Journal of Asia-Pacific Entomology 15 (2012) 375–381

Contents lists available at SciVerse ScienceDirect

Journal of Asia-Pacific Entomology

journal homepage: www.elsevier.com/locate/jape

Individual haplotype assembly of *Apis mellifera* (honeybee) using a practical branch and bound algorithm

Hyeong-Seok Lim [a], In-Seon Jeong [b], Seung-Ho Kang [c,*]

[a] *Department of Computer Science, Chonnam National University, 77 Yongbong-ro, Buk-gu, Gwangju 500-757, Republic of Korea*
[b] *Department of Agricultural Bio-resources, National Academy of Agricultural Science, Rural Development Administration, 150 Suin-ro, Seodun-dong Gwonseon-gu, Suwon 441-855, Republic of Korea*
[c] *Division of Fusion Convergence of Mathematical Sciences, National Institute for Mathematical Sciences, KT Daeduk 2 Research Center, 463-1 Jeonmin-dong, Yuseong-gu, Daejeon 305-811, Republic of Korea*

ARTICLE INFO

Article history:
Received 2 November 2011
Revised 8 May 2012
Accepted 14 May 2012

Keywords:
Apis mellifera (honeybee)
Single nucleotide polymorphism (SNP)
Haplotype assembly
Local search algorithm
Branch and bound algorithm

ABSTRACT

A haplotype is a single nucleotide polymorphism (SNP) sequence and a representative genetic marker describing the diversity of biological organs. Haplotypes have a wide range of applications such as pharmacology and medical applications. In particular, as a highly social species, haplotypes of the *Apis mellifera* (honeybee) benefit human health and medicine in diverse areas, including venom toxicology, infectious disease, and allergic disease. For this reason, assembling a pair of haplotypes from individual SNP fragments drives research and generates various computational models for this problem. The minimum error correction (MEC) model is an important computational model for an individual haplotype assembly problem. However, the MEC model has been proved to be NP-hard; therefore, no efficient algorithm is available to address this problem. In this study, we propose an improved version of a branch and bound algorithm that can assemble a pair of haplotypes with an optimal solution from SNP fragments of a honeybee specimen in practical time bound. First, we designed a local search algorithm to calculate the good initial upper bound of feasible solutions for enhancing the efficiency of the branch and bound algorithm. Furthermore, to accelerate the speed of the algorithm, we made use of the recursive property of the bounding function together with a lookup table. After conducting extensive experiments over honeybee SNP data released by the Human Genome Sequencing Center, we showed that our method is highly accurate and efficient for assembling haplotypes.

Introduction

The genomes of various insects such as the fruit fly (Adams et al., 2000) and malaria mosquito (Holt et al., 2002) have been sequenced. Among them the genome of the highly social species, honey bee (*Apis mellifera*) has important meaning to another highly social species, *Homo sapiens*. With the complete genome sequences for honey bee now available (Beye and Moritz, 1995; Weinstock et al., 2006), investigating genetic differences is one of the main topics in insect genomics (Zayed and Whitfield, 2008). Single nucleotide polymorphisms (SNP) are believed to be the most frequent form of genetic variability, and understanding SNPs will help treat human disease and contribute to agricultural production (Schumacher and Egen, 1995; Morse and Calderone, 2000; Qin et al., 2006). Furthermore, understanding SNPs will deliver some important answers to the question about differences in social behaviors of humans at the molecular level by experimental molecular analysis (Breed and Rogers, 1991; Robinson et al., 1997; Krieger and Ross, 2002).

The sequence of SNPs in a specific chromosome is called a haplotype. In diploid organisms such as queen or worker bees, genomes are organized into pairs of chromosomes, so there are two copies of the haplotypes for each of the SNP sequences. The haplotype plays a more important role than the genotype, which is the conflation of two haplotypes on homologous chromosomes, in genetic association studies (Liu et al., 2005; Chen et al., 2010; Hussin et al., 2010). Although current sequencing techniques can detect the presence of SNP sites, they cannot tell which copy of a pair of chromosomes the alleles belong to. Hence, it is more difficult to determine haplotypes than to determine genotypes.

Haplotype assembly is the process of determining a pair of haplotypes, one for each copy of a given chromosome that provides information for an individual. This process requires cloning and sequencing of chromosomes as a preprocess (Fig. 1). However, due to the current restrictive DNA sequencing techniques, only individual fragments or pairs of fragments that unavoidably contain sequencing errors are obtained. Furthermore, these fragments may come from both copies of a pair of chromosomes and it is difficult to associate

[*] Corresponding author. Tel.: +82 42 717 5738; fax: +82 42 717 5769.
E-mail address: shkang@nims.re.kr (S.-H. Kang).

1226-8615/$ – see front matter © Korean Society of Applied Entomology, Taiwan Entomological Society and Malaysian Plant Protection Society, 2012. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.aspen.2012.05.012

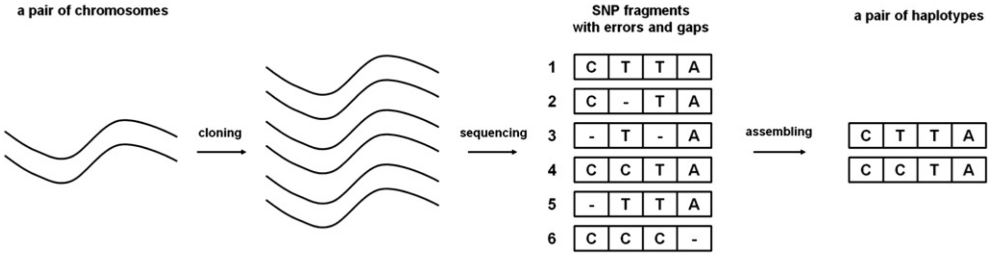

Fig. 1. The process of individual haplotype assembly.

them with the copy that they belong to. Therefore, several models and algorithms for each model have been proposed to overcome this problem (Lippert et al., 2002; Zhang et al., 2006a, 2006b).

First introduced by Lancia et al. (2001), depending on the types of errors and constraints, several different models have been proposed to address the haplotype assembly problem (also called the individual haplotyping problem) such as minimum fragment removal (MFR), minimum SNP removal (MSR) and minimum error correction (MEC). Among these models, the minimum error correction (also called minimum letter flips) model is considered important due to its practical assumptions in environments where errors occur (Wang et al., 2005). Variant models that add some other constraints and information to the MEC model have been subsequently suggested (Zhao et al., 2005; Zhang et al., 2006a, 2006b; Kang et al., 2008).

However, the haplotype assembly problem with MEC as the object function is NP-hard even for the gapless (Zhang et al., 2006a, 2006b). This means that there is no polynomial time algorithm for this problem guaranteeing an optimal solution. Many heuristic approaches have been proposed to overcome the difficulties of the MEC problem (Panconesi and Sozio, 2004; Bansal et al., 2008; Genovese et al., 2008; Wu et al., 2009a, 2009b; Geraci, 2010). However, although they achieve reasonably good results, they cannot guarantee optimal solutions.

A branch and bound algorithm and a fixed parameterized algorithm have been proposed to guarantee an optimal solution for the MEC problem. Wang et al. (2005) designed a naive branch and bound algorithm. The running time of this algorithm largely depends on the number of fragments. As another exact approach, Xie et al. (2008) proposed a fixed parameterized algorithm (called K-Mec) to solve the MEC problem. Although the K-Mec algorithm avoids dependency on the number of fragments, the running time of K-Mec rapidly increases when the number of SNP sites becomes large. In addition, it does not allow for insertion of gaps inside the fragments or insertion of mate pair sequences (Geraci, 2010). Mate pair sequences are generated by modern sequencing technologies such as shotgun sequencing, and they usually span a long fragment, which can be as long as a few hundred positions (Jeong et al., 2010). Thus, K-Mec is impractical when mate pair sequences are considered. In contrast, the branch and bound algorithm does not require any conditions to solve the MEC problem such as those of the fixed parameterized algorithm.

We observed that a simple local search algorithm produces near optimal solutions. Furthermore, we found that the bounding function hinted by Wang et al. (2005), which was originated by Koontz et al. (1975), has a recursive property. From these observations, we improved the branch and bound algorithm using the local search algorithm as an initial upper bound and by exploiting the recursive property of the bounding function. These modifications allowed the branch and bound algorithm to tackle the MEC problem in practical time without assuming special conditions.

In this study, we tested the validity of our method by simulation with honeybee haplotype data released by the Human Genome Sequencing Center (HGSC, 2006). Our results showed that the proposed algorithm could be used as an algorithm for real applications such as the HapAssembler due to its efficiency and accuracy (Kang et al., 2010).

Materials and methods

SNP fragments

We performed haplotype assembly simulations using the honeybee SNP data from the Human Genome Sequence Center (HGSC, 2006). These data sets consist of 16 haplotypes obtained from 16 linkage groups. Sequencing reads were generated from random Africanized honeybee shotgun libraries mapped to genome assembly. We generated 16 pairs of haplotypes with 100 SNP sites from 16 haplotypes using the first 100 SNPs with 30% of heterozygous sites.

Actual SNP fragment data are not available to the public domain. Therefore, we generated simulated fragments with errors and gaps from the obtained haplotypes in the same way as the previous studies (Wang et al., 2005; Zhao et al., 2005; Kang et al., 2008, 2010). Every test for a pair of haplotypes consisted of 30 SNP fragments, each of which was generated by randomly copying either of the two seed haplotypes or indicating it as missing according to the parameters. The gaps in every SNP fragment were randomly produced at a missing rate (*GR*), 0.1, 0.3, and 0.5, respectively. The SNP error in a correct SNP fragment was simulated by turning the nucleotide type at a certain site into another nucleotide type at error rates (*ER*) from 0.1 to 0.3. In this way, we generated 10 examples for each of the 16 pairs of haplotypes according to the parameters.

MEC model

An individual diploid organism possesses a pair of haplotypes from a pair of chromosomes for the given sequences of SNPs. In general, each SNP site shows variability of only two possible alleles: wild type (denoted by 0) and mutant type (denoted by 1); hence, a haplotype can be represented as a string over {0, 1, -}. The symbol "-" represents a SNP site in which the sequencing machine failed to read the allele. If a SNP site has the same allele on both haplotypes, then it is called a homozygous site; otherwise, it is called a heterozygous site. Suppose that there are m SNP fragments coming from a pair of chromosomes and that the corresponding haplotype length is n. We represent a set of SNP fragments as an $m \times n$ matrix M as shown in Fig. 2A, whose every entry f_{ij} has values 0, 1, or -. Each row corresponds to a SNP fragment and each column corresponds to a SNP site.

If we define the distance between two SNP sites as

$$d\left(f_{ik}, f_{jk}\right) = \begin{cases} 1, \text{if} f_{ik} \neq -, f_{jk} \neq -, \text{ and} f_{ik} \neq f_{jk} \\ 0, \text{otherwise} \end{cases}$$