

Efficient sampling of ancestral states in the infinite site model

5.7.09

Background and Motivation. Given a sample of sequences from a population, it is important to calculate the probability that they evolved from a known ancestor in order to make estimates of mutation rates, and to produce accurate models of population structure, selection and recombination. The infinite sites model is commonly used to model mutations in sequences. Together with the coalescent model, it describes the evolution of infinitely long sequences. Between 1987 and 1995, Ethier, Griffiths and Tavaré published a series of recursions for calculating these probabilities under the infinite sites model. However, exact implementation of these recursions is extremely slow for large data sets, and approximations must be made in order to accelerate calculations whilst maintaining reliability of results.

Previous work on approximation techniques for these models focused on a classical Monte Carlo approximation scheme, importance sampling (Liu, 2001). Importance sampling requires a proposal distribution to be defined over histories, that is easy to sample from. The efficacy of the algorithm (measured in terms of the so called effective samples size, ESS) critically depends on the similarity between the proposal and the true distribution of histories. Until recently series of proposal distributions have been published (Felsenstein et al., 1999; Stephens and Donnelly, 2000; Hobolt et al., 2008, to appear; Andreassen and Okholm, 2008), but even with using the most advanced methods importance sampling cannot provide satisfying results for large datasets.

Another widely used class of Monte Carlo sampling techniques are Markov chain Monte Carlo (MCMC) techniques (Liu, 2001), which usually provide somewhat more reliable results than importance sampling. We are not aware of any published work investigating MCMC sampling for the models concerned here.

The infinite site model clearly falls into the class of self-reducible counting problems: the likelihood of each potential history can be calculated quickly, the hard counting problem is the summation of the likelihoods over all possible history. Griffiths and Tavaré (1995) gave a recursion for the exact calculation of the likelihood, however, this algorithm runs in exponential time in worst case. Since the problem is self reducible, any FPAES (Fully Polynomial Almost Exact Sampler) for the possible evolutionary hystories would provide an FPRAS (Fully Polynomial Randomized Approximation Scheme) for the total likelihood and *vice versa*, a FPAES could be constructed from any FPRAS. An FPAES is an algorithm whose input is a data sequence x and an $\epsilon > 0$ and generates random samples from a distribution p_ϵ such that

$$d_v(p_\epsilon, \pi_x) \leq \epsilon$$

where d_v is the variational distance, π_x is a distribution defined by x , and the algorithm's running time is polynomial in both $|x|$ and $-\log(\epsilon)$. An FPRAS is an algorithm with inputs x , δ and ϵ , and generates a random estimate $\hat{\Theta}_{x,\delta,\epsilon}$ for some Θ_x such that

$$P\left(\frac{\Theta_x}{1+\epsilon} \leq \hat{\Theta}_{x,\delta,\epsilon} \leq \Theta_x(1+\epsilon)\right) \geq 1-\delta$$

and the running time is fully polynomial with $|x|$, $1/\epsilon$ and $-\log(\delta)$. For more details see the introductory lecture by vigoda on the web: http://www.cc.gatech.edu/~vigoda/MCMC_Course/Sampling-Counting.pdf

Goals of the project. The main aim of the project is to give lower and upper bounds on the goodness of techniques that try to estimate the total likelihood of some data in the infinite site model (or equivalently, samples histories and ancestral states from a distribution that is proportional to their likelihood). Recently Miklos (<http://ramet.elte.hu/~miklosi/msis.pdf>) gave upper and lower bounds on the relaxation time of an MCMC that samples from a distribution that is described by a derivation tree (Sinclair & Jerrum, 1989). The central property here is the hiddenness of the derivation tree. If the hiddenness of the derivation tree might grow exponentially with the data size, then the Markov chain will have torpid mixing, on the other hand, if the hiddenness rate has an upper bound that is polynomial with the data size, then the relaxation time of the Markov chain is subexponential. Therefore the first central question is what the hiddenness rate of the derivation tree is under some particular auxiliary distributions. There are several natural candidates for these auxiliary distributions: the partial likelihoods of the prefixes of histories, the number of possible histories with the so far observed prefixes, any other auxiliary distribution coming from the existing scientific literature on Sequential Importance Sampling.

Miklos *et al.* (2009) recently proved that in case of genome rearrangement, there might be arbitrary big gaps in the network spanned by the evolutionary histories, and hence describing the solutions with a

derivation tree instead of a network does not help for a particular Markov chain. However, Miklos and Darling (2009) showed that additional transition kernels in the Markov chain might help in the mixing even if the hiddenness rate grows exponentially with the input data size. There are natural transition kernels that perturb a small part of the actual history, see for example Mithani *et al.* (2009). It is a natural question how big gaps exist in the space of evolutionary histories under the infinite site model? If the size of the gap grows at most logarithmically with the data size, then a transition kernel with small perturbations could yield an irreducible chain. If the diameter of the Markov chain is small, then a Markov chain with such transition kernels would be a candidate for a rapidly mixing Markov chain since the backproposal probabilities would remain large due to the small perturbations. Note that the small backproposal probabilities cause the slow convergence of a Markov chain on genome rearrangement histories (Miklos *et al.*, 2009).

Plan.

- Read the corresponding literature
- Implement the MSIS Markov chain, infer the mixing via how quickly a path is grown from the root to a leaf of the derivation tree
- Give theoretical upper and lower bounds on the hiddenness rate of the derivation tree
- Design and implement an irreducible Markov chain on the evolutionary histories. Infer the mixing of the Markov chain based on some traditional measurement (loglikelihood trace, autocorrelation, etc.)
- Try to give lower and upper bounds on the relaxation time of the designed Markov chain

References:

- Andressen CM, Okholm A. 2008. From exact marginals to better importance sampling. Technical report, Genome Analysis and Bioinformatics Group, Department of Statistics, University of Oxford.
- Felsenstein J, Kuhner MK, Yamato J, Beerli P. 1999. Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. In: Seillier-Moisewitsch F (ed.), *Statistics in Molecular Biology and Genetics*, Institute of Mathematical Statistics and American Mathematical Society, pp. 163–185.
- Griffiths RC. 1989. Genealogical-tree probabilities in the infinitely-many-sites model. *Journal of Mathematical Biology* 27:667–680.
- Griffiths RC, Ethier S. 1987. The infinitely-many-sites model as a measure valued diffusion. *Ann Prob* 15:515–545.
- Griffiths RC, Tavaré S. 1995. Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Mathematical Biosciences* 127:77–98.
- Hein J, Schierup MH, Wiuf C. 2005. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press.
- Hobolt A, Uyenoyama MK, Wiuf C. 2008(to appear). Importance sampling for the infinite sites model. *Statistical Applications in Genetical Molecular Biology* 7:electronic only.
- Lyngsø RB, Song YS, Hein J. 2008. Accurate computation of likelihoods in the coalescent with recombination via parsimony. In: Vingron M, Wong L (eds.), *RECOMB2008*. Springer, volume 4955 of *Lecture Notes in Computer Science*, pp. 463–477.
- Miklós, I., Mélykúti, B., Swenson, K. (2009) The Metropolized Partial Importance Sampling MCMC mixes slowly on minimum reversal rearrangement paths *ACM/IEEE Transactions on Computational Biology and bioinformatics*, accepted.
- Miklós, I., Darling, A. (2009) Efficient sampling of parsimonious inversion histories with application to genome rearrangement in *Yersinia* *Genome Biology and Evolution*, advance published, doi:10.1093/gbe/evp015
- Sinclair A, Jerrum M. 1989. Approximate counting, uniform generation and rapidly mixing Markov chains. *Inf Comput* 82:93–133.
- Song YS, Lyngsø R, Hein J. 2006. Counting all possible ancestral configurations of sample sequences in population genetics. *IEEEACM Trans Comput Biol Bioinformatics* 3:239–251.
- Stephens M, Donnelly P. 2000. Inference in molecular population genetics. *Journal Of The Royal Statistical Society Series B* 62:605–635.