Random trees in molecular genetics

Simon Tavaré
Program in Molecular Biology
Department of Biological Sciences
University of Southern California
Los Angeles, CA 90089-1340, U.S.A.
stavare@gnome.usc.edu

1. Introduction

This paper describes three examples of statistical inference problems involving molecular data and branching processes. The three examples illustrate branching on widely diverse time scales: reconciling molecular and fossil estimates of divergence times, ancestral inference in population genetics, and reconstruction of tumor history within an individual. The common theme is the development of novel computational inference methods for branching processes, in particular for non-standard sampling schemes.

2. Molecular clocks and divergence times

Evolutionary biologists have long used molecular clocks to date the divergence times of taxonomic groups. Often these estimates predate the oldest fossil found in the groups, leading to a flurry of interest in assessing why fossil-based estimates and molecular-based estimates differ; see Foote *et al.* (1999) for a discussion. For statisticians interested in branching processes this problem raises a number of interesting questions. Suppose then that (on some convenient time scale) the species of interest have been branching according to a Galton-Watson branching process $\{Z_n, n \geq 0\}$ with $Z_0 = 1$, and a given offspring distribution. The simplest inference problem, discussed by Stigler (1970), is to estimate the age of the process given an observation on the current number, s, of extant species. Stigler's approach uses $P(Z_n = s | Z_n > 0)$ as the likelihood, and maximizes over n.

The sampling in the fossil context is more complicated: imagine time divided into a number of geologically defined bins $B_1, \ldots, B_r$, with the rth bin ending g generations before present. We search for fossils in each bin, recording fossils from Y_1 distinct species in the first bin, $\ldots$, Y_r in the rth bin, and no fossils older than g generations before present. The problem is to estimate the time back from g to the founder species using the observations $Y_1, \ldots, Y_r$, the number of species currently alive, and the fact that no species older than g generations was observed. Assuming even simple models for how the sampling distribution of the Y_i depends on the Z_i, it seems to be very difficult to find analytical results for the estimator of the age of the process, but a number of computational likelihood approaches seem promising. To address some of the realities of speciation we have used a non-homogeneous Markov branching process. For such models, we have developed a simpler fitting method that does not require detailed assumptions about the structure of the offspring distributions over time; see Tavaré *et al.* (1999).

3. Molecular population genetics

A number of authors have addressed questions concerning ancestral inference based on samples of DNA sequence data from natural populations. A typical problem is: given a model for the ancestral relationships among the sequences in the sample, what can be said about the time to the most recent common ancestral sequence of that sample, or of the population that was

sampled? Once more, such problems pose a number of interesting computational and statistical issues; for the flavor of this, see for example Part I of Donnelly and Tavaré (1997).

Recently, several authors have addressed the problem of estimating the age of a neutral mutation. There are two rather different approaches to this problem. In the first, the age of the mutation is treated as a parameter in a model for the evolution of that mutation in the population that was sampled (cf. Thompson (1976), Slatkin and Rannala (1997)); the age is then estimated by (for example) likelihood methods. In the second, the age of the mutation is an unobserved random variable; what is then reported is the conditional distribution of the age given data from the sample. For an analysis of this problem in the coalescent setting, when the data is the number of copies of the mutant observed in a random sample of n chromosomes, see Griffiths and Tavaré (1998), Wiuf and Donnelly (1999) and Stephens (1999). Slatkin and Rannala (1997) suppose that further molecular information is obtained *just for those individuals carrying the mutation*, and the estimate of the age of the mutation is then modified to account for this extra information. In the coalescent setting, this problem raises some interesting computational issues. When the individuals carrying the mutation are sequenced in the region of the mutation, a Markov chain Monte Carlo approach can be used to approximate the conditional distribution of the age of the mutation.

4. Tumor cell lines

The third example involves CA repeat microsatellite (MS) loci on the X chromosome of male patients with hereditary nonpolyposis colorectal cancer (HNPCC). Since individual cells cannot readily be isolated from fixed tissue, the experimental approach isolates DNA from specific tumor regions on a microscope slide. The isolated DNA is fragmented in such a way that the MS loci are physically separated from each other. DNA is sampled at random from this pool, and single loci are typed using the polymerase chain reaction. In this setting, typing results in a measurement of the length of the CA repeat at each locus. The process is repeated from the same pool until MSs at several loci are typed; see Tsao *et al.* (1999a) for further details.

Since we use X chromosome MS loci from males, we can use the measurements of CA-repeat lengths at the different loci to reconstruct aspects of the history of the cell lineage that led to the sample. The length of each CA repeat region of interest in the initial cell of the lineage is known. This cell has undergone loss of mismatch repair (MMR), leading to rapid changes in the lengths of the MS loci as the cell line evolves. We model the evolution of the cell line by a branching process in which each cell is replaced by 0, 1 or 2 cells at division. The cells present after g divisions are sampled, and the lengths of the CA repeat loci are measured. The problem is to estimate g, the time since loss of MMR, using the variation in the MS lengths observed in the sample.

The sampling scheme used in these experiments can be summarized as follows: a random sample of the cells at generation g is taken, r cells being sampled for each of the p different MS loci of interest. For each locus, a summary statistic of the MS length is found using the approach sketched above. In practice, it is easiest to estimate a modal repeat length, but in principle the mean length can also be obtained. We want to estimate g using the variation in the modal repeat lengths observed in the p loci. This problem is reminiscent of the ancestral inference problems in the previous section, and indeed MS loci have been used to trace the origins of modern humans; cf. Donnelly and Tavaré (1997).

The nature of the sampling scheme and the mutation mechanism at MS loci makes this problem ideally suited to computational inference methods. A key ingredient is an algorithm for simulating the genealogical history of a sample of individuals (cells) from a branching process. This is sketched in the next section.

5. The genealogy of a branching process

There have been many theoretical treatments of ancestral processes in the area of branching processes, dating back at least to the late 60s. For a flavor of this, see O'Connell (1997) for example. Much of the theory treats the asymptotic shape of the ancestry for either critical or just-supercritical processes which have simple reproduction mechanisms over time. Little seems to have been written about the genealogy of samples, particularly when the processes are inhomogeneous. Here we sketch a computational device that has proved very useful for simulating the genealogical history of a sample of n individuals from an arbitrary discrete-time branching process.

The idea is to generate the history of the sample without having to simulate the relationships among all the individuals in the population. To do this, we first simulate the family-size counts in each of generations 1, 2, ..., g. For the cell-splitting problem, this requires recording the number of families (N_{m1}, N_{m2}) of sizes 1 and 2 in generations $m = 1, \ldots, g$. Once the family-size process is determined, we work back from generation g by allocating individuals to families in the familiar 'balls-in-boxes' fashion. Thus the n individuals in generation g are chosen at random from the $Z_g = N_{g1} + 2N_{g2}$ individuals in generation g, each being assigned a family label from among the $N_{g1} + N_{g2}$ families present. The set of family labels identifies the distinct ancestors in generation $g-1$. These individuals are in turn assigned randomly to their parents, consistent with the counts $N_{g-1,1}, N_{g-1,2}$, and so on. Once these allocations have stepped back to generation 1 the resulting genealogical tree has the distribution of the ancestral tree of a sample of n individuals taken at generation g, conditional on there being at least n individuals at time g.

It is worth noting that complicated branching mechanisms, in particular non-homogeneous Galton-Watson processes and those with offspring distributions that depend on the history of the process (for example, density dependent reproduction laws), can be treated the same way. The same method applies for any number of founding individuals, although then the history of the sample may be a forest rather than a tree. This approach can also be used when there are different types of individual in the population, making it useful for studying the ancestral history of populations with selectively different individuals, as well as migration and population splitting processes.

6. Application to tumor cell lines

We return briefly to the problem of estimating the time g to loss of MMR from the tumor cell line data. As in many MS studies, the idea is to relate the observed variance $S^2_{\rm obs}$ of the CA repeat measurements to g; cf. Goldstein and Pollock (1997). To do this we assume the cell line evolves according to a branching process with known offspring distributions, and we use the ancestral simulation algorithm to generate a genealogy for $n = rp$ cells. The effects of mutation are then superimposed on this genealogy, starting from the initial cell and proceeding through the ancestral cell population to the sample, to produce a collection of MS lengths at each of the p loci. The modal length is found for each sample of r cells, and the variance in these lengths over the p loci is calculated. Repeating this gives an estimate of the (monotone) relationship between g and $E(S^2)$, say $E(S^2) = f(g)$. To get a moment estimator $\tilde{g}$ of g, we solve $f(g) = S^2_{\rm obs}$ for g. Further details may be found in Tsao *et al.* (1999b). The ancestral simulation algorithm can also be used in a parametric bootstrap approach to examine the precision in the estimate of g. Simulate b trajectories using $g = \tilde{g}$, and re-estimate g for each run to get a series of estimates $g^*_1, \ldots, g^*_b$; the distribution of values of $g^*_i - \tilde{g}$ is taken to reflect that of $\tilde{g} - g$.

REFERENCES

Donnelly, P. and Tavaré, S. (eds) (1997). *Progress in population genetics and human evolution.* IMA Volumes in Mathematics and its Applications, #87. Springer Verlag, Berlin.

Foote, M., Hunter, J.P., Janis, C.M. and Sepkoski, J.J. Jr. (1999). Evolutionary and preservational constraints on origins of biologic groups: divergence times of eutherian mammals. *Science*, 283:1310-1314.

Goldstein, D.B. and Pollock, D.D. (1997). Launching microsatellites: a review of mutation processes and methods of phylogenetic inference. *J. Heredity*, 88:335-342.

Griffiths, R.C. and Tavaré, S. (1998). The age of a mutation in a general coalescent tree. *Stochastic Models*, 14:273-295.

O'Connell, N. (1997). Branching and inference in population genetics. *Progress in Population Genetics and Human Evolution*, eds. P. Donnelly and S. Tavaré. IMA Volumes in Mathematics and its Applications, 87:165-182. Springer Verlag, Berlin.

Slatkin, M. and Rannala, B. (1997). Estimating the age of alleles by use of intraallelic variability. *Am J Hum Genet*, 60:447-458.

Stephens, M. (1999). Times on trees and the age of an allele. Submitted.

Stigler, S.M. (1970). Estimating the age of a Galton-Watson branching process. *Biometrika*, 57:505-512.

Tavaré, S., Marshall, C.R., Will, O., Soligo C. and Martin, R.D. (1999). Molecular and fossil estimates of primate divergence times: a reconciliation? Submitted.

Thompson, E.A. (1976). Estimation of the age and rate of increase of rare variants. *Am. J. Hum. Genet.*, 28:442-452.

Tsao, J-L., Tavaré, S., Salovaara, R., Jass, J.R., Altonen, L.A. and D. Shibata (1999a). Colorectal adenoma and cancer divergence: evidence of multi-lineage progression. *Amer. J. Path.*, in press.

Tsao, J-L., Salovaara, R., Altonen, L.A., Tavaré, S. and D. Shibata (1999b). Genetic estimates of colorectal tumor ages. Submitted.

Wiuf, C. and Donnelly, P. (1999). Conditional genealogies and the age of a neutral mutant. *Theor. Popul. Bio.*, in press.

RÉSUMÉ

This paper describes three examples of statistical inference problems involving molecular data and branching processes. The three examples illustrate branching on widely diverse time scales: reconciling molecular and fossil estimates of divergence times, ancestral inference in population genetics, and reconstruction of tumor history within an individual. The common theme is the development of novel computational inference methods for branching processes, in particular for non-standard sampling schemes.