1
The Age of a Rare Mutation

Carsten Wiuf
University of Oxford

1 Introduction

Within the last 10 years coalescent theory has become an integrated part of population genetics and the advances in computational statistics have made coalescent theory a standard component in the analysis of genetic variation data. The coalescent and its modifications (Griffiths, Weiss in this book) are highly structured stochastic systems and offer a framework in which many problems in population genetics have been and can be casted and solved.

This discussion deals with one such problem; the problem of dating the age of a mutation, in particular that of a rare mutation. This is not a new problem and dates back before the emergence of the standard coalescent process (Kingman 1982), back to papers by Kimura and Ohta (1973) and Thompson (1976). Here, I will discuss two solutions to the problem, one that springs from the work of Thompson (1976) and treats the age as a fixed parameter in a likelihood framework, and one based on coalescent theory where the age is treated as a random variable. It will be argued that the latter provides the most satisfactory solution.

The age of a mutation is the time since it arose in the population. Recent interest in estimating mutation age stems from the extensive DNA sequencing and marker typing being done to map mutations that cause genetic diseases. Estimating mutation age is partly done out of curiosity and partly to make further use of data that has been gathered for other purposes, though it relates to other interesting and important problems. To get a sense of this, imagine we have a sample from a population in which we find k copies of the mutation (Fig. 1). One approach to estimate the age would be to trace the ancestry of the k genes that bear the mutation and estimate when the mutation happened prior to the event of a most recent common ancestor of the k mutant genes; that is the genealogy that relates the k mutants provides information about the age. (In the chapter by Griffiths and Tavaré the genealogical structure of a sample of both mutants and non-mutants is discussed more generally.)

However, this genealogy is useful in other contexts. It will be shaped by the demographic history of the population and by forces of selection; as such, inference on the age relates closely to inference on selection and demography.

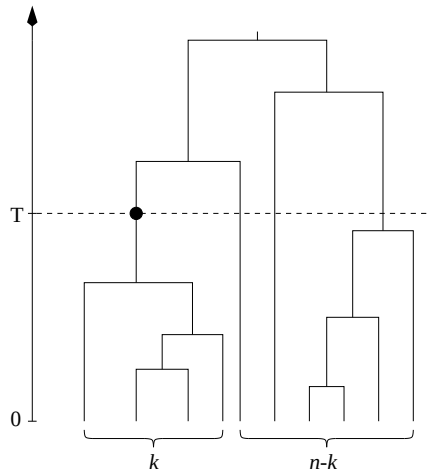

FIG. 1. Genealogy of a sample of size n with k mutants. The mutants find a most recent common ancestor before sharing ancestor with any other gene in the sample. The mutation arose at time T ago.

In the context of gene mapping, modelling of the genealogy of a mutation also provides insight into the genealogy and the genetic variation at linked loci; insight that can be used in development of gene mapping methods.

Examples of such mutations include mutations in the CF locus causing cystic fibrosis and mutations in the BRCA1 and BRCA2 loci causing breast cancer. These mutations are all found in low frequencies, $q = \lim_n k/n < 2\%$.

2 Defining the Problem

It is assumed that the mutation is unique in the history of the entire population, otherwise the age cannot be unambiguously defined. On the other hand, a stationary stochastic evolution process is required for the age to be defined properly as a stochastic variable (Watterson 1976). If only one mutation is allowed this cannot be achieved.

Several solutions have been suggested. Kimura and Ohta (1973) were first to consider the relationship between the frequency of a mutant gene and the age, T, of the mutant. It was assumed that the mutant gene arose initially at frequency $1/2N$ (where $2N$ is the number of genes in a population of size N), and that its subsequent evolution to its present frequency, q, was a diffusion process. In the absence of selection, they found that T, conditional on q, has expectation

$$E(T \mid q) \;=\; -\frac{2q\log(q)}{1-q} \tag{2.1}$$

(T is measured in units of $2N$ generations). The mutation process giving rise to the mutant gene was not explained.

We will later see that (2.1) can be justified if the rate, θ, by which new mutants arise is low. In contrast, Thompson (1976) found it more natural to consider T to be a fixed parameter, thereby avoiding the issue of defining the age distribution.

3 The Likelihood Approach

Thompson's (1976) work is closely related to work by Slatkin and Rannala (1997) and it suffices to consider their work. They formulate the problem in the following way. Assume the mutation arose time T ago, that is at time T there is one single copy of the mutation. Due to drift and selection, this copy is multiplied over the years and reaches frequency q today.

The number of copies present at any time between T and the present is modelled by a branching process. Slatkin and Rannala (1997) use a continuous birth-death process, whereas Thompson (1976) uses its discrete counterpart, a branching process with a modified geometric offspring distribution. Both these processes are commonly used in genetics and appear as approximations to the evolution of mutations in low frequency. The birth-death process allows certain realistic demographic features to be introduced in the model, e.g., population growth.

The form of the likelihood function of the data (here q) allows the maximum likelihood estimator, $\hat{T}$, of T to be derived analytically,

$$\hat{T} \;=\; \frac{1}{\xi}\log\{\,2\xi q \,+\, 1\,\} \;\text{ for }\; \xi > 0, \;\text{ and }\; \hat{T} = 2q \;\text{ for }\; \xi > 0. \tag{3.1}$$

Here, ξ is a combined measure of population growth and selection. In the absence of selection, $\xi = 2Nr$, where r is growth rate per generation. This approach has a number of drawbacks;

- The mutation process is not modelled,
- T is a parameter, all other times in the genealogy are stochastic,
- Copies of the mutation could have existed before time T; it is only assumed that there is one at time T.

4 The Coalescent Approach

The genealogy, G, of the sample or the entire population is modelled by the coalescent process. On top of the genealogy mutations are imposed according to a Poisson process with intensity $\theta t/2$, where t is the length of a branch (see also Griffiths, this book). The genealogy we seek is G conditional on exactly one mutation in the sample's (population's) history. Further, we assume $\theta \approx 0$ (formally, we consider $\theta \to 0$), so that mutations are rare.

In this setting, T is an unobserved random variable and it is natural to report the posterior density of T given the data. As such, the coalescent approach is Bayesian. Further, we see that all three points of concern mentioned above have been taken into account.

The general form of the age distribution, $f(\,t\,|\,q\,)$, conditional on its current frequency, is mathematical untractable, but simulations from the distribution can, in general, be performed (Griffiths and Tavaré 1998, Wiuf 2000). If the total population size is constant over time and the mutation is neutral, then the expectation is given by (2.1). If we compare (2.1) to $\hat{T} = 2q$ we find that $E(\,T\,|\,q\,)/\hat{T} \to \infty$ for $q \to 0$, that is the ratio between the two quantities become arbitrary large for small (and realistic) q. For example, if $q = 1\%$, $E(\,T\,|\,q\,) = 0.09$ and $\hat{T} = 0.02$.

If the frequency is less than, say, $q < 10\%$, very accurate and simple approximations exist (Wiuf 2000); for $\xi > 0$,

$$f(\,t\,|\,q\,) \;\propto\; \frac{1}{(e^{\xi t}-1)^2}\,\exp\left\{-\frac{2\xi q}{e^{\xi t}-1}\right\}, \tag{4.1}$$

where $\xi = 2Nr$ measures population growth. If ξ is large ($\xi > 10$), $f(\,t|\,q)$ is centred around the mode, T_{mode}, and

$$E(T\,|\,q) \;\approx\; T_{\text{mode}} \;=\; \frac{1}{\xi}\,\log(\,\xi q \,+\, 1\,). \tag{4.2}$$

Considering T_{mode} as a point estimate of the age, we find $T_{\text{mode}}/\hat{T} \to 1$ for $\xi \to \infty$ and the two approaches give estimates of the same order of magnitude.

5 Conclusion

It has been demonstrated that the likelihood approach has a number of features that makes it inappropriate. The coalescent approach is natural in that it considers the genealogy to be wholly stochastic. For mutations in low frequency, the difference between $\hat{T}$ and $E(\,T\,|\,q\,)$ can be substantial.

References

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

Kimura, M. and Ohta, T. (1973) The age of a neutral mutant persisting in a finite population. *Genetics* **75**, 199-212.

Kingman, J. F. C. (1982) The coalescent. *Stochastic Processes Applied* **13**, 235-248.

Slatkin, M. and Rannala, B. (1997) Estimating the age of alleles by use of intraallelic variability. *American Journal of Human Genetics* **60**, 447-458.

Thompson, E. A. (1976) Estimation of age and rate of increase of rare variants. *American Journal of Human Genetics* **28**, 442-452.

Watterson, G. A. (1976) Reversibility and the age of an allele. I. Moran's infinitely many neutral alleles model. *Theoretical Population Biology* **10**, 239-253.

Wiuf, C. (2000) On the genealogy of a sample of rare alleles. *Theoretical Population Biology* **58**, 61-75.