Cell Research (2009) 19:519-523.

www.nature.com/cr

LETTER TO THE EDITOR

A partition-ligation-combination-subdivision EM algorithm for haplotype inference with multiallelic markers: update of the SHEsis (http://analysis.bio-x.cn)

Cell Research (2009) **19**:519-523. doi: 10.1038/cr.2009.33; published online 17 March 2009

Dear Editor,

Haplotypic information in diploid organisms provides valuable information on human evolutionary history and plays an important role in identifying a candidate gene in the etiology of complex genetic diseases. However, haplotypes of diploid individuals cannot be acquired easily. Molecular haplotyping methods are very costly and have low throughput, and current genotyping and sequencing methods do not provide information on the linkage phase in diploid organisms. The application of statistical methods to infer the haplotype phase in samples of diploid sequences is a very cost-effective approach. Several computational and statistical methods have been developed for haplotype inference, including Clark's algorithm [1], the Expectation Maximization (EM) algorithm [2], and Gibbs sampler [3]. Because of its interpretability and stability, the EM algorithm has become one of the most widely used statistical algorithms. However, the standard EM algorithm has several weaknesses, including the inability to handle a large number of markers and convergence to the local optimum. To overcome these problems, various derivative methods have been developed, such as the Partition-Ligation EM (PLEM) algorithm to handle many more linked loci [4], the Optimal Step Length EM (OSLEM) algorithm to accelerate the calculations [5], and the Stochastic EM (SEM) algorithm to deal with missing genotypic data and to avoid convergence to local maxima [6]. However, most packages are intended for use with single-nucleotide polymorphism (SNP) data in a biallelic manner.

More and more researchers are analyzing both multiallelic and biallelic markers in the linkage and/or association studies of complex diseases. The analysis of linkage disequilibrium (LD) between multiallelic loci and haplotype inference of many loci (including bi- and multiallelic markers) present a number of common problems. The major difficulty for the haplotype inference problem is exploring the large number of haplotype pairs that are consistent with the observed genotypes. One solution is to limit the number of candidate haplotypes. Ideally, a condensed haplotype set would make the process of haplotype inference quick and accurate. Our approach is to construct the haplotype space gradually rather than beginning with the set of all possible haplotypes.

SHEsis [7] (http://analysis.bio-x.cn) is a robust and user-friendly software platform for the analysis of linkage disequilibrium, haplotype construction, and genetic association at polymorphism loci [7]. However, the haplotype construction element is based on the standard EM algorithm (the so-called Full-Precise-Iteration algorithm referred to in a previous article), and it therefore comes with the deficiencies mentioned above. In this study, we have developed a new improved algorithm, the PL-CSEM (Partition-Ligation Combination-Subdivision EM), designed for efficient estimation of haplotypes constructed from large numbers of biallelic or multiallelic loci in diploid individuals.

We applied the CS strategy, which deals with the number of alleles for each locus, to construct an optimum set of candidate haplotypes. The essential steps of the CS strategy are as follows: we first combine some alleles to reduce the total for each multiallelic locus (the combination step), and then use the EM algorithm to construct haplotypes with the new alleles, i.e. from the combination step, and subdivide the phase hierarchically, through a bottom-up approach (the subdivision step). For instance, given a pair of multiallelic loci with k and l alleles, respectively, there are $k \times l$ possible haplotypes (Figure 1). The standard EM algorithm has to consider all the possible haplotypes, but using the CS strategy, we first combine alleles to make both loci biallelic, and then use the EM algorithm to estimate haplotypes with the new alleles. In the EM step, only four possible haplotypes are taken into consideration. Then, only the haplotypes whose frequencies are greater than a threshold value (*e.g.*

10^{-5}) are retained in the next step. In the subdivision step, every allele (except the initial allele) is subdivided into two parts, so the remaining haplotypes are broken up to form the set of candidate haplotypes for the next EM step. The subdivision and EM steps are repeated until all the alleles are returned to the initial state. In the program, the alleles are combined randomly in the combination step. To test whether the random combination is robust, we generated 10 simulated data sets (5 multiallelic loci of 500 individuals). For each data set, we estimated the haplotype frequencies 10 times (they can be combined in different ways), and calculated the performance indexes for comparison. The results show that despite the different combination possibilities, the groups of significant haplotypes are very close, and only those haplotypes with very low frequencies are different. We believe, therefore, that the random method of combination is robust (for details, see Supplementary information, Table S1).

Niu and Qin *et al*. first implemented the Partition-Ligation (PL) strategy together with Gibbs sampling and the EM algorithm (i.e. PLEM) to estimate haplotype phases for a large number of SNPs [4, 8]. The PL strategy is a divide-conquer-combine technique, and it is useful in dealing with a large number of linked loci. We also implemented the PL strategy in our program. It can be described as follows: first break down all of the marker loci into stretches of "atomistic" units, construct haplotypes for each unit and then rebuild the phase hierarchically, through a bottom-up approach. In each unit, haplotypes are generated through the EM algorithm or the CSEM algorithm (for the ones that include multiallelic loci). In addition, we have devised another method for use in the ligation step. Since every unit can be considered as a multiallelic locus, the CS strategy could be useful. If the set of candidate haplotypes space is too large, the CS strategy would be utilized in this step. More than two segments could be combined at each ligation step.

To avoid eliminating haplotypes too readily, in each EM/CSEM step the threshold value for eliminating haplotypes is adaptable. However, a backup-buffering strategy is also available. The user can set the threshold value and the size of a buffer for retaining some partial haplotypes whose estimated frequencies are below the threshold value. Thus, appropriate values can be set to retain more independent haplotypes.

To illustrate the use of our algorithm in practice, we analyzed real data sets of HLA (5 multiallelic loci (15, 32, 14, 13 and 7, respectively) of 420 unrelated individuals) and GH1 (14 biallelic loci and 1 triallelic locus of 154 unrelated individuals). We also derived real SNP data sets (100 replicates of 20 and 40 SNPs of 60 individuals) from the International HapMap Project and generated simulated data sets (100 replicates of 20 and 40 SNPs of 60 individuals, and 20, 40, 80, and 160 SNPs of 500 and 1000 individuals) using a coalescence model [9]. We evaluated the performance of our method, compared with PLEM. The PLEM program was downloaded from J. S. Liu's web site. The parameters of the program were set up as recommended. We computed the I_H and I_F scores previously used by Excoffier and Slatkin [2], and the average error rates based on either individual phase call (INDI) or the proportion of incorrectly inferred loci (LOCI) [4]. I_H is an index of performance in terms of haplotype identification, and its value can vary between 1, when the identified haplotypes are exactly those present in the true sample, and 0, when none of the true haplotypes has been identified. I_F is defined as one minus half the sum of absolute differences between estimated and true haplotype frequencies, and it varies from 0 to 1. The more accurate the estimation, the closer to 1 the I_F value. In addition, we recorded the running time (RT), using Inter (R) Xeon(TM) CPU 2.80 GHz, 2.79 GHz, and 2.00 GB RAM on the Microsoft Windows XP operating system with 1 s as the time unit.

Tables 1 and 2 show the performance of our program and PLEM for the real data sets and the simulated data sets, with the better performance shown in bold. Because the real phase information of the HLA data were unknown, we could not calculate its I_H, I_F, and error rates. The results of GH1 and HLA for the PLEM program were also unavailable, as the PLEM program cannot handle multiallelic loci. For most of the comparisons, the PL-CSEM program performs better than the PLEM program, obtaining higher IH and IF scores, with lower error rates in terms of both the individual phase calls and the incorrectly inferred loci, as well as requiring less running time.

In summary, the most important improvement in our program, relative to PLEM, is that for large data sets, our system is much less time-consuming and has a lower, or at least comparable, error rate. In addition, our program can deal with loci with dozens of alleles, which is beyond the scope of the PLEM program. The PL-CSEM program, which is integrated into the SHEsis [7], an existing web-based platform, is freely available through the internet (http://analysis.bio-x.cn). (A test version is currently available. 〈http://analysis.bio-x.cn/myAnalysis0.php〉) However, reference 7 and this paper should be cited whenever it is used in publication.

Acknowledgments

This work was supported by grants from the Hi-Tech Research and Development Program of China (2006AA02A407), Shang-

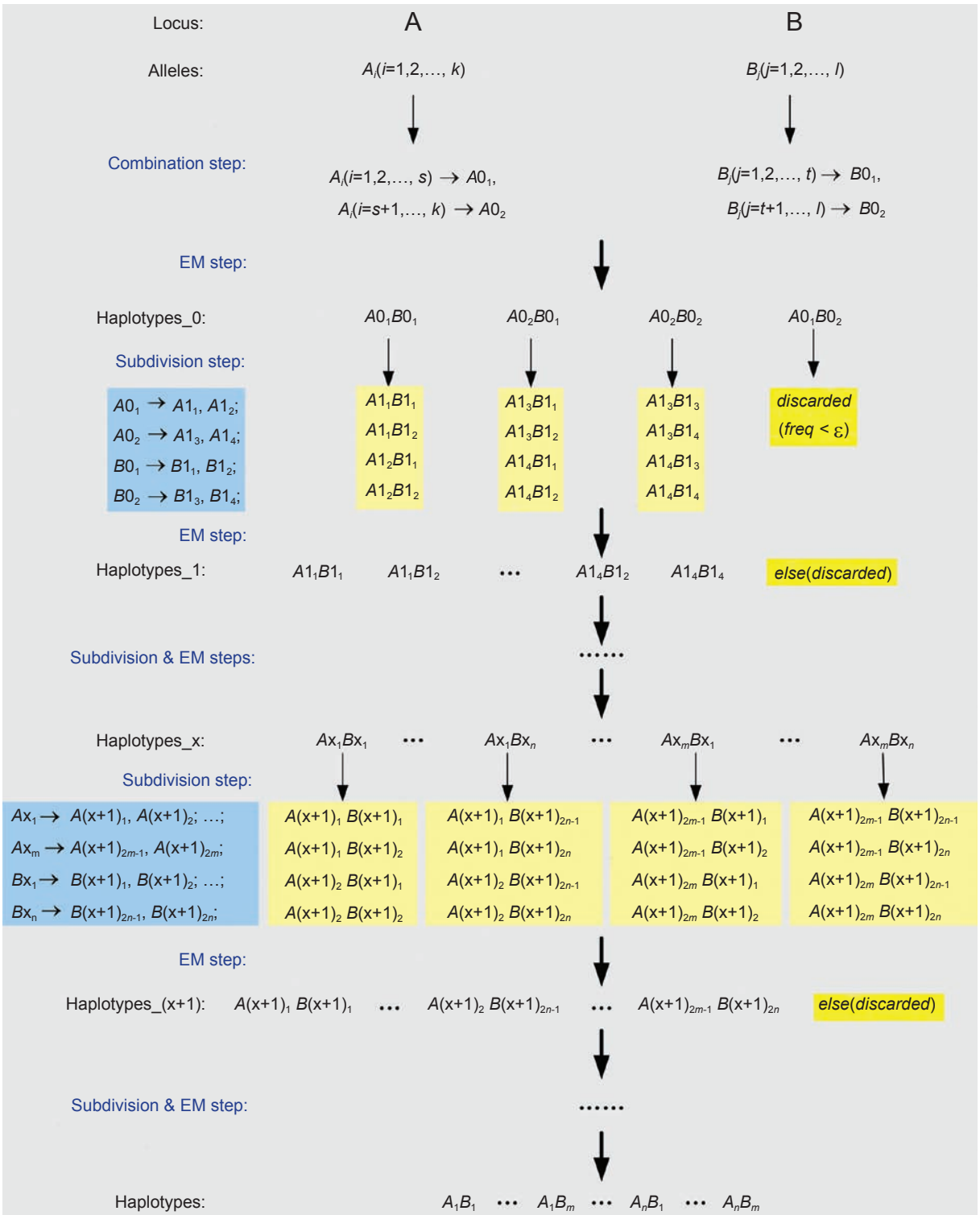

Figure 1 A flow chart of the CS-EM algorithm. A and B are two multiallelic loci with *k* and *l* alleles, respectively. *x* is the level of the CS pyramidal hierarchy; *Ax* and *Bx* denote the temporary alleles, which are made up of the initial alleles. To prove the random combination is robust, we generated 10 simulated data sets (5 multiallelic loci of 500 individuals). For each data set, we estimated the haplotype frequencies 10 times (there are different combination possibilities), and calculated the I_H and I_F scores for comparison. The results show that although there are different possible combinations, the groups of significant haplotypes are very close, and only those haplotypes with very low frequencies are different. We therefore consider that the random method of combination is robust.

Table 1 Performance of programs for the real data sets

OUR \ PLEM	HLA	GH1	Chr12-HapMap_CEU	
	5 Loci	15 Loci	20 SNPs	40 SNPs
I_H	– \ –	**0.784** \ –	**0.877** \ 0.867	**0.770** \ 0.764
I_F	– \ –	**0.927** \ –	**0.954** \ 0.948	**0.887** \ 0.867
INDI	– \ –	**9.09** \ –	**4.00** \ 4.55	**11.32** \ 13.23
LOCI	– \ –	**0.82** \ –	**0.36** \ 0.39	**0.75** \ 0.81
RT (s)	**104.2** \ –	**0.13** \ –	**0.05** \ 0.11	**0.20** \ 0.23

The unavailable value is shown in '–'. The better performance is shown in bold.

Table 2 Performance of programs for the simulated data sets

	Sample Num	60		500				1000			
	Loci Num	20	40	20	40	80	160	20	40	80	160
I_H	OUR	**0.806**	**0.710**	**0.931**	**0.915**	**0.905**	**0.895**	**0.976**	**0.972**	**0.977**	**0.979**
	PLEM	0.752	0.665	0.911	0.877	0.827	0.761	0.964	0.941	0.913	0.880
I_F	OUR	**0.899**	**0.812**	**0.986**	**0.980**	**0.972**	**0.955**	**0.993**	**0.990**	**0.990**	**0.987**
	PLEM	0.864	0.747	0.983	0.966	0.951	0.932	0.991	0.984	0.975	0.965
INDI	OUR	**8.05**	**18.15**	**0.80**	**1.57**	**2.73**	5.97	**0.49**	**0.81**	**1.47**	2.75
	PLEM	11.28	24.25	1.08	2.61	3.70	**4.13**	0.60	1.29	1.97	**2.05**
LOCI	OUR	**0.63**	**1.03**	**0.06**	**0.09**	**0.11**	0.16	**0.03**	**0.04**	**0.06**	0.08
	PLEM	0.92	1.51	0.08	0.15	0.17	**0.15**	0.04	0.07	0.09	**0.07**
RT	OUR	**0.10**	0.45	**0.37**	**3.04**	**19.1**	133.8	**0.49**	**4.9**	**46.3**	**304.9**
(s)	PLEM	0.13	**0.33**	0.51	6.31	41.8	**117.9**	0.94	19.2	850.9	2518.0

The better performance is shown in bold.

hai Changning Health Bureau Program (2008406002), Shanghai Municipal Health Bureau Program (2008095), Shanghai Program (07DZ22917), Shanghai Leading Academic Discipline Project (B205), National S973 Program (2007CB947303), National Basic Research Program of China (2006CB910601), Shanghai-Unilever Research and Development Fund (06SU07007), and National Key Technology R&D Program (2006BAI05A05). We are grateful to Amy De'ath of the Welsh Transplantation and Immunogenetics Laboratory for providing the Sinhala Data, and Horan *et al.* of the Institute of Medical Genetics, Cardiff University, for providing the GH1 data set.

Zhiqiang Li[1, 2, 3,*], Zhao Zhang[2, 3,*], Zangdong He[1, 2, 3], Wei Tang[2, 3], Tao Li[1, 2, 3], Zhen Zeng[1, 2, 3], Lin He[2, 3, 4], Yongyong Shi[1, 2, 3]

[1]Changning Mental Health Center, Affiliated Hospital of Bio-X Center, Shanghai Jiao Tong University, 299 Xiehe Road, Shanghai 200042, China; [2]Bio-X Center, Shanghai Jiao Tong University, 1954 Huashan Road, Shanghai 200030, China; [3]Institute for Nutritional Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 294 Taiyuan Road, Shanghai 200031, China; [4]Institutes of Biomedical Sciences, Fudan University, Shanghai 200032, China

*These two authors contributed equally to this work.
Correspondence: Yongyong Shi[a], Lin He[b]
Tel: +86-21-6293-3338; Fax: +86-21-3226-0640
E-mail: [a]shiyongyong@gmail.com; [b]helinhelin@gmail.com

References

1 Clark AG. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol Biol Evol* 1990; **7**:111-122.
2 Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 1995; **12**:921-927.
3 Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001; **68**:978-989.
4 Qin ZS, Niu T, Liu JS. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am J Hum Genet* 2002; **71**:1242-1247.
5 Zhang P, Sheng H, Morabia A, Gilliam TC. Optimal step length EM algorithm (OSLEM) for the estimation of haplotype frequency and its application in lipoprotein lipase genotyping. *BMC Bioinformatics* 2003; **4**:3.
6 Tregouet DA, Escolano S, Tiret L, Mallet A, Golmard JL. A new algorithm for haplotype-based association analysis: the Stochastic-EM algorithm. *Ann Hum Genet* 2004; **68**:165-177.

7 Shi YY, He L. SHEsis, a powerful software platform for analyses of linkage disequilibrium, haplotype construction, and genetic association at polymorphism loci. *Cell Res* 2005; **15**:97-98.
8 Niu T, Qin ZS, Xu X, Liu JS. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet* 2002; **70**:157-169.
9 Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 2002; **18**:337-338.

(**Supplementary information** is linked to the online version of the paper on the Cell Research website.)