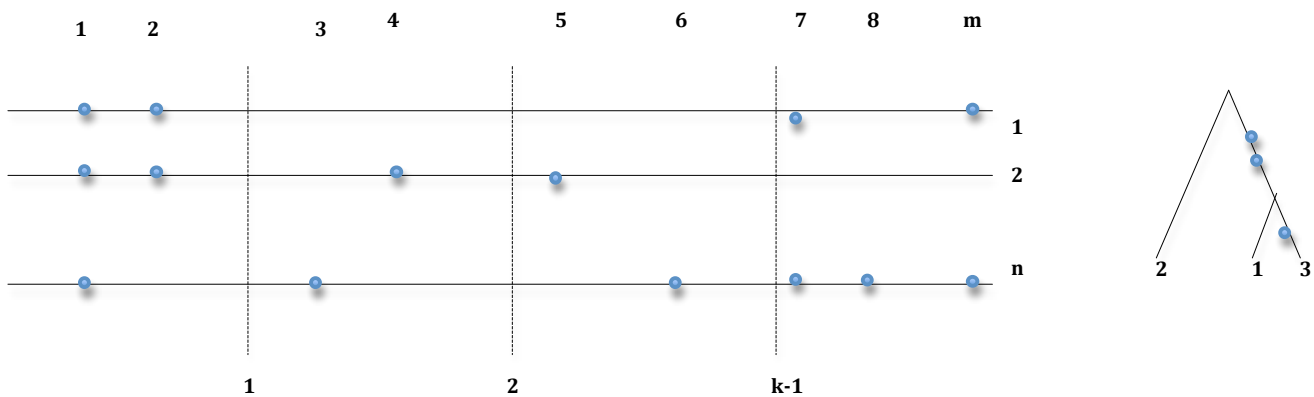


From exact marginals to good importance sampling

15.6.08

Motivation and Background. DNA sequences from a population has an unobservable genealogical history. In analyzing such data, a crucial stepping stone is to be able to integrate over evolutionary histories according to their probability according to a model and given the data. Doing this has been the focus of research for more than 2 decades. The basic probability model for genealogical histories without recombination was given Watterson (1975) and Kingman (1982). Until 1994 (Griffiths and Tavaré), this was solely used as a tool for simulating genealogical histories without knowing the content (mutational configuration) of the sequences. Since late 90s there has been a string of attempts to apply stochastic integration methods (Importance Sampling, MCMC,...) to do this. In the absence of recombination it is a hard, but doable problem. Due to the enormous increase in DNA sequences from populations and the importance of this problem in genetic mapping, the problem remains as important as ever. Major contributions to this problem can be found in Stephens and Donnelly (2000) and Hobolt et al.(2008).

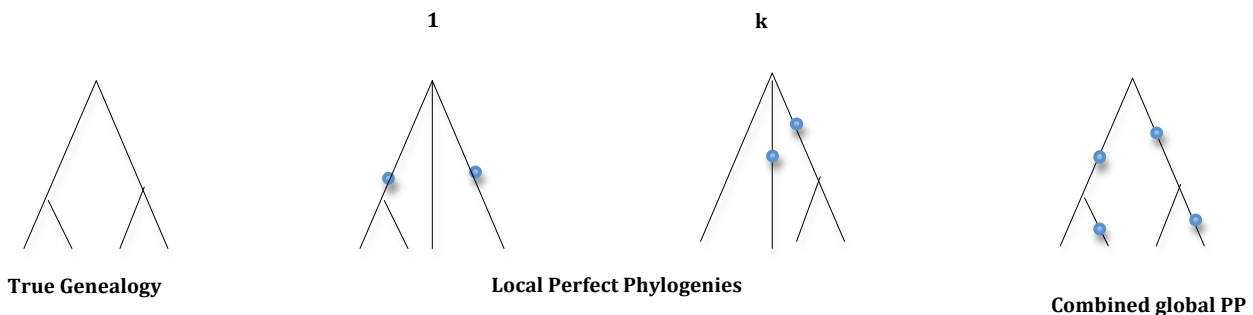
Recursions (Ethier and Griffiths, 1987) can be written to calculate the likelihood of a data set in terms of possible ancestral configurations (Song, Lyngsø and Hein, 2006). The number of ancestral configurations grows as a function of number of sampled sequences and segregating sites but also depends on the exact configuration of segregating sites. A data set could look like below and it was known that the most recent ancestral sequence to all the sequences was “blue ball” free, then all possible histories could be obtained by taking the data set and removing blue balls that was alone on their position (mutations happens in a single sequence) and merging (coalescing) identical sequences.



In this illustration (left part) we have n sequences (3 shown), m segregating are observable and the sequences have been cut into k (4) pieces by 3 vertical lines. Mutations are shown as blue balls and it is often assumed that “blue ball” is derived state and the ancestor sequence did not have this at this position. It is also possible allow more than two states at each position. The relationship of the sequences is the same for all positions and the positions of the mutations are without information. It is thus possible to summarize the data in a “gene-tree” as illustrated to the right for the right most segment as if the data only had three sequences ($n=3$). Sequences 1 and 3 share two mutations that are place on branch above them. Sequence 3 has a single private mutation on its own branch. Sequence 2 has no mutation on it and is thus identical to the ancestral sequence (root) within this segment. Sequence data with only two states at each position and no *incompatibilities* (no two positions where all four combinations of (ball, no ball)² have been observed) can always be brought on gene-tree form. This is not surprising: a mutation can only happen once and will thus be associated 1 branch with the blue ball states below the mutation event. Since all positions are compatible, there will be no conflict between different positionings of mutations.

Proposed Project. The basic idea of this project is to use cases where exact computations can be performed maximally to construct either a pseudo-likelihood function for the data (Hudson, 2001) or an importance sampler. We will only consider the case without recombination. In this case, we think this approach could have important potential and lead to the ability to analyze data sets of fully realistic size, for instance 100 segregating sites and many hundred samples.

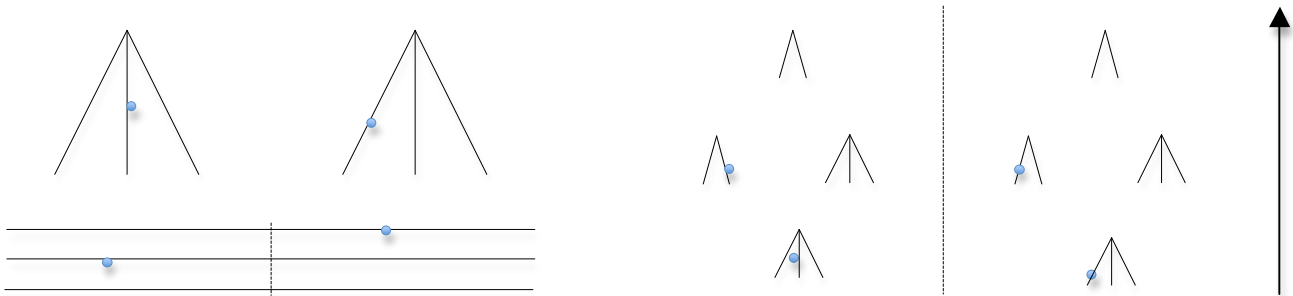
The factors/parameters important for the difficulty in calculating the likelihood of a data set are: the sample size, n , the number of segregating sites, m , the ratio scale mutation rate to length of sequence, θ/L and the “complexity” of the data set.



There is one common globally correct genealogy (left) for the complete sequences that is unobservable. Methods exists that infer and characterize distributions for perfect phylogenies PP (phylogenies with mutations perfectly assigned to internal branches) for data sets of moderate size and the distributions of branch lengths and mutation positions. Such methods can be used on segments of suitable size giving local perfect phylogenies. The local perfect phylogeny distributions can be used to evaluate the global genealogy and place all mutations on the branches of this global genealogy.

Say that we want to calculate the likelihood for the above data, but the set of ancestral configurations are too large, but we can calculate the likelihood of k equally large sub-regions, then we could calculate exactly the probability of the genealogical histories for these sub-regions. The sub-regions should be chosen independently of the data. This cannot be done perfectly, but for instance θ could be estimated using Watterson's estimator (Watterson, 1975) and k chosen so the regions would only have analyzable size. The basic events in evolutionary histories – mutations and coalescent events – are different in their roles. Mutations are private to a segment, while coalescent events are common to all segments (in absence of recombination/gene conversion). Calling mutations private to a segment is not strictly true as the bipartition (edge in tree) corresponding to a mutation will be present on the global tree, but it not possible in an easy way to calculate the distributional consequences of a mutation on one segment for the branch length distribution of gene trees on other segments. In the limit (k big) one could imagine that one had perfect knowledge of the coalescent events and mutation events were placed on branches of known length. In the infinite site model, the mutations could unambiguously be placed on branches and would have uniform distributions.

Combining local exact distributions to a global distribution is simpler when using the infinite site assumptions, since mutations in each segments gives perfect information on the global genealogy. Without the infinite site assumption, the problem becomes slightly more difficult and will not be considered at present. To define the IS a few issues will have to be solved:



A toy example illustrating how to combine calculations from different segments. To the left a data example that we pretend is too big to be calculated exhaustively and is cut into two segments. Each segment can be represented as a gene tree. To the right. The gene trees for each segment represent symmetric cases and only one will be considered. Such a gene tree has three leaves and one branch with a mutation. In tracing its history, three events must occur – one mutation (removal of blue ball) and 2 merging of ball free edges. Starting at the bottom gene tree, there are two choices: merge two edges into one edge or remove blue ball leading to two possible ancestral states (middle set to the right). The next step will be either coalescent or mutation and will lead to the same ancestral state (two identical sequences). These will coalesce, leading to the tree consisting of a single point. In the segments, the probability of the next step can be calculated perfectly, but for the whole sequences this is not possible. A natural choice of the probability for the next step of the complete sequence would be to approximate the probability, by the product of the probabilities calculated at the segments. For a complete sequence, a set of events are possible and we will let them be weight proportionally to the probabilities they would get in segments they would happen.

How to solve EGT recursion at different intervals, when they all have to obey to total set of observed bi-partitions. This can be done in two ways: The easiest way is to prohibit certain paths/ancestral states at each segment if they would postulate bipartitions conflicting with the total set. This corresponds to observing the data at the segment and additionally some bipartitions (edges in tree). Alternatively, one can condition on the observed bipartition and then the fundamental EGT recursions will have to be rederived.

Plan

1. Read Hein, Schierup and Wiuf (2005) chapt 1-3 and Griffiths (2001). Design data structure for AC. Decide programming language.
2. Present project. Implement EGT recursions when given a set of known bi-partitions.
3. Build importance samplers based on approximate marginals
4. Experiment with software.
5. Rederive EGT with known bi-partitions and check how sampling is affected.
6. Finish report and program.

Comments. It is natural to study this problem in its simplest form first (ancestor state known, infinite site assumption, no recombination), but relaxing the first two assumptions will make a much more practical model and is in principle straightforward. Allowing the presences of recombination could possibly be done using methods presently being developed by Paul Jenkins and Bob Griffiths. This project ties in well with the project "Corner Cutting Approaches to the EGT recursions"

References:

- Ethier and Griffiths (1987) "The infinitely many sites model as a measure valued diffusion" *Ann. Prob.* 15:2.515-545.
 Griffiths, R.C. (1987). "Counting genealogical trees" *J. Math. Biol.* 25, 422-432.
 Griffiths, R.C. (1989). "Genealogical-tree probabilities in the infinitely-many-sites model". *J. Math. Biol.* 27, 667-680.
 Griffiths, R.C. and Tavaré, S. (1995). "Unrooted genealogical tree probabilities in the infinitely-many-sites model". *Mathematical Biosciences* 127, 77-98.
 Griffiths, R.C. (2001). "Ancestral inference from gene trees" In: Donnelly, P. and Foley, R. (Eds.), *Genes, Fossils, and Behaviour: an Integrated Approach to Human Evolution*, IOS Press, pp.137-172.
 Hein, Schierup and Wiuf (2005) "Gene Genealogies, Variation and Evolution" OUP.
 Hobolth, Uenoyama and Wiuf (2008) "Importance Sampling for the Infinite Sites Model" manuscript in preparation
 Hudson (2001) Two-locus sampling distributions and their application. *Genetics* 159: 1805-1817
 Kingman (1982) "The Coalescent" *Stochastic Processes and their Applications*, 13:235-248
 Lyngso, Song and Hein (2008) "Accurate calculations of likelihoods in the coalescent with recombination using parsimony" *Recomb 2008 Singapore*
 Song, Y., R. Lyngso & J. Hein (2006) "Counting Ancestral States in Population Genetics" *Bioinformatics and Computational Biology* vol.3.3.239-252
 Stephens, M. and Donnelly, P. (2000). Inference in Molecular Population Genetics. *Journal of the Royal Statistical Society, Series B*, 62, 605–655
 Watterson (1975) 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7, pp. 256–276.