



Estimating primate divergence times by using conditioned birth-and-death processes

Richard D. Wilkinson^{a,*}, Simon Tavaré^b

^a Department of Probability and Statistics, University of Sheffield, Hicks Building, Hounsfield Road, Sheffield S3 7RH, United Kingdom

^b DAMTP, University of Cambridge, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WA, United Kingdom

ARTICLE INFO

Article history:

Received 5 January 2009

Available online 6 March 2009

Keywords:

Conditioned birth-and-death process

Size-biased trees

Primate divergence time

Inference from the fossil record

ABSTRACT

The fossil record provides a lower bound on the primate divergence time of 54.8 million years ago, but does not provide an explicit estimate for the divergence time itself. We show how the pattern of diversification through the Cenozoic can be combined with a model for speciation to give a distribution for the age of the primates. The primate fossil record, the number of extant primate species, and information about the structure of the primate phylogenetic tree are combined to provide an estimate for the joint distribution of the primate and anthropoid divergence times. To take this information into account, we derive the structure of the birth-and-death process conditioned to have a subtree originate at a particular point in time. This process has a size-biased law and has an immortal line running from the root of the tree to the root of the subtree, with species on the spine having modified offspring and length distributions. We conclude that it is not possible, with this model, to rule out a Cretaceous origin for the primates.

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

Sam Karlin made numerous contributions to the theory of stochastic processes. Among the earliest of these is the now-classical Karlin–McGregor integral representation of the transition function of birth-and-death and related processes (Karlin and McGregor, 1958) and its application to coincidence probabilities (Karlin and McGregor, 1959a,b). For applications of the latter to combinatorics, see Karlin (1988). Many of Sam's results arose in the study of evolutionary or population genetics, beginning with a mathematical analysis of Moran's model (Karlin and McGregor, 1962), a continuous-time analogue of the Wright–Fisher model of gene frequency change in a finite population. A more general class of discrete-time models obtained by conditioning branching processes on a fixed total size was described in Karlin and McGregor (1964). This provided motivation for Cannings' exchangeable models (Cannings, 1974) and their recent developments (Möhle and Sagitov, 2001).

Several of Sam's papers exploited compound stochastic process arguments, in particular to study population models in which new populations arise at the points of non-homogeneous Poisson processes (Karlin and McGregor, 1967); we use a similar approach in the present paper. Karlin and McGregor (1972) gave a prescient

argument that we now recognize as a “coalescent method” to derive the celebrated Ewens Sampling Formula (Ewens, 1972). Kingman's elegant formulation of the ancestral structure of neutral population genetics models, the coalescent (Kingman, 1982), appeared in 1982 and its uses are now commonplace in population genetics (cf. Tavaré (2004), Hein et al. (2005) and Wakeley (2008)).

One of us (ST) was a postdoc of Sam's at the time computational molecular biology was coming into its own. We still worked on stochastic problems in population genetics (such as Karlin and Tavaré (1982)), but DNA sequencing had become a reality (although for a while the data came in *books*—ST remembers typing all 48,502 basepairs of bacteriophage lambda into a text file!) and there were new problems to think about. Sam's interests moved towards statistical issues in sequence analysis, resulting first in Karlin et al. (1983), and remained there for the rest of his life. Nonetheless, he still had time for questions about stochastic processes. With this in mind, we think Sam would have liked the problem (and perhaps even the approach) we describe in our paper, which we dedicate to his memory.

2. Estimating divergence times by using the fossil record

The crown divergence time of a monophyletic group of species is the most recent time at which all the species shared a common ancestor. Informally, one can think of the divergence time as the point at which a single ancestor species first diverged into two or more distinct species. Thinking in terms of phylogenetic trees, estimating divergence times is essentially a problem of how to

* Corresponding author.

E-mail addresses: r.d.wilkinson@sheffield.ac.uk (R.D. Wilkinson), st321@cam.ac.uk (S. Tavaré).

learn the depth of a tree from an incomplete snapshot of its various parts.

In this paper we show how to date the divergence time of a taxonomic group using the fossil record. This is an important problem as fossil evidence is the only direct source of information about the age of a group of species. Genetic data do not explicitly contain any information about age; dating methods that use DNA rely on one or more dates estimated from the fossil record in order to calibrate the speed of mutation in a dating model (the so-called molecular clock). Fossils, on the other hand, can be dated to provide tangible evidence of the existence of a species at a particular point in time. However, fossil evidence only provides a lower bound on the age of a group, with the divergence time of a taxon bounded above by the age of the oldest fossil. For well-sampled taxa with relatively complete fossil records, such as marine invertebrates, it is likely that this lower bound will be close to the true divergence age (Raup and Sepkoski, 1982). However, for poorly sampled taxa, a category including most terrestrial vertebrates, we intuitively expect the temporal gap between the divergence time and the oldest fossil discovery to be more variable and potentially much longer than for well-sampled taxa.

While fossil data do not explicitly provide an upper bound on the age of a clade, the pattern of fossil finds can provide information about how the diversity (number of species) of the clade varied through time. This signal will often be highly noisy, and using it to infer the true diversity is complicated by not knowing the completeness of the fossil record and by the belief that the fossil sampling and discovery rate varies over the geologic time scale (Raup, 1979). However, by modelling diversification and fossil preservation we can use the fossil record, along with other information such as the modern diversity and the known phylogenetic structure, to estimate the divergence time of a clade. We can then give a probability distribution for the divergence time which represents our remaining uncertainty given the data, giving a credibility interval for the range and estimating the most likely divergence time.

2.1. The primate fossil record

We extend the work of Tavaré et al. (2002), estimating the joint distribution of the primate and anthropoid divergence times. The estimation of these divergence times merits special care and attention because of the debate about the primate divergence time that has taken place in recent years. The argument has been characterized by Benton (1999) as ‘molecules versus morphology’ and concerns whether the primates coexisted with the dinosaurs during the Cretaceous over 65 million years (My) ago. Direct readings of the fossil record tend to place the divergence time in the Cenozoic (Gingerich and Uhen, 1994; Kay et al., 1997), whereas molecular dates tend to place the divergence time in the Cretaceous (Kumar and Hedges, 1998; Arnason et al., 1996; Hedges et al., 1996; Bininda-Emonds et al., 2007). There are sound reasons for why some disparity is expected between the two dating methods, as genetic dates record when inter-breeding ceased, whereas fossils date when morphological difference arose. However, this cannot account for the magnitude of the difference and there is reason to believe that date estimation from fossil evidence can be improved (Martin, 1993). The completeness of the primate fossil record (the proportion of species preserved as fossils) has been estimated to be less than 10% by Martin (1990) and as noted above, for incomplete taxa the temporal gap between oldest fossil and divergence time will be stochastically large.

Table 1 shows the available primate fossil data. It consists of a collection of counts of the number of distinct primate species in each of the past 14 geologic epochs, along with the number of extant primate species (reported in Groves (2001, 2005)). It

also gives the number of anthropoid species known from the fossil record. These data are an unpublished updated version of the data given in Tavaré et al. (2002). The anthropoids are an infraorder of the primates consisting of the new and old world monkeys and the apes, and they form a monophyletic subtree in the primate phylogeny. Further information on the primates, along with information about the data, is available in Martin et al. (2007). There are two important points to note from the data: no primate fossil predating the Eocene has been found, with the oldest primate fossil being at most 54.8 My old and no anthropoid fossil has been found before the Late-Eocene, with the oldest anthropoid fossil being at most 37 My old. Throughout this paper, we let τ denote the temporal gap between the oldest primate fossil and the primate crown divergence time, so that the primate divergence occurred $54.8 + \tau$ My ago. We similarly define τ^* to be the temporal gap between the oldest anthropoid fossil and the anthropoid divergence time, so that the anthropoid divergence time was $37 + \tau^*$ My ago. Fig. 1 is a simple illustration showing this structure.

Aside from the fossil data, there are other sources of information that can be utilized. Firstly, the modern diversity can inform us about fossil sampling rates and the completeness of the record. Secondly, morphological considerations can allow for the identification of some phylogenetic structure. For example, it is known that the anthropoids are a monophyletic subgroup of the primates, so that the anthropoid phylogenetic tree is a subtree of the primate tree, as shown in Fig. 1. Knowing this structure can inform our beliefs about the placement of the root and shape of the tree. Molecular evidence can also provide information about the divergence time, although we do not explore that route here.

The focus in this paper is on combining the information in the fossil record with the modern diversity and on using the known phylogenetic structure to date multiple divergence times simultaneously. By using this structure we hope to date two divergence times with more accuracy than is possible in dating a single divergence time. Also, by giving the joint distribution, it is possible to quantify the joint distribution of the error terms, offering a potential improvement in accuracy if these dates are used as calibration nodes in subsequent molecular analyses. We model both the primate and anthropoid divergence times with the aim of learning how these times can be constrained given our model and the data. We take a forwards modelling approach, giving a model for speciation and fossil discovery, and then fit the model to the data to learn about the temporal gaps τ and τ^* .

3. Modelling speciation

In order to combine the fossil record, the number of extant species, and the known phylogenetic structure to estimate divergence times, we need a model which incorporates all three aspects. We take a forwards modelling approach, explicitly modelling speciation using a simple stochastic birth-and-death process. Although it is easy to criticize the model, it should be borne in mind that this is an advance over previous approaches to dating using the fossil record, which tend to have been statistical approaches relying on correlations, rather than process models. It is also unclear, due to the limited data available, whether a more complex modelling approach is feasible. Our model can then be used to assess the range of uncertainty one can expect for the temporal gap between the oldest fossil and the divergence time.

We now describe the basic model, which is then conditioned to account for the known phylogenetic structure. The notation and development follow that given in Harris (1963). We consider the birth-and-death process to be an evolving tree process, with each lineage in the tree representing a different species. In order to describe the dynamics of the process, it will be useful to have the following definition of an exponential distribution with time-varying rate.

Table 1
A summary of the number of primate and anthropoid species known from the fossil record (Martin et al., 2007). Time during the Cenozoic is divided into 14 geologic epochs, with the dates for each epoch given in the table in millions of years (My). Also given is the modern diversity (Groves, 2005).

Epoch	k	Time at base of interval k (My)	Primate fossil counts, \mathcal{D}	Anthropoid fossil counts, \mathcal{A}
Extant	0		376	281
Late-Pleistocene	1	0.15	22	22
Middle-Pleistocene	2	0.9	28	28
Early-Pleistocene	3	1.8	30	30
Late-Pliocene	4	3.6	43	40
Early-Pliocene	5	5.3	12	11
Late-Miocene	6	11.2	38	34
Middle-Miocene	7	16.4	46	43
Early-Miocene	8	23.8	34	28
Late-Oligocene	9	28.5	3	2
Early-Oligocene	10	33.7	22	6
Late-Eocene	11	37.0	30	2
Middle-Eocene	12	49.0	119	0
Early-Eocene	13	54.8	65	0
Pre-Eocene	14		0	0

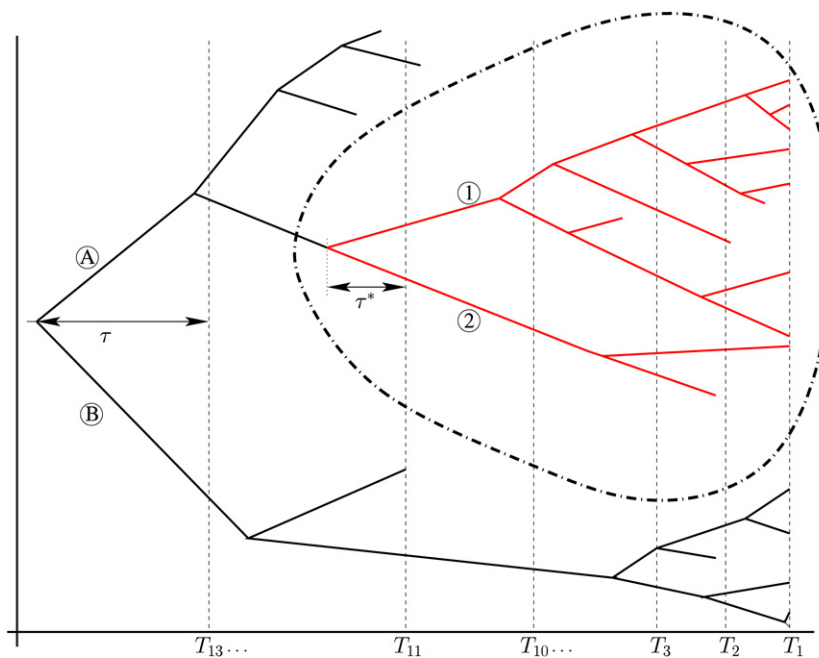


Fig. 1. A sample primate speciation tree with anthropoid-subtree highlighted. Here, tree A represents the haplorhini and tree B the strepsirrhini, while subtrees 1 and 2 represent the platyrrhine and catarrhine species. Parameter τ records the distance between the base of the Eocene and the base of the primate tree, whereas τ^* records the distance between the base of the Late-Eocene and the anthropoid subtree.

Definition. Let $b(\cdot)$ denote a positive integrable function with $\int_s^\infty b(t)dt = \infty$ for all s . We say the random variable X has an inhomogeneous exponential distribution begun at time s , and write $X \sim \text{Exp}_s(b(\cdot))$, if X has the probability density function

$$\pi_s(x) = b(s+x) \exp\left(-\int_s^{s+x} b(t)dt\right), \quad x > 0.$$

We consider the inhomogeneous birth-and-death process. Each lineage lives for an inhomogeneous exponential period of time with variable rate $b(t) = \lambda(t) + \mu(t)$. Upon the death of a species at time t , it is replaced with either zero (a death) or two (a birth) new species with probabilities

$$p_0(t) = \frac{\mu(t)}{\lambda(t) + \mu(t)} \quad \text{and} \quad p_2(t) = \frac{\lambda(t)}{\lambda(t) + \mu(t)} \quad (1)$$

respectively. If we denote the number of species alive at time t by $Z(t)$, the process can be described in terms of the infinitesimal

change equations

$$Z(t+h) = \begin{cases} Z(t) + 1 & \text{w.p. } Z(t)\lambda(t)h + o(h) \\ Z(t) - 1 & \text{w.p. } Z(t)\mu(t)h + o(h) \\ Z(t) & \text{w.p. } 1 - Z(t)(\lambda(t) + \mu(t))h + o(h), \end{cases}$$

completing the description of the basic model.

In order to date two or more divergence times simultaneously, we must be able to include any known phylogenetic structure into the model. The type of information that is typically available is that a subgroup of the species form a subtree within the main tree. One approach to using this information is to find post-hoc within the tree, the most likely subtree; we can simulate a sample tree for the complete phylogeny, then exhaustively search all subtrees to find the subtree that most closely matches the data for the subgroup. We then measure the divergence time from this optimal subtree. The problem with this approach is that it is difficult to interpret the results. The optimal subtree is in some sense the closest match to the data, but is not interpretable as a posterior distribution or in any other standard way.

A more satisfactory approach to modelling subtree origination is to condition the birth-and-death process to have a subtree

