

Gene genealogies and the coalescent process

RICHARD R. HUDSON

1. INTRODUCTION

When a collection of homologous DNA sequences are compared, the pattern of similarities between the different sequences typically contains information about the evolutionary history of those sequences. Under a wide variety of circumstances, sequence data provide information about which sequences are most closely related to each other, and about how far back in time the most recent common ancestors of different sequences occurred. If the sequences were obtained from distinct species, then the information is frequently extracted and displayed in the form of an inferred phylogenetic tree, which may represent the evolutionary relationships of the species from which the sequences were sampled. If, instead of being from different species, the sequences are from different individuals of the same population, the information is genealogical, and in this case gene trees can sometimes be inferred. A gene tree shows which sampled sequences are most closely related to each other and perhaps the times when the most recent common ancestors of different sequences occurred. A hypothetical gene tree, or genealogy, of five sampled sequences is shown in Fig. 1. In the absence of recombination, each sequence has a single ancestor in the previous generation. (It is important to distinguish a gene tree of sampled sequences from the pedigree of a sample of diploid individuals, in which the number of ancestors grows as one proceeds back in time, because each diploid individual has two parents.) The possibility of obtaining detailed information about the genealogy of sampled genes dramatically changes the situation for molecular population geneticists.

Before the DNA era, molecular polymorphism data were primarily in the form of frequencies of electromorphs, alleles distinguished by their mobility on electrophoretic gels. With protein electrophoresis, two homologous copies of a gene could be classified as being the same or different. If they were different, one could not measure how different; if the two copies were the same, one could not with confidence distinguish whether

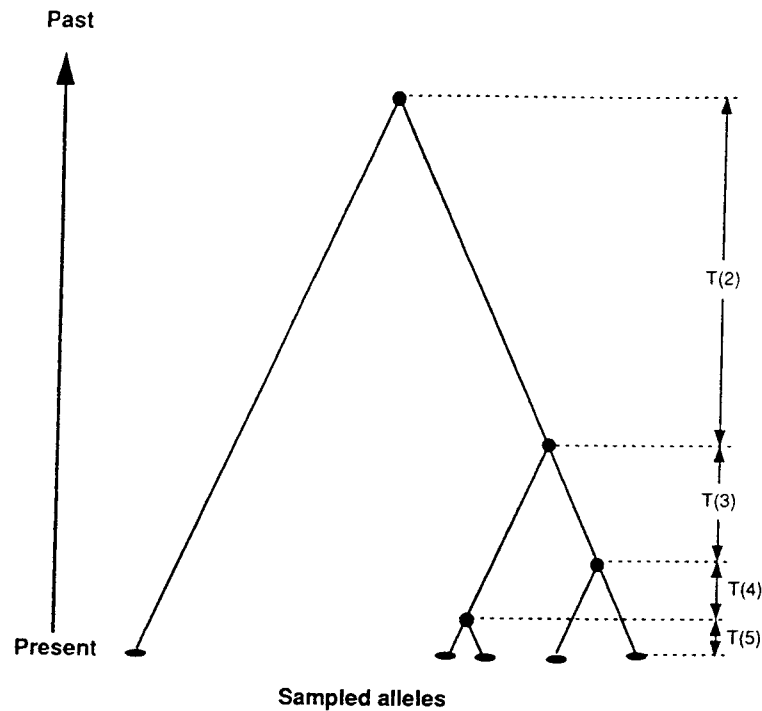


Fig. 1. An example of a genealogy of a sample of five alleles, showing the time intervals between coalescent events. In this figure, the intervals, $T(i)$, are shown with lengths proportional to their expected values as given by eqn (5).

they were really the same or simply convergent in certain physical properties leading to similar electrophoretic mobility. Thus detailed information about the genealogies of genes could not be extracted from data on electromorph frequencies. With modern DNA techniques, sequences of homologous regions of many individuals are obtainable and detailed information about the genealogy of sampled genes will be obtained. Examples of genealogies inferred from sampled alleles are given in Stephens and Nei (1985), Aquadro *et al.* (1986), Bermingham and Avise (1986), Avise *et al.* (1987) and Cann *et al.* (1987).

The obvious challenge for molecular population geneticists is: How can we utilize this information to increase our understanding of the forces acting on molecular variation in natural populations? From the theory side, we can begin by examining the properties of genealogies that arise under a variety of population genetic models. It is important to ask: Are genealogies expected to be very different under different competing models? Can we devise statistical tests that take advantage of the different genealogies expected? To proceed with this task, one needs to examine