European Journal of Human Genetics (2004) 12, 971–974

www.nature.com/ejhg

SHORT REPORT

Cox proportional hazards survival regression in haplotype-based association analysis using the Stochastic-EM algorithm

D-A Tregouet*,[1] and L Tiret[1]

[1]*INSERM U525, 75634 Paris, France*

It is now widely recognized that haplotype information inferred from genotypes can be of great interest to better characterize the role of a candidate gene in the etiology of a complex trait in the context of association studies. Several works have recently advocated the simultaneous estimation of haplotype frequencies and haplotype effects in order to get a better efficiency in parameter estimation. Most of the available models can deal with a binary or a quantitative phenotype, but none has yet discussed the application of haplotype-based association analysis to a survival outcome. We describe how the recently proposed Stochastic-EM (SEM) algorithm can be applied to estimate haplotype effects in censored data analysis using a standard Cox proportional hazards formulation. This model has been implemented in the THESIAS software freely available at http://www.genecanvas.org
European Journal of Human Genetics (2004) **12,** 971–974. doi:10.1038/sj.ejhg.5201238
Published online 7 July 2004

Keywords: haplotype; survival analysis; prospective data

Introduction

It is now widely recognized that haplotype information inferred from genotypes can be of great interest to better characterize the role of a candidate gene in the etiology of a complex trait.[1–4] Haplotype-based analysis may help in differentiating the true effect of a polymorphism from what is due to its linkage disequilibrium with other variant(s). Haplotypes may serve as better markers for unknown functional variants than single polymorphisms. Lastly, they may define functional units whose effects cannot be predicted from what is known of the individual effect of each variant. This explains the large amount of work that has been devoted to the development of statistical tools for making haplotype inference.[4–14] It is now widely admitted that haplotype frequencies and haplotype effects have to be estimated simultaneously in order to get a better efficiency in parameter estimation.[4,7,8,11–14] To our knowledge, available models allowing this joint estimation can deal with a binary and/or a quantitative phenotype, but none has yet discussed the application of haplotype-based analysis to a survival outcome. The objective of this work is to describe how our recently proposed Stochastic-EM (SEM) algorithm[13] can be extended to apply to an haplotype-based analysis of censored data using a standard Cox proportional hazards formulation.[15]

*Correspondence: Dr D-A Tregouet, INSERM U525, 91 boulevard de l'Hôpital, 75634 Paris, France. Tel: +33 1 40 77 96 93;
Fax: +33 1 40 77 97 28; E-mail: david.tregouet@chups.jussieu.fr
Received 24 February 2004; revised 4 May 2004; accepted 5 May 2004

System and methods

Consider a sample of N unrelated individuals and let ($\tilde{T}_i$, D_i, G_i) denote the ith individual's triplet where $\tilde{T}_i = T_i \wedge C_i$ with T_i being his/her failure time or C_i his/her censoring time, $D_i = I(T_i \leq C_i)$ and G_i being his/her genotypic vector at k different loci. For ease of presentation, only the case of di-allelic polymorphisms will be addressed here and we assume that G_i does not include any missing genotype even if these assumptions can be easily relaxed.[13] The number of possible haplotypic pairs compatible with G_i is

2^{c_i-1} where c_i is the number of loci where the ith individual is heterozygous. Except when $c_i \leq 1$, the true haplotypic pair of the ith individual cannot be unambiguously deduced from G_i. Would the haplotypic pair $H_i = (h_{i1}, h_{i2})$ of the ith individual be observed, the contribution of this individual to the likelihood of the sample under the standard Cox formulation would be

$$[\lambda_{T/H_i}(\tilde{T}_i, \beta)]^{D_i} S_{T/H_i}(\tilde{T}_i, \beta) = [\lambda_0(\tilde{T}_i) \exp(\beta_{i1} + \beta_{i2})]^{D_i} S_{T/H_i}(\tilde{T}_i, \beta)$$

where $\lambda_0(t)$ is an unspecified baseline hazard function and $S_{T/H_i}(\tilde{T}_i, \beta)$ is the survival function at time $\tilde{T}_i$. In this modeling, $e^{\beta i1}$ and $e^{\beta i2}$ represent the hazard risk ratios (HRRs) for the survival outcome associated with haplotypes h_{i1} and h_{i2}, respectively, by comparison to a reference haplotype (which can be taken as the most frequent haplotype, for example), under the assumption of additive haplotype effects. $S_{T/H_i}(\tilde{T}_i, \beta)$ is defined by

$$\exp\left(-\int_0^{\tilde{T}_i} \lambda_{T/Hi}(s, \beta)\mathrm{d}s\right) = \exp\left(-\exp(\beta_{i1} + \beta_{i2}) \int_0^{\tilde{T}_i} \lambda_{0T}(s)\mathrm{d}s\right) = \exp(-\exp(\beta_{i1} + \beta_{i2})\Lambda(\tilde{T}_i))$$

where $\Lambda(\tilde{T}_i)$ is the cumulative hazard function at time $\tilde{T}_i$ whose estimation will be detailed thereafter.

Algorithm

The SEM algorithm whose general description for haplotype-based association analysis has been given previously[13] is an iterative algorithm where, at each iteration, any ambiguous haplotypic pair, considered as missing data, is replaced by a simulated value drawn from its conditional distribution given the observed data and the parameters obtained from the previous iteration.

The vector of parameters to be estimated, θ, is composed of the haplotype frequencies $f(h_l)$ $(l = 1 \ldots s)$ $(s \leq 2^k)$ and the logarithm of the haplotypic HRRs, β_l $(l = 1 \ldots s)$. The $(m+1)$th iteration consists of two steps, the stochastic imputation step and the maximization step that take the following forms in the context of a Cox survival haplotype analysis.

The Stochastic-Imputation step

The unobserved haplotypic pair of an ambiguous individual i is set at a single draw from the distribution of haplotypic pairs H specified by $P(H/\tilde{T}_i, D_i, G_i)$ evaluated at $\theta^{(m)}$, the current vector estimated parameter at the mth iteration, and defined by:

$$\frac{[\exp(\beta^{(m)}H)]^{D_i} \exp(-\exp(\beta^{(m)}H)\hat{\Lambda}(\beta^{(m)}, \tilde{T}_i))P(H)}{\sum_{j \in S(G_i)} [\exp(\beta^{(m)}H_j)]^{D_i} \exp(-\exp(\beta^{(m)}H_j)\hat{\Lambda}(\beta^{(m)}, \tilde{T}_i))P(H_j)} \quad (1)$$

where $S(G_i)$ is the set of all haplotypic pairs H_j such that $H_j = (h_{j1}, h_{j2})$ is compatible with G_i and where $P(H_j)$ is a function of the current estimated haplotype frequencies $f^{(m)}(h_l)$.

The Maximisation step

With the pseudo-completed sample, a likelihood maximization routine is then used to obtain updated parameters $\theta^{(m+1)}$. This can be decomposed into two parts. First, haplotype frequencies are obtained by counting the pseudo-observed haplotypic pairs $H_i = (h_{i1}, h_{i2})$ under the assumption of Hardy–Weinberg equilibrium (HWE). Then, the logarithm of the haplotypic HRRs are independently updated by the standard maximum likelihood (ML) estimates obtained from the partial Cox likelihood performed on the pseudo-completed data where the haplotypic pair of any individual is now considered to be observed, that is, by maximizing the following likelihood:

$$\prod_{i=1}^{N} \left[\frac{\exp(\beta_{i1} + \beta_{i2})}{\sum_{n=1}^{N} \exp(\beta_{n1} + \beta_{n2}) I(\tilde{T}_i \leq \tilde{T}_n)}\right]^{D_i} \quad (2)$$

Given the updated $\beta^{(m+1)}$, the cumulated hazard function is then updated according to the Breslow estimates[16] used in the context of Cox proportional hazards analysis.

To initialize the algorithm, a starting value $\theta^{(0)}$ must be provided. For example, all β_l's can be set to 0 and haplotype frequencies can be calculated assuming that all polymorphisms are in linkage equilibrium, that is, are the product of allele frequencies.

Let M be the total number of iterations of the SEM algorithm. The properties of the generated sequence of $\{\theta^{(m)}\}$, $m = 1 \ldots M$, are detailed elsewhere.[17,18] The main results are that the sequence of $\{\theta^{(m)}\}$ does not converge pointwise but composes a Markov chain that rapidly converges, under regularity conditions, to a stationary distribution.[18] The stationarity is obtained after a sufficiently long 'burn-in' period and the point estimate $\tilde{\theta}$ is then simply the mean of the $\theta^{(m)}$ within this stationary distribution. The resulting SEM estimate $\tilde{\theta}$ has been shown to be asymptotically equivalent to the ML estimate θ in the exponential family case[17] and this equivalence has been observed in many other situations.

Once the SEM estimate $\tilde{\theta}$ is obtained, we propose as parameter variance estimates those obtained by inverting the Fisher information matrix derived from the following

likelihood expression evaluated at $\tilde{\theta}$:

$$\prod_{i=1}^{N}\left[\frac{\left(\sum_{j\in S(G_i)}P(H_j)\exp(\beta_{j1}+\beta_{j2})\right)\Big/\sum_{j\in S(G_i)}P(H_j)}{\sum_{n=1}^{N}\left(I(\tilde{T}_i\leq\tilde{T}_n)\left(\sum_{j\in S(G_n)}P(H_j)\exp(\beta_{j1}+\beta_{j2})\right)\Big/\sum_{j\in S(G_n)}P(H_j)\right)}\right]^{D_i} \quad (3)$$

Finally, evaluating (3) at $\tilde{\theta}$ provides an estimation of the partial Cox likelihood of the sample that can then be used for hypothesis testing by means of the likelihood ratio test statistics.

Discussion

In this report, we proposed a flexible model allowing the joint estimation of haplotype frequencies and haplotype effects in a context of survival analysis. This model is based on the Cox formulation[15] that is considered as a standard in proportional hazard analysis. The estimates provided by the proposed SEM algorithm are expected to be close to the ML estimates even though the theoretical equivalence between the SEM and ML estimates has not been fully demonstrated in the case of the partial Cox likelihood. We compared on two real data sets,[19,20] the results provided by the proposed SEM algorithm to those obtained by a standard ML method for survival data analysis. However, since the implementation of a partial Cox likelihood with missing data (ie ambiguous haplotypes) is not easily tractable and can be quite computationnally cumbersome by use of the standard Newton–Raphson (NR) algorithm,[13] we implemented a parametric Weibull model[21] in our previous NR-based method for haplotype-association analysis,[4,22] and we compared estimates obtained by the two methods. Results of these comparisons are available online (httt://www.genecanvas.org). Even though the Cox and Weibull models are quite different in terms of the mathematical formulations and assumptions, they have been shown to produce similar results in many situations and the similarity between the parameter estimates provided here by both methods strengthened our confidence about the validity of the SEM algorithm. The limitations of the current model are the assumption of HWE at the haplotypic level and that of proportional hazards. Note, however, that the assumption of HWE is less questionable and more reasonable here in the whole population of a cohort than in a case–control design. It would also be interesting to develop a statistical tool to assess the goodness-of-fit of the Cox proportional hazards assumption under the framework of a haplotype-based association analysis.

While this manuscript was reviewed, a similar approach based on the EM algorithm was proposed.[23] Even though the SEM and EM algorithms are expected to be asymptotically equivalent, it would be interesting to compare them in situations where asymptotic properties may not be valid, in particular in the case of rare haplotypes. Ambiguous haplotypes can be considered as variables observed with measurement error that would be a function of the LD pattern between the studied polymorphisms. Application of statistical methods dealing with errors in variables in Cox regression analysis[24–26] may then be envisaged in the context of haplotype analysis and would deserve further attention.

This model has been implemented in the THESIAS program that can also deal with a quantitative or a binary phenotype, both under a standard and a matched (using a similar partial likelihood as that described above) case–controls designs. Our model is general enough to incorporate information on additional covariates and to test for the deviation from the hypothesis of additivity of the haplotypic effects. THESIAS is written in ANSIC and is available free of charge from http://www.genecanvas.org. THESIAS has already been used by different groups for real data analysis, either for a binary, a quantitative or a survival outcome and appears to be a tool of great usefulness for haplotype-based association study.

Acknowledgements

We wish to thank JL Golmard for his helpful comments on a earlier draft of this article and the AtheroGene Group for kindly providing us the data used for illustration.

References

1 Drysdale CM, McGraw DW, Stack CB *et al*: Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict *in vivo* responsiveness. *Proc Natl Acad Sci USA* 2000; **97**: 10483–10488.
2 Klerkx AH, Tanck MW, Kastelein JJ *et al*: Haplotype analysis of the CETP gene: not TaqIB, but the closely linked −629C→A polymorphism and a novel promoter variant are independently associated with CETP concentration. *Hum Mol Genet* 2003; **12**: 111–123.
3 Soubrier F, Martin S, Alonso A *et al*: High-resolution genetic mapping of the ACE-linked QTL influencing circulating ACE activity. *Eur J Hum Genet* 2002; **10**: 553–561.
4 Tregouet DA, Barbaux S, Escolano S *et al*: Specific haplotypes of the P-selectin gene are associated with myocardial infarction. *Hum Mol Genet* 2002; **11**: 2015–2023.
5 Stephens M, Smith NJ, Donnelly P: A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001; **68**: 978–989.
6 Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG: Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum Hered* 2002; **53**: 79–91.
7 Epstein MP, Satten GA: Inference on haplotype effects in case-controls studies using unphased genotype data. *Am J Hum Genet* 2003; **73**: 1316–1329.
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 Zhao LP, Li SS, Khalid N: A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes

and environmental variables in case–control studies. *Am J Hum Genet* 2003; **72**: 1231–1250.
10 Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA: Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* 2002; **70**: 425–434.
11 Tanck MW, Klerkx AH, Jukema JW, De Knijff P, Kastelein JJ, Zwinderman AH: Estimation of multilocus haplotype effects using weighted penalised log-likelihood: analysis of five sequence variations at the cholesteryl ester transfer protein gene locus. *Ann Hum Genet* 2003; **67**: 175–184.
12 Lake SL, Lyon H, Tantisira K *et al*: Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. *Hum Hered* 2003; **55**: 56–65.
13 Tregouet DA, Escolano S, Tiret L, Mallet A, Golmard JL: A new maximum likelihood algorithm for haplotype-based association analysis: the SEM algorithm. *Ann Hum Genet* 2004; **68**: 165–177.
14 Stram DO, Pearce CL, Bretsky P *et al*: Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case–control study of unrelated individuals. *Hum Hered* 2003; **55**: 179–190.
15 Cox DR: Regression models and life-tables (with discusssion). *J Roy Statist Soc B* 1972; **34**: 187–220.
16 Breslow NE: Contribution to the discussion on the paper by DR Cox, Regression models and life tables. *J Roy Statist Soc B* 1972; **34**: 216–217.
17 Diebolt J, Ip EHS: Stochastic EM: method and application; in Gilks WR, Richardson S, Spiegelhalter DJ (eds): *Markov Chain Monte Carlo in practice*. London: Chapman & Hall, 1996, pp 259–273.
18 Diebolt J, Celeux G: Asymptotic properties of a stochastic EM algorithm for estimating mixing proportions. *Comm Statist B: Stoch Model* 1993; **9**: 599–613.
19 Tregouet DA, Barbaux S, Poirier O *et al*: SELPL gene polymorphisms in relation to plasma SELPLG levels and coronary artery disease. *Ann Hum Genet* 2003; **67**: 504–511.
20 Ninio E, Tregouet D, Carrier JL *et al*: Platelet-activating factor-acetylhydrolase (PAF-AH) and PAF-receptor gene haplotypes in relation to future cardiovascular event in patients with coronary artery disease. *Hum Mol Genet* 2004, doi:10.1093/hmg/ddh145.
21 Cox D, Oakes D: *Analysis of Survival Data*. London, UK: Chapman & Hall, 1984.
22 Roussel R, Tregouet D, Hadjadj S, Jeunemaitre J, Marre M: Investigation of the human ANP gene in type 1 diabetic nephropathy: case–control and follow-up studies. *Diabetes* 2004; **53**: 1394–1398.
23 Lin DY: Haplotype-based association analysis in cohort studies of unrelated individuals. *Genet Epidemiol* 2004; **26**: 255–264.
24 Spiegelman D, McDermott A, Rosner B: Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *Am J Clin Nutrition* 1997; **65**: 1179S–1186S.
25 Wang CY, Hsu L, Feng ZD, Prentice RL: Regression calibration in failure time regression. *Biometrics* 1997; **53**: 131–145.
26 Nakamura T: Proportionnal hazards model with covariates subject to measurement error. *Biometrics* 1992; **48**: 829–838.