Haplotype Reconstruction In Half-Sib Pedigree Using A Combination Of EM Algorithm And Rule-Based Methods

Ding Xiang Dong [*], Zhang Qin[*] and H. Simianer[†]

Introduction

Recently, haplotype based inference has become highly relevant in studies of evolution, gene mapping and other applications in statistical genetics (Hirschhorn and Daly 2005; The International HapMap 2005). Considerable research effort has been devoted to algorithms that infer haplotype phase from genotype data. Due to the explosion of the numbers of genotyped markers in genetic studies in humans (The International HapMap 2005), experimental animals (Blake *et al* 2003), and farm animals (Fadiel *et al* 2005), the availability of accurate and computationally efficient haplotyping algorithms is of high relevance.

Large half-sib families are widely occurring in many species like cattle, sheep, poultry, fish, and many lab animals, where the common parent of a family may or may not be genotyped. Several programs are available to perform haplotype reconstruction in half-sib design, such as MRH (Qian and Beckman 2002), PedPhase (Li and Jiang 2003), GENEHUNTER (Kruglyak *et al* 1996), SimWalk2 (Sobel and Lange 1996), and others. All approaches have their own strengths and weaknesses. In this paper, we present an approach using a statistical approach based on the maximum likelihood principle expectation maximization (EM) algorithm jointly with rule-based method for haplotype reconstruction in half-sib pedigree. Simulations are provided to indicate that our new approach is more robust and accurate compared to the established methods.

Methods

Definitions. A haplotype is defined as the ordered series of alleles on one of the homologous chromosomes of one individual, The diplotype is defined as a particular combination of two haplotypes. A half-sib family with m offspring can be decomposed into m parent-offspring pairs. For one parent-offspring pair, the offspring inherits a haplotype from the common parent and they will share at least one common haplotype, assuming that there are no recombination events. Thus, one haplotype combination of this parent-offspring pair consists of the haplotype common to both parent and offspring (H_C) together with haplotypes unique to the parent (H_P) and to the offspring (H_O). One haplotype combination of a parent-offspring

[*] Key Laboratory for Animal Breeding and Genetics of Ministry of Agriculture of China, College of Animal Science and Technology, China Agricultural University, Beijing, 100094, China

[†] University of Goettingen, Department of Animal Science, Goettingen, 37075, Germany

pair can be represented as *(H_C,H_P,H_O)*, and is termed as parent-offspring haplotype trio (POHT). For one parent-offspring pair, there should be several possible parent-offspring haplotype trios.

The general idea of our approach can be decomposed into three steps: (1) the missing genotype of common parents at each locus is inferred using population and offspring's information in accordance with Bayes' theorem; (2) under the assumption of no recombination events between tightly linked markers, a maximum likelihood method via EM algorithm is proposed to obtain the possible diplotypes of all parent-offspring pairs in the families; (3) a rule-based approach is then implemented to find the minimum recombinant haplotype configuration of common parents, where the recombination events among the considered loci are taken into account. The minimum recombinant haplotype configurations will be considered as the final diplotypes of common parents. The general ideal of our approach is illustrated in Figure 1.

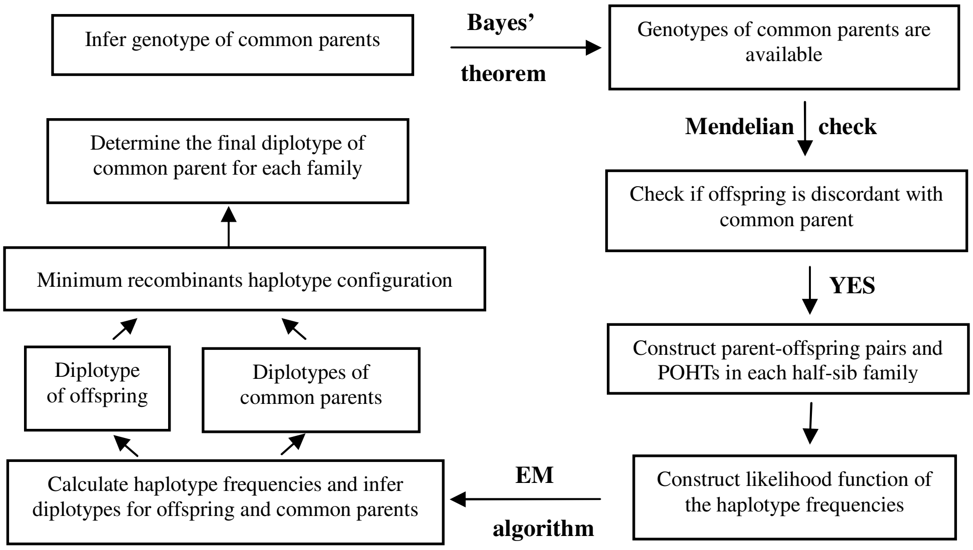
Infer genotype of common parents
Bayes'
theorem
Genotypes of common parents are available
Mendelian
check
Check if offspring is discordant with common parent
YES
Construct parent-offspring pairs and POHTs in each half-sib family
Construct likelihood function of the haplotype frequencies
EM
algorithm
Calculate haplotype frequencies and infer diplotypes for offspring and common parents
Diplotype of offspring
Diplotypes of common parents
Minimum recombinants haplotype configuration
Determine the final diplotype of common parent for each family

Figure 1: A flow-chart defining the algorithm

Simulation study

Following the arguments as presented by Weller *et al.* (Weller *et al.* 1990). We carried out a series of simulation studies to evaluate our approach, termed HSHAP, with the widely used programs PedPhase (Li and Jiang 2003) and Simwalk2 (Sobel and Lange 1996).

Error rate are used as criteria to evaluate the efficiency of haplotype reconstruction of our approach. An individual will be considered as correctly haplotyped if its most likely diplotype is the same as the simulated true diplotype. The error rate is the proportion of not correctly haplotyped individuals in the population, and is given separately for common parents and offspring.

Results and Discussion

The results of our comparisons with respect to the performance of HSHAP, PedPhase and SimWalk2 are shown in Table 1. Our method (HSHAP) performs significantly better than PedPhase and SimWalk2. HSHAP produces the lowest error rate in the haplotype reconstruction of common parents and offspring. As mentioned by Li and Jiang (2003), PedPhase only has simple missing allele imputation capabilities, it can not completely infer the common parent's genotypes for all loci at all as HSHAP and SimWalk2 did, which deduce the error rate in common parent from PedPhase dramatically higher, as shown in Table 1.

Table 1: The comparison of our approach HSHAP with PedPhase and SimWalk2 in half-sib design*

	HSHAP	PedPhase	SimWalk2
Error rate in common parents	0.007^{c}	0.682^{a}	0.126^{b}
Error rate in offspring	0.021^{c}	0.0330^{b}	0.0705^{a}
Running time (m)	8.211	0.006	60.17

* 10 half-sib families with 30 offspring each, 10 SNPs with equal marker distance of 0.5 centiMorgan are simulated, 100 replicates are generated and analyzed.

[a,b,c] Means with different superscripts within same scenario differ significantly at the $\alpha<0.01$ significance level

Compared to PedPhase and HSHAP, SimWalk2 runs quite slowly as shown in Table 1. For large scale haplotyping projects based on a half-sib design structure this characteristic of Simwalk2 may quickly become prohibitive. PedPhase runs fastest, even when dealing with large numbers of SNPs, which is a typical advantage for most of rule-based methods. Although our approach does not run as fast as PedPhase, it can complete the haplotype reconstruction in a few minutes.

HSHAP performs better than SimWalk2 and PedPhase as well (Table 2), HSHAP can infer the genotypes at all the missing loci. While PedPhase and SimWalk2 are severely affected by missing alleles, the genotypes at the missing locus can not be given, resulting in the dramatically high error rate in offspring and further reducing the accuracy of haplotype reconstruction in common parent.

Table 2: The comparison of our approach with PedPhase and SimWalk2 in the scenario of 30% individuals with one random missing locus*

	HSHAP	PedPhase	SimWalk2
Error rate in common parents	0.0310^{c}	0.8910^{a}	0.0930^{b}
Error rate in offspring	0.0159^{c}	0.7607^{a}	0.2164^{b}
Running time (m)	0.01	0.002	10.65

* 5SNPs with equal marker distance of 0.5 centiMorgan and 10 half-sib families with 20 offspring each are simulated.

The program is available on request from the authors (xding@cau.edu.cn).

Acknowledgements

This research was supported by the Functional Genome Analysis in Animal Organisms (FUGATO) program， the Foerderverein Biotechnologieforschung e.V. (FBF), National Natural Science Foundation of China (30800776), State High-Tech Development Plan (2008AA101002), the National Key Basic Research Program of China (2006CB102104).

References

Blake, J.A., Richardson, J.E., Bult, C.J. et al. (2003). *Nucleic Acids Res* 31: 193–195.

Fadiel, A., Anidi, I. and D. EK (2005). *Nucleic Acids Res* 33: 6308–6318.

Hirschhorn JN, and Daly MJ (2005). *Nature Rev Genet* 6: 95-108

The International HapMap (2005). *Nature* 437: 1299-1320.

Kruglyak, L., Daly, M., Reeve-Daly, M. et al. (1996). *Am J Hum Genet* 58: 1347–1363.

Li, J., and Jiang, T. (2003). *J Bioinform Comput Biol* 1: 41–69.

Qian, D., and Beckman, L. (2002). *Am J Hum Genet* 70: 1434–1445.

Sobel, E., and Lange, K. (1996). *Am J Hum Genet* 58: 1323-1337.

Weller, J.I., Kashi Y. And Soller,M. (1990) J Dairy Sci 73:2525-2537.