A model of higher accuracy for the individual haplotyping problem based on weighted SNP fragments and genotype with errors

Minzhu Xie1,2, Jianxin Wang1,* and Jianer Chen1,3

1School of Information Science and Engineering, Central South University, Changsha 410083, 2College of Physics and Information Science, Hunan Normal University, Changsha 410081, China and 3Department of Computer Science, Texas A&M University, College Station, TX 77843, USA

ABSTRACT
Motivation: In genetic studies of complex diseases, haplotypes provide more information than genotypes. However, haplotyping is much more difficult than genotyping using biological techniques. Therefore effective computational techniques have been in demand. The individual haplotyping problem is the computational problem of inducing a pair of haplotypes from an individual’s aligned SNP fragments. Based on various optimal criteria and including different extra information, many models for the problem have been proposed. Higher accuracy of the models has been an important issue in the study of haplotype reconstruction.

Results: The current article proposes a highly accurate model for the single individual haplotyping problem based on weighted fragments and genotypes with errors. The model is proved to be NP-hard even with gapless fragments. Based on the characteristics of Single Nucleotide Polymorphism (SNP) fragments, a parameterized algorithm of time complexity \( O(nk^2m^2+\log n+\log k) \) is developed, where \( n \) is the number of fragments, \( m \) is the number of SNP sites, \( k \) is the maximum number of SNP sites that a fragment covers (no more than \( n \) and usually smaller than 10) and \( k_2 \) is the maximum number of the fragments covering a SNP site (usually no more than 19). Extensive experiments show that this model is more accurate in haplotype reconstruction than other models.

Availability: The program of the parameterized algorithm can be obtained by sending an email to the corresponding author.

Contact: jxwang@mail.csu.edu.cn

1 INTRODUCTION
The different DNAs between two individuals’ genomes account for about 0.5% of the whole genome sequence (Levy et al., 2007), and these differences make the two individuals different from each other in figures, diseases susceptibilities and other phenotypes. A single nucleotide polymorphism (SNP) is a change of a single nucleotide in a given position of the genome sequence with a frequency not <1% in a given population. There are millions of SNPs in the human genome (The International HapMap Consortium, 2005; Venter et al., 2001). SNPs are believed to be the major genetic cause to human phenotypic variability. It is widely accepted that at a given SNP site, there are only two possible nucleotides, one usually occurs in more than 90% of the human genome (International HapMap Consortium, 2005; Levy et al., 2001). SNPs are believed to be the major genetic cause of human phenotypic variability.

The human genome is made up of 23 pairs of chromosomes. A sequence of SNP alleles on one of a pair of chromosomes is called a haplotype, which can be denoted by a string over \([0, 1]\). A sequence of conflated (unordered pair of) SNP alleles at each SNP site of a pair of homologous chromosomes is called a genotype. A genotype can be represented by a string over \([0, 1, 2]\), where ‘0’ (resp. ‘1’) indicates that both SNPs are ‘0’ (resp. ‘1’) at the same SNP site of the pair of chromosomes, and ‘2’ indicates that at the same SNP site, the SNP on one of the pair of chromosomes is ‘0’, while on the other of the pair is ‘1’.

In Figure 1, the haplotypes of the individual are \((A, C, G, T)\) and \((G, C, C, T)\), which can be denoted by ‘0100’ and ‘1110’. The genotype is \((A/G, C/C, G/C, T/T)\), which can be denoted by ‘2120’. In finding susceptibility loci for complex diseases, haplotype-based methods are more powerful and robust than the methods based on individual SNPs (Akey et al., 2001). However, determining haplotypes using biological techniques is both time consuming and expensive, and is much harder than determining individual SNPs or genotypes. Therefore, to reduce the cost of determining haplotypes, effective computational methods have been in demand.

There have been many computational models for the haplotyping problem (Adkins, 2004; Bonizzoni et al., 2003; Zhang et al., 2006), and they generally fall into two classes: individual haplotyping and population haplotyping. The individual haplotyping problem is concerned with assembling a pair of haplotypes from an individual’s aligned DNA fragments, while the population haplotyping problem is to infer the haplotypes of a sample of people in a population from their genotypes.

This current article is focused on the individual haplotyping problem and aimed at studying highly accurate model and developing effective computational algorithms for the problem. The article is organized as follows. In Section 2, we introduce the individual haplotyping problem and propose a new model based on weighted SNP fragments and genotype with errors. We prove that the individual haplotyping problem under this model is NP-hard. To effectively solve the problem under this model, we develop in Section 3 an exact parameterized algorithm for the problem. Section 4 presents experimental results and discussions. The article is concluded by Section 5.

© 2008 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
2 FORMULATION AND PROBLEMS

2.1 The individual haplotyping problem

For large-scale haplotyping, a set of aligned SNP fragments from a pair of chromosomes can be generated by DNA shotgun sequencing or other sequencing experiments. The individual haplotyping problem (Lancia et al., 2001) is aimed at partitioning the set of SNP fragments into two subsets, each determining a haplotype.

To formulate the problem, we introduce some notations and concepts, similar to those used in Xie and Wang (2007).

In the following discussion, m aligned SNP fragments coming from a pair of chromosomes with n SNP sites are represented by an \( m \times n \) SNP matrix \( M \) over the alphabet \{0, 1, −, \}, in which each row corresponds to a fragment and each column corresponds to a SNP site. The SNP allele of the \( i \)th fragment at the \( j \)th SNP site is denoted by \( M_{ij} \). The SNP allele of the \( i \)th fragment at the \( j \)th SNP site corresponds to a fragment and each column corresponds to a SNP site. The SNP allele of the \( i \)th fragment at the \( j \)th SNP site is denoted by \( M_{ij} \).

As an example, in the SNP matrix in Figure 2, row 2 covers columns 2, 3, 4, 5, and 6, and \( M_{2, i} \) and \( M_{2, j} \) are not ‘−’.

The set of (ordered) rows covering the \( j \)th column is denoted by \( R_j(i) \). The first column that the \( i \)th row covers is denoted by \( h(i) \), and the last column that the \( i \)th row covers is denoted by \( r(i) \).

As an example, in the SNP matrix in Figure 2, row 2 covers columns 2, 3, 4, 5, and 6, and \( R_5(5) = \{1, 2, 4, 7\} \).

If \( M_{ij} \) takes the value ‘0’ and \( M_{ij} \) takes the value ‘1’, or \( M_{ij} \) and \( M_{ij} \) take the value ‘0’, we say that the \( i \)th row and the \( j \)th row of \( M \) are compatible at the column \( j \). If two rows of \( M \) do not conflict at any column, we say that they are compatible.

If the rows of a SNP matrix \( M \) can be partitioned into two subsets such that the rows in each subset are all compatible, then we say that the SNP matrix \( M \) is feasible.

Obviously, a SNP matrix \( M \) is feasible if and only if there are two haplotypes such that every row of \( M \) is compatible with one of the two haplotypes. In this case, we say that the SNP matrix \( M \) can be derived from the two haplotypes.

Since a row of \( M \) comes from one of a pair of chromosomes, if there are no DNA sequencing errors, we can always derive \( M \) from the haplotypes of the pair of chromosomes. However, DNA sequencing errors are unavoidable and it is hard to decide which copy of chromosome a SNP fragment comes from. This has made the individual haplotyping problem complicated.

Based on different optimal criteria, various computational models have been proposed for the problem. Some typical models include (Lancia et al., 2001; Lippert et al., 2002): Minimum Fragment Removal (MFR), Minimum SNPs Removal (MSR) and Minimum Error Correction (MEC). Among these models, MEC is considered to have most biological meaning. MEC is also called Minimum Letter Flips (MLF) (Greenberg et al., 2004) and has been extended to include different extra information.

A DNA sequencer can provide a confidence level for each base, which is the probability that the base was correctly read (Zhao et al., 2005). The confidence levels for an \( m \times n \) SNP matrix \( M \) can be represented as an \( m \times n \) weight matrix \( W \), in which the element \( W_{ij} \) of \( W \) at row \( i \) and column \( j \) is the confidence level of the base \( M_{ij} \).

We also define \( W_{ij} = 0 \) if \( M_{ij} = '−' \). By including a weight matrix, Greenberg et al. (2004) introduced the weighted minimum letter flips (WMLF) model, and Zhao et al. (2005) formulated it as follows:

\text{Weighted Minimum Letter Flips (WMLF): Given a SNP matrix} \( M \) \text{ and the corresponding weight matrix} \( W \), \text{flip a number of elements in} \( M \) \text{ from}'0' \text{into}'1' \text{and vice versa} so that the resulting matrix is feasible and the sum of the elements in} \( W \) \text{corresponding to the flipped elements in} \( M \) \text{is minimized.}

An individual’s genotype can be phased relatively easily. Wang et al. (2005) extended the MEC model as follows:

\text{MEC with Genotype Information (MEC/GI): Given a SNP matrix} \( M \) \text{ and a genotype} \( G \), \text{flip a minimum number of elements in} \( M \) \text{from}'0' \text{into}'1' \text{and vice versa} so that the resulting matrix can be derived by two haplotypes that make up the genotype} \( G \).

In general, WMLF and MEC/GI are superior to MEC (or MLF) in the accuracy of haplotype reconstruction. However, WMLF does not take genotype information into account, and MEC/GI does not consider the errors in a genotype and the confidence levels of DNA bases. In fact, there are always genotyping errors. In order to improve the accuracy of haplotype reconstruction, we propose a new model in the next subsection.

2.2 WMLF incorporating genotyping uncertainty

There are two broad categories of genotyping errors (Kang et al., 2004): operational errors and genotype scoring errors. Recently, the operational errors have decreased significantly in high-throughput genotyping due to technological advances. On the other hand, genotype scoring errors are still a significant challenge for automated scoring softwares (Kang et al., 2004), and research on kernel algorithms for genotype scoring remains a hot spot (Carvalho et al., 2007; Xiao et al., 2007). Limited by scoring softwares, genotyping errors are unavoidable (Kang et al., 2004; Zhu, 2006).

Therefore, incorporating genotyping uncertainty will help the process of reconstructing haplotypes.

To incorporate genotyping uncertainty, Kang et al. (2004) introduced GenoSpectrum. Considering \( n \) SNPs of an individual, a GenoSpectrum \( F \) is a \( 3 \times n \) matrix as given in Figure 3, where for each \( j \), \( 1 \leq j \leq n \), \( f_{1j}, f_{2j}, f_{3j} \) are the likelihood that the individual’s genotype at the \( j \)th SNP site is 0, 1, and 2, respectively.
A model of higher accuracy for WMEC/GS

Let $H[j]$ and $G[j]$ denote the $j$th character of a haplotype $H$ and a genotype $G$, respectively.

**Definition 1.** Let $M$ be an $m \times n$ SNP matrix and let $F$ be a $3 \times n$ GenoSpectrum. The most likely genotype $G_M(F)$ of $M$ with respect to $F$ is defined to be the genotype whose values satisfy the following conditions for all $j$:

$$G_M(F)[j] = \begin{cases} k: k \text{ maximizes } f_{k,j}, & \text{if } M_{i,j} = 0 \text{ and } M_{j,1} = 1; \\ k: M_{i,j} \neq 0, & \text{otherwise.} \end{cases}$$

Roughly speaking, $G_M(F)[j]$ gives an index $k$ so that the value $f_{k,j}$ shows the degree of consistency of the $j$th column of the SNP matrix $M$ and the $j$th column of the GenoSpectrum $F$. In particular, if all elements in the $j$th column of $M$ are $0$, the SNP matrix $M$ provides no information for the $j$th SNP site. Thus, the largest value $f_{k,j}$ in the $j$th column of $M$, where $k = \arg \max_G G_M(F)[j]$, shows the degree of confidence that the most likely genotype of the $j$th SNP site is $k$. If there are both 0’s and 1’s in the $j$th column of $M$, then $G_M(F)[j] = 2$ and the value $f_{2,j}$ is the degree of confidence that the $j$th column in the SNP matrix $M$ and the $j$th column in the GenoSpectrum $F$ are consistent. Finally, if the $j$th column in the SNP matrix $M$ contains exactly one value $q$ from $[0, 1]$, then $G_M(F)[j] = q$ and the value $f_{q,j}$ is the degree of confidence that the $j$th column in the SNP matrix $M$ and the $j$th column in the GenoSpectrum $F$ are consistent.

**Definition 2.** Let $M$ be an $m \times n$ SNP matrix and let $F$ be a $3 \times n$ GenoSpectrum. The distance $d(M, F)$ between $M$ and $F$ is defined as follows:

$$d(M, F) = \sum_{j=1}^{n} (1 - f_{GM}(j)).$$

By the above discussion on the most likely genotype $G_M(F)$ of $M$ with respect to $F$, we can see, intuitively, that the distance $d(M, F)$ measures the degree of inconsistency between the SNP matrix $M$ and the GenoSpectrum $F$.

**Definition 3.** Let $M$ be an $m \times n$ SNP matrix, $W$ be an $m \times n$ weight matrix corresponding to $M$, and $F$ be a $3 \times n$ GenoSpectrum, $g_w$ be a weighted coefficient, $S$ be an element subset of $M$ (i.e. $S \subseteq [M_{i,j} | 1 \leq i \leq m, 1 \leq j \leq n]$), and $M'$ be the SNP matrix derived from $M$ by flipping the elements in $S$. The flipping cost of $S$ based on $g_w$, $W$, $M$, and $F$ is defined as follows:

$$C(S) = g_w \cdot d(M', F) + \sum_{i,j \in S} W_{i,j}.$$

Since the value $d(M', F)$ measures the degree of inconsistency between the SNP matrix $M'$ and the GenoSpectrum $F$, the flipping cost $C(S)$ of a subset $S$ of elements in the SNP matrix $M$ is measured in terms of two metrics: the degree of inconsistency between the resulting SNP matrix $M'$ and the GenoSpectrum $F$ (i.e. $d(M', F)$), and the degree of confidence that is decreased during flipping the elements in $S$ (i.e. $\sum_{i,j \in S} W_{i,j}$). The weighted coefficient $g_w$ is used to adjust the relationship between these two metrics.

Based on Definitions 2 and 3, the following equation holds true:

$$C(S) = g_w \sum_{j=1}^{n} (1 - f_{GM}(j)) + \sum_{i,j \in S} W_{i,j}$$

$$= ng_w + \sum_{j=1}^{n} (g_w f_{GM}(j) + \sum_{i,j \in S} W_{i,j})$$

(1)

In the following, we propose a new computational model for the individual haplotyping problem.

**WMLF/GS:** (Weighted Minimum Letter Flips with GenoSpectrum)

Given an $m \times n$ SNP matrix $M$, an $m \times n$ weight matrix $W$, a $3 \times n$ GenoSpectrum $F$ and a weighted coefficient $g_w$, find a subset $S$ of elements in $M$ and flip the elements in $S$ (‘0’ into ‘1’ and vice versa) so that the resulting matrix is feasible and the flipping cost of the subset $S$ based on $g_w$, $W$, $M$, and $F$ is minimized.

It is easy to see that when the weighted coefficient $g_w$ is set to 0, the WMLF/GS problem degenerates into the WMLF problem.

**Theorem 1.** The WMLF/GS problem is NP-hard even if its SNP matrix is gapless.

**Proof.** The WMLF problem can be reduced from the weighted max-cut problem, which is a well-known NP-hard problem (Zhao et al., 2005). Similarly, the WMLF/GS problem can be reduced from the weighted max-cut problem.

The weighted max-cut problem is defined as follows: given a directed graph $G=(V,E)$ and a positive edge weight function $w: E \rightarrow \mathbb{R}$, find a partition $(V_1, V_2)$ of $V$ such that the sum of the weights of the edges with one end in $V_1$ and the other end in $V_2$ is maximized. For brevity, the sum of the weights of the edges with one end in $V_1$ and the other end in $V_2$ will be called the cut weight of $(V_1, V_2)$, and the weight of an edge $(i,j)$ is denoted by $w(i,j)$.

As an example, consider the graph $G$ in Figure 4(a). Let $V_1 = \{1, 4\}$ and $V_2 = \{2, 3\}$. Then the cut weight of $(V_1, V_2)$ is 0.5. 0.2 + 0.3 + 0.6 = 1.6. It can be easily verified that this cut has a maximum weight.

Given a directed graph $G=(V,E)$ with $n$ vertices and a positive edge weight function $w$, where, without loss of generality, the vertices are named as 1, 2, ..., $n$. We can construct an $n \times n$ SNP matrix $M$ and an $n \times n$ weight matrix $W$ as follows:

$$M_{i,j} = \begin{cases} 0 & \text{if } i = j, \\ 1 & \text{if } i \neq j. \end{cases}$$

$$W_{i,j} = \begin{cases} w(i,j) & \text{if } i \neq j \text{ and } (i,j) \notin E, \\ 0 & \text{if } i \neq j \text{ and } (i,j) \in E, \\ 1 + \sum_{k \neq j \in E} w(k,j) & \text{if } i = j. \end{cases}$$
The 3 × n GenoSpectrum F can be constructed as follows: for j = 1, ..., n, f_{1j} = f_{2j} = f_{3j} = 1. Let g_{0j} be any positive number.

According to the rules above, the SNP matrix M and the corresponding matrix W constructed from the undirected weighted graph G in Figure 4a are given in Figure 4b (with the weight values given in the brackets). The GenoSpectrum F is given in Figure 4c.

In general, for the SNP matrix M constructed above, the ith row represents the ith vertex of G, and a partition of V corresponds to a partition (S_1 = V_1, S_2 = V_2) of the rows of M. Let E = E_{12} be 1 or 2. To make the rows in S_2 compatible with each other and minimize the sum of the weights of the flipped elements, for each row i ∈ S_2, all the rows of other rows in S_2 at column i should flip from 1 to 0, in order not to conflict with row i at column i. Therefore, the minimum sum of the weights of the flipped elements is equal to ∑_{i ∈ S_2, j ≠ i} w(i, j). Consequently, we have:

\[
\sum_{i,j \in S_2, i \neq j} W_{ij} + \sum_{i,j \in V \setminus S_2, i \neq j} W_{ij} = \sum_{i,j \in V_1 \setminus S_2, i \neq j} w(i, j) + \sum_{i,j \in V_2, i \neq j} w(i, j)
\]

Let the set of the flipped elements above be S. After flipping the elements in S the distance between M and F is 0. Therefore, C(S) is minimized if and only if the cut weight of (V_1, V_2) is maximized. This completes the proof of the theorem.

In fact, most models for the individual haplotyping problem are NP-hard. To obtain exact solutions to these models, some dynamic programming and branch-and-bound algorithms have been proposed. For larger problem instances, these exact algorithms become infeasible, and heuristic algorithms have been used instead. However, heuristic algorithms usually cannot ensure accuracy. Recently, fixed-parameter tractability theory has been used to design practical exact algorithms for certain NP-hard problems with great success. For the MFR and MSR problems, Xie et al. (2007) and Xie and Wang (2007) have proposed parameterized algorithms by taking the advantage of the characters of SNP fragments. In the following section, we will use the technique to develop a parameterized algorithm for the WMLF/GFS problem.

3 METHODS

In Xie and Wang (2007) and Xie et al. (2007), we have observed the following properties of DNA sequence fragment data. Due to technical limits, the sequencing instruments such as ABI 3730 and MagelBACE in most big sequencing centers can only sequence DNA fragments whose length is usually not more than 1200 nucleotide bases. Since the SNP density is about 1 SNP per 1 kb, the maximum number of SNP sites that a fragment can cover is small.

Moreover, in DNA sequencing experiments, the fragment coverage is also small. In both Celera’s whole-genome shotgun assembly of the human genome and the human genome project of the International Human Genome Sequencing Consortium, the fragment average coverage is about five (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001). Although the fragment covering rate varies along the whole genome, the fragment coverage plot in Huson et al. (2001) about the fragment data of the human genome project of Celera’s shows that the number of fragments covering a site is bounded by 19. Therefore, compared with the total number of fragments, the number of fragments that cover a SNP site is very small.

Based on the observations above, as in Xie and Wang (2007), we introduce the (k_1, k_2) parameterized condition.

Definition 4. The (k_1, k_2) parameterized condition: a single fragment covers at most k_1 SNP sites, and the number of fragments that cover a SNP site is bounded by k_2.

As to the corresponding SNP matrix M, the (k_1, k_2) parameterized condition means that each row of M covers at most k_1 columns and each column of M is covered by at most k_2 rows.

For an m × n SNP matrix M, the parameters k_1 and k_2 can be obtained by scanning all rows of M. In the worst case, k_1 = n and k_2 = m. But as to the fragment data of Celera’s human genome project, k_2 is no more than 19 (Huson et al., 2001).

3.1 A parameterized algorithm for WMLF/GFS

In this subsection, the SNP matrix M is preprocessed as follows: sort the rows in M in ascending order such that for any two rows i_1 and i_2, if i_1 < i_2, then h_{1i_1} ≤ h_{1i_2}, which is similar to the preprocessing in Xie et al. (2007) and Xie and Wang (2007). For a solution to the WMLF/GFS problem for a SNP matrix M, after flipping the corresponding elements, all the rows of M can be partitioned into two classes H_0 and H_1, such that every two rows in the same class are compatible.

Definition 5. Let R be a subset of rows in a SNP matrix M. A partition function P on R maps each row in R to one of the values (0, 1).

Suppose that R contains h > 0 rows, a partition function P on R can be denoted by an h-digit binary number in (0, 1), where the ith digit is the P value of the ith row in R. If P is 0, we also define a unique partition function P, which is denoted by −1.

Fig. 4. (a) A weighted graph. (b) A SNP matrix corresponding to the weighted graph in (a), with weighted values given in brackets. (c) A GenoSpectrum corresponding to the weighted graph in (a).
Recall that \( R(j) \) denotes the set of rows in the SNP matrix \( M \) that cover the \( j \)th column. For briefness, a partition function defined on \( R(j) \) is called a partition function at column \( j \). For a SNP matrix \( M \) satisfying the \((k_1, k_2)\) parameterized condition, there are at most \( 2^{k_2} \) different partition functions at column \( j \).

Let \( R \) be a set of rows of the matrix \( M \), and \( P \) be a partition function on \( R \). For a subset \( R' \) of \( R \), the partition function \( P' \) on \( R' \) obtained by restricting \( P \) on the subset \( R' \) is called the projection of \( P \) on \( R' \), and \( P' \) is called an extension of \( P \) on \( R' \).

For briefness, let \( M(i) \) denote the set of all non-empty elements from column 1 to column \( j \) in \( M \), and \( M(j) \) (\( W(j), P(j) \)) or \( F(j) \) denotes the SNP matrix (resp. the weighted matrix, or the GeneSpectrum consisting of the first \( j \) columns of \( M \) (resp. \( W \), or \( F \)).

**Definition 5.** Fix a column index \( j \). Let \( g_w \) be a weighted coefficient, \( M' = M(j) \), \( W \subseteq W(j) \), \( F \subseteq F(j) \). \( P \) be a partition function at column \( j \), and the collection of all possible subsets \( V_q[P(j)] \) be \( Y \). Defined \( E[P(j)] = \min_{v \in Y} C(S) \) and \( S_2[P(j)] \) to be any subset in \( Y \) that satisfies \( C(S_2[P(j)]) = E[P(j)] \), where \( C(S) \) and \( C(S_2) \) are the flipping costs of \( S \) and \( S_2(P(j)) \), respectively, based on \( g_w, M', W' \) and \( F' \).

From Definitions 6 and 7, it is easy to verify that the following equation holds true:

\[
\text{WMLFGS}(g_w, M', W, F') = S_2[P(n)], \tag{2}
\]

where \( P \) is a partition function at column \( n \) that minimizes \( E[P(n)] \).

Given a \( V_2[P(j)] \), let \( M' \) be the SNP matrix derived from \( M(j) \) by flipping the elements in \( V_q[P(j)] \), and \( V_2[P(j)] \) be the set of the elements of \( M' \) that are in column 1 and not in \( V_2[P(j)] \). Let \( C(V_2[P(j)]) \) denote \( \sum_{i \in M(j) \setminus \{j\}} |V_2[P(j)] \cap W_i| = g_w \sum_{i \in M(j) \setminus \{j\}} |V_2[P(j)] \cap W_i| \), where \( g_w \) is the value at column 1 of the most likely genotype of \( M' \), and \( V_2[P(j)] \cap W_i \) is the value of the element in \( P(j) \) at row \( g_w \). \( g_2[P(j)] = V_2[P(j)] \) and \( F_2[P(j)] = F_2 \).

Given a partition function \( P \) at column \( j \), the rows covering column \( j \) can be partitioned into \((H_0, H_1)\) by \( P \) according to the following rule: for each row \( r \) in \( H_0 \), \( r \in H_0 \) if and only if \( P(r) = 0 \). In order to make the rows in the same class not conflict at column \( j \), the values at column \( j \) of some rows may have to be flipped. For \( q \in \{0, 1\} \), let \( v_0 = 0 \). For any row \( r \in H_0 \), if \( M_r = 1 \), \( M \) is to be flipped. In consequence, we obtain a set \( F \) of flipped elements, and \( C(F) = \sum_{i \in M(j) \setminus \{j\}} |V_2[P(j)] \cap W_i| = g_w \sum_{i \in M(j) \setminus \{j\}} |V_2[P(j)] \cap W_i| \), where \( M' \) is the SNP matrix derived from \( M(j) \) by flipping the elements in \( F \). Obviously, \( M' \) is feasible.

\[
\text{CompFlips}(j, P, Flips, C) \label{eq:compflips}
\]

\[
\begin{align*}
\text{Flips} &\text{ denotes Flips}(F_2[P(j)], J) \text{ and } C \text{ denotes } \sum_{i \in M(j) \setminus \{j\}} |V_2[P(j)] \cap W_i| = g_w \sum_{i \in M(j) \setminus \{j\}} |V_2[P(j)] \cap W_i| \\
\text{where } &M_0 = \{ i \in M(j) \setminus \{j\} : (v_0 = 0, v_1 = 1) \} \text{ and } M_1 = \{ i \in M(j) \setminus \{j\} : (v_0 = 1, v_1 = 0) \}.
\end{align*}
\]

Fix a \( P \) and \( j \), it is obvious that \( C(V_2[P(j)]) \) is the minimum if and only if \( V_2[P(j)] = \text{Flips}(P(j)) \). Therefore, the following equations hold true.

\[
S_2[P(j)] = \text{Flips}(P(j)) \tag{3}
\]

\[
E[P(j)] = g_w + C(\text{Flips}(P(j))) \tag{4}
\]

Figure 5 gives a function \( \text{CompFlips}(j, P, Flips, C) \) to compute \( \text{Flips}(P(j)), \) and \( C(\text{Flips}(P(j))) \), whose time complexity is \( O(2^k) \) for a SNP matrix \( M \) satisfying the \((k_1, k_2)\) parameterized condition.

In order to present our algorithm, we need to extend the above concepts from one column to two columns as follows. Let the set of all rows that cover both columns \( j_1 \) and \( j_2 \) be \( R_{(j_1,j_2)} \).

**Definition 8.** Fix a \( j \). Let \( g_w \) be a weighted coefficient, \( M' = M(j) \), \( W = W(j) \), \( F = F(j) \). \( P \) be a partition function on \( R_{(j_1,j_2)} \), and the collection of all possible subsets \( V_q[P(j)] \) be \( Y \). Defined \( E[P(j)] = \min_{v \in Y} C(S) \) and \( S_2[P(j)] \) to be any element in \( Y \) that satisfies \( C(S_2[P(j)]) = E[P(j)] \), where \( C(S) \) and \( C(S_2) \) are the flipping costs of \( S \) and \( S_2(P(j)) \), respectively, based on \( g_w, M', W' \) and \( F' \).

Given a \( j \) and a partition function \( P \) on \( R_{(j_1,j_2)} \). If \( E[P(j)] \) and \( S_2[P(j)] \) are known for each extension \( P \) on \( R_{(j_1,j_2)} \), \( B[P(j)] \) and \( S_2[P(j)] \) can be calculated by the following equations:

\[
B[P(j)] = \min_{P \text{ an extension of } P \text{ on } R_{(j_1,j_2)}} E[P(j)] \tag{5}
\]

\[
S_2[P(j)] = S_2[P(j) \setminus j_2] \tag{6}
\]

\[
E[P(j)] = B[P(j) \setminus j_2] + g_w + C(\text{Flips}(P(j))) \tag{7}
\]

Based on Equations (2)-(8), the solution to the WMLFGS problem for a SNP matrix \( M \) can be obtained as follows: first, \( E[1, P] \) and \( S_2[P] \).
are obtained according to Equations (3) and (4) for all partition functions \( P \) at column 1; second, \( E[P_1] \) and \( S_q[P] \) can be obtained by using Equations (5) and (6) for all partition functions \( P \) on \( R_2(1,2) \); third, \( E[P_2] \) and \( S_q[P_2] \) can be obtained by using Equations (7) and (8) for all partition functions \( P \) on \( R_2(2) \); and so on. Finally, \( E[P, n] \) and \( S_q[P, n] \) can be obtained for all partition functions \( P \) at column \( n \). Once \( E[P, n] \) and \( S_q[P, n] \) for all possible \( P \) are known, a solution to the WMLF/GS problem for \( M \) can be obtained by using Equation (2). See Figure 6 for the details of our P-WMLF/GS algorithm.

**Algorithm P-WMLF/GS**

**Input:** an \( m \times n \) SNP matrix, \( M \); an \( m \times n \) weight matrix \( W \); a \( 3 \times n \) Geno Spectrum \( F \); and a weighted coefficient \( q_w \).  

**Output:** a solution to the WMLF/GS problem for \( M \).

1. **Initialization:** sort the rows in \( M \) in ascending order such that for any two rows \( i \) and \( j \), if \( i < j \), then \( l(i) \leq l(j) \); for each column \( j \), calculate an ordered set \( R_j(1) \) and the number \( H(j) \) of the rows that cover column \( j \); \( l = 1 \);

2. For \( P = 0 \) do

   2.1. CompFlipW(W, P, Flips, C);

   2.2. If \( E[P] \geq E[P_1] \) and \( E[P] \geq E[P_2] \), then \( E[P] = E[P_1] \) and \( S_q[P] = S_q[P_1] \).

3. While \( j < n \) do

   3.1. If column \( j \) is empty, then \( R_j(1) \) covers column \( j + 1 \); and a vector Bits such that Bits\((i) = 1\) denotes the \( i \)th row of \( R_j(1) \) covers column \( j + 1 \);

5. For \( P' = 0 \) to \( 2^{H(j)} - 1 \) do

   5.1. For \( q_w = 0 \) to \( 2^{l(j)} - 1 \) do

   5.1.1. CompFlipW(W, P, Flips, C);

   5.2. Calculate \( P' \) from \( P \) by \( P' = 2^{H(j)} \cdot P + 2^{l(j)} \cdot Flips + C \);

   5.3.1.1. If \( E[P'] \geq E[P] \), then \( E[P] = E[P'] \) and \( S_q[P] = S_q[P'] \).

   5.3.1.2. If \( E[P'] > E[P] \), then \( E[P] = E[P'] \) and \( S_q[P] = S_q[P'] \).

   5.3.1.3. If \( E[P] > E[P_1] \) then \( E[P] = E[P_1] \) and \( S_q[P] = S_q[P_1] \).

   5.4. \( j = j + 1 \); next column

6. For \( P' = 0 \) to \( 2^{H(j)} - 1 \) do

   6.1. For each extension \( P \) on \( R_j(1) \) do

   6.1.1. \( P = P' + C \); Flips = Flips + C;

   6.2. If \( E[P] > E[P'] \) for \( P' \) on \( R_j(1) \) do

   6.2.1. \( E[P] = E[P'] \) and \( S_q[P] = S_q[P'] \);

   6.3. \( E[P] \neq E[P'] \) and the corresponding \( S_q[P] \); Flips, \( C \);

   6.4. \( P = P' + C \); Flips = Flips + C;

   6.5.1.1. If \( E[P] > E[P_1] \), then \( E[P] = E[P_1] \) and \( S_q[P] = S_q[P_1] \);

   6.5.1.2. \( E[P] = E[P_1] \) and \( S_q[P] = S_q[P_1] \).

7. \( \sum_{i=0}^{n} \sum_{j=0}^{m} \) of the \( M \) is the minimal value \( \sum_{i=0}^{n} \sum_{j=0}^{m} \) for all possible \( P \) and \( q_w \).

8. For an error SNP site, \( \mu \) and \( \sigma \) is obtained according to Equations (3) and (4) for all partition functions \( P \) at column \( j \), but do not cover column \( j - 1 \) are all behind the rows in \( R(j - 1, j) \). In all, Step 3 takes time \( O(k^2 \log k + nk) \). Then Step 3 is iterated \( n - 1 \) times and takes time \( O(k^2 \log k) \). Step 4 takes time \( O(k^2) \). In summary, the time complexity of the algorithm is \( O(k^2 \log k + nk) \).

This completes the proof of the theorem.

**4 EXPERIMENTAL RESULTS**

We compare three models WMLF/GS, MEC/GI and WMLF for the individual haplotyping problem. For WMLF/GS, we adopt our P-WMLF/GS algorithm; for MEC/GI, we adopt Wang et al.’s genetic algorithm GA-MEC/GI (Wang et al., 2005); and for WMLF, we adopt Zhao et al.’s dynamic clustering algorithm DC-WMLF (Zhao et al., 2005). In the experiments, we compare the running time and the reconstruction rate of haplotypes (Wang et al., 2005) of these three algorithms. The reconstruction rate of haplotypes is defined as the ratio of the number of the SNP sites that are correctly inferred out by an algorithm to the total number of the SNP sites of the haplotypes.

The haplotype data can be obtained by two methods (Wang et al., 2005): the first is to get real haplotypes from public domain, and the second is to generate simulated haplotypes by computers. In our experiments, the real haplotypes were obtained from the file genotypes_chr1_CEU_r21_m_rwd phased.gz,1 which was issued in July 2006 by the International HapMap Project (2005). The file contains 120 haplotypes on chromosome 1 of 60 individuals of the CEU with each haplotype containing 193,333 SNP sites. From the 60 individuals, select a individual at random. Then beginning with a random SNP site, a pair of haplotypes of a given length can be obtained from the haplotypes of the selected individual.

The simulated haplotypes can be generated as follows (Panconesi and Sozio, 2004; Wang et al., 2005). At first a haplotype \( h_2 \) of length \( n \) is generated at random, then another haplotype \( h_2 \) of the same length is generated by flipping every character of \( h_1 \) with a probability of \( d \).

As to fragment data, to the best of our knowledge, real DNA fragments data in the public domain are not available, and references Wang et al. (2005) and Panconesi and Sozio (2004) used computer-generated simulated fragment data. After obtaining a pair of real or simulated haplotypes, in order to make the generated fragments have the same statistical features as the real data, a widely used shotgun assembly simulator Celsum (Myers, 1999) is invoked to generate \( m \) fragments whose lengths are between \( \min \) and \( \max \). At last the output fragments are processed to plant reading errors with probability \( e_1 \) and empty values with probability \( p \).

In our experiments, the parameters are as follows: fragment coverage rate \( e = 10 \), the difference rate between two haplotypes \( d = 20 \), the minimal length of fragment \( \min = 3 \), the maximal length of fragment \( \max = 7 \) and empty values probability \( p = 2 \% \).

The weight matrix \( W \) corresponding to the fragments is generated by the method of Zhao et al. (2005): the entries of \( W \) are normally distributed with mean \( \mu \) and variance \( \sigma^2 = 0.05 \). For a correct SNP site, \( \mu = 0.9 \), and for an error SNP site, \( \mu = 0.8 \).

The genotype without error can be obtained from the pair of haplotypes produced above. Since the genotyping error in biological

---

All experiments are repeated 100 times with Table 1. The data not enclosed in brackets are the experiment results on the real haplotype data, and the data enclosed in brackets are the experiment results on the simulated haplotype data. In the following experiments, we set $P$-$WLMF/GS$ achieves the highest haplotype reconstruction rate. In $\sigma$ and variance brackets. As shown in Table 1, the haplotype reconstruction rate of experiment results on the simulated haplotype data are enclosed in brackets. The experiment results on the real haplotype data, and Figure 8 shows the experiment results on the simulated haplotype data. When $n$ increases, the haplotype reconstruction rate of the three algorithms decreases, and their running time increases accordingly. In Figure 8a, when $n=20$, the haplotype reconstruction rates of $P$-$WLMF/GS$, $GA$-$MECGI$ and $WMLF$ are 96.8, 93.0 and 84.9%, respectively; and their running time are 1.3, 0.4 and 0.0007 s. When $n$ increases to 120, their haplotype reconstruction rates decrease to 93.5, 88.1 and 79.8%; and their running time increases to 10.2, 6.5 and 0.008 s. The experiment results in Figure 8b are similar. Although $P$-$WLMF/GS$ is the slowest among the three

Fig. 7. The performance of $P$-$WLMF/GS$ with $g_a$ varying.

Table 1. Comparison of the haplotype reconstruction rate of the algorithms with $c_e$ and $e_g$ varying

<table>
<thead>
<tr>
<th>$e_g$</th>
<th>$c_e=3%$</th>
<th>$c_e=5%$</th>
<th>$c_e=7%$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WMLF/GS</td>
<td>MEC/GI</td>
<td>WMLF</td>
</tr>
<tr>
<td>0</td>
<td>94.6 (94.7)</td>
<td>90.5 (90.7)</td>
<td>80.5 (80.4)</td>
</tr>
<tr>
<td>3</td>
<td>94.3 (93.8)</td>
<td>90.2 (89.5)</td>
<td>81.1 (80.3)</td>
</tr>
<tr>
<td>5</td>
<td>95.0 (93.8)</td>
<td>89.6 (88.7)</td>
<td>80.0 (80.5)</td>
</tr>
<tr>
<td>7</td>
<td>93.9 (93.3)</td>
<td>87.5 (86.7)</td>
<td>79.8 (80.9)</td>
</tr>
</tbody>
</table>

The data not enclosed in brackets are the experiment results on the real haplotype data, and the data enclosed in brackets are the experiment results on the simulated haplotype data. All experiments are repeated 100 times with $n=100$ and $n=200$. 

For using the genotype information, MEC/GI gets higher haplotype reconstruction rate than WMLF, which is illustrated in Table 1. Anyway, the experiment results on both the real and the simulated haplotype data show consistently that, in the reconstruction rate of haplotypes, WMLF/GS model is about 4% more accurate than MEC/GI model and is about 12% more accurate than WMLF model.

As shown in Figure 8, when $e_g=c_e=5\%$, the experiments on both the real and the simulated haplotype data with $n$ increasing from 20 to 120 also show that P-$WMLF/GS$ is the most accurate algorithm in haplotype reconstruction rate. In Figure 8, the left $Y$-axis shows the reconstruction rate of haplotypes, and the right $Y$-axis shows the running time. Figure 8a shows the experiment results on the real haplotype data, and Figure 8b shows the experiment results on the simulated haplotype data. When $n$ increases, the haplotype reconstruction rate of the three algorithms decreases, and their running time increases accordingly. In Figure 8a, when $n=20$, the haplotype reconstruction rates of $P$-$WMLF/GS$, $GA$-$MECGI$ and $WMLF$ are 96.6, 93.0 and 84.9%, respectively; and their running time are 1.3, 0.4 and 0.0007 s. When $n$ increases to 120, their haplotype reconstruction rates decrease to 93.5, 88.1 and 79.8%; and their running time increases to 10.2, 6.5 and 0.008 s. The experiment results in Figure 8b are similar. Although $P$-$WMLF/GS$ is the slowest among the three

assay is about 1% (Xiao et al., 2007), we plant some errors in the genotype (Zhu, 2006) by changing the value of every char of the genotype with a probability $e_g$ as follows: if the value is ‘0’ or ‘1’, it is changed to ‘2’; if the value is ‘2’, it is changed to ‘0’ or ‘1’ randomly. The GenoSpectrum $F$ is generated as follows. For the column $j$ at which the genotype is $k$: when $k$ is correct, if $k=0$ (or 1), $f_{0,j}$ (or $f_{1,j}$) is normally distributed with mean $\mu=0.9$ and variance $\sigma^2=0.05$, $f_{0,j}$ (or $f_{1,j}$) = 0, $f_{2,j}=\max(0,1-f_{0,j})$ (or $\max(0,1-f_{1,j})$); if $k=2$, $f_{2,j}$ is normally distributed with mean $\mu=0.9$ and variance $\sigma^2=0.05$, and $f_{0,j}$ and $f_{1,j}$ are normally distributed with mean $\mu=0.05$ and variance $\sigma^2=0.05$; when $k$ is error, suppose the correct genotype at column $j$ is $i$, then $f_{0,j}$ and $f_{1,j}$ are normally distributed with variance 0.05 and means 0.8 and 0.2, respectively. For the remaining element of $F$ at column $j$, let it be $\max(0,1-f_{0,j})-f_{2,j}$.

We ran our experiments on a Linux server (4 Intel Xeon 3.6GHz CPU and 4GByte RAM) with the length of haplotype $n$, the number of fragments $m$ ($m=2n+e$/c(IMax+4Min)), the reading error probability $e_r$ and the genotype error $e_g$ varied. The data of the following tables and figures are the average over 100 repeated experiments.

To select a appropriate weighted coefficient $g_a$, we change $g_a$ from 0 to 6.5, and examine the haplotype reconstruction rate of $P$-$WMLF/GS$ on the simulated data with $n=80$, $e_r$ and $e_g$ be randomly selected from 0 to 5%. The experiment result is illustrated by Figure 7. Figure 7 shows that when $g_a=2.5$, i.e. 1/4 of $e_r$, $P$-$WLMF/GS$ achieves the highest haplotype reconstruction rate. In the following experiments, we set $g_a=2.5$.

When $e_r$ varies from 3 to 7% and $e_g$ varies from 0 to 5%, we test the three algorithms on both the real haplotype data and the simulated haplotype data with $n=100$ and $n=200$. The experiment results are presented in Table 1, in which the experiment results on the real haplotype data are put outside brackets, and the experiment results on the simulated haplotype data are enclosed in brackets. As shown in Table 1, the haplotype reconstruction rate of WMLF/GS and MEC/GI are decreasing with the increasing of $e_r$ and $e_g$. Since WMLF does not consider the genotype information, its haplotype reconstruction rate is not related with $e_g$, but is decreasing with the increasing of $e_r$. Although there are errors planted in the genotype information, the false-genotype alleles are much less than the true-genotype alleles, and the positive impact suppresses the negative impact of the genotype information.
WMLF/GS problem. With the fragment maximum length proposed a parameterized algorithm P-WMLF/GS to solve the maximum number of fragments covering a SNP site is small (usually proves it to be NP-hard. Based on the fact that the maximum sequencer provides and the genotyping uncertainty, the current variations of MEC. By including the confidence levels a DNA individual haplotyping problem are MFR, MSR, MEC and the forensic applications. The typical computational models of genetics such as locating of genes, designing of drugs and microarrays.

5 CONCLUSION

Haplotyping plays an increasingly important role in some regions of genetics such as locating of genes, designing of drugs and forensic applications. The typical computational models of individual haplotyping problem are MFR, MSR, MEC and the variations of MEC. By including the confidence levels a DNA sequencer provides and the genotyping uncertainty, the current article proposes a new computational model WMLF/GS, and proves it to be NP-hard. Based on the fact that the maximum number of fragments covering a SNP site is small (usually no more than 19, Huson et al. (2001)), the current article proposed a parameterized algorithm P-WMLF/GS to solve the WMLF/GS problem. With the fragment maximum length $k_1$ and the maximum number $k_2$ of fragments covering a SNP site, the P-WMLF/GS algorithm can solve the WMLF/GS problem in time $O(nk_2^{2k_2} + m_{k_1} + nk_3)$ and in space $O(k_2^{2k_2} + nk_2)$. Extensive experiments show that WMLF/GS has higher haplotype reconstruction rate than other models, and that P-WMLF/GS algorithm is practical.

ACKNOWLEDGEMENTS

The authors are very grateful to the anonymous referees for comments and help in improving the presentation of the earlier version of the paper.

Funding: This research was supported in part by the National Natural Science Foundation of China under Grant Nos. 60433020 and 60773111, the National Basic Research Program of China 2008CB317107, the Program for New Century Excellent Talents in University No. NCET-05-0683, the Program for Changjiang Scholars and Innovative Research Team in University No. IRT0661 and the Scientific Research Fund of Hunan Provincial Education Department under Grant No. 06C526.

Conflict of Interest: none declared.

REFERENCES


