The Thirteenth International Symposium on Artificial Life and Robotics 2008(AROB 13th '08),
B-Con Plaza, Beppu, Oita, Japan, January 31-February 2, 2008

Haplotype Estimation from Genotypical Data by Genetic Algorithm

Ryuichi Azuma
Faculty of Engineering
University of Miyazaki
Miyazaki City, 889-2192

Makoto Sakamoto, Hiroshi Furutani
Faculty of Engineering
University of Miyazaki
Miyazaki City, 889-2192

Abstract

The study of a disease using genetic identification becomes possible by using haplotype information. The Expectation-Maximization algorithms are the standard approach in the haplotype analysis. These approaches maximize the likelihood function of genotypic distribution assuming Hardy-Weinberg equilibrium. However, these methods are time consuming when applied to sequence of many loci. In this study, we used a genetic algorithm to obtain the haplotype frequencies from frequencies of genotypes.

1 Introduction

The DNA sequence of the human genome was almost decoded in 2003, and the examination of genetic arrangements and the relation between DNA sequences and properties of proteins became the main interest of recent studies. For example, it becomes possible to make an identification of the disease gene from DNA sequence data of patients. We compare genes from a physically normal person with genes from a patient owning a certain disease, and are able to find a gene related to the disease. In this kind of analysis, haplotype information is an essential factor for fine-scale molecular genetic research. When it was assumed that a specific disease was caused by ancestral mutation, base sequence of the neighborhood of disease gene may resemble that of the people having this disease. Thus the study of a disease using genetic identification becomes possible by using haplotype information. Haplotype information can be obtained by special experiments at considerable cost, or by using information of additional family members. Therefore these methods restrict the applicability of haplotype information. Some researchers developed other methods which use varieties of statistical techniques. Clark's algorithm [1] can be viewed as an attempt to minimize the total number of haplotypes observed in the sample. The Expectation-Maximization (EM) algorithm is the approach to maximize the likelihood function of genotypic distribution assuming Hardy-Weinberg equilibrium. However, these methods are time consuming when applied to sequence of many loci. In this study we performed the haplotype estimation by applying genetic algorithm (GA) to the artificial data. We used a GA of very simple form, and checked the precision of the results obtained by this method. As an artificial chromosome, we consider the genes corresponding to the relative frequencies of assumed haplotypes.

2 Methods

2.1 Haplotype

Homo sapiens has two chromosomes. A genetic constitution on one chromosome is called as haplotype. Haplotype does not change as far as there is no recombination. If there is recombination, new haplotypes is formed, and it is conveyed to the next generation by a gamete and does not change till there occurs a next recombination.

Haplotypes of a gamete conveyed to the next generation depends on gametes of parents and a probability of occurring recombination between loci. The ratio of occurring recombination is called a recombination ratio. Therefore the heredity of haplotype is regarded as a stochastic phenomenon determined by a recombination ratio.

It is important to get haplotypes information for genetic analysis. However, haplotype information is very difficult to obatins both experimentally and theoretically. Because we cannot distinguish alleles from a locus of two chromosomes which are in the mixed state, even if all the genotypes of the individual are observed experimentally.

Haplotypes of the individual has a case to be decided if we can obtain the genotype of the relative. If there is no other information, we assume Hardy-Weinberg equilibrium and can estimate the haplotype

©ISAROB 2008

frequency of the population by means of EM algorithm. In this study, we take other approach of statistical haplotype estimation by applying GA of simple structure.

2.2 EM Algorithm

We introduce here the steps of EM algorithm to help understanding of the estimation process.

Let $\mathbf{g}_i = \{\mathbf{g}_1, \dots, \mathbf{g}_r\}$ denote the observed genotypes and $f_{\mathbf{g}_i}$ denote frequencies of genotype $\mathbf{g}_i$. When the population size is $\mathbf{N}$, let $n_{\mathbf{g}_i}$ denote the number of $\mathbf{g}_i$. Namely,

$$\mathrm{N} f_{\mathbf{g}_i} = n_{\mathbf{g}_i}, \qquad \mathrm{N} = \sum_{i=1}^{r} n_{\mathbf{g}_i}.$$

Let $\mathbf{g}_{ij}$ denote the locus j of genotype $\mathbf{g}_i$. Let $\mathbf{h}\ _j$ denote locus j of haplotype $\mathbf{h}\ = \{\mathbf{h}_1, \dots, \mathbf{h}_s\}$ likewise. The u stands for the number of loci, and the locus j takes values of $1, \dots, u$. When an allele has the possibility to take two different values at each locus, the number of such loci is u_0. Let s denote the number of haplotypes with the possibility to obtain from observed genotypes, $s = 2^{u_0}$.

We set the repetition time of the EM algorithm $k = 0$. Initially we assume haplotype frequencies

$$f_{\mathbf{h}_\ell}^{(0)} = \frac{1}{s}.$$

Under the assumption of Hardy-Weinberg equilibrium, we derive the haplotype-likelihood

$$\hat{\mathrm{P}}^{(k)}(\mathbf{g}_i|\mathbf{h}_x, \mathbf{h}_y) = 2^c \prod f_{\mathbf{h}_x}^{(k)} f_{\mathbf{h}_y}^{(k)} \tag{1}$$

When $\mathbf{h}_x$ and $\mathbf{h}_y$ are different, $c = 1$. When $\mathbf{h}_x$ and $\mathbf{h}_y$ are identical, $c = 0$.

Let L denote the probability that the number of $\mathbf{g}_i$ is $n_{\mathbf{g}_i}$,

$$\mathrm{L}^{(k)} = \sum_{i=1}^{r} n_{\mathbf{g}_i} \hat{\mathrm{P}}^{(k)}(\mathbf{g}_i|\mathbf{h}_x, \mathbf{h}_y) \tag{2}$$

The form of the logarithm likelihood can express L with the polynomial distribution[4],

$$\log \mathrm{L}^{(k)} = \sum_{i=1}^{r} n_{\mathbf{g}_i} \log \hat{\mathrm{P}}^{(k)}(\mathbf{g}_i|\mathbf{h}_x, \mathbf{h}_y) \tag{3}$$

We treat a problem to maximize this $\log \mathrm{L}^{(k)}$. Expectation $\hat{\mathrm{E}}^{(k)}(n_{\mathbf{h}_i})$ of haplotype is

$$\hat{\mathrm{E}}^{(k)}(n_{\mathbf{h}_\ell}) = \sum_{i=1}^{r} n_{\mathbf{g}_i} \hat{\mathrm{P}}^{(k)}\left(\mathbf{h}_x, \mathbf{h}_y|\mathbf{g}_i, \mathbf{g}_i; (\mathbf{h}_x, \mathbf{h}_y)\right) \tag{4}$$

Then, we estimate it by

$$\hat{\mathrm{P}}^{(k)}\left(\mathbf{h}_x, \mathbf{h}_y|\mathbf{g}_i, \mathbf{g}_i; (\mathbf{h}_x, \mathbf{h}_y)\right) = \frac{2 f_{\mathbf{h}_x}^{(k)} f_{\mathbf{h}_y}^{(k)}}{\hat{\mathrm{P}}^{(k)}(\mathbf{g}_i|\mathbf{h}_x, \mathbf{h}_y)}. \tag{5}$$

Updating the value $k \to k+1$, we have the haplotype frequency

$$f_{\mathbf{h}_\ell}^{(k+1)} = \frac{n_{\mathbf{h}_\ell}^{(k)}}{2\mathrm{N}}. \tag{6}$$

When this value converges, the estimate of the haplotype frequency is $f_{\mathbf{h}_\ell}^{(k)}$[2, 5].

2.3 Genetic Algorithm

Instead of using EM algorithm, we adopt the GA approach. In this study, we use an artificial data of haplotype frequencies shown in table 1.

Number	Haplotype	Population
1	GGG	50
2	GGA	6
3	GAG	13
4	GAA	2
5	AGG	22
6	AGA	2
7	AAG	4
8	AAA	1
Total		100

Table 1: Haplotypes and the population.

In this example, the number of loci is 3 and there are two alleles 'G' and 'A'. Therefore there are eight haplotypes.

The first step is to construct genotypical data from these haplotypes. From these eight haplotypes, we reorder this haplotype population randomly by using random numbers, and make 50 haplotype pairs.

Table 2 shows the frequencies of genotypical data obtained from Table 1. The genotype is shown by (GG, GA, GA). These are genotypes constructed from the haplotypes. The first genotype (GG, GG, GG) means that the sequence has only one haplotype 'GGG'.

The third genotype (GA, GG, GA) has a quite different property. From this genotype, we cannot predict the haplotype pair uniquely. In other words, two haplotype pairs ('GGG', 'AGA') and ('GGA', 'AGG') reproduce the same genotype (GA, GG, GA).

Genotype	Population
(GG, GG, GG)	11
(GA, GG, GG)	14
(GA, GG, GA)	5
(GG, GA, GG)	7
(AA, GG, GG)	2
(GA, GA, GG)	4
(GG, AA, GG)	2
(GG, GA, GA)	1
(GG, GG, GA)	1
(GG, GG, AA)	1
(GG, AA, GA)	1
(AA, AA, GA)	1
Total	50

Table 2: A genotype and the population which it was demanded from by haplotype.

In the second step, we estimate haplotype frequencies from these genotype data by means of GA. The ambiguity of haplotypes appears in the third (GA, GG, GA) with $n = 5$, the sixth (GA, GA, GG) with $n = 4$, and (GG, GA, GA) with $n = 1$. In this study, we consider the haplotype pairs of having ambiguity. Genotype (GA, GG, GA) may get haplotype pairs ('GGG', 'AGA') and ('GGA', 'AGG'). We consider the gene of the length $\ell = 10 = 5 + 4 + 1$. The locus of this gene takes a value of 0 and 1. In the case of 0, it is assumed that haplotype pair gets 'GGG', 'AGA', and, in the case of 1, it is assumed that haplotype pair gets 'GGA', 'AGG'.

This step is illustrated in Figure 1 schematically. All haplotype frequencies can be obtained from this gene.

We find genotype frequencies from these haplotype frequencies assuming Hardy-Weinberg equilibrium. We subtract estimated genotype frequencies and observed genotype frequencies in each genotype. We add them and add 1 and we make it a reciprocal number and use it as the fitness of a population. Let F_m denote fitness of mth population,

$$\mathrm{F}_m = \frac{1}{\sum_{i=1}^{r}(\mathrm{P}_{\mathbf{g_i}} \quad \hat{\mathrm{P}}_{\mathbf{g_i}})^2 + 1},\ m = 1, \ldots, v. \qquad (7)$$

3 Results

In this study, we set population size of $N = 20$. The crossover rate is 0.5, and the mutation rate is 0.05. Table 3 shows the list of calculated haplotype frequencies. The top shows the true frequency. We notice that there is a solution completely identical to the true solution. However, the fitness value of this solution is low, which suggests the stochastic nature of the present problem.

	Haplotype Number								Fitness
	1	2	3	4	5	6	7	8	
OHF	50	6	13	2	22	2	4	1	
EHF	46	8	15	2	26	0	2	1	0.03501
	49	6	14	2	23	2	3	1	0.03774
	48	6	15	2	24	2	2	1	0.04218
	50	4	15	2	22	4	2	1	0.04223
	48	6	15	2	24	2	2	1	0.04218
	48	6	15	2	24	2	2	1	0.04218
	49	5	15	2	23	3	2	1	0.04341
	48	6	15	2	24	2	2	1	0.04218
	50	4	15	2	22	4	2	1	0.04223
	48	6	15	2	24	2	2	1	0.04218
	50	5	14	2	22	3	3	1	0.03796
	48	6	15	2	24	2	2	1	0.04218
	48	7	14	2	24	1	3	1	0.03585
	48	6	15	2	24	2	2	1	0.04218
	47	7	15	2	25	1	2	1	0.03909
	50	4	15	2	22	4	2	1	0.04223
	50	6	13	2	22	2	4	1	0.03267
	49	6	14	2	23	2	3	1	0.03774
	48	6	15	2	24	2	2	1	0.04218
	49	4	16	2	23	4	1	1	0.04731

Table 3: The genotypes and the population which were obtained from haplotype data. OHF is Observed haplotype frequencies. EHF is Estimated haplotype frequencies.

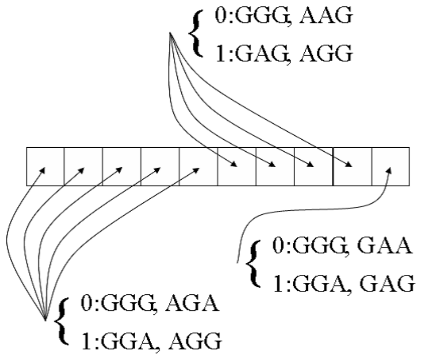

Figure 1: The relations of the value of each locus of the gene and haplotype pairs.

4 Summary

In this study, we treat the genotypical data of length 3 and with only two alleles 'G' and 'A'. Therefore it has comparatively little computational complexity. However, in practical situations, we must estimate haplotypes from the genotypes of many loci. We must consider the genotypes whose data at a specific locus were missing. Then the computational complexity may increases. As for the number of haplotype candidates, there are two kinds of alleles in each locus. When we considered ℓ loci, computational complexity becomes O(2). Therefore computational complexity becomes the vast quantity as the number of loci increases. Actually, the analysis beyond $\ell = 30$ is difficult on the scale of the current computer. Thus we have to treat such problems by improving and mixing existing algorithms including GA proposed here.

We are now studying the more complex data of actual DNA sequences, and the results will be reported in other occasion[1].

References

[1] Clark AG: Inference of haplotypes from PCR-amplified samples of diploid populations. Molecular Biology and Evolution, 7:111–122, 1990.

[2] Connie M. Drysdale et al: Complex promoter and coding region $_2$-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. Proceedings of National Academy of Sciences of the United States of America. 10483–10488, 2006.

[3] Andreas Ziegler and Inke R. König: A Statistical Approach to Genetic Epidemiology. WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim. 243–250, 2006.

[4] Matthew Stephens et al: A New Statistical Method for Haplotype Reconstruction from Population Data. Am. J. Hum. Genet. 68:978–989, 2001.

[5] Tianhua Niu et al: Bayesian Haplotype Inference for Multiple Linked Single-Nucleotide Polymorphisms. Am. J. Hum. Genet. 70:157–169, 2002.

[6] Jian Zhang et al: Haplotype Reconstruction for Diploid Populations. Hum Hered 59:144–156, 2004.