BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 12 2002
Pages 1694–1695

GENECOUNTING: haplotype analysis with missing genotypes

Jing Hua Zhao[1,*], Sebastien Lissarrague[2], Laurent Essioux[3] and Pak Chung Sham[4]

[1]Department of Epidemiology and Public Health, University College London, 1–19 Torrington Place, London WC1E 6BT, UK, [2]Genset SA, Site SNECMA RN7, 91030 Evry, France, [3]ValiGen SA, Tour Neptune, 92086 Paris-La-Défense, France and [4]Section of Genetic Epidemiology, PO Box 80, Institute of Psychiatry, London SE 8AF, UK

Received on November 22, 2001; revised on March 25, 2001; accepted on May 21, 2001

ABSTRACT

Summary: A general algorithm is described for haplotype analysis of unrelated individuals with missing genotypes. It can handle problems involving multiple polymorphic markers with missing data.

Availability: GENECOUNTING is available from http://www.iop.kcl.ac.uk/IoP/Departments/PsychMed/GEpiBSt/software.stm

Contact: j.zhao@public-health.ucl.ac.uk; p.sham@iop.kcl.ac.uk

Maximum likelihood estimation of haplotype frequencies from unphased, multi-locus genotype data can be carried out by a special case of EM algorithm (Dempster *et al.*, 1977) that involves iterative counting of haplotypes. In principle, this algorithm can be extended to take account of missing genotypes, but current implementations of the algorithm only deal with limited amounts of missing data or do not deal with missing data appropriately.

The standard gene-counting algorithm for haplotype frequency estimation formulates the problem of uncertain phase as incomplete data, and consists of an E-step where the expected counts of the phased genotypes are calculated using current haplotype frequency estimates, and an M-step where the expected counts are summed over all individuals to provide revised haplotype frequency estimates. The treatment of missing genotype data requires a generalization of the E-step to consider all possible phased genotypes that are consistent with the non-missing genotypes of each individual. We have implemented this algorithm in a program called GENECOUNTING.

Algorithm G (*gene counting with missing genotypes*). We classify genotypic configurations into those with and without missing data; the counts of these are denoted as n and m, respectively, so that the total sample size is $N = \sum_{p=1}^{P} n_p + \sum_{q=1}^{Q} m_q$ if there are P configurations without missing data and Q configurations with missing data. If the haplotype frequencies are denoted as h, then the probability of each confguration without missing genotype, g, is a function of h, under the assumption of random mating. Furthermore, the probability of each configuration with missing genotypes, t, is a 'marginal' probability defined as the sum of all the g's which have the same genotypes at the non-missing markers. For clarity below let c, c' and c'' be the haplotype counts from complete data, data with ambiguous phase but no missing genotype, and data with with missing genotype, respectively.

G_1 [Initialize] set c, c' and c'' to be zero, set haplotype frequencies h (e.g. at random or the product of allele frequencies), calculate genotype probabilities from haplotype frequencies and obtain log-likelihood l.

G_2 [save log-likelihood] $l_s \leftarrow l$

G_3 For each configuration with no missing data and at most one heterozygous marker, deduce the two haplotypes and count the $2n_p$ haplotypes into c

G_4 For other configurations, do iterative counting through steps G_5 to G_8

G_5 [test for missing genotype] If there is missing genotype goto step G_7

G_6 [count using data without missing genotypes]

- count number of heterozygotes m
- obtain phase probabilities for 2^{m-1} phases
- count for each phase the two haplotypes each by (phase probability) $\times\, n_p$ into c'

G_7 [count using data with missing genotypes]

- list all possible genotypes for each configuration
- for each possible genotype calculate its probability (g)

*To whom correspondence should be addressed.

Table 1. Genotype counts for biallelic markers

Marker 1	Marker 2 1/1	1/2	2/2	Missing
1/1	n_1	n_2	n_3	m'_1
1/2	n_4	n_5	n_6	m'_2
2/2	n_7	n_8	n_9	m'_3
Missing	m_1	m_2	m_3	

Table 2. Genotypic probabilities for two biallelic markers

Marker 1	Marker 2 1/1	1/2	2/2	t'
1/1	h_{11}^2	$2h_{11}h_{12}$	h_{12}^2	t'_1
1/2	$2h_{21}h_{11}$	$2h_{21}h_{12}+2h_{22}h_{11}$	$2h_{22}h_{12}$	t'_2
2/2	h_{21}^2	$2h_{21}h_{22}$	h_{22}^2	t'_3
t	t_1	t_2	t_3	

- accumulate total probabilities for configuration (t)
- perform G_6 using $m_q g/t$ as n_p to obtain (c'')

G_8 [obtain haplotype frequencies and log-likelihood] set $h \leftarrow (c + c' + c'')/(2N)$ and calculate log-likelihood l

G_9 [test for convergence] if $l - l_s > \epsilon$ save log-likelihood] $l_s \leftarrow l$ and goto step G_4

Algorithm G is suitable when data are missing at random. In GENECOUNTING the initialization of h at step G_1 (when assuming linkage equilibrium) and enumeration of phases at steps G_5 and G_7 are recursive, making it easy to accommodate different numbers of loci. Full details of the implementation has been described elsewhere (Zhao and Sham, 2002).

We illustrate the algorithm for the simple case of two biallelic markers. Table 1 defines the possible configurations and their counts, while Table 2 gives the probabilities of these configurations in terms of haplotype frequencies.

For haplotype 11, the count c is $2n_1 + n_2 + n_4$ for all iterations, while the counts c' and c'' changes from one iteration to the next and are given by $c' = n_5(2h_{22}h_{11})/g_5$ and $c'' = 2m_1g_1/t_1 + m_2g_2/t_2 + m_1g_4/t_1 + m_2 2h_{22}h_{11}/t_2 + 2m'_1g_1/t'_1 + m'_1g_2/t'_1 + m'_2g_4/t'_2 + m'_2 2h_{22}h_{11}/t'_2$. The counting of the other three haplotypes (12, 21 and 22) proceeds in a similar fashion. The log-likelihood contains contributions from both complete and incomplete data, i.e.

$$l = \sum_{i=1}^{9} n_i \ln(g_i) + \sum_{j=1}^{3} m_j \ln(t_j) + \sum_{k=1}^{3} m'_k \ln(t'_k)$$

GENECOUNTING is implemented in C with dynamic memory allocation and is able to run on both Unix and Windows systems. It has been tested with simulated samples (unpublished results from Sebastien Lissarrangue). It can handle both SNPs and microsatelite markers without restriction on missing data pattern. We recommend it for problems of moderate size (say 10–15 loci) with not-so-heavy missing data. For such data it provides a useful tool for haplotype association analysis. For bigger problems other approaches such as Markov chain Monte Carlo (Stephens *et al.*, 2001; Niu *et al.*, 2002) or heuristic approximations (e.g. SNPHAP, http://www-gene.cimr.cam.ac.uk/clayton) are necessary.

ACKNOWLEDGEMENTS

We thank Dr Andrew Pakstis for his willingness to analyse a SNP data using HAPLO. We also thank editor Dr Babara Cox and reviewers for their help and comments during preparation of the manuscript. This work is supported by Wellcome project grant 055379.

REFERENCES

Dempster,A.P., Laird,N. and Rubin,D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc.* B, **39**, 1–38.

Niu,T., Qin,Z., Xu,X. and Liu,J.S. (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am. J. Hum. Genet.*, **70**, 157–169.

Stephens,M., Smith,N.J. and Donnelly,P. (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, **68**, 978–989.

Zhao,J.H. and Sham,P.C. (2002) Generic number system and haplotype analysis. *Comput. Meth. Prog. Biomed.*, in press.