A survey of current software for haplotype phase inference

Michael E. Weale

Bloomsbury Analytical Services, 28/30 Little Russell Street, London WC1A 2HN, UK;
Tel: +44 020 7404 3040; Fax: +44 020 7404 2083; Email: mw@bloomsburyanalytical.com

Date received (in revised form): 9th November 2003

Abstract

In the past two years, tracking the explosion in data due to ever-improving single nucleotide polymorphism (SNP) maps and cheaper high-throughput genotyping technologies, a bewildering array of new algorithms and relevant software have appeared for haplotype phase inference. The alternatives to haplotype inference are to resolve haplotypes completely, either by *in vitro* methods or by typing close pedigrees, which is expensive and is not guaranteed in pedigrees, or to ignore haplotype-level analysis in favour of genotype-level analysis, which avoids the danger of treating inferred haplotypes as real but denies the researcher, potentially, any valuable analytic insights. This review attempts a snapshot of this rapidly moving field as it stands at present, and is mainly restricted, given the current predominance of SNP genotyping, to the consideration of diallelic data. For completeness, the review will occasionally refer to algorithms for which no software exists.

Keywords: *haplotype phase inference, algorithms, software, parsimony, maximum likelihood, Bayesian analysis*

Introduction

Haplotype phase algorithms can be conveniently split into three main types: parsimony, maximum likelihood and Bayesian. The researcher may either want to infer haplotype frequencies in the population, impute the haplotypes possessed by given individuals, or both. In general, parsimony methods most naturally estimate individual haplotypes, maximum likelihood methods most naturally estimate population frequencies and Bayesian methods can do both.

Parsimony algorithms avoid explicit likelihood calculations by minimising a 'costly' constraint. The grandfather of all haplotype phase algorithms (an elderly 13 year old) is Clark's method,[1] a simple iterative procedure inspired by the constraint 'minimise the number of new haplotypes you have to invent'. (To obtain 'HAPINFERX' software, apply to ac347@cornell.edu.) The method can either suffer from having too many solutions or from having none (although the general problem of convergence is a common issue with all haplotype inference algorithms). There is also no guarantee that the global minimum for the 'minimise haplotype number' constraint is reached by Clark's algorithm. This latter problem is fixed in a more recent algorithm[2] ('HAPAR'; apply to lwang@cs.cityu.edu.hk). Phylogenetic parsimony methods have been explored by Daniel Gusfield and colleagues ('GPPH', 'DPPH' and 'BPPH'; http://wwwcsif.cs.ucdavis.edu/~gusfield/). The constraint here is 'minimise the number of ancestral recombination events required to link the new invented haplotypes'. As one might expect, this constraint works well in small, tightly-linked genomic regions and less well in bigger regions.[3]

Because parsimony algorithms avoid explicit likelihood calculations, they do not provide any natural way to measure uncertainty in the estimates. Maximum likelihood and Bayesian methods provide a way around this problem.

Maximum likelihood estimation is predominantly undertaken via Expectation–Maximisation (EM) algorithms. These use an explicit but very simple likelihood model for the data (the so-called 'gene counting' model). Observed (or partially observed) haplotype counts follow a multinomial distribution conditional on the haplotype population frequencies. Random assortment of haplotypes to individuals is assumed (a standard assumption for all algorithms, whether maximum likelihood or Bayesian, working with likelihood functions). The EM algorithm avoids making assumptions about the mutational and recombinatorial relationships of the final set of inferred haplotypes, which some see as an advantage and others as a disadvantage. The original EM algorithm citation here is usually Excoffier and Slatkin,[4] but see also Hawley and Kidd[5] ('HAPLO'; http://krunch.med.yale.edu/haplo/). Some well-used implementations of the algorithm are: 'EM-decoder'[6] (http://www.people.fas.harvard.edu/~junliu/em/em.htm), 'EH+'[7] (http://www.iop.kcl.ac.uk/IoP/Departments/PsychMed/GEpiBSt/software.shtml), 'GENECOUNTING'[8] (same website as 'EH+') and 'snphap' (D. Clayton; http://www-gene.cimr.cam.ac.uk/clayton/software/). Of these,

'GeneCounting' and 'snphap' have the added refinement of allowing for missing data. 'snphap' has the additional refinement that, once haplotype frequencies are estimated, the program swaps from likelihood-based to posterior probability-based imputation and calculates haplotype-pair probabilities — conditional on the estimated haplotype frequencies — for all pairs consistent with an individual's genotype. 'snphap' works only on diallelic data. The extension 'hap' (J. H. Zhao, same website as 'EH+' and 'GeneCounting') runs the same algorithm but accepts multiallelic data.

Bayesian algorithms have the potential to address the issue, missing from EM algorithms, of how to guide the haplotype inference process so as to favour solutions which make sense in terms of an underlying genealogy connecting the haplotypes, via manipulation of the prior. The first proposed Bayesian algorithm, and still one of the best, is that implemented in the 'PHASE' program[9] (http://www.stat.washington.edu/stephens/phase.html). The proposed prior is derived, approximately, from coalescent theory, and ensures that new 'invented' haplotypes look mutationally similar to the others at any one stage of the iterative (Gibbs sampler) stochastic convergence process. The main disadvantage of the original version of 'PHASE' was its plodding speed of convergence for datasets of any reasonable size.

Extensions

The key developments have been towards improving speed as datasets increase in size, and coping with ever larger genomic regions, where it becomes impossible to infer unbroken haplotypes over the entire region because their estimated frequencies become vanishingly small. For parsimony algorithms, Gusfield shows how to implement a speeded up version of Clarke's algorithm.[3] Eskin and colleagues also illustrate the considerable speed advantages of this simple algorithm in cases where a simple 'block'-like structure of the genome is observed[10] ('HAP'; http://www1.cs.columbia.edu/compbio/hap/).

For Bayesian algorithms, one key idea that has since been implemented in several new extensions is the Partition-Ligation strategy proposed by Niu and colleagues.[6] Here, the genome region is split into a number of smaller regions (either arbitrarily or by some process that attempts to maximise linkage disequilibrium within each region). The haplotype inference method is then applied separately to two adjacent sub-regions and allowed to converge separately. Larger haplotypes are then formed by allowing haplotypes to merge at random across the boundary, using current estimates of their respective frequencies. The haplotype inference method is then applied to the new larger region (and allowed to converge), and separately to another adjacent sub-region. The process repeats until all sub-regions are merged. Niu and colleagues also speeded up convergence steps by 'prior annealing', in which jumps in posterior probability space are allowed to be larger at first, then progressively smaller.

Stephens and Donnelly have implemented these ideas in a new faster 'PHASE' program,[11] and Niu and colleagues have implemented them in a new Bayesian algorithm 'HAPLOTYPER'[6] (http://www.people.fas.harvard.edu/~junliu/Haplo/docMain.htm). This latter algorithm abandons the idea of a prior favouring mutational similarity among inferred haplotypes, and instead applies a Dirichlet prior. This prior functions in a similar way to the multinomial in EM algorithms, in that it avoids making assumptions about mutational and recombinatorial relationships among inferred haplotypes. Lin and colleagues have also proposed a separate Bayesian algorithm with a Dirichlet prior.[12] The issue of what constitutes a good prior for a Bayesian model remains unresolved.[13] While the Dirichlet is computationally convenient, there is valuable extra information in the mutational and recombinatorial relationships that should lead to more accurate inferences of haplotypes, provided that the models dealing with both of these phenomena are reasonable. Eronen and colleagues propose a new prior allowing for recombination, designed explicitly for long-range genotype data[14] ('MC-VL'; http://www.cs.helsinki.fi/group/genetics/haplotyping.html). Another promising new algorithm that explicitly incorporates recombination into its strategy and is also designed specifically for long-range genotype data is the 'ELB' algorithm proposed by Excoffier and colleagues.[15] The latest version of 'PHASE' also optionally incorporates a recombination model.[16]

For EM algorithms, Qin and colleagues have implemented the above partition-ligation ideas into an EM context[17] ('PL-EM', http://www.people.fas.harvard.edu/~junliu/plem/). A very similar algorithm has been proposed by Li and colleagues[18] ('HPlus'; http://qge.fhcrc.org/hplus/). Zhang and colleagues propose an improvement to the speed of the E-step in the EM algorithm[19] ('OSLEM'; http://genome3.cpmc.columbia.edu/cgi-bin/GENOME/oslem/doHaplo.cgi), and Thomas proposes other approximations to increase EM algorithm speed[20] ('GCHap'; http://episun7.med.utah.edu/~alun/gchap/). David Clayton's 'snphap' tackles the large data set problem by starting with two-locus haplotypes, extending the haplotype one locus at a time, and culling low-frequency haplotypes at an early stage. The effect of these short cuts on the optimality of the final solution is unclear.

A number of researchers have proposed EM algorithms that take advantage of the increased (but not complete) certainty in haplotype phase afforded by simple pedigree data, especially trios. These include Rohde and Fuerst[21] (apply to rohde@mdc-berlin.de for software), Li and Jiang[22] ('PedPhase'; http://www.cs.ucr.edu/~jili/haplotyping.html), Dudbridge,[23] and Weale and colleagues[24] ('EMtrio', part of the 'TagIT' package; http://popgen.biol.ucl.ac.uk/software.html). 'EMtrio' is designed to cope with partially missing genotype data, in which one homologous chromosome may be phase-resolved and the other not. The Bayesian 'PHASE'

program also allows input of phase-resolved data, but does not handle the above partially-missing situations. A front-end to 'PHASE', called 'PHamily', automatically resolves trio data (H. Ackerman and M. Stephens; http://archimedes.well.ox.ac.uk/pise/). Opinion is divided on whether it is worth the extra genotyping effort to type close relatives to help resolve phase.[25–28] Interest has also focused recently on the use of EM algorithms to infer haplotypes from pooled DNA data.[29–31]

Regardless of which method of haplotype inference is used, it is generally recognised that any subsequent analyses using such haplotypes (eg association tests against phenotype) should ideally take account of the uncertainty associated with these inferred haplotypes. There has also been a considerable amount of recent literature on this subject, which is not reviewed here. One promising program that allows for this is 'BLADE'[32,33] (Version 2: http://www.fas.harvard.edu/~junliu/TechRept/03folder/bladev2.tgz).

Despite the assertions of some, it is currently not clear which one of these alternative methods and their extensions will provide the most reliable estimates. All the rival algorithms tend to do well when datasets and genomic regions are small; all do badly when they are large. One prudent measure is to check the results of different methods against each other for consistency. The program 'HIT' brings together four well-used algorithms for this purpose (including two EM algorithms, 'PHASE', and 'HAPLOTYPER'; apply to wangx@udel.edu). The 'HapScope' package[34] (ftp://ftp1.nci.nih.gov/pub/HapScope) incorporates versions of both 'PHASE' and 'snphap'. When consistency breaks down in the larger datasets, the way forward is still unclear. The key issue will not be to find a better haplotype inference method *per se*, but rather to find a better strategy for partitioning large genomic regions into manageable sub-regions without losing useful linkage disequilibrium information along the way.

References

1. Clark, A.G. (1990), 'Inference of haplotypes from PCR-amplified samples of diploid populations', *Mol. Biol. Evol.* Vol. 7, pp. 111–122.
2. Wang, L. and Xu, Y. (2003), 'Haplotype inference by maximum parsimony', *Bioinformatics* Vol. 19, pp. 1773–1780.
3. Gusfield, D. (2003), 'Haplotype inference by pure parsimony'. *Combinatorial Pattern Matching, Proceedings: Lecture Notes in Computer Science.* Vol. 2676, pp. 144–155.
4. Excoffier, L. and Slatkin, M. (1995), 'Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population', *Mol. Biol. Evol.* Vol. 12, pp. 921–927.
5. Hawley, M.E. and Kidd, K.K. (1995), 'HAPLO: A program using the EM algorithm to estimate the frequencies of multi-site haplotypes', *J. Hered.* Vol. 86, pp. 409–411.
6. Niu, T., Qin, Z.S., Xu, X. *et al.* (2002), 'Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms', *Am. J. Hum. Genet.* Vol. 70, pp. 157–169.
7. Zhao, J.H., Curtis, D. and Sham, P.C. (2000), 'Model-free analysis and permutation tests for allelic associations', *Hum. Hered.* Vol. 50, pp. 133–139.
8. Zhao, J.H., Lissarrague, S., Essioux, L. *et al.* (2002), 'GENECOUNTING: Haplotype analysis with missing genotypes', *Bioinformatics* Vol. 18, pp. 1694–1695.
9. Stephens, M., Smith, N.J. and Donnelly, P. (2001), 'A new statistical method for haplotype reconstruction from population data', *Am. J. Hum. Genet.* Vol. 68, pp. 978–989.
10. Eskin, E., Halperin, E. and Karp, R. (2003), 'Large scale reconstruction of haplotypes from genotype data', *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology,* (RECOMB-2003), pp. 104–113, ACM Press, New York, NY.
11. Stephens, M. and Donnelly, P. (2003), 'A comparison of Bayesian methods for haplotype reconstruction from population genotype data', *Am. J. Hum. Genet.* Vol. 73, pp. 1162–1169.
12. Lin, S., Cutler, D.J., Zwick, M.E. *et al.* (2002), 'Haplotype inference in random population samples', *Am. J. Hum. Genet.* Vol. 71, pp. 1129–1137.
13. Morris, A., Pedder, A. and Ayres, K. (2003), 'Linkage disequilibrium assessment via log-linear modeling of SNP haplotype frequencies', *Genet. Epidemiol.* Vol. 25, pp. 106–114.
14. Eronen, L., Geerts, F. and Toivonen, H. (2003), 'A Markov Chain approach to reconstruction of long haplotypes', *Pacific Symposium on Biocomputing*, see http://www-smi.stanford.edu/projects/helix/psb04.
15. Excoffier, L., Laval, G. and Balding, D. (2003), 'Gametic phase estimation over large genomic regions using an adaptive window approach', *Human Genomics* Vol. 1, pp. 7–19.
16. Li, N. and Stephens, M. (2003), 'Modelling linkage disequilibrium and identifying recombination hotspots using SNP data', *Genetics*, In press.
17. Qin, Z.S., Niu, T. and Liu, J.S. (2002), 'Partition-ligation-expectation-maximization algorithm for haplotype inference with single nucleotide polymorphisms', *Am. J. Hum. Genet.* Vol. 71, pp. 1242–1247.
18. Li, S.S., Khalid, N., Carlson, C. *et al.* (2003), 'Estimating haplotype frequencies and standard errors for multiple single nucleotide polymorphisms', *Biostatistics* Vol. 4, pp. 513–522.
19. Zhang, P., Sheng, H., Morabia, A., *et al*, (2003), 'Optimal step length EM algorithm (OSLEM) for the estimation of haplotype frequency and its application in lipoprotein lipase genotyping', *BMC Bioinformatics* Vol. 4, p. 3.
20. Thomas, A. (2003), 'GCHap: Fast MLEs for haplotype frequencies by gene counting', *Bioinformatics* Vol. 19, pp. 2002–2003.
21. Rohde, K. and Fuerst, R. (2001), 'Haplotyping and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information', *Hum. Mutat.* Vol. 17, pp. 289–295.
22. Li, J. and Jiang, T. (2003), 'Efficient inference of haplotypes from a genotype on a pedigree', *J. Bioinf. Comp. Bio.* Vol. 1, pp. 41–69.
23. Dudbridge, F. (2003), 'Pedigree disequilibrium tests for multilocus haplotypes', *Genet. Epidemiol.* Vol. 25, pp. 115–121.
24. Weale, M.E., Depondt, C., Macdonald, S.J. *et al.* (2003), 'Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: Implications for linkage-disequilibrium gene mapping', *Am. J. Hum. Genet.* Vol. 73, pp. 551–565.
25. Fallin, D. and Schork, N.J. (2000), 'Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data', *Am. J. Hum. Genet.* Vol. 67, pp. 947–959.
26. Becker, T. and Knapp, M. (2002), 'Efficiency of haplotype frequency estimation when nuclear family information is included', *Hum. Hered.* Vol. 54, pp. 45–53.
27. Schaid, D.J. (2002), 'Relative efficiency of ambiguous vs. directly measured haplotype frequencies', *Genet. Epidemiol.* Vol. 23, pp. 426–443.
28. Cheng, R., Ma, J.Z., Wright, F.A. *et al.* (2003), 'Nonparametric disequilibrium mapping of functional sites using haplotypes of multiple tightly linked single-nucleotide polymorphism markers', *Genetics* Vol. 164, pp. 1175–1187.
29. Wang, S., Kidd, K.K. and Zhao, H.Y. (2003), 'On the use of DNA pooling to estimate haplotype frequencies', *Genet. Epidemiol.* Vol. 24, pp. 74–82.
30. Ito, T., Chiku, S., Inoue, E. *et al.* (2003), 'Estimation of haplotype frequencies, linkage-disequilibrium measures and combination of

haplotype copies in each pool by use of pooled DNA data', *Am. J. Hum. Genet.* Vol. 72, pp. 384–398.
31. Yang, Y.N., Zhang, J.S., Hoh, J. *et al.* (2003), 'Efficiency of single-nucleotide polymorphism haplotype estimation from pooled DNA', *Proc. Natl. Acad. Sci. USA* Vol. 100, pp. 7225–7230.
32. Liu, J.S., Sabatti, C., Teng, J. *et al.* (2001), 'Bayesian analysis of haplotypes for linkage disequilibrium mapping', *Genome Res.* Vol. 11, pp. 1716–1724.
33. Lu, X., Niu, T.H. and Liu, J.S. (2003), 'Haplotype information and linkage disequilibrium mapping for single nucleotide polymorphisms', *Genome Res.* Vol. 13, pp. 2112–2117.
34. Zhang, J., Rowe, W.L., Struewing, J.P. *et al.* (2002), 'HapScope: A software system for automated and visual analysis of functionally annotated haplotypes', *Nucleic Acids Res.* Vol. 30, pp. 5213–5221.