African Journal of Biotechnology Vol. 10(7), pp. 1095-1098, 14 February, 2011
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB10.1754
ISSN 1684–5315 © 2011 Academic Journals

Theoretical Research Paper

An approximate confidence interval for recombination fraction in genetic linkage analysis using a two stage Monte Carlo method Gibbs sampling

Gholamreza Jandaghi

Qom College, University of Tehran, Iran. E-mail: jandaghi@ut.ac.ir.

Accepted 3 January, 2011

One of the important parameters in genetic linkage analysis is recombination fraction. In this paper, we proposed a two stage Markov Chain Monte Carlo (MCMC) method to calculate an approximate confidence interval (ACI) for the recombination fraction. We also presented a formula for calculation of simulation size namely: outer and inner Gibbs sample sizes.

Key words: Markov Chain Monte Carlo (MCMC), Gibbs sampler, approximate confidence interval, simulation size.

INTRODUCTION

The Markov Chain Monte Carlo (MCMC) Gibbs sampler (Geman and Geman, 1984) is an iterative procedure for drawing multiple dependent realizations from a distribution known only to be up to a proportionality constant. In genetics, the Gibbs sampler provides realizations from the distribution of genotypes $P_\theta(g)$, beginning from any initial realization of genotypes of the pedigree that is compatible with observed phenotypes. An individual's genotype is updated in turn by sampling from local conditional distributions at parameter values θ, given the observed data (phenotypes) and the genotypes of all other members of the pedigree. The general theoretical justification for the Gibbs sampler was developed by Geman and Geman (1984).

The calculation of the likelihood plays an important role in the analysis of genetic data. In many instances, the likelihood can be written as a product of probabilities summed over all possible genotype configurations. The sum over genotypes can be computed easily along the lines of the Elston-Stewart algorithm (Elston and Stewart, 1971) and its extensions (Cannings et al., 1978; Janss et al., 1995; Lange and Boehnke, 1983; Lang and Elston, 1983; Stricker et al., 1995; Thomas, 1986a, b). If exact peeling over all genotypic configurations is not possible, one alternative is to use MCMC procedures to sample genotypic configurations according to the posterior distribution. The known difficulties with using Gibbs sampler such as the determination of initial configuration, advantages of random scans versus fixed scans and other aspects of the procedure, has been investigated by Sheehan et al. (1993), Jandaghi (1994) and Abraham et al. (2007).

One of the important parameters of interest in genetic linkage analysis is recombination fraction which plays an important role in analysis of pedigree data. In this paper, we proposed a two stage Monte Carlo Gibbs sampling procedure for calculation of an approximate confidence interval (ACI) for recombination fraction.

TERMINOLOGY AND NOTATIONS

Consider a pedigree with n individuals. Let $g = (g_1, g_2, ..., g_n)$ be the vector of genotypes of the

Abbreviations: MCMC, Markov Chain Monte Carlo; **ACI,** approximate confidence interval.

individuals in the pedigree, where g_j is the genotype of the *j-th* individual. Let $g_{-j} = (g_1, g_2, ..., g_{j-1}, g_{j+1}, ..., g_n)$ and $x = (x_1, x_2, ..., x_n)$ be the vector of observed phenotypes where x_j is the phenotype of the *j-th* individual. Let F be the set of founders (individuals whose parents are not in the pedigree) and θ be the recombination fraction. $P(x \mid g)$ is the penetrance probability, that is, the probability that an individual with genotype g has phenotype x. Let $P(g_k \mid g_{f_k}, g_{m_k})$ be the transmission probability, that is, the probability that an individual having genotype g_k is given the parental genotypes g_{f_k} and g_{m_k}. The likelihood function for the pedigree will be as follows:

$$L(\theta) = \sum_{feasiblegenes} P(x \mid g) P(g, \theta)$$

Where

$$P(x \mid g) = \prod_{j=1}^{n} P(x_j \mid g_j)$$

and

$$P(g, \theta) = \prod_{j \notin F} P_\theta(g_j \mid g_{f_j}, g_{m_j}) \prod_{j \in F} P(g_j)$$

Jandaghi (2010) has proposed a two stage Gibbs sampling procedure for calculating the distribution of score and likelihood ratio statistics through the following steps:

Step 1. Generating n_O outer Gibbs samples from the distribution of pedigree genotypes unconditional on their phenotypes.

Step 2. Assigning the phenotypes consistent with each set of genotypes generated in step, so that we have n_O sets of phenotypes for the pedigree.

Step 3. For each set of phenotypes produced in step 2, n_I inner Gibbs samples are generated from the conditional distribution of genotypes on phenotypes.

Step 4. Using the realizations generated in previous steps to calculate n_O values of score and likelihood ratio statistics to estimate the distribution of those statistics.

APPROXIMATE CONFIDENCE INTERVAL FOR RECOMBINATION FRACTION

Approximations to the distribution of an estimate of test statistic may be used for generating tests of hypotheses and confidence intervals. When calculating likelihoods on pedigrees, we can approximate the distribution of an estimate of θ, the parameter of interest and an ACI for it. Assuming we are interested in calculating the distribution of an estimate of recombination (θ) and its ACI, it can be done based on the following stepwise procedure; also assuming the true value of θ has been estimated by $\hat{\theta}_m$.

Step 1. Calculating the likelihood ratio at two more points $\hat{\theta}_l$ and $\hat{\theta}_r$, where $\hat{\theta}_l$ and $\hat{\theta}_r$ stand for the left and right equidistant values of $\hat{\theta}_m$.

Step 2. Since the likelihood ratio at $\hat{\theta}_m$ is 1, we can fit a quadratic polynomial to the three points $\left(\hat{\theta}_l, \frac{L(\hat{\theta}_l)}{L(\hat{\theta}_m)}\right), (\hat{\theta}_m, 1)$ and $\left(\hat{\theta}_r, \frac{L(\hat{\theta}_r)}{L(\hat{\theta}_m)}\right)$. So, the equation would be of the form:

$$A\theta^2 + B\theta + C. \quad (1)$$

Step 3. Maximizing the quadratic polynomial with respect to θ yielding:

$$\theta^* = \frac{\alpha_1 \hat{L}_l + \alpha_m \hat{L}_m + \alpha_r \hat{L}_r}{\beta_1 \hat{L}_l + \beta_m \hat{L}_m + \beta_r \hat{L}_r}.$$

Since we have n_O different sets of

$$\{(\theta_l, l(\theta_l, \theta_m)), (\theta_m, l(\theta_m, \theta_m)), (\theta_r, l(\theta_r, \theta_m))\},$$

where $l(.,.)$ denotes the likelihood ratio, Equation (1) yields n_O values of θ^* since the distribution of $\hat{\theta}$ can be estimated. Once the approximate empirical distribution of

$\hat{\theta}$ is built, we can use it to calculate its quantiles and hence an ACI.

SIMULATION SIZE FOR CALCULATION OF THE DISTRIBUTION OF TEST STATISTICS

Based on the procedures shown earlier, we can provide some ideas of the simulation sizes required for both inner and outer Gibbs samples. Suppose:

$$\hat{l} = l + e \qquad (2)$$

Where, l is the true value of likelihood ratio and e is the random error. Substituting (2) in (1) and after some algebra, we have:

$$\theta^* = \left(1 + \frac{\alpha_l e_l + \alpha_m e_m + \alpha_r e_r}{\alpha_l l_l + \alpha_m l_m + \alpha_r l_r}\right)\left(1 + \frac{\beta_l e_l + \beta_m e_m + \beta_r e_r}{\beta_l l_l + \beta_m l_m + \beta_r l_r}\right)^{-1} \qquad (3)$$

Using the expansion of $(1+x)^{-1}$ for the second term in the right hand of Equation (3):

$$\left(1 + \frac{\beta_l e_l + \beta_m e_m + \beta_r e_r}{\beta_l l_l + \beta_m l_m + \beta_r l_r}\right)^{-1} \approx 1 - \frac{\beta_l e_l + \beta_m e_m + \beta_r e_r}{\beta_l l_l + \beta_m l_m + \beta_r l_r}$$

we will have:

$$\theta^* = \hat{\theta} + \eta l$$

Where $\eta = (\eta_l, \eta_m, \eta_r)$ and $l = (l_l, l_m, l_r)'$

Assuming the quadratic approximation to l is plausible and assuming the approximate independence of $\hat{\theta}$ and ηl:

$$\mathrm{var}(\hat{\theta}) = \mathrm{var}(\theta^*) + \mathrm{var}(\eta l)$$

Since

$$\mathrm{var}(\eta l) = \eta' \sum_l \eta$$

Where $\sum_l$ is the variance-covariance matrix of l_l, l_m and l_r and again assuming the independence of $\mathrm{var}(\theta^*)$ and $\eta' \sum_l \eta$ we have:

$$\hat{\mathrm{var}}\left(\hat{\mathrm{var}}(\hat{\theta})\right) \approx \frac{S^2_{\theta^*}}{n_O} + \frac{\eta' \sum_l \eta}{n_I} \qquad (4)$$

To minimize Equation (4) with an additional condition of $n_O n_I = K$, where K is an arbitrary total number of Gibbs samples, we will have:

$$n_O \approx \sqrt{\frac{S^2_{\theta^*}.K}{\eta' \sum_l \eta}} \qquad (5)$$

and

$$n_I \approx \sqrt{\frac{\eta' \sum_l \eta.K}{S^2_{\theta^*}}}. \qquad (6)$$

So, if we want to generate a total K Gibbs realizations to estimate the distributions of score and likelihood statistics as well as building an ACI for recombination fraction, Equations (5) and (6) determine how many inner and outer Gibbs samples will be suitable. Of course, before determining these sample sizes, we require a trial outer and inner Gibbs samples based on which the calculation of n_O and n_I is made.

Conclusion

Recombination fraction is an important parameter in genetic linkage analysis. To have an estimate of the recombination fraction, one needs to use simulation analysis due to existence of the huge number of gene configurations and the amount of computation involved. The two stage Monte Carlo method proposed in this paper can aid researchers to calculate an approximate confidence limit for recombination fraction. The size of simulation is another issue in the computations. This paper gives some idea about the size of inner and outer Gibbs samples. This area still needs more work since there are many more complicated problems in pedigree analysis such as looped pedigrees in which one needs to do simulations and the simulation size may increase drastically.

REFERENCES

Abraham KJ, Totir LR, Fernando RL (2007). Improved techniques for sampling complex pedigrees with the Gibbs Sampler. Genet. Sel. Evol. 39: 27-38.

Cannings C, Thompson E A, Skolnick MH (1978) Probability functions on complex pedigrees. Adv. Appl. Prob. 10: 26-61.

Elston RC, Stewart J (1971). A general model for genetic analysis of pedigree data. Hum. Hered. 21: 523-542.

Geman A, Geman D (1984).Stochastic relaxation, Gibbs distributions and the bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence. 6: 721-741.

Jandaghi GH (1994) Monte carlo estimation of the pedigree likelihood statistics using Gibbs Sampler. Ph.D. thesis, Department of Preventive Medicine and Biostatistics, University of Toronto, Canada.

Jandaghi Gh(2010) A new algorithm for setting initial values for Markov Chain Monte Carlo in genetic linkage analysis via Gibbs sampling, Sci. Res. Essays, 5(22): 3447–3454.

Janss LLG, Van Arendonk JAM, Van Der Werf JHJ (1995) Computing approximate monogenic model likelihoods in large pedigrees with loops. Genet. Sel. Evol. 27: 567-579.

Lange K, Boehnke L (1983) Extensions to pedigree analysis V. optimal calculations of Mendelian likelihoods. Hum. Hered. 33: 291-301.

Sheehan N, Thomas A (1993). On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. Biometrics. 49:163-175.

Stricker C., Fernando RL, Elston RC (1995). An algorithm to approximate the likelihood for pedigree data with loops by cutting. Theor. Appl. Gen. 91: 1054-1063.

Thomas A (1986a). Approximate computations of probability functions for pedigree analysis. IMJ J. Math. Appl. Med. Biol. 3: 157–166.

Thomas A (1986b). Optimal computations of probability functions for pedigree analysis. IMJ J. Math Appl. Med. Biol. 3: 167-178.