, we also expect to evaluate their confidence. We define the confidence level of a haplotype segment $s_{1,h}s_{1,h+1}\dots s_{1,h+k}$ and $s_{2,h}s_{2,h+1}\dots s_{2,h+k}$ by

$$\Pr(S_{k,j} = s_{k,j}, k = 1, 2, j = h, \dots, h+k | X_{ij} = x_{ij},\ i = 1, \dots, m, j = h, \dots, h+k). \tag{5}$$

To obtain this confidence level, we need the joint probability

$$\Pr(S_{k,j} = s_{k,j}, k = 1, 2, j = h, \dots, h+k; X_{ij} = x_{ij},\ i = 1, \dots, m, j = h, \dots, h+k),$$

which can be computed using formula (1) and the marginal probability $\Pr(\mathbf{X}_{ij} = x_{ij},\ i = 1, \dots, m, j = h, h+1 \dots, h+k)$, which is the sum of joint probabilities over all haplotypes. The computation gets more intensive as the segment size gets larger. We note that most of these joint probabilities are close to zero. Thus, an approximate solution could be obtained by considering a reduced set of haplotypes. Implicitly, this assumes that those probabilities of haplotypes not in the candidate list are zero. This would overestimate the confidence of the most-likely haplotype.

2.4. Reduction of computational complexity by redundant bases

An accurate solution can be obtained with a little more cost. For example, if we only observe multiple "G"s other than "N" at a SNP site, then the most likely genotype must be a homozygote "GG." For the sake of completeness, we assume that each of the two letters in the genotype is from the alphabet $\{G, \overline{G}\}$, where $\overline{G}$ represents the redundant set $\{A, T, C, -\}$. At the next SNP site, if we observe only two letters A and C other than "N," then we assume that the two letters of the genotype are from the alphabet $\{A, C, \overline{AC}\}$, where $\overline{AC}$ represents $\{G, T, -\}$. This generates 6 genotypes including 3 homozygotes and 3 heterozygotes in comparison to 15 genotypes from the complete alphabet; see Table 1. By doing so for each SNP site, the list of "significant" candidate haplotypes is reduced. Of course, we now need to compute the sequencing error rates for the new alphabet by Bayes's rule. For example,

$$\Pr(X_{ij} = G|Y_{ij} = \overline{AC}) = \Pr(X_{ij} = G|Y_{ij} = G,\ T,\ -) = \frac{p(G)p(G|G) + p(T)p(G|T) + p(-)p(G|-)}{p(G) + p(T) + p(-)}.$$

We summarize our steps in the following algorithm.

Algorithm 1. *We start with an assembly* $\mathbf{X} = \{X_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n\}$.

1. *For* $j = 1$ *to* n*, consider the fragments that cover the* j*-th SNP. Apply formula (2) to compute*

$$\Pr((S_{1,j}, S_{2,j}) = (s_{1,j}, s_{2,j})|X_{ij},\ i = 1, \ldots, m),$$

for all the possible genotypes; see Table 1.

2. *For* $j = 1$ *to* $n - 1$*, consider the fragments that cover either the* j*-th or* $(j+1)$*-th SNP. Apply formula (4) to compute*

$$\Pr(S_{k,l} = s_{k,l}, k = 1, 2, l = j, j+1|X_{il} = x_{il},\ i = 1, \ldots, m, l = j, j+1),$$

for all the possible haplotypes; see Table 1 and 2. For each pair of SNPs, we either report the most probable haplotype phase or leave it open based on odds ratio or confidence score.

3. *Link the haplotype phases obtained in Step 2 and construct haplotype segments. For those segments containing only one SNP, we report the most likely genotypes obtained in step 1. If inconsistent adjacent pairs occur, then we consider these sites jointly. We select the most probable haplotypes or leave the region open based on odds ratio or confidence score.*
4. *Evaluate the confidence levels of haplotype segments according to formulas (5) and (1).*

3. SIMULATION STUDIES

In our simulation, we assume the following stochastic model originally proposed in Lander and Waterman (1988) for physical mapping. For our convenience and to a good approximation, we can regard nucleotide positions on a genome as integerized locations on a continuous time scale. Denote the clone length by H. According to the random model, the inter-arrival times between adjacent SNPs are independent and follow an exponential distribution with expectation $1/\lambda$. Consequently, the number of SNPs in the clone, denoted by $N(H)$, is a Poisson random variable with the parameter λH. Its expectation and standard deviation are, respectively, given by $E[N(H)] = \lambda H$ and $SD[N(H)] = \sqrt{\lambda H}$. Conditional on the total number, the positions of the SNPs are uniformly distributed along the interval $[0, H]$. In shotgun sequencing, we generate many random fragments with average length L along the clone. The number of fragments D relates to the average coverage κ via the equality $\kappa H = DL$. Conditional on the total number of fragments, the starting positions of these random fragments are uniformly distributed along the clone. The parameter values of the sequencing design in our simulation are specified in Table 3. The composition probability of the SNPs is shown in Table 4. The sequencing error probabilities are shown in Table 5. The clone size is 180,000 bp. The average coverage is 7. Each clone has 599 SNPs and therefore has 598 adjacent pairs.

To determine the significance of pairwise comparison, we set thresholds for both the confidence and odds ratio of the two most probable cases. We show results under seven thresholds for confidence and four

Table 3. Parameter Values in the Simulation

Clone length H	180,000 bp
Average distance between SNPs	300 bp
Average fragment length L	650 bp
Average coverage κ	6–10

Table 4. Composition Probability in the Simulation

A	*G*	*C*	*T*	—
0.24	0.24	0.24	0.24	0.04

Table 5. Sequencing Error Probabilities in the Simulation

	Observation					
Truth	*A*	*G*	*C*	*T*	—	*N*
A	0.96	0.005	0.01	0.015	0.005	0.005
G	0.005	0.96	0.015	0.01	0.005	0.005
C	0.005	0.015	0.96	0.01	0.005	0.005
T	0.015	0.005	0.01	0.96	0.005	0.005
—	0.003	0.003	0.003	0.003	0.985	0.003

thresholds for odds ratio. Under each of the twenty eight configurations, 1,000 replicates were simulated. The results are shown in Table 6. Each subtable corresponds to a constant threshold for the odds ratio. We explain one case in detail. That is, we take the thresholds for the confidence and odds ratio to be 0.25 and 1.3, respectively. Among all the 598 adjacent pairs, 97 pairs are not connected by overlapping fragments. In terms of physical maps, they are separated by oceans; see Waterman (1995). In fact, an average of 0.57 SNPs are in the oceans. Among the rest of the 501 adjacent pairs, we missed 45 pairs due to low scores and detected 456 connected pairs using the above decision rule. Among the detected pairs, 419 of them are consistent with the true haplotype, and the false positive rate is 8.04%. By checking consistency, we found that 354 loci out of the 456 pairs do not conflict with others. In addition, the procedure reported genotypes for singletons, namely, those sites that cannot be connected to any other sites. The consistent pairs together with the singletons cover 526 positions. Among them, 498 positions are correct. The true positive rate is 94.67%. In total, 83.05% of all SNP sites are reported without error, and 70.17% of all 598 adjacent pairs are correctly reconstructed. The average size of haplotype segments including singletons is 4.7. From Table 6, we can choose the thresholds by trading off sensitivity and specificity. In fact, the results are quite robust with respect to selection of thresholds.

The statistics depend on the size of average coverage. The larger the coverage, the longer the reconstructed haplotype segments. Each replicate takes ten seconds of CPU time on a Dell PC equipped with an Intel Pentium-4 processor (2.26 GHz CPU speed). Our simulation also shows that considering triples instead of pairs offers slight gain in lengths of haplotype segments.

4. DISCUSSION

4.1. Haplotype membership of fragments

After the reconstruction of haplotype segments, the fragments are also grouped in accordance with these segments. The determination of haplotype memberships of these fragments is the dual problem of reconstructing haplotypes. For fragments falling in the range of a haplotype segment, the complete haplotyping information is in $\Pr(\mathbf{F}|\mathbf{X}, \mathbf{S})$. Maximizing this probability gives one solution to the imputation of the missing membership $\mathbf{F}$. This joint way of haplotyping becomes computationally intensive when many fragments are involved. A simple solution is to consider one fragment at a time.

4.2. Gibbs sampling

One solution to the computationally intensive problem of considering many SNPs at one time is the conditional version of the Gibbs sampling technique. As a matter of fact, the conditional distribution

Table 6. Simulation Results[a]

	Threshold for odds ratio: 1						
Threshold for confidence	*0.2*	*0.225*	*0.25*	*0.275*	*0.3*	*0.325*	*0.35*
False positive rate (pairs)	9.91%	9.50%	8.65%	7.92%	7.08%	6.31%	5.65%
# SNP reported (including singleton)	553	548	535	524	510	498	486
True positive rate (all reported)	92.1%	92.6%	94.3%	94.8%	95.4%	95.9%	96.3%
Percentage of correctly detected pairs	71.7%	71.4%	70.7%	70.0%	69.2%	68.3%	67.4%
Average segment length	5.31	5.17	4.78	4.67	4.56	4.47	4.37

	Threshold for odds ratio: 1.1						
Threshold for confidence	*0.2*	*0.225*	*0.25*	*0.275*	*0.3*	*0.325*	*0.35*
False positive rate (pairs)	9.19%	8.90%	8.47%	7.82%	7.02%	6.30%	5.65%
# SNP reported (including singleton)	540	537	531	522	509	497	486
True positive rate (all reported)	93.9%	94.1%	94.4%	94.9%	95.5%	95.9%	96.3%
Percentage of correctly detected pairs	70.9%	70.8%	70.5%	69.9%	69.1%	68.2%	67.4%
Average segment length	4.85	4.80	4.75	4.66	4.55	4.46	4.37

	Threshold for odds ratio: 1.3						
Threshold for confidence	*0.2*	*0.225*	*0.25*	*0.275*	*0.3*	*0.325*	*0.35*
False positive rate (pairs)	8.59%	8.37%	8.04%	7.58%	6.83%	6.13%	5.54%
# SNP reported (including singleton)	532	530	526	519	507	496	485
True positive rate (all reported)	94.2%	94.4%	94.7%	95.0%	95.5%	96.0%	96.4%
Percentage of correctly detected pairs	70.5%	70.4%	70.2%	69.8%	69.0%	68.2%	67.3%
Average segment length	4.77	4.74	4.70	4.64	4.54	4.46	4.37

	Threshold for odds ratio: 1.5						
Threshold for confidence	*0.2*	*0.225*	*0.25*	*0.275*	*0.3*	*0.325*	*0.35*
False positive rate (pairs)	8.40%	8.21%	7.91%	7.49%	6.83%	6.13%	5.54%
# SNP reported (including singleton)	529	527	524	517	507	496	485
True positive rate (all reported)	94.4%	94.5%	94.8%	95.1%	95.5%	96.0%	96.4%
Percentage of correctly detected pairs	70.3%	70.2%	70.0%	69.6%	69.0%	68.2%	67.3%
Average segment length	4.74	4.71	4.67	4.62	4.54	4.46	4.37

[a] To determine the significance of pairwise comparison, we set thresholds for both confidence and odds ratio of the two most probable cases. We show results under seven thresholds for confidence and four thresholds for odds ratio. Each subtable corresponds to a constant threshold for odds ratio. Under each configuration, 1,000 replicates were simulated. The false positive rates are for pairwise comparison. Some of these positive pairs were later either modified or discarded by checking consistency. The SNP number in the final report includes singletons, namely, those single sites that cannot be connected to others. In this case, we report their genotypes. The true positive rates are for those reported sites, either genotypes or haplotypes. The last two accounts are the percentage of correctly detected pairs among all and average lengths of haplotype segments.

$\Pr(\mathbf{S}|\mathbf{F}, \mathbf{X})$ is essentially the one-chromosome problem solved in Churchill and Waterman (1992). We can compute $\Pr(\mathbf{S}, \mathbf{F}|\mathbf{X})$ by alternating the following steps:

1. generate $s^{(l+1)}$ from $\Pr(\mathbf{S}|\mathbf{F} = f^{(l)}, \mathbf{X} = x)$;
2. generate $f^{(l+1)}$ from $\Pr(\mathbf{F}|\mathbf{S} = s^{(l+1)}, \mathbf{X} = x)$.

As we mentioned earlier, the distribution of $\Pr(\mathbf{F}|\mathbf{S} = s, \mathbf{X} = x)$ is also difficult to obtain. A remedy is to replace Step 2 by a series of sampling. That is, we update one fragment membership while keeping other fragment memberships unchanged and carry out the operation through all the fragments.

4.3. Parameter estimation

We can apply the E-M method given in Churchill and Waterman (1992) to estimate the composition probabilities and sequencing error rates for conserved regions. If these parameters for SNPs regions are about the same as those in conserved regions, then no further consideration is necessary. If not, then we have to estimate the parameters using only data from the SNP assembly. To describe the real data, we need to consider the orientation issue. For each fragment, we define the reverse complement indicator to be

$$r_i = \begin{cases} 1 & \text{fragment i is direct} \\ 0 & \text{fragment i is reverse complemented.} \end{cases}$$

Also we define the complement operator $\sim$ on $\mathcal{B}$ by $\tilde{A} = T$, $\tilde{T} = A$, $\tilde{G} = C$, $\tilde{C} = G$, $\tilde{-} = -$, $\tilde{N} = N$, and $\tilde{\phi} = \phi$. Given the complete data, **S**, **F**, and **X**, we calculate the sufficient statistics:

$$\begin{aligned} n_a &= \sum_{i=1}^{m}\sum_{j=1}^{n}\{r_i[\mathbf{1}(F_i = 1)\mathbf{1}(S_{1j} = a) + \mathbf{1}(F_i = 2)\mathbf{1}(S_{2j} = a)] \\ &\quad + (1 - r_i)[\mathbf{1}(F_i = 1)\mathbf{1}(\tilde{S}_{1j} = a) + \mathbf{1}(F_i = 2)\mathbf{1}(\tilde{S}_{2j} = a)]\}, \\ n_{ab} &= \sum_{i=1}^{m}\sum_{j=1}^{n}\{r_i\mathbf{1}(X_{ij} = b)[\mathbf{1}(F_i = 1)\mathbf{1}(S_{1j} = a) + \mathbf{1}(F_i = 2)\mathbf{1}(S_{2j} = a)] \\ &\quad + (1 - r_i)\mathbf{1}(X_{ij} = b)[\mathbf{1}(F_i = 1)\mathbf{1}(\tilde{S}_{1j} = a) + \mathbf{1}(F_i = 2)\mathbf{1}(\tilde{S}_{2j} = a)]\}. \end{aligned}$$

The maximum likelihood estimate of sequencing error rates is given by

$$\hat{p}(b|a) = n_{ab}/n_a.$$

In order to apply the E-M algorithm—see McLachlan and Krishnan (1996) for references and details—we need to impute the missing information by taking expectations conditional on the data; that is

$$\begin{aligned} \hat{n}_a &= \sum_{i=1}^{m}\sum_{j=1}^{n}\{r_i[\Pr(F_i = 1, S_{1j} = a|\mathbf{X}) + \Pr(F_i = 2, S_{2j} = a|\mathbf{X})] \\ &\quad + (1 - r_i)[\Pr(F_i = 1, \tilde{S}_{1j} = a|\mathbf{X}) + \Pr(F_i = 2, \tilde{S}_{2j} = a|\mathbf{X})]\}, \\ \hat{n}_{ab} &= \sum_{i=1}^{m}\sum_{j=1}^{n}\{r_i\mathbf{1}(X_{ij} = b)[\Pr(F_i = 1, S_{1j} = a|\mathbf{X}) + \Pr(F_i = 2, S_{2j} = a|\mathbf{X})] \\ &\quad + (1 - r_i)\mathbf{1}(X_{ij} = b)[\Pr(F_i = 1, \tilde{S}_{1j} = a|\mathbf{X}) + \Pr(F_i = 2, \tilde{S}_{2j} = a|\mathbf{X})]\}. \end{aligned}$$

The E-M algorithm is a special case of the following more general iterative scheme.

1. Imputation: given a set of parameter values, we impute the haplotype segments as well as the fragments' memberships.
2. Parameter estimation: update the parameter estimates based on current imputed values.

Various methods proposed in the literature can be applied to our problem. One such example is the data augmentation procedure; see Tanner and Wong (1987).

4.4. Indexing haplotypes

In order to evaluate $\Pr(\mathbf{S}|\mathbf{X}) = \Pr(\mathbf{X}, \mathbf{S})/\Pr(\mathbf{X})$, we need to compute $\Pr(\mathbf{X}) = \sum \Pr(\mathbf{X}, \mathbf{S})$, where the summation is taken over all possible haplotypes. For K SNPs, this number is 5^{2K} if we count without treating homozygous sites differently. Next, we describe a more efficient way of indexing haplotypes. We use numeric code 0, 1, 2, 3, 4 to represent -, A, G, C, T, respectively. At one locus, there are

25 states for $\{s_{1,j}, s_{2,j}\}$, which can be indexed by one number defined by $s_{1,j} + s_{2,j} * 5$. To eliminate the redundancy, we add a constraint that $s_{1,j} \leq s_{2,j}$. This results in 15 genotypes. At two loci, there are 625 states for $\{s_{1,j_1}, s_{2,j_1}, s_{1,j_2}, s_{2,j_2}\}$, which can be indexed by $s_{1,j_1} + 5 * s_{2,j_1} + 5^2 * s_{1,j_2} + 5^3 * s_{2,j_2}$. To eliminate the redundancy, we impose the following constraints: 1. $s_{1,j_1} \leq s_{2,j_1}$; 2. if $s_{1,j_1} = s_{2,j_1}$, then $s_{1,j_2} \leq s_{2,j_2}$. In general, if we consider K SNPs simultaneously, there are 25^K states for $\{s_{1,j_1}, s_{2,j_1}, s_{1,j_2}, s_{2,j_2}, \ldots, s_{1,j_K}, s_{2,j_K}\}$, which can be indexed by $\sum_{k=1}^{K}[s_{1,j_k} * 5^{2k-2} + s_{2,j_k} * 5^{2k-1}]$. To eliminate the redundancy, we impose the constraint defined by the following iteration, which can be implemented in programming.

```
Let $t = 1$;
while $t < K + 1$ do
    if $s_{1,j_t} < s_{2,j_t}$ then
        exit and the index is legitimate;
    end
    if $s_{1,j_t} > s_{2,j_t}$ then
        exit and the index is illegitimate;
    end
    if $s_{1,j_t} = s_{2,j_t}$ then
        $t = t + 1$;
    end
end
if $t = K + 1$ then
    The index is legitimate;
end
```

The cases satisfying the last "**If**" condition are homozygotes at all sites. This definition leads to the following account.

Proposition 1. *The number of different haplotypes is* $(5^K + 5^{2K})/2$.

Proof. It is easy to check that the total number is given by $5^K + 10\sum_{j=0}^{K-1} 5^j(5^2)^{K-1-j}$. When $K = 2$, the number is 325, which is consistent with the number in Table 2. ■

4.5. Future work

We will extend the approach from one person to a group of people, such as a library of sequences from a family. Then we need to estimate the haplotype frequencies. For a large scale genome sequencing project, the assumptions A6 and A7 are not suitable for the model. A more careful treatment is to divide the whole genome into different regions and compute the sequencing error rates and compositions adaptively. Then the starting point or the initialization of the E-M algorithm will be an interesting topic. If we have the quality values for each fragment or each position in each fragment such as Phred scores (see Ewing and Green, 1998), then we can remove A3 and A4 and take the quality values into consideration. PolyPhred uses Phred scores to detect SNPs; see Nickerson (1997). Our method is based on a statistical-model, and we emphasize the reconstruction of haplotypes instead of genotypes. So far, we deal only with the SNP fragment alignment. If we keep the alignment of non-SNP sites, we can remove A5 and consider the local context of each site within a fragment.

ACKNOWLEDGMENTS

The authors were supported by NIH Grant R01 HG02366. Lei Li also received support from NSF DMS 9971698 and Michael Waterman received support from Celera Genomics/Applied Biosystems. We are grateful to Dr. Sorin Istrail for suggesting the problem.

REFERENCES

Churchill, G.A., and Waterman, M.S. 1992. The accuracy of DNA sequences: Estimating sequence quality. *Genomics* 14, 89–98.

Cornish-Bowden, A. 1985. Nomenclature for incompletely specified bases in nucleic acid sequences: Recommendations 1984. *Nucl. Acids Res.* 13, 3021–3030.

Ewing, B., and Green, P. 1998. Basecalling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* 8, 186–194.

Irizarry, K., Kustanovich, V., Li, C., Brown, N., Wong, W., Nelson, S., and Lee, C. 2000. Genome-wide analysis of single nucleotide polymorphisms in human expressed sequences. *Nature Genet.* 26, 233–236.

Lancia, G., Bafna, V., Istrail, S., Lippert, R., and Schwartz, R. 2001. SNPs problems, complexity, and algorithms, *European Symposium on Algorithms*, 182–193.

Lander, E.S., Waterman, M.S. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* 2, 231–239.

McLachlan, J.G., and Krishnan, T. 1996. *The EM Algorithm and Extensions*, John Wiley, NY.

Nickerson, D.A., Tobe, V.O., and Taylor, S.L. 1997. Polyphred: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucl. Acids Res.* 25, 2745–2751.

Venter, J.C., *et al.*. 2001. The sequence of human genome. *Science* 291, 1304–1351.

Tanner, M., and Wong, W.H. 1987. The calculation of posterior distributions by data augmentation. *J. Am. Statist. Assoc.* 82, 528–547.

Waterman, M.S. 1995. *Introduction to Computational Biology*, Chapman and Hall, London.

Address correspondence to:
Lei M. Li, Jong Hyun Kim, and Michael S. Waterman
Molecular and Computational Biology Program
Department of Biological Sciences
University of Southern California
1042 West 36th Place, DRB 289
Los Angeles, CA 90089

E-mail: lilei@hto.usc.edu