Cladogenesis, coalescence and the evolution of the three domains of life

Olga Zhaxybayeva and J. Peter Gogarten

Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269, USA

In this article, we explore the large-scale structure of the tree of life by using a simple model with a constant number of species and rates of speciation that equal the rates of extinction. In addition, we discuss the consequences of horizontal gene transfer for the concept of a most recent common ancestor of all living organisms (cenancestor). A simple null hypothesis based on coalescence theory explains some features of the observed topologies of the tree of life. Simulations of genes and organismal lineages suggest that there was no single common ancestor that contained all the genes ancestral to those shared among the three domains of life. Each contemporary molecule has its own history that traces back to an individual molecular cenancestor. However, these molecular ancestors were likely to be present in different organisms and at different times.

The tracing and timing of the most recent common ancestor of all organismal lineages [i.e. CENANCESTOR (see Glossary)] from the molecular record remains one of the debated issues in biology. Since Charles Darwin's time it has been assumed that there was a single organism that gave rise to all known life. With the availability of many molecular markers, it became clear that different molecules have different histories and there is disagreement on the location of the root of the tree of life (e.g. different studies place the root: (i) within the bacterial domain or on the branch that leads to the bacterial domain [1,2]; (ii) within the eukaryotic domain [3–5]; (iii) within the archaeal domain [6]; or (iv) yield inconclusive results [7]). The timing of the organismal cenancestor is another unresolved question [8–11]. The time estimates vary dramatically, and in many instances the proposed age of the most recent common ancestor exceeds the estimated age of the Universe ([12] and J.P. Gogarten, unpublished). The probable absence of an adequate amount of time for life to evolve on Earth is used by the proponents of the directed panspermia hypothesis [13], which states that life originated elsewhere and was deliberately transported to Earth. Tracing back the history of molecules usually reveals little about extinct lineages. Reconstructed phylogenies, because of their steadily furcating nature, often give the impression that there were fewer species in the past than exist today. Many analyses only focus on the 'lucky ancestors' whose offspring survived to the present

day (or left a recognizable imprint in the fossil record). However, it is reasonable to assume that in addition to these 'lucky ancestors' there were many other lineages that coexisted in the past and occupied all ecological niches that were available: most of those lineages became extinct, some gave rise to new lineages, some fused with others and some have contributed genes via horizontal gene transfer (HGT) to the lineages that still exist today.

As lineages coalesce to their common ancestors, they form clades. The rate of cladogenesis is an interesting problem [14]. How often do new clades arise? Is the extinction rate equal to the speciation rate? What are the driving forces for cladogenesis? Several authors have described mathematical models for cladogenesis [15,16] and goodness-of-fit tests to ascertain if models accurately explain the data [17,18], and have attempted to infer extinction and speciation rates from phylogenetic trees [18,19]. Two simple models of cladogenesis can be used as null hypotheses: (i) a pure-birth model (i.e. a model based

---

Glossary

Cenancestor: from the Greek 'kainos' meaning recent and 'koinos' meaning common – the most recent common ancestor of all the organisms that are alive today. The term was proposed by Fitch in 1987 [52]. The cenancestor should not be confused with the progenote, which denotes a hypothetical organism in which genotypes and phenotypes were not strictly coupled [6]. The cenancestor might have been a progenote but in many scenarios for the evolution of life the progenotic stage occurred much earlier in evolution.

Coalescence: the process of tracing lineages backwards in time to their common ancestors. Every two extant lineages coalesce to their most recent common ancestor. Eventually, all lineages coalesce to the cenancestor.

Organismal lineage: can be defined by the majority of genes passed on over short time intervals. Provided the time intervals are sufficiently short, this definition only fails in the rare event of two organisms making co-equal contributions to a new line of descent. Gary Olsen (University of Illinois, Urbana-Champaign) used the metaphor of a rope to illustrate this concept – no single cellulose fiber (representing the genes) might persist throughout a rope (representing the organismal lineage) from beginning to end; nevertheless, the rope has continuity. The presented definition of organismal lineage is theoretical and its usefulness in studying organismal lineages in practice remains under debate. This gene-centered definition also can be faulted for ignoring epigenetic traits that might have been important, particularly during the early evolution of life.

Effective number of species (N_e): this is analogous to the definition of effective population size – the theoretical number of species in an idealized constant species model that results in the same coalescence time as observed with the actual number of species, which might have varied during evolutionary history.

Molecular lineage: evolutionary history of a single gene. Although genes can be mosaic, this definition still holds true because the evolutionary history of a gene does not necessarily have to be a bifurcating tree.

Clade: from the Greek 'klados' meaning branch or twig – a group of organisms that includes all of the descendants of an ancestral taxon. In a rooted phylogeny every node defines a clade as the lineages originating from this node, including those that arise in successive furcations.

Corresponding author: J. Peter Gogarten (gogarten@uconn.edu).

on a Yule process [20]), where there is no extinction; and (ii) a birth–death model, where speciation and extinction are considered. The pure-birth model is probably inadequate for the time-scale of the tree of life because there were many extinction events during the evolution of life.

Explanations for the long and empty branches that connect the three domains of life

A common feature of most phylogenetic trees that are calculated from different markers is the observation that long and empty branches connect the bacterial, archaeal and eukaryotic domains to the cenancestor. The absence of side branches can be explained by several hypotheses.

One group of hypotheses assumes that early evolution, before the three domains split, used different mechanisms. Zillig and colleagues [21] proposed a progenote population that exchanged genes efficiently between members of the population, thereby preventing speciation. This phase was followed by geographic isolation of two subpopulations that later gave rise to bacteria and archaea. In Zillig's model the eukaryotes are proposed to have appeared from a fusion of archaea and bacteria. Kandler suggested a similar hypothesis, which stated that before the diversification of the domains of life there was a population of pre-cells that at successive times gave rise to the three domains of life; each of the domain ancestors recruited different sub-sections from the genes present in the pre-cell populations [22]. At the early state, while the separation of domains was taking place, the members of the population of pre-cells were frequently exchanging and recombining their genetic material. After the three domains had emerged, the genetic exchange among them had decreased dramatically. Woese [23] proposed a similar model in which he described 'genetic annealing' as the crystallization of cells and cellular functions within pre-cell populations. In Woese's model the early evolution had a different tempo, characterized by extensive HGT and higher mutation rates. Replication and translation evolved in this phase and formed the core around which the genomes of the domain ancestors formed. Koch [24] proposed a different mode for early evolution, which he termed the 'monophyletic epoch'. During that period many independent lineages existed (but they lacked the mechanisms of transfer) and evolution occurred by mutation and selection. Koch's idea suggests that the long and bare branches at the root of the tree of life might be due to the absence of mechanisms that are necessary for speciation (Figure 1a).

Another group of hypotheses assumes that there was a major catastrophic event in the past that led to a bottleneck with few survivors (e.g. [25–27]). The observed bottleneck might have been caused by the tail of the early heavy meteorite bombardment [26]. The only survivors were the ancestors of the three domains of life (Figure 1b). These hypotheses also can explain the apparent distribution of thermophiles on the tree of life.

In this article, we draw attention to a third type of hypothesis that is based on the COALESCENCE model. Coalescence theory was first introduced to genetics by Wright for two alleles [28] and was later generalized for many alleles by Kingman [29]. Kingman's coalescence theory of different alleles in a population was widely popularized and applied by Felsenstein [30]. This theory discusses the coalescence of alleles in a population to their common ancestral allele. Analogously, simple coalescence also provides a mathematical framework for cladogenesis, if one assumes that splitting of a lineage into two new lineages and extinction of a lineage occurs with equal frequency [31]. A noteworthy characteristic of the coalescence processes is the long time interval needed for the two lineages leading to the most recent common ancestor to meet [29]. This might explain why the branches leading to the cenancestor of the three domains are long and empty.

Justification for a simplifying null hypothesis

Figure 1c depicts a possible scenario for the origin and early evolution of life on Earth. Although there might have been different autocatalytic reaction cycles that emerged during prebiotic evolution, the emergence of an aperiodic biopolymer that functioned as genetic material might have occurred only once, giving rise to the RNA or pre-RNA world [32,33]. Some prebiotic reaction cycles might have survived independently and later contributed to the formation of cells. The emergence of genetic material, considered by some to mark the origin of life, was followed by a phase of diversification in which the first self-replicating systems evolved to occupy different environments. Possibly only a single lineage emerged from the origin of life and this lineage is ancestral to all extant life; however, this lineage does not represent the cenancestor. All extant cellular organisms that are known today share a multitude of characteristics including the use of DNA as genetic material, energy-coupling membranes and template-directed protein synthesis using ribosomes. These shared complex properties suggest that the cenancestor lived at a stage much later than the origin of life. It seems reasonable to assume that at this point in time multiple lines of decent coexisted.

If the cenancestor existed after the diversification of life into many separate lineages had taken place, it seems justified to assume a simple model for cladogenesis that is based on coalescence. Clearly, the assumption of an exact balance between speciation and extinction is unrealistic. During >3.5 billion years, life has diversified into increasingly complex ecological patterns. The modern environment provides many more ecological niches than the early Earth. The value of a simple model that assumes a constant number of species over time is that it provides a null hypothesis enabling a comparison with more-complex scenarios. In addition, it might be considered as a first approximation where the number of species that is assumed to be constant over time represents the effective number of species, which is similar to the concept of an effective population size.

Simulation of cladogenesis using a model based on coalescence

Simulation model for the evolution of organismal lineages

We performed simulations of the evolution of organismal lineages under a simple model. Starting with the population
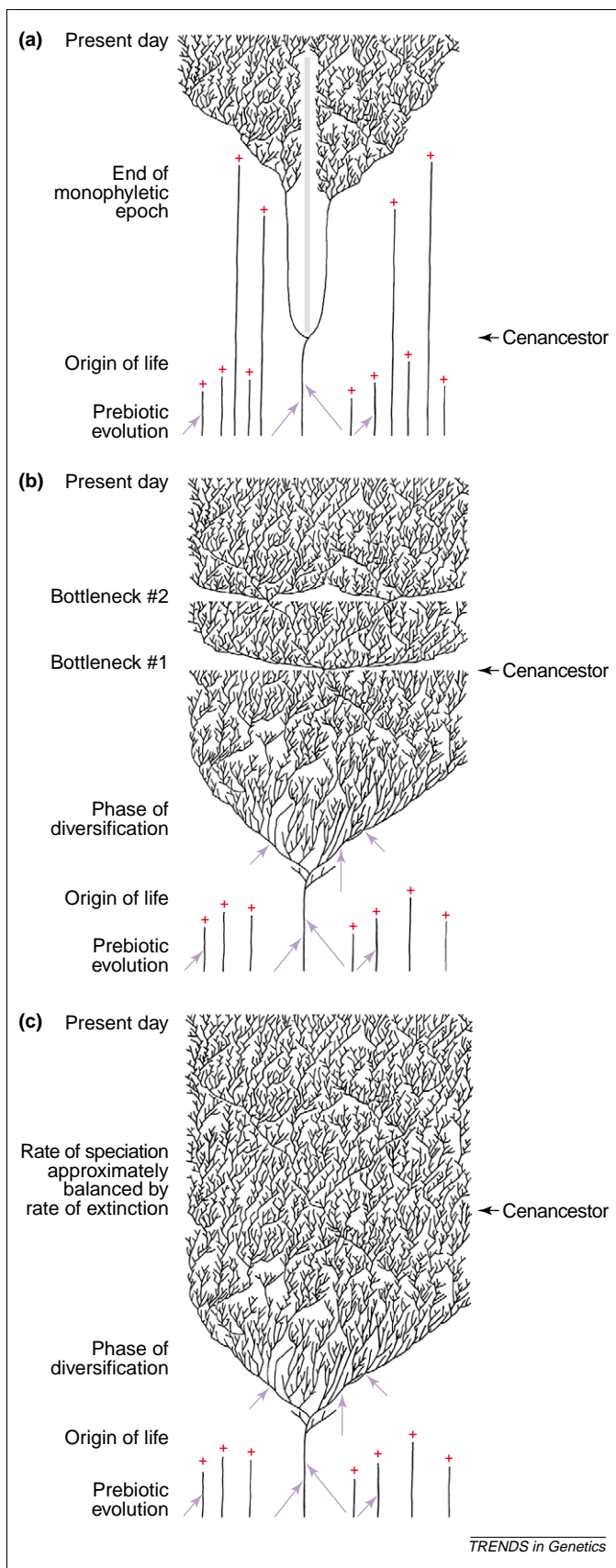
Figure 1. Scenarios summarizing possible histories of the origin and early evolution of life on Earth. During prebiotic evolution, there might have been many independently arising lineages of autocatalytic chemical reactions and networks; those that became extinct are indicated by crosses. Although the use of an information-carrying aperiodic biopolymer as genetic material might have been a unique event, natural selection is likely to have generated different species or quasi-species that occupied the ecological niches that were available. Arrows indicate the prebiotic reaction networks that might have contributed to the diversifying lineages. **(a)** Owing to an existing geographic barrier and a lack of ability for speciation, only two separate lineages might have given rise to all present-day organisms [24]. **(b)** The major bottlenecks that occurred as a result of catastrophic events in Earth's history (e.g. meteorite impacts or snowball Earth events) might explain why we only have three domains of life [25–27]. **(c)** If initial diversification was followed by the period of balanced speciation and extinction, then extant lineages coalesce to a cenancestor that is not necessarily located at the root of the tree, but might have lived much later, in the phase of balanced extinctions and speciations.
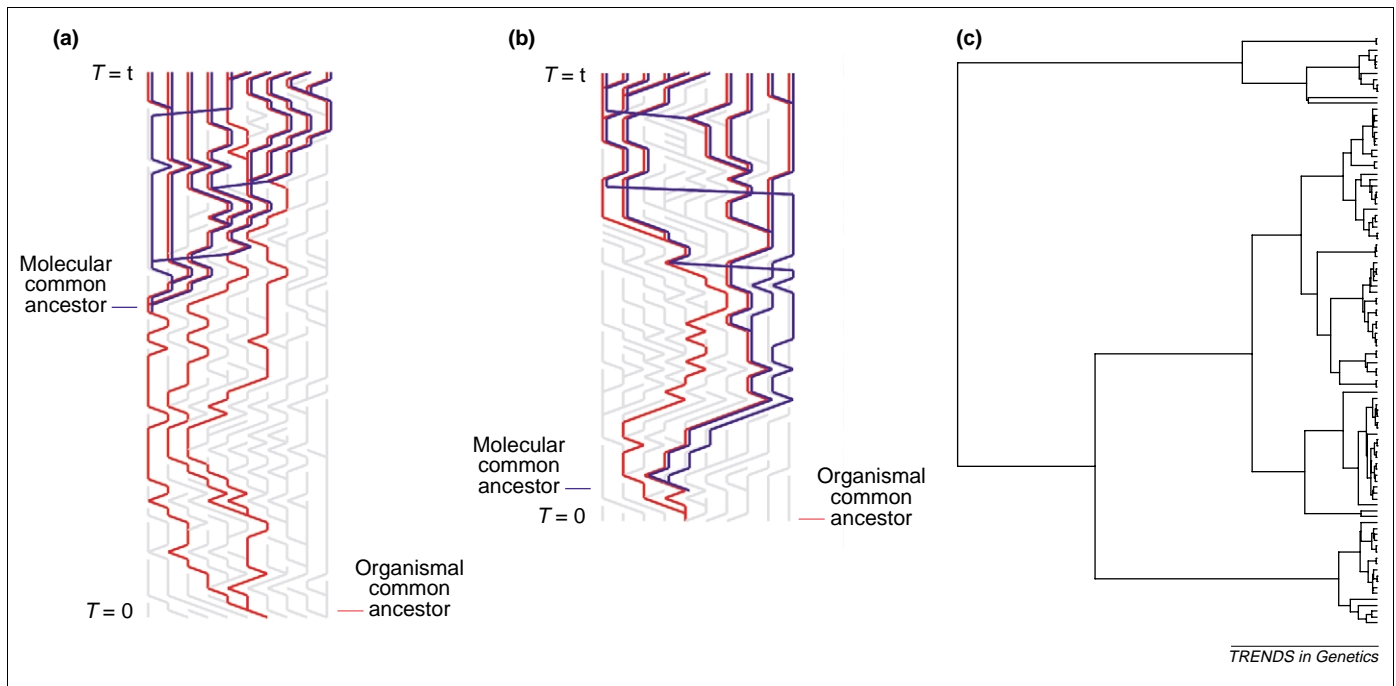
with n lineages at the time $T = 0$, at each successive time interval one randomly chosen lineage becomes extinct and another splits into two new lineages. The process is terminated at time $T = t$ when all of the lineages extant at that time have a common ancestor. The extant lineages are then traced back to their most recent common ancestor (Figure 2).

Introduction of HGT into the model
To simulate the evolution of genes in organismal lineages, we introduced HGT events by allowing one transfer event per m time intervals between two randomly chosen lineages. The following restrictions on the choice of lineages are imposed: (i) the recipient lineage should not become extinct during the same time interval; and (ii) the donor lineage should not undergo a simultaneous speciation event, which simplifies tracing the molecular histories. The extant genes are again traced back to their most recent common ancestor.

Molecular and organismal cenancestors do not necessarily coincide

Because genes can and have been transferred between divergent organisms [34–38] and genomes are mosaics where different parts of the genome have different histories [39,40], the usefulness of the concept of an ORGANISMAL LINEAGE has been questioned [41,42]. Conceptually, an organismal lineage can be defined by the majority of genes passed on over short time intervals.

Tracing evolution by using molecules from extant organisms restricts the analyses to the lucky ancestors, which produced offspring that survived and propagated into the present time. Figure 2a,b depicts the simulation of cladogenesis through coalescence where $n = 10$ lineages. Figure 2a demonstrates that the introduction of rare HGT events can lead to a molecular cenancestor that does not reside in the cenancestor of the organismal lineages. Figure 2a also illustrates that the coalescence time for a single gene to its common ancestor can be different from the time of the organismal cenancestor. Each gene in an organismal lineage has its own pattern of HGT and therefore its own history. The genes histories are not necessarily congruent with one another or with the history of the organism in both topology and time estimates.

Coalescence leads to long branches at the root

The two branches that connect to the cenancestor often are long (e.g. Figure 2b). According to Kingman [29,30] their coalescence takes on average half of the total time between the present and the cenancestor. Although there are many lineages present at each time step and the speciation rate

Figure 2. Illustration of the coalescence of lineages to their most recent common ancestor. Two separate runs of simulations for $n = 10$ species and one run for $n = 100$ species are shown. **(a)** The most recent common ancestor for the molecular phylogeny with horizontal gene transfer (HGT) is different from the cenancestor of the organismal lineage, and the time at which these last common ancestors (molecular and organismal) exists also differs. **(b)** The branches that lead to the last common ancestor are long compared with the other branches, illustrating that coalescence is a stochastic process (Figure 3). Coalescence of organismal lineages or molecular lineages that do not experience any HGT is shown in red. The coalescence of a gene present in the organismal lineages but with HGT events incorporated into the simulation (one transfer event per ten speciation events) are shown in blue. The extinct lineages are shown in gray. **(c)** The neighbor-joining tree of 100 extant lineages obtained through simulations. The extinct lineages are not depicted. The distance matrix was generated from the pairwise distances measured in time intervals between extant lineages. The Neighbor-Joining tree was calculated from the distance matrix using the NEIGHBOR program of the PHYLIP (Phylogeny inference) package v.3.6a2.1 (distributed by J. Felsenstein, Department of Genetics, University of Washington, Seattle). The tree was visualized using the NJPLOT program [53]. Abbreviation: *T*, time.

is assumed to be constant throughout the simulation, when considering only molecular data from extant lineages the branches leading to the cenancestor appear long and empty (Figure 2b,c). This is similar to many molecular phylogenies (e.g. rRNA [6], vacuolar ATPases, F type ATPases [35] and many other proteins that are sufficiently conserved for tree of life studies [43]). The speciation rate observed only through inspection of the surviving lineages increases as one moves towards more recent times. Figure 3 summarizes the increase in the number of lineages over time observed in repeated simulations. The semi-logarithmic plot reveals that the number of lineages increases faster than exponential. This is because only extant lineages were considered. If all species present at a particular time were considered and constant rates of speciation and extinction are assumed, the number of species would follow a simple exponential, with a growth rate equal to the difference between speciation and extinction rates [16]. However, if only the lucky ancestors are considered, as shown here, the number of lineages grows faster than exponential.

Plotting the number of lineages through time provides a means to assess the deviations between a reconstructed phylogeny and a model [18,19,44]. The shape of the lineages-through-time plot reflects the relative contribution of extinctions, the extent to which the diversity of a group is sampled and the increase of coexisting species. At one extreme, if there are no extinctions, there is a simple exponential growth curve where all branches are on average the same length. The occurrence of long and

empty branches at the base of a tree depends on non-negligible rates of extinction [14]. In the past, the shape of phylogenies (as captured in lineage-through-time plots) was used mainly to characterize evolutionary processes of macroscopic eukaryotes; however, these approaches are equally useful to study the evolution of prokaryotes [45].

Analogy between allele coalescence within a species and cladogenesis
The study of human evolution provides a powerful illustration of the coalescence principle applied to the derivation of most recent common ancestors: on the basis of analyses of the Y chromosome and mitochondrial genes, Y-chromosome Adam and mitochondrial Eve never met. The Y chromosome of all extant human males is traced back to a most recent common ancestor who lived ~50 000 years ago [46,47], whereas the mitochondrial genes trace back to a most recent common ancestor who lived ~166 000–249 000 years ago [48,49]. Therefore, there are thousands of years that separate 'Adam' and 'Eve'. These findings are not restricted to the Y chromosome and the mitochondrial genome. Biparentally inherited genes also are not expected to coalesce to the same individual; however, owing to recombination between the two parental alleles their histories are more difficult to reconstruct.

Outlook
Our simple coalescence model can potentially provide an estimate of the EFFECTIVE NUMBER OF SPECIES N_E of all
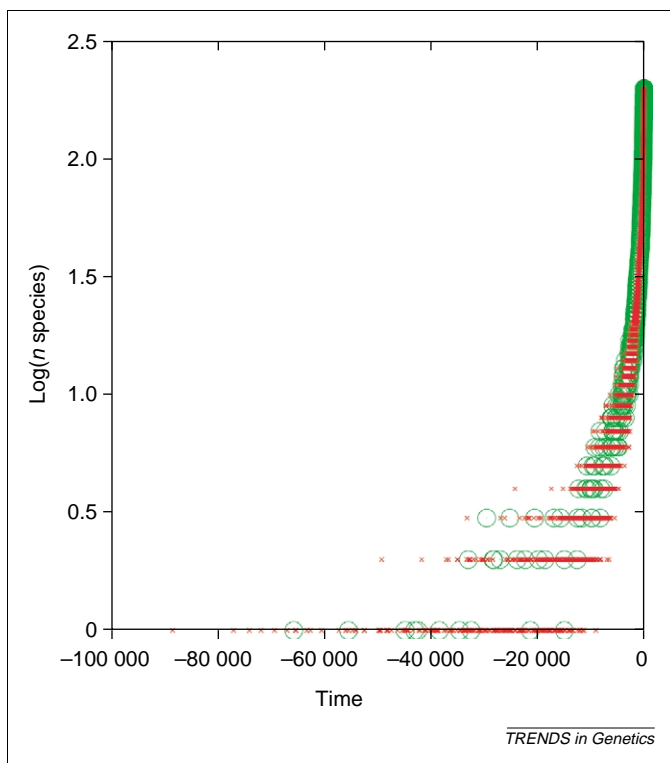
Figure 3. The number of extant species at each generation that have progeny among the contemporary species. The simulations were performed for 200 species. Time 0 indicates the present time. Data from 10 independent simulations of organismal evolution are shown in green, and for each organismal simulation 25 simulations of gene evolution were performed [one horizontal gene transfer (HGT) event per 10 generations] and are shown in red. Note the semi-logarithmic coordinate system and the faster than exponential increase in lineages. Comparison between red and green points shows that HGT can shorten or extend the time needed for coalescence

living organisms during the phase of balanced speciation and extinction. Because the current number of species is large, the coalescence to the cenancestor occurred $\sim 2N_e$ 'generations' ago [30]. In the context of species evolution 'generation' refers to the expected time interval between two successive speciation events (or between speciation and the consecutive extinction event) occurring in a lineage. If this time interval and the time of the cenancestor was known, one could estimate the effective number of species.

Many complex models for the coalescence of alleles in populations have been developed. Some of these promise to be useful in studying organismal lineages. An example is provided by the models that consider the migration between two subpopulations [30,50]. Applied to the evolution of organismal and MOLECULAR LINEAGES this coalescence model enables incorporating non-random HGT. If the transfer between two groups of organisms is much smaller than within groups, the molecular phylogenies will quickly coalesce within each group. But the coalescence between two groups is longer and mainly determined by the gene-transfer rate between groups (analogous to two populations with low levels of migration between them). This suggests that some phyletic patterns that are frequently found in molecular phylogenies might have been due to preferential gene transfer and not shared ancestry.

Consideration of a simple symmetrical model as a null hypothesis will be most valuable in those instances where

the data deviate from the prediction [51]. For example, under the simplifying conditions of our model, Kingman's coalescence can be applied to any CLADE of sufficient size. For each of these clades, the coalescence of the two deepest branches extends, on average, over half the time the clade was in existence. In the case of the bacterial domain this is clearly not true: the different bacterial phyla all appear to emerge in a single radiation, suggesting that during this period actual evolution deviated from the assumptions of our null hypothesis, possibly because of an actual increase in species number.

Concluding remarks

The hypothesis of constant extinction and speciation provides a reasonable null hypothesis for cladogenesis. This hypothesis alone explains some features of the observed topology of the tree of life. Therefore, it does not appear warranted to invoke more complex hypotheses – or to derive complex scenarios involving bottlenecks and extinction events – to explain features of the tree of life that are compatible with the null hypothesis.

Using single genes as phylogenetic markers, it is difficult to trace organismal phylogeny in the presence of HGT. Combining the simple coalescence model of cladogenesis with rare HGT events suggests that there was no single last common ancestor that contained all of the genes ancestral to those shared among the three domains of life. Each contemporary molecule has its own history and traces back to an individual molecular cenancestor. However, these molecular ancestors were likely to be present in different organisms at different times.

References

1 Gogarten, J.P. *et al.* (1989) Evolution of the vacuolar H + -ATPase: implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 86, 6661–6665
2 Iwabe, N. *et al.* (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. U. S. A.* 86, 9355–9359
3 Forterre, P. and Philippe, H. (1999) Where is the root of the universal tree of life? *BioEssays* 21, 871–879
4 Lopez, P. *et al.* (1999) The root of the tree of life in the light of the covarion model. *J. Mol. Evol.* 49, 496–508
5 Philippe, H. and Forterre, P. (1999) The rooting of the universal tree of life is not reliable. *J. Mol. Evol.* 49, 509–523
6 Woese, C.R. (1987) Bacterial evolution. *Microbiol. Rev.* 51, 221–271
7 Caetano-Anolles, G. (2002) Evolved RNA secondary structure and the rooting of the universal tree of life. *J. Mol. Evol.* 54, 333–345
8 Doolittle, R.F. *et al.* (1996) Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science* 271, 470–477
9 Feng, D.F. *et al.* (1997) Determining divergence times with a protein clock: update and reevaluation. *Proc. Natl. Acad. Sci. U. S. A.* 94, 13028–13033
10 Adkins, R.M. and Li, W-H. (1998) Dating the age of the last common ancestor of all living organisms with a protein clock. In *Horizontal Gene Transfer* (Syvanen, M. and Kado, C.I.   pp. 380–391, Academic Press

11 Hedges, S.B. *et al.* (2001) A genomic timescale for the origin of eukaryotes. *BMC Evol. Biol.* 1, 4

12 Gogarten, J.P. *et al.* (1996) Dating the cenancestor of organisms. *Science* 274, 1750–1751

13 Crick, F.H.C. (1981) *Life Itself: its Origin and Nature*, Simon and Schuster

14 Barraclough, T.G. and Nee, S. (2001) Phylogenetics and speciation. *Trends Ecol. Evol.* 16, 391–399

15 Watson, H.W. and Galton, F. (1875) On the probability of the extinction of families. *Journal of the Anthropological Institute of Great Britain and Ireland* 4, 138–144

16 Raup, D.M. (1985) Mathematical models of cladogenesis. *Paleobiology* 11, 42–52

17 Hey, J. (1992) Using phylogenetic trees to study speciation and extinction. *Evolution* 46, 627–640

18 Nee, S. (2001) Inferring speciation rates from phylogenies. *Evolution Int. J. Org. Evolution* 55, 661–668

19 Nee, S. *et al.* (1994) Extinction rates can be estimated from molecular phylogenies. *Philos. Trans. R. Soc. Lond. Ser. B* 344, 77–82

20 Yule, G.U. (1924) A mathematical theory of evolution based on conclusions of Dr. J.C. Willis. *Philos. Trans. R. Soc. Lond. Ser. B* 213, 21–87

21 Zillig, W. *et al.* (1992) A model of the early evolution of organisms: the arisal of the three domains of life from the common ancestor. In *The Origin and Evolution of the Cell* (Hartman, H. and Matsuno, K., eds), pp. 163–182, World Scientific Publishing

22 Kandler, O. (1994) The early diversification of life. In *Early Life on Earth, Nobel Symposium* (Vol. 84) (Bengston, S., ed.), pp. 152–509, Columbia University Press

23 Woese, C. (1998) The universal ancestor. *Proc. Natl. Acad. Sci. U. S. A.* 95, 6854–6859

24 Koch, A.L. (1994) Development and diversification of the last universal ancestor. *J. Theor. Biol.* 168, 269–280

25 Gogarten-Boekels, M. *et al.* (1995) The effects of heavy meteorite bombardment on the early evolution – the emergence of the three domains of life. *Orig. Life Evol. Biosph.* 25, 251–264

26 Nisbet, E.G. and Sleep, N.H. (2001) The habitat and nature of early life. *Nature* 409, 1083–1091

27 Miller, S.L. and Lazcano, A. (1995) The origin of life - did it occur at high temperatures? *J. Mol. Evol.* 41, 689–692

28 Wright, S. (1931) Evolution in Mendelian populations. *Genetics* 16, 97–159

29 Kingman, J.F.C. (1982) The coalescent. *Stochastic Processes and their Applications* 13, 235–248

30 Felsenstein, J. (2003) *Inferring Phylogenies*, Sinauer Associates

31 Moran, P.A.P. (1958) Random processes in genetics. *Proc. Camb. Philol. Soc.* 54, 60–71

32 Gilbert, W. (1986) The RNA world. *Nature* 319, 618

33 Joyce, G.F. (1989) RNA evolution and the origins of life. *Nature* 338, 217–224

34 Hilario, E. and Gogarten, J.P. (1993) Horizontal transfer of ATPase genes – the tree of life becomes a net of life. *BioSystems* 31, 111–119

35 Olendzenski, L. *et al.* (2000) Horizontal transfer of archaeal genes into the Deinococcaceae: detection by molecular and computer-based approaches. *J. Mol. Evol.* 51, 587–599

36 Gogarten, J.P. (1995) The early evolution of cellular life. *Trends Ecol. Evol.* 10, 147–151

37 Woese, C.R. *et al.* (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol. Mol. Biol. Rev.* 64, 202–236

38 Koonin, E.V. *et al.* (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.* 55, 709–742

39 Nesbo, C.L. *et al.* (2001) Phylogenetic analyses of two 'archaeal' genes in thermotoga maritima reveal multiple transfers between archaea and bacteria. *Mol. Biol. Evol.* 18, 362–375

40 Zhaxybayeva, O. and Gogarten, J.P. (2002) Bootstrap, Bayesian probability and maximum likelihood mapping: Exploring new tools for comparative genome analyses. *BMC Genomics* 3, 4

41 Gogarten, J.P. *et al.* (2002) Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* 19, 2226–2238

42 Doolittle, W.F. (1999) Phylogenetic classification and the universal tree. *Science* 284, 2124–2129

43 Brown, J.R. *et al.* (2002) Horizontal gene transfer and the universal tree of life. In *Horizontal Gene Transfer* (Syvanen, M. and Kado, C.I., eds), pp. 305–349, Academic Press

44 Mooers, A.O. and Heard, S.B. (1997) Inferring evolutionary process from phylogenetic tree shape. *Q. Rev. Biol.* 72, 31–54

45 Martin, A.P. (2002) Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl. Environ. Microbiol.* 68, 3673–3682

46 Thomson, R. *et al.* (2000) Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc. Natl. Acad. Sci. U. S. A.* 97, 7360–7365

47 Underhill, P.A. *et al.* (2000) Y chromosome sequence variation and the history of human populations. *Nat. Genet.* 26, 358–361

48 Cann, R.L. *et al.* (1987) Mitochondrial DNA and human evolution. *Nature* 325, 31–36

49 Vigilant, L. *et al.* (1991) African populations and the evolution of human mitochondrial DNA. *Science* 253, 1503–1507

50 Hudson, R.R. (1990) Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology* (Vol. 7) (Futuyma, D. and Antonovics, J., eds), pp. 1–44, Oxford University Press

51 Hubbell, S. (2001) *The Unified Neutral Theory of Biodiversity and Biogeography*, Princeton University Press

52 Fitch, W.M. and Upper, K. (1987) The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. *Cold Spring Harb. Symp. Quant. Biol.* 52, 759–767

53 Perriere, G. and Gouy, M. (1996) WWW-Query: an on-line retrieval system for biological sequence banks. *Biochimie* 78, 364–369