BMC Bioinformatics

Methodology article

Optimal Step Length EM Algorithm (OSLEM) for the estimation of haplotype frequency and its application in lipoprotein lipase genotyping

Peisen Zhang*[1], Huitao Sheng[1], Alfredo Morabia[3] and T Conrad Gilliam[1,2]

Address: [1]Columbia Genome Center, Columbia University, New York NY 10032, USA, [2]Departments of Genetics & Development, and Psychiatry, Columbia University, New York NY 10032, USA and [3]Division d'Epidémiologie Clinique, Hôpitaux Universitaires de Genève, Geneva, Switzerland

Email: Peisen Zhang* - pz6@columbia.edu; Huitao Sheng - hs734@columbia.edu; Alfredo Morabia - Alfredo.Morabia@hcuge.ch; T Conrad Gilliam - tcg1@columbia.edu

* Corresponding author

Published: 15 January 2003

BMC Bioinformatics 2003, **4**:3

Received: 16 September 2002
Accepted: 15 January 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/3

Abstract

Background: Haplotype based linkage disequilibrium (LD) mapping has become a powerful and cost-effective method for performing genetic association studies, particularly in the search for genetic markers in linkage disequilibrium with complex disease loci. Various methods (e.g. Monte-Carlo (Gibbs sampling); EM (expectation maximization); and Clark's method) have been used to estimate haplotype frequencies from routine genotyping data.

Results: These algorithms can be very slow for large number of SNPs. In order to speed them up, we have developed a new algorithm using numerical analysis technology, a so-called optimal step length EM (OSLEM) that accelerates the calculation. By optimizing approximately the step length of the EM algorithm, OSLEM can run at about twice the speed of EM. This algorithm has been used for lipoprotein lipase (LPL) genotyping analysis.

Conclusions: This new optimal step length EM (OSLEM) algorithm can accelerate the calculation for haplotype frequency estimation for genotyping data without pedigree information. An OSLEM on-line server is available, as well as a free downloadable version.

Background

Estimation of haplotype frequencies from routine genotyping data plays an important role in LD analysis, and can be achieved by varied methods, including Monte-Carlo (Gibbs sampling [1], PHASE method [2]), EM [3-5] and Subtraction [6]. At least two studies have tested and compared some of these programs. Xu and his collaborators [7] have empirically evaluated and compared the accuracy of the Subtraction method [6], the expectation-maximization (EM) method, and the PHASE method [2] for estimating haplotype frequency and for predicting haplotype phase. Summarizing from the studies of Xu and his collaborators in [7]: "Where there was near complete linkage disequilibrium (LD) between SNPs (the NAT2 gene), all three methods provided effective and accurate estimates for haplotype frequencies and individual haplotype phases. For a genomic region in which marked LD was not maintained (the chromosome X locus), the computational methods were adequate in estimating overall haplotype frequencies. However, none of the methods were accurate in predicting individual haplotype phases. The EM and the PHASE methods provided better estimates for overall

haplotype frequencies relative to the Subtraction method for both genomic regions." The PHASE algorithm is extremely slow. A comparison of run-times was reported for five SNPs from the NAT2 gene [7]: Subtraction method, 0.01 second; EM method, 13.48 seconds; and PHASE method, over 128 minutes. Zhang and his collaborators [8] pointed out that, " The PHASE method did not yield significantly different results from a simple maximum-likelihood procedure." From the two comparative studies [7,8] and from one simulation study [9], it was shown convincingly that the expectation maximization (EM) algorithm is accurate for estimation of haplotype frequencies.

The EM algorithm is much faster than Monte-Carlo based algorithms. Whereas it is fast with single runs for a relative small number of SNPs, it can be slow with multiple runs and for large number of SNPs. The EM algorithm is an optimization algorithm. In order to obtain a global maximization, EM should run many times from varied starting points. This can be very time consuming. Although in the standard EM algorithm procedure, the step length is the optimal length under the conditional expectation, the step length is not optimal in general. In order to run faster and more accurately, we have developed a new algorithm using numerical analysis technology, a so-called optimal step length EM (OSLEM) to accelerate the calculation. By optimizing approximately the step length of the EM (expectation maximization) algorithm, OSLEM can run at about twice the speed of EM.

Algorithm

We start with the same premises and notations as Stephen et al [2]. Given n diploid individuals from a population, let $G = (G_1,..., G_n)$ denote the (known) genotypes for the individuals, let $H = (H_1,..., H_n)$ denote the (unknown) corresponding haplotype pairs, let $F = (F_1,...,F_M)$ denote the set of (unknown) population haplotype frequencies (the M possible haplotypes are arbitrarily labeled $1,...,M$). Here H_i, a random variable depends on F. Let H_i be the set of all (ordered) haplotype pairs consistent with the multilocus genotype G_i, and suppose the distribution of H_i on H_i will follow the Hardy-Weinberg equilibrium.

The EM and OSLEM algorithms attempt to find the haplotype F that maximizes the likelihood.

$$L(F) = \Pr(G \mid F) = \prod_{i=1}^{n} \Pr(G_i \mid F).$$

Here,

$$\Pr(G_i \mid F) = \sum_{(b_1,b_2)\in \mathsf{H}_i} F_{b_1} F_{b_2},$$

where H_i is the set of all (ordered) haplotype pairs consistent with the multilocus genotype G_i. Note that this likelihood is just the probability of observing the sample genotypes, as a function of the population haplotype frequencies, under the assumption of Hardy-Weinberg equilibrium (HWE).

Before running the iteration, for each genotype, find all possible haplotype pairs that are consistent with the genotype. Given k markers, there are 2^{k-1} possible haplotype pairs per genotype. In our implementation, if the haplotype was already generated, we will create a link to connect the haplotype and the genotype. Otherwise, a new haplotype will be generated and linked to the genotype.

We outline the EM and OSLEM algorithms as follows:

Step1: Obtain an initial distribution for genotype observed to corresponding haplotype pairs. For example, equal distribution is commonly used but random generation is also possible.

Step2: Gene-Counting [11,12] calculating haplotype frequencies from the haplotype pair distribution.

Step3: Recalculate distributions for genotypes by Hardy-Weinberg equilibrium condition

$\mathbf{D_{preN}}$

Step 4: Recalculate distribution by optimal step length:

$\mathbf{D_N = D_{N-1} + \lambda\ (D_{preN} - D_{N-1})}$ **where** $\lambda \geq 1$

Step 5: Go to step 2 until step size becomes less then a given small value (precision).

Where $\mathbf{D_{preN}}$, $\mathbf{D_N}$ and $\mathbf{D_{N-1}}$ are array variables and λ **is** a constant.

The EM algorithm jumps over Step 4, so it always takes λ **= 1**. In order to generate global maximization, the EM (or OSLEM) procedures are usually repeated 100 or more times for different initial distributions. In our implementation, we generate the initial distributions randomly from the first one as the equal distributions.

We use the following procedure to calculate λ in step 4.

Table 1: Single Run Comparisons of OSLEM and EM:

# SNPs	Precision	Number of Loops			CPU time		
		OSLEM	EM	OSLEM: EM	OSLEM	EM	OSLEM: EM
12	1e-9	305	621	0.49	0.45	0.80	0.56
13	1e-7	305	579	0.53	0.95	1.69	0.56
14	1e-7	183	343	0.53	0.78	1.62	0.48
15	1e-7	182	342	0.53	0.87	1.78	0.49
16	1e-9	523	1149	0.45	4.92	11.28	0.44

EM = Expectation maximization algorithm, OSLEM = Optimal Step Length EM algorithm The unit for CPU time is millisecond. The precision is the sum of the absolute differences of the haplotype frequencies between two loops.

Do loop for every ambiguous genotype G_i (genotypes with more than one heterozygous locus),

Do loop for every haplotype pair (b_k, b_l) that belongs to G_i,

If $(D_{preN\,(k,l)} - D_{N-1(k,l)}) < 0$

Calculate the upper bound for $B_{(k,l)} = -D_{N-1(k,l)} / (D_{preN(k,l)} - D_{N-1(k,l)})$

else

$\lambda_{(k,l)} = 1 + C_i/[S_i/(A_k + A_l) - C_i]$

if $\lambda_{(k,l)} < 0$

$\lambda_{(k,l)} = 1$,

(where C_i is the counting number of genotype G_i, S_i is the sum of the products of haplotype counts for all haplotype pairs (b_k, b_l) that belongs to G_i, A_k is the haplotype count for b_k and A_l is the haplotype count for b_l).

After the above two nested loops, calculate the average of $\lambda_{(k,l)}$. If the average is bigger than the minimum of the bounds $B_{(k,l)}$, take the minimum upper bound as λ. We use 2 as an upper bound for λ. If the average is less than 1, we set the average of $\lambda_{(k,l)} = 1$.

This iteration procedure can be viewed as searching for a fixed-point. By trying to solve the fixed-point equation approximately, we obtain an almost optimal step length formula for λ.

Results

By our tests using real data and tailored data, this new algorithm runs about twice as fast and obtains the same results as the EM algorithm. To test the performance, we generated the data in Table 1 for a single run (equal initial distribution) and Table 2 for multiple runs (initial distribution generated randomly). The whole run procedure includes three steps: the input/output step, data manipulation step, and the haplotype frequency estimation step. In the following tables, we only consider the haplotype frequency estimation step. The tailored data is edited from our epidemiological data. The data set is available on our website.

We have applied OSLEM to reconstruct haplotypes for epidemiological data. Lipoprotein lipase (LPL) is a glycoprotein involved in the transformation of dietary lipids into sources of energy for peripheral tissues (e.g., heart, muscle, adipose tissue) [10]. We performed an exhaustive analysis of genotypes and haplotypes spanning the LPL gene in 186 subjects whose blood lipid levels conferred a high risk of atherosclerosis (hereby referred to as "cases") and in 185 controls with non-atherogenic blood lipid profiles. Those subjects, ages 35 to 74, are representative of the general population of Geneva, Switzerland, in 1999 and 2000. Lipoprotein lipase sequence variants were surveyed by first re-sequencing its 10 exons and introns/flanking regions in a selected subgroup of the case-control sample, followed by measurement of the most common SNPs in all cases and controls. Haplotypes were reconstructed from the individual SNPs separately for cases, controls, and the total sample. The relative frequencies of the estimated haplotypes in cases and controls are shown in table 3.

Discussion

By optimizing the step length of the EM (expectation maximization) algorithm, we have developed an accurate and faster algorithm for haplotype frequency estimation. This algorithm has been used successfully for lipoprotein lipase (LPL) genotyping analysis. The genetic analysis of lipoprotein lipase (LPL) gene-variants and their relation to population based variance in lipid profiles is been published separately [13].

The theoretical analysis of global optimization is, in general, a challenging mathematical target. Thus, a rigorous

Table 2: Multiple-Run Comparisons of OSLEM and EM:

Table 2.1 Multiple-Run Maximum λ = 2 Precision: 1e-7 Run: 1000 times

SNPs #	OSLEM	EM	OSLEM : EM
12	3m10.30s	5m4.43s	62.51%
13	8m20.61s	14m33.71s	57.30%
14	11m29.68s	20m38.47s	55.69%
15	12m50.12s	23m38.03	54.31%
16	37m58.03s	1h17m41.10s	48.87%

Table 2.2 Multiple-Run Maximum λ = 2 Precision: 1e-7 Run: 100 times

SNPs #	OSLEM	EM	OSLEM : EM
12	19.59s	28.51s	65.21%
13	54.18s	88.05s	61.53%
14	65.65s	118.96s	55.19%
15	75.65s	130.90s	57.79%
16	228.39s	426.47s	53.55%

Table 3: OSLEM-reconstructed haplotypes:

Haplotype (LPL exons)*		Frequencies within subgroups		
	3 4 5 6 8 9 10	All	Cases	Controls
0:	0 0 000 000 00 00 00	0.4582	0.4933	0.4459
1:	0 0 000 0**v**0 00 00 00	0.1191	0.1224	0.1000
2:	0 0 **vvv v**0**v 0v vv v**0	0.0767	0.0474	0.1077
3:	0 **v** 000 000 **vv v**0 0**v**	0.0389	0.0361	0.0363
4:	0 0 000 000 **vv v**0 0**v**	0.0320	0.0324	0.0271
5:	0 0 000 000 0**v v**0 0**v**	0.0363	0.0279	0.0443
6:	0 0 **vvv v**0**v vv v**0 0**v**	0.0239	0.0250	0.0223
7:	0 0 000 000 0**v vv v**0	0.0240	0.0200	0.0248
8:	**v** 0 000 **v**00 00 00 00	0.0200	0.0186	0.0152
9:	0 0 **vvv v**0**v** 00 00 00	0.0211	0.0172	0.0225
Totals:		0.8502	0.8403	0.8461

On 14 Single Nucleotide Polymorphisms of the lipoprotein lipase gene in 163 atherogenic cases and 157 non-atherogenic controls. Geneva, Switzerland, 1999–2000. * SNPs are divided into subgroups according to their locations on exons, for example the first SNP is located on exon3, the second on exon 4, the third, fourth, and fifth are on exon 5, and so on. In the haplotype sequences, v means minor variation and 0 means major variation.

analysis of the rate of convergence for OSLEM may be quite difficult, but is very important. For this reason, there is a need for both analytic work and further computer simulation work.

It may be the case that there are multiple local maximization points for the mathematical formulation of haplotype frequency estimation. In this case, it would be possible to devise an algorithm to determine the extent of local maximization, which, in turn, would allow one to determine whether the globe optimal solution has been obtained.

We have set up a web server to provide haplotype frequency estimation service. The URL is http://genome3.cpmc.columbia.edu/~genome/HDL/.

Author's contributions

CG and AM conceived the Lipoprotein Lipase Genotyping project. PZ developed the new optimal step length EM

(OSLEM) algorithm. PZ and HS coded the algorithm and developed the web server. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Dr. Joseph Terwilliger, Dr. Hank Juo and Dr. Haghighi for helpful discussion, and Dr Michael C Costanza for performing the application. We would like to thank the reviewers and the editors for their comments and suggestions.

References

1. Niu T, Qin ZS, Xu X and Liu JS **Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms.** *Am J Hum Genet* 2002, **70(1):**157-69
2. Stephens M, Smith NJ and Donnelly P **A new statistical method for haplotype reconstruction from population data.** *Am J Hum Genet* 2001, **68:**978-989
3. Excoffier L and Slatkin M **Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population.** *Mol Biol Evol* 1995, **12:**921-927
4. Hawley M and Kidd K **HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes.** *J Hered* 1995, **86:**409-411
5. Long JC, Williams RC and Urbanek M **An E-M algorithm and testing strategy for multiple locus haplotypes.** *Am J Hum Genet* 1995, **56:**799-8103
6. Clark AG **Inference of haplotypes from PCR-amplified samples of diploid populations.** *Mol Biol Evol* 1990, **7:**111-122
7. Xu CF, Karen Lewis, Kathryn Cantone L, Parveen Khan, Christine Donnelly, Nicola White, Nikki Crocker, Pete Boyd R, Dmitri Zaykin V and Ian Purvis J **Effectiveness of computational methods in haplotype prediction** *Hum Genet* 2002, **110(2):**148-156
8. Zhang S, Andrew Pakstis J, Kenneth Kidd K and Hongyu Zhao **Comparisons of Two Methods for Haplotype Reconstruction and Haplotype Frequency Estimation from Population Data** *Am J Hum Genet* 2001, **69:**906-912
9. Fallin D and Schork NJ **Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data.** *Am J Hum Genet* 2000, **67(4):**947-59
10. Murthy V, Julien P and Gagne C **Molecular pathobiology of the human lipoprotein lipase gene.** *Pharmacol Ther* 1996, **70:**101-135
11. Ceppellini RM, Siniscalco M and Smith CAB **The estimation of gene frequencies in a random mating population.** *Ann Hum Genet* 1955, **20:**97-115
12. Smith CAB **Counting methods in genetical statistics.** *Ann Hum Genet* 1957, **21:**254-276
13. Morabia A, Cayanis E, Costanza MC, Ross BM, Bernstein MS, Flaherty MS, Alvin GB, Das K, Morris MA, Penchaszadeh GK, Zhang P and Gilliam TC **Association between the lipoprotein lipase (LPL) gene and blood lipids: a common variant for a common trait.** *Genetic Epidemiol* 2003,