

BIOINFORMATICS APPLICATIONS NOTE

GeneRecon—A coalescent based tool for fine-scale association mapping

Thomas Mailund^{a,b,d,*} Mikkel H. Schierup^{a,b} Christian N.S. Pedersen^{a,b,c}
 Jesper N. Madsen^b Jotun Hein^d and Leif Schauer^{a,b}

^aBioinformatics Research Center, University of Aarhus, Høegh-Guldbergs Gade 10, DK-8000 Århus C, Denmark ^bBioinformatics ApS, Høegh-Guldbergs Gade 10, DK-8000 Århus C, Denmark ^cDept. of Computer Science, University of Aarhus, IT-Parken, DK-8200 Århus N, Denmark ^dDept. of Statistics, University of Oxford, Oxford, OX1 3TG, UK

Associate Editor: Martin Bishop

ABSTRACT

Summary: GeneRecon is a tool for fine-scale association mapping using a coalescence model. GeneRecon takes as input case-control data from phased or unphased SNP and micro-satellite genotypes. The posterior distribution of disease locus position is obtained by Metropolis Hastings sampling in the state space of genealogies. Input format, search strategy, and the sampled statistics can be configured through the Guile Scheme programming language embedded in GeneRecon, making GeneRecon highly configurable.

Availability: The source code for GeneRecon, written in C++ and Scheme, is available under the GNU General Public License (GPL) at <http://www.birc.au.dk/~mailund/GeneRecon>.

Contact: mailund@birc.au.dk

1 INTRODUCTION

We have implemented a software package, *GeneRecon*, based on extensions of the shattered coalescence model of Morris et al. (2002) for Bayesian Markov-chain Monte Carlo (MCMC) fine-scale linkage-disequilibrium (LD) gene mapping. GeneRecon uses the coalescent model (Hein et al., 2005) to explicitly model the genealogy of a sample of case chromosomes. The location of the mutation influencing the disease is inferred based on the observed linkage disequilibrium at multiple genetic markers. Given the computational complexity of the problem, a Metropolis-Hastings algorithm is deployed to integrate over unknown population genetic parameters of the coalescence model and sample the marginal posterior probability density for the parameter(s) of interest.

2 THE MODEL

GeneRecon handles both SNP and microsatellite marker genotype or haplotype data from case/control design studies. Phenocopies and locus- and allele-heterogeneity are modeled in two ways. Firstly, the “shattered” coalescent allows genealogical independence of coalescent subtrees (Morris et al., 2002). Secondly, cases are partitioned into two clusters, a “null”-cluster which is not evaluated by the model, and hence greatly reduces the search space, and a “mutation”-cluster of cases which is evaluated by the model (Liu et al., 2001).

3 IMPLEMENTATION

GeneRecon can be obtained from its homepage, where instructions for the installation are provided. The MCMC engine of GeneRecon is written in C++ and is available as a command-line executable for Linux. A “Getting started” document provides an introduction to the functionality of GeneRecon, whereas a users manual provides examples of more advanced uses, including examples of using Guile Scheme.

4 FLEXIBILITY OF USING SCHEME

Using the Guile Scheme programming language as a front-end for input specifications and execution control allows a highly flexible interaction with the MCMC engine of GeneRecon. A collection of Guile modules allow easy changes to functionality specifications. Input file format, population genetic parameters, MCMC sampling strategy, and output options can be configured. Prior knowledge of population genetic parameters, such as effective population size (N_e) or local recombination rates (ρ) may be explicitly defined, if available from independent sources such as HapMap or the DeCODE genetic map. Presently, sampling of the likelihood, disease location, effective population size, coalescent tree and cluster indices are supported. The MCMC sampling strategy is defined by the number of iterations, the burn-in period, and the proposal densities of the sampled parameters. The choice of strategy will strongly affect the mixing properties of the Markov Chain and convergence to a stationary distribution (for details on MCMC strategies see Gilks et al. (1995) or Liu (2001)).

5 PERFORMANCE

To evaluate the prediction capabilities of GeneRecon, we have conducted a large simulation study (Mailund et al., 2005a) where sequence data was simulated under various parameters using the CoaSim tool (Mailund et al., 2005b), and then analysed using GeneRecon, with four independent runs for each dataset. Results for a Mendelian scenario i.e. all cases carry the disease causing mutation and all controls are wild-types, are shown in Table 1.

We have also tested GeneRecon on the $\Delta F508$ mutation for cystic fibrosis data from Kerem et al. (1989). The results from this analysis is shown in Figure 1. GeneRecon compares favorably in comparison with other fine-mapping tools (Table 2).

*to whom correspondence should be addressed

Setup	All diseased			50% diseased		
	50%	75%	95%	50%	75%	95%
20 Markers, 1cM	0.034 cM	0.114 cM	0.358 cM	0.052 cM	0.190 cM	0.777 cM
40 Markers, 1cM	0.058 cM	0.241 cM	0.564 cM	0.137 cM	0.377 cM	0.727 cM
20 Markers, 2cM	0.052 cM	0.154 cM	0.499 cM	0.113 cM	0.348 cM	0.721 cM
20 Markers, 0.1cM	0.0043 cM	0.0073 cM	0.0532 cM	0.0062 cM	0.0021 cM	0.0372 cM
40 Markers, 0.1cM	0.0072 cM	0.0269 cM	0.0532 cM	0.0126 cM	0.0302 cM	0.0650 cM
20 Markers, 0.2cM	0.0057 cM	0.0283 cM	0.0759 cM	0.0084 cM	0.0267 cM	0.0618 cM

Table 1. GeneRecon error, as measured by distance from the inferred to the true position of the disease locus of simulated haplotype SNP data from 100 cases and controls. The table shows the number of chains inferring a disease locus in the given distance from the true locus at the 50%, 75% and 95% quantile, respectively. Six setups were explored in two GeneRecon settings. In setting one, all cases were considered by the coalescence model, whereas in setting two the fraction of cases evaluated by the model was 50%. In the first setup a region of 1 cM is covered by 20 markers of minor allele frequency (MAF) > 10% and a trait locus with disease allele frequency of 20%, all placed at random, in the second setup the density of markers is doubled and in the third setup region is twice the size. Setup four to six follow this pattern, but in a smaller region (0.1 for setup four and five and 0.2 cM for setup six). That the accuracy is decreased when the number of markers are doubled, as seen in setup two and five, may seem counter-intuitive, but is a consequence of the increase in the search space. GeneRecon explored the search spaces of all setups for four CPU-days each, regardless of the number of markers, so doubling of this number leads to the evaluation of a smaller fraction of the search space, causing the reduction in accuracy.

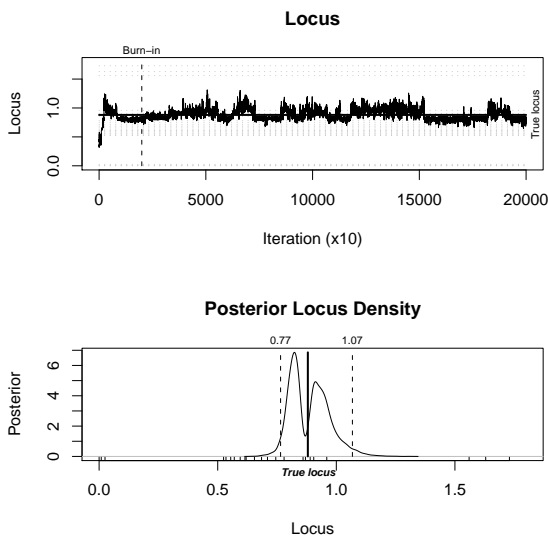


Fig. 1. Example of GeneRecon inferring the position of the $\Delta F508$ mutation for cystic fibrosis with data from Kerem et al. (1989). The MCMC was allowed to burn in for 20,000 iterations and the posterior was sampled from the following 180,000 iteration. In the posterior plot, the true position is indicated by the solid vertical line and the 95% credibility interval is indicated by dashed vertical lines. Ticks at the x-axis indicate the position of SNP markers.

6 CONCLUSION

GeneRecon is designed to allow flexible multimarker LD mapping using coalescent model. Adaptations and extensions of the various Scheme modules provided allow users to accommodate a wide range of scenarios, data types, sampling strategies and convergence

diagnostics, without much loss of user friendliness compared to competing software.

Method	Estimate	95% credibility interval
Liu et al. (2001)	0.87	0.82–0.93
Morris et al. (2002)	0.85	0.65–1.00
GeneRecon	0.82	0.73–1.03

Table 2. Comparison of location estimates of the $\Delta F508$ mutation for cystic fibrosis data Kerem et al. (1989) by GeneRecon and other coalescent-based fine mapping tools. The mutation is located at 0.88.

REFERENCES

- W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, (1995).
- J. Hein, M. Schierup, and C. Wiuf. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, (2005).
- B. Kerem, J. M. Rommens, J. A. Buchanan, D. Markiewicz, T. K. Cox, A. Chakravarti, M. Buchwald, and L. C. Tsui, (1989), Identification of the cystic fibrosis gene: genetic analysis. *Science*, 245, 1073–1080.
- J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, (2001).
- J. S. Liu, C. Sabatti, J. Teng, B. J. Keats, and N. Risch, (2001), Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res.*, 11, 1716–1724.
- T. Mailund, C. N. S. Pedersen, J. Bardino, B. Vinter, and H. H. Karlsen, (2005), Initial experiences with GeneRecon on MiG. In *Proceedings of The 2005 International Conference on Grid Computing and Applications (GCA'05)*.
- T. Mailund, M. Schierup, C. N. S. Pedersen, P. Mechlenborg, J. Madsen, and L. Schauer, (2005), CoaSim: A flexible environment for simulating genetic data under coalescent models. *BMC Bioinformatics*, 6(252).
- A. P. Morris, J. C. Whittaker, and D. J. Balding, (2002), Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *Am J Hum Genet*, 70, 686–707.

ACKNOWLEDGEMENTS

TM is funded by the Danish Research Agency, FNU grant 272-05-0283 and FTP grant 274-05-0365. The project is supported by the ISIS project 123 to LS and CNSP.