J. theor. Biol. (2002) **218,** 395–408
doi:10.1006/yjtbi.3086, available online at http://www.idealibrary.com on IDEAL®

Testing the Hypothesis of Common Ancestry

ELLIOTT SOBER†‡ AND MICHAEL STEEL*§

†*Department of Philosophy, University of Wisconsin, Madison, WI* 53706, *U.S.A.*, ‡*London School of Economics and Political Science, London WC2A, 2AE, U.K. and* §*Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand*

(*Received on* 11 *November* 2001, *Accepted in revised form on* 15 *May* 2002)

The hypothesis that all life on earth traces back to a single common ancestor is a fundamental postulate in modern evolutionary theory. Yet, despite its widespread acceptance in biology, there has been comparatively little attention to formally testing this "hypothesis of common ancestry". We review and critically examine some arguments that have been proposed in support of this hypothesis. We then describe some theoretical results that suggest the hypothesis may be intrinsically difficult to test. We conclude by suggesting an approach to the problem based on the Aikaike information criterion.

1. Introduction

It is a central tenet of modern evolutionary theory that all living things now on earth trace back to a single common ancestor. Not only do plants share a common ancestor and animals do the same; in addition, plants and animals derive from a common progenitor. This proposition, which we call the *Hypothesis of Common Ancestry*, does not assert that life on earth arose just once. Multiple start-ups are allowed, if all of them, except one, went extinct. It also does not exclude the possibility that some genetic material has been exchanged between primitive ancestral species.

This proposition is central because it is presupposed so widely in evolutionary research. When biologists attempt to reconstruct the phylogenetic relationships that link a set of species, they usually *assume* that the taxa under study are genealogically related. Whether one uses cladistic parsimony, distance measures, or maximum likelihood methods, the typical question is *which* tree is the best one, not *whether* there is a tree in the first place.[||] The same pre-supposition is at work in the pattern of reasoning that biologists often use to develop adaptive hypotheses. When biologists consider the

*Corresponding author. Tel.: +64-3-366-7001, +64-3-366-7688; fax: +64-3-364-2587.
E-mail address: m.steel@math.canterbury.ac.nz (M. Steel).

[||] In saying that parsimony pre-supposes the existence of a common ancestor, and so does not permit a test of that hypothesis, we are claiming that the following inference is invalid. In that argument, being alive (L) is treated as a derived character and non-life (N) as ancestral, this on the grounds that life arose from nonlife. One then uses parsimony to conclude that the tree in Fig. 1(a) is more plausible than the tree in Fig. 1(b). The problem with this line of reasoning is that it pre-supposes that non-living things form a genealogical tree, subject to descent with modification. Although this may be true of *some* non-living things, it surely is not true of *all*. If reticulate genealogies produced by hybridization render parsimony inapplicable in the case of organisms, surely this causal pattern (wherein effects have more than one direct cause) undermines the applicability of parsimony in the domain of non-living things as well.

0022-5193/02/$35.00/0

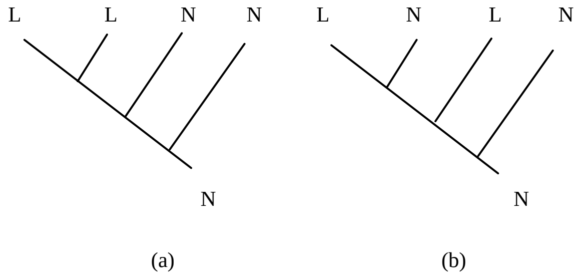


FIG. 1.

possible adaptive reasons why a species exhibits some trait, they usually think about the trait as evolving against a background of biological features already in place. They infer what that ancestral condition was by assuming that there is a phylogenetic tree that unites the species of interest with other species. Traits of sister groups are then "read back" into the past (using parsimony or some other method of inference), thereby providing an estimate of the trait values of ancestors.

In view of the importance within evolutionary biology of the Hypothesis of Common Ancestry, it is worth reviewing what evidence there is that the hypothesis is true. In this paper, we assess the arguments that have been made in the biological literature and discuss a methodology that has not been applied to this problem before.

2. Previous Arguments and Analyses

Perhaps the most frequently cited argument for a single common ancestor is Crick's (1968) idea that the genetic code is a "frozen accident", meaning that the pattern by which nucleotide triplets code amino acids is functionally arbitrary. There is no adaptive reason why UUU should code for phenylalanine (which it does) rather than isoleucine (which it does not). Given this hypothesis, the (near) universality of the genetic code among the earth's organisms provides strong evidence that all those organisms trace back to a common ancestor. If the Hypothesis of Common Ancestry were true, we would expect the code to be universal; if lineages arose separately, we would expect them to exhibit different codes. The crucial idea in this argument is that different codes constitute different adaptive peaks between which there are deep valleys. It is not essential that all peaks have the same height. This is fortunate, since it has been argued that the code we now observe is optimal (Freeland *et al.*, 2000). What is fundamental is the idea of stabilizing selection—once an organism deploys a given code, selection will tend to retain that trait in the organism's descendants.

Although we do not object to Crick's argument as far as it goes, we think it is worthwhile to consider whether other lines of evidence support the Hypothesis of Common Ancestry. For one thing, Crick's argument is based on a single shared trait; although that trait has obvious biological importance, it would be interesting to see if other traits can be brought to bear on the question. Few biologists would be content to use a single trait in reconstructing the best phylogenetic tree. Why, then, is a single trait sufficient to settle the question about common ancestry? In fact, Crick's argument does not stand on its own; the same form of argument can be applied to other universals of biochemistry. For example, the fact that all amino acids found in proteins are left-handed likewise can be viewed as a frozen accident, since right-handed amino acids evidently would work just as well. These features, taken together, provide a more powerful argument for common ancestry than any one of them does singly.

Another feature of Crick's argument that leads us to think that the question is worth pursuing further is the fact that the argument is not quantitative. It asserts that there are multiple adaptive peaks with deep valleys between, but how deep are those valleys? We know from evolutionary theory that valleys can be traversed. It would be worthwhile to develop a quantitative version of Crick's qualitative argument.

Finally, it is worth observing that Crick's argument depends on the correctness of a certain functional analysis of the genetic code. Yet, the standard methods for determining which tree is best supported by a body of data—for example, parsimony and maximum likelihood—do not require a prior assessment of function. Can this approach be applied to the Hypothesis of Common Ancestry itself?

Another argument for the Hypothesis of Common Ancestry can be constructed by coupling an idea in the theory of the origin of life with an idea in the theory of stochastic processes. Oparin (1953) suggested that when life first arises from non-living materials, it alters the environment so as to make subsequent start-ups much less probable.¶ Generalizing this point, let us suppose that each new start-up makes the next one less probable. If we add to this idea the process of the "extinction of family names", we have the skeleton of an argument—it is highly probable that all life will *eventually* trace back to just one ancestor, if enough time elapses from the date of the first start-up. This argument, like Crick's, is limited by the fact that it is merely qualitative (Sober, 1999). What is the expected number of start-ups? How probable is it, given the time that has actually elapsed, that all but one of those start-ups should have disappeared?**

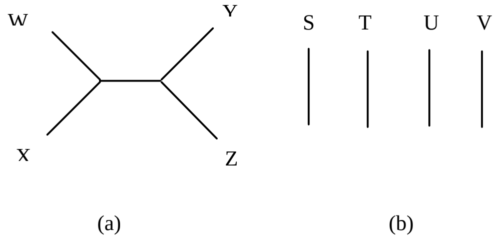


FIG. 2.

A third proposal for testing the Hypothesis of Common Ancestry was formulated by Penny *et al.* (1982). They used parsimony as their method of phylogenetic reconstruction, however their approach (and our comments on it) would apply to other tree reconstruction methods, such as maximum likelihood, or distance-based methods. In addition, they relied on the concept of *character congruence*. A method of phylogenetic inference judges that two sets of characters are congruent when the method says that they point to the same phylogenetic tree. Penny *et al.* reasoned that if there is a single tree uniting a set of species, then the different characters that evolve in that tree should be highly congruent (when parsimony is used to identify the tree that is best supported by each character); on the other hand, if the species originated separately, the characters should be much less congruent. The five sets of characters they considered —each of which consisted of nucleotide sequence data for a particular protein across a group of mammals and a marsupial—were highly congruent.††

While this test procedure may have some merit, it is flawed in two ways. The first problem is that a tree can generate *non*-congruent characters with high probability; the second is that lineages that arise independently of each other can generate highly congruent characters. The first possibility can be visualized by considering the unrooted tree (WX)(YZ) depicted in Fig. 2(a). Suppose we obtain two data sets from this tree, and that the characters in the two data sets evolved according to different rules. For characters in the first set, the rules of evolution make it overwhelmingly probable that the tip species will occupy states W = 1, X = 1, Y = 0, and Z = 0 (1100, for short); in contrast, characters in the second set evolve according to rules that make it overwhelmingly probable that the

¶ In fact, this argument is presented by Darwin: "It is often said that all the conditions for the first production of a living organism are now present, which could ever have been present. But if (and oh! What a big if!) we could conceive in some warm little pond, with all sorts of ammonia and phosphoric salts, light, heat, electricity, etc. present, that a proteine compound was formed, ready to undergo still more complex changes, at the present day such matter would be instantly devoured or absorbed, which would not have been the case before living creatures were formed (Darwin 1887, vol. 3, p. 18)."

** There is a further detail. The process of the extinction of family names, strictly speaking, pertains to the survival dynamics of variants that arise in tree topologies. The traditional pattern that the label was invented to describe has surnames passing from a single parent (the father) to children. The dynamics of the process are different under other topologies. For example, suppose that life began with *two* start-ups, which go extinct right after they hybridize to produce a daughter species, and that subsequent descent from that daughter is strictly tree-like [as in Fig. 5(d)]. One no longer can predict that if one waits long enough, one of those two start-ups must eventually have no descendants.

†† The idea behind this test is not that the existence of a tree should lead us to expect that the most parsimonious tree constructed from all the data at hand will contain few homoplasies. Rather, the idea is that different trees constructed from different data sets should largely agree if and only if all the species are genealogically related, regardless of whether those different trees each contain lots of homoplasy.

tip species will exhibit the pattern 1010. The two data sets, though produced by evolution in the same tree, will therefore be highly incongruent. The second problem for Penny *et al.*'s congruence test is illustrated by the four unrelated lineages leading to the tip species S, T, U, and V depicted in Fig. 2(b). Suppose that all characters evolve according to rules that make it overwhelmingly probable that the tip species will exhibit the 1100 pattern, thus leading parsimony to erroneously infer that either ST is monophyletic, or UV is, or both are.

We conclude that the congruence test is subject to type-1 and type-2 errors, the first arising from rate differences among traits, the second arising from rate differences among lineages. Without some assurance that the probabilities of these errors are small, we conclude that character congruence does not provide a good test of the Hypothesis of Common Ancestry. Notice that both these sources of error would disappear if a trait followed the same rules of evolution in all lineages and if all traits followed the same rules. However, this reply raises the question of why this restrictive model of character evolution should be thought to apply to the five proteins that Penny *et al.* considered. We note that the probability of a type-1 error would be reduced if the data sets were constructed by randomly sampling from a larger data set (though this was not part of the approach of Penny *et al.*).

Even if one were willing to assume that different traits evolve according to the same rules, there still is no unconditional guarantee that a tree will generate data sets that are more congruent than the data sets that would be generated if taxa arose separately. As Penny *et al.* note at the end of their article, the Hypothesis of Common Ancestry can be statistically indistinguishable from the hypothesis of multiple originations if lineages are sufficiently old. If the past events of interest occurred sufficiently long ago, they may constitute a *temps perdu*, as we now discuss.

3. Why a Past Event May be Unknowable

It is worth pausing to consider the possibility that there may be no way to find out whether the Hypothesis of Common Ancestry is true. To describe this further we recall some concepts from information theory, in particular the *mutual information* between two random variables X and Y, written $I(X,Y)$ and defined by

$$I(X,Y) = \sum_{x,y} P(X = x \,\&\, Y = y) \log\left(\frac{P(X = x \,\&\, Y = y)}{P(X = x)P(Y = y)}\right).$$

Informally, $I(X,Y)$ is a nonnegative number that measures the degree to which knowledge of the value that Y takes conveys information about the value X takes (and vice versa). In particular, $I(X,Y) = 0$ if and only if X and Y are independent. Moreover, when $I(X,Y)$ is close to zero, no method can reliably predict X from knowledge of Y. That is, no method can do much better than just ignoring Y and guessing X; this idea can be formalized by reference to "Fano's Inequality" (Cover & Thomas, 1991).

A simple analytical encapsulation of the concept of time as an information destroying process is provided by the classical "data processing inequality" from information theory, one form of which states:

If $X \rightarrow Y \rightarrow Z$ forms a Markov chain,

then $I(X,Z) \leqslant I(Y,Z)$.

That is, in the chain that goes from distal cause to proximate cause to effect, the effect provides at least as much information about the proximate cause as it does about the distal cause; information fades as we make inferences that go deeper and deeper into the past. For example, suppose we are considering a phylogeny, where X is the branching pattern early on, Y is the more recent branching pattern, and Z summarizes the character states of tip species. Even if the data on tip species provide substantial information about relatively recent branching events, they may fail to do so about branching events that took place in the more distant past. It is useful to quantify this and we do so now, using some very recent results from probability theory.

In the following discussion, we will suppose that a group of species has a common ancestor and that these have evolved according to a tree from this common ancestor (this does not necessarily entail the Hypothesis of Common Ancestry for all life, but just for the species we are considering). Thus, we will suppose we have a phylogenetic tree T whose leaf vertices comprise a set S of extant species. Given a discrete character that assigns to each species from S a corresponding character state we wish to use this information to infer an ancestral state at some interior node in the tree. Relating this to $I(X,Y)$, the random variable X will specify this unknown ancestral state, and Y will specify the joint assignment of states to the extant species. We wish to use Y to infer X, and we will be helped in this task if we know the underlying model of character evolution, the underlying phylogeny T, and its corresponding branch durations with high precision. Yet, even in this optimistic setting, if $I(X,Y)$ is small it is impossible to reliably infer the ancestral state from the extant states (and this is true regardless of whether one's preferred methodology is maximum likelihood, parsimony, Bayesian-based or otherwise). We will see shortly that this has further consequences for the resolution of deep divergences in tree reconstruction, and for testing the Hypothesis of Common Ancestry.

When will $I(X,Y)$ be small? Essentially, it is when the character is "saturated" by having undergone too many substitutions due to some combination of a high rate of substitution and large time-scales. However, the story is complicated by the fact that trees that last longer may have more tip species, and larger numbers of tip species provide a larger inferential basis from which to infer the state of the root of the tree (Salisbury & Kim, 2001). Thus, increasing the duration of the tree has both a negative and a positive effect—it moves lineages closer to saturation, but it also gives rise to more lineages, and hence to more sources of information. For this reason, it is sometimes possible to have more information about a node (A_1) in the tree that is further back in time from the present than one has concerning the character state at a more recent node (A_2), as shown in Fig. 3.

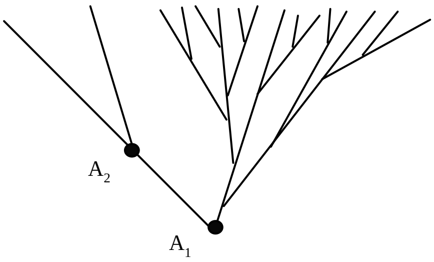


FIG. 3.

To quantify the loss of information we can use some very recent results from Evans *et al.* (2000). To simplify the discussion we will assume that the character is 2-state (binary) and that it undergoes substitution according to a symmetric (Poisson) model with a constant rate of r substitutions (on average) per year. The restriction to symmetric models on binary characters is for convenience, and some of the underlying theory has recently been generalized to multi-state characters by Mossel (2001). Furthermore, for certain non-symmetric models of multi-state character evolution, some similar bounds can be derived (Mossel, in prep.) although the technical details are slightly more involved.‡‡

Suppose that the most recent common ancestor of the species in S lived t years ago, as in Fig. 4(a). Then from Evans *et al.* (2000) the following result can easily be deduced, as shown in the Appendix:

Theorem 1. *Under the* 2*-state, constant rate symmetric Poisson model, operating on any phylogeny T, let Y denote the joint states on the set S of leaves of T and let X denote the state at*

‡‡ In contrast to the binary symmetric case, there exist non-symmetric random models of multi-state character evolution that do not lead to the eventual loss of information (regarding an ancestral state) over time, even when we have just a single lineage. For example, consider three states α, β, γ that are initially equally probable, but evolve in such a way that α may randomly mutate into either β or γ (but β or γ may not mutate to any other state). Observing the state after any period of time allows us to correctly estimate the root state with probability at least 1/2, in contrast to the prior probabilities of 1/3, so the information about the root state is never lost completely.

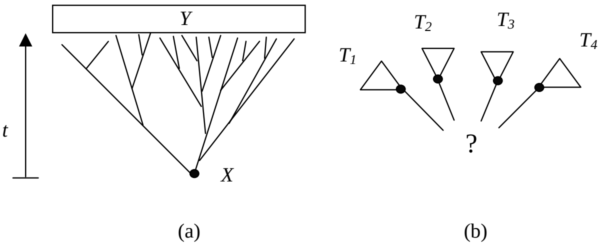


FIG. 4.

the root of T. Then, the mutual information I(X,Y) satisfies the bound

$$I(X,\, Y) \leqslant ne^{-4rt},$$

where n is the number of species in S, t is the time to the most recent common ancestor of S, and r is the substitution rate.

Notice that the statement of the theorem is very general in that although $I(X,Y)$ depends on the underlying tree, T, and its branch durations, these parameters do not enter explicitly into the inequality shown. Note also that the term rt is the expected number of substitutions that the character has undergone on the path from the root of T to any species in S. If this expected number is large, the only hope for using Y to estimate X is if the number of species in S is large, but the relationship in the theorem shows that n must grow exponentially with rt to keep $I(X,Y)$ away from 0.

For example, suppose we have a group of 1000 species, and their most recent common ancestor lived at least 20 million years ago. Then for a binary character that follows a symmetric substitution model with an average rate of one substitution per 2 million years we have $I(X,\, Y) < 1000\mathrm{e}^{-40}$. Fano's inequality (Cover & Thomas, 1991) then guarantees that no method can infer X from Y with a probability that is (to many decimal places) any better than simply ignoring Y and blindly guessing X.

3.1. SIGNIFICANCE FOR TREE RECONSTRUCTION AND TESTING THE HYPOTHESIS OF COMMON ANCESTRY

We now describe the significance of this result for two further questions. Firstly, Theorem 1 has a direct bearing on the limits of tree reconstruction, in particular, for the resolution of "deep divergences" in a tree. Suppose we have four monophyletic groups, and we wish to use character data to infer which of the three possible (unrooted) topologies describes the relationship between the four groups, as depicted in Fig. 4(b). Suppose our character data consist of aligned binary sequences of length k. As before, we will be helped in our task (of determing the relationships between the groups) if we know the underlying model of character evolution, the underlying phylogenies of the four groups (T_1–T_4) and the corresponding branch durations with high precision. So we will assume that trees T_1–T_4 are known and that the sequence sites evolve independently and identically according to a 2-state, constant rate Poisson model with known branch durations. How large then must k be in order to determine the relationship between the four groups? As before, this depends on the interaction between the factors affecting saturation (substitution rates and divergence times) and the underlying phylogenies (T_1–T_4). Here, the data processing inequality is useful. Let X denote the unknown topological relationship of the four trees T_1–T_4, that is, the branching pattern, indicated by a question mark in Fig. 4(b) that we seek to determine. Let Y denote the four sequences at the roots of the four trees [each of which is indicated by a dot in Fig. 4(b)] and let Z denote the collection of sequences at the tips of the four trees (our data). We wish to use Z to infer information about X. Since $X \rightarrow Y \rightarrow Z$ forms a Markov chain, the data processing inequality gives

$$I(X,Z) \leqslant I(Y,Z).$$

Let us write $Y = [Y_{ij}]$, where Y_{ij} is the state at the i-th site of the sequence at the root of tree T_j, and similarly let us write $Z = [Z_{ij}]$, where Z_{ij} is the aligned column of states at the i-th site of the sequences at the tips of tree T_j. By the assumption that the sites evolve independently, and according to a Markov process on a tree, the Z_{ij} are conditionally independent once we specify Y. Consequently, by the subadditivity property,

$$I(Y,Z) \leqslant \sum_{\substack{i=1,\ldots,k\\ j=1,\ldots,4}} I(Y,Z_{ij}).$$

Furthermore,

$$I(Y,Z_{ij}) = I(Y_{ij},Z_{ij}) = I(Y_{1j},Z_{1j}),$$

where the first equality applies since, given Y_{ij}, the random variable Z_{ij} is conditionally independent of all the other components of Y; the second equality follows from the assumption that the sites evolve identically. Combining these (in)equalities, and applying Theorem 1 to bound the term $\mathrm{I}(Y_{1j},Z_{1j})$ gives immediately the following result:

Corollary. *The mutual information between the branching pattern* (X) *of the four trees in Fig.* 4*b and the sequences* (Z) *at the tips of the tree satisfies*

$$I(X,Z) \leqslant 4k \max_i \{n_i \mathrm{e}^{-4rt_i}\},$$

where n_i is the number of tip species in T_i, t_i is the time to the most recent common ancestor of T_i, r is the substitution rate and k is the length of the sequences.

This corollary sets explicit bounds on the limits to which deep divergences can be resolved with any reasonable accuracy in terms of some basic quantities (n_i, t_i, k, r). For certain data sets there can be no realistic possibility of resolving deep divergences. For example, suppose we were interested in some deep divergence in the tree of life, such as the relationship between four groups, each of which has a common ancestor that dates back at least one billion years ago. Suppose further that, for each group, we have binary sequence data of length 10 000 perfectly aligned across say 100 species. Suppose the substitution rate is at least one substitution per 100 million years. Then the bound given by the corollary (together with Fano's inequality) implies that no method can reliably determine from this data how these four groups are related historically.

The corollary is also significant for the question of testing the Hypothesis of Common Ancestry. For, referring to our example depicted in Fig. 4(b), if we cannot determine which of the three possible (unrooted) topologies describes the relationship between the four groups because of the saturation effect described, then a similar argument shows that we also cannot determine from the data whether the four groups have evolved independently from separate common ancestors. In short: no tree, no test of the hypothesis of common ancestry.

The above discussion is relevant to primary sequence data under simple models of character evolution. It would be illuminating to extend this analysis to more elaborate models of sequence evolution, such as covarion-type models, where sites switch "on" and "off" over time (see Huelsenbeck, 2002 or Penny *et al.* 2001). Similarly, it would be useful to investigate models for the evolution of secondary and tertiary structure, and to compare the information-theoretic loss for this data with that of the primary sequence data.

The earth came into existence about 4.5 billion years ago, and life made its first appearance (as far as we know) about 0.7 billion years thereafter. Suppose that all organisms now alive trace back to a single ancestor that existed 3.5 billion years ago. Some species (such as the ones that Penny *et al.* studied) have their most recent common ancestor much more recently; they are very closely related. Others have their most recent common ancestor much longer ago. Closely related species can be expected to retain evidence of their common ancestry. Is the same true of more remotely related species? Is it true of *all* the species now alive? The Hypothesis of Common Ancestry needs to be looked at more closely, especially since evolutionary theory suggests the possibility that information about very ancient events may have been whited out by the passage of time.

4. The Competing Hypotheses

The Hypothesis of Common Ancestry says that there exists a single ancestral origin to which all present-day living things trace back. This hypothesis competes with alternative hypotheses that say that the number of ancestors is 2, 3, ..., or n. More precisely, the hypotheses we want to consider have the following form for $i = 1, 2, \ldots n$:

CA$-i$ There existed a set A consisting of i species, and no set with fewer than i species, such that:

(i) none of the species in A are ancestral to any other species in A,

(ii) each of the current species (S_1, S_2, ..., S_n) has at least one ancestor in A, and

(iii) each species in A is ancestral to at least one S_k.

Notice that, given a set of n current species, the n hypotheses CA-1,CA-2, ..., CA-n are mutually exclusive, and one of them must be true. Notice also that if the hypothesis CA-i correctly describes a set of species, then for any subset of those species the hypothesis CA-j applies for some $j \leqslant i$.

One possible criticism of our formulation is that it involves the concept of an ancestral "species" which may be problematic or ambiguous for certain origin-of-life models. For example, the progenote theory of Woese (1998, 2000), postulates an ancestral community of "primitive types of cellular entities" engaged in extensive horizontal gene transfer as the precursor to the three major domains of life (Archae, Bacteria and Eukaryotes). If one regards this ancestral community as constituting a single "species" then this theory may be compatible with CA-1. However, certain formulations of this theory might also support CA-i for $i > 1$. We will not explore this issue further here, but note that this issue is relevant to any formulation or testing of the hypothesis.

The Hypothesis of Common Ancestry thus occupies one extreme on a scale; it sets $i = 1$. Located at the other end of this scale is the hypothesis that each present day species is the result of a separate origination event. This last hypothesis, which sets $i = n$, has been defended by creationists, but it need not be given a theistic formulation. Notice that we can consider these n competing hypotheses in terms of what they say about *all* the species now in existence, or with respect to what they say about a more limited set of current species.

Figure 5 depicts four genealogies, each consistent with the hypothesis of common ancestry (CA-1). Figure 5(a) exhibits the familiar bifurcating tree topology. Figure 5(b) is a star phylogeny. Figure 5(c) contains reticulation; branches merge as well as split. In all three of these genealogies, A is the first progenitor from which all tip species derive. In topology 4d, however, there are two start-ups (A_1 and A_2) and neither goes extinct; however the two lines deriving from A_1 and A_2 at some point merge, and all tip species derive from the resulting bottleneck (A). Figure 5(b) illustrates the fact that the hypothesis of common ancestry does not require that there be a nested hierarchy. Fig. 5(c) shows that the hypothesis of common ancestry does not require a strict tree topology. And from Fig. 5(d) we see that the hypothesis does not require that only one of the many start-ups that life may have had has survived to the present. The hypothesis requires this if all genealogies are strictly tree-like, but not if there are reticulations, and so Figure 5(d) supports CA-1.

To test these n hypotheses against each other, we must determine what each predicts about the observable features of organisms. However, this is far from straightforward, since each of these hypotheses covers a range of cases, and these different cases confer different probabilities on the data. Each hypothesis (CA-i) is consistent with several *kinds* of topology, and each kind of topology is consistent with several *specific* topologies. For example, the Hypothesis of Common Ancestry (CA-1) is consistent with a strict tree-like genealogy, but it also is consistent with reticulations, as we have just seen. And within these two types of topology, there are numerous specific topologies. For example, for n tip species, there are $(2n-3) \times (2n-5) \times \cdots \times 5 \times 3$ rooted trees. The same situation obtains,

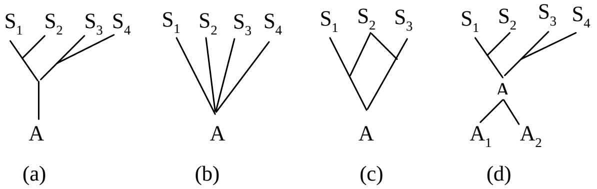
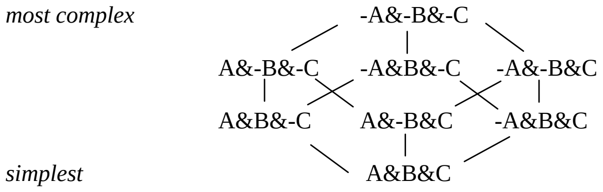


FIG. 5.

of course, for (CA-2). Here, we have to consider the possibility that there are two trees, that there is one tree and one reticulate structure, and that there are two reticulate graphs.

However, we still are not finished, in that even a specific topology does not, by itself, confer a probability on the traits we observe. What we need next is a model of the processes of character evolution that occur in the different branches in a genealogy. Here again, there are many possibilities. A model may require that a trait follow the same rules of evolution on different branches, or allow it to obey different rules ("branch homogeneity/heterogeneity"). And the model may constrain different traits on a given branch to obey the same rules, or allow them to follow different rules ("trait homogeneity/heterogeneity"). And the model may stipulate, of a single trait on a given branch, that all changes are equiprobable (i.e. the probability of changing from *i* to *j* is the same as the probability of changing from *k* to *l*, for all states *i*, *j*, *k*, *l*), or it may allow different changes to have different probabilities. There are further distinctions that might be drawn among different process models—for example, a model need not assume that different traits evolve independently of each other on a given branch—but we will ignore these complications here.

These different features of the process model's parameterization—the rules governing a single trait on different branches, the rules governing different traits on the same branch, and the rules governing different changes in a single trait on a single branch—are logically independent of each other. The set of all process models will be the logical product of these different elements. It will comprise a partially ordered set of models, with a maximal element (one that is most complex) and a minimal element (one that is simplest). The simplest model says that all traits follow the same rules (A), that each trait follows the same rules on all branches (B), and that all changes that a single character can experience on a given branch must have the same probability (C); the Jukes–Cantor (1969) model for nucleotide evolution is of this form. At the opposite extreme is a model that allows traits to follow different rules (-A), and allows a single trait to follow different rules on different branches (-B), and also allows each possible change of a single trait on a single branch to have its own probability (-C). The partial ordering is as follows:



Each of these process models contains adjustable parameters, with simpler models containing fewer. When we conjoin a graph G_i with a process model M_j, we still have not obtained a hypothesis that confers a probability on the observations. However, we can estimate the values of the parameters in the model G_i & M_j by using maximum likelihood estimation. Once the adjustable parameters in G_i & M_j are fixed at their likeliest values, we obtain a statement that contains no adjustable parameters, namely L(G_i & M_j). This, finally, is a hypothesis of the type we have been seeking. Note that G_i & M_j and G_i & M_k may be nested, but that G_i & M_j and G_k & M_j are not. The set of topology/model pairs we are considering consists of disjoint subsets, each of whose members forms a partial ordering.

At long last we have obtained hypotheses that confer probabilities on the observed features of tips species. We now face two questions: How are these different conjunctive hypotheses to be evaluated? And how does their evaluation bear on the question with which we began—that of assessing the hypothesis of Common Ancestry (CA-1) against its competitors (CA-*i*, for $i > 1$)?

5. The Problem of Realism and the Mimic Theorem

Since there are so many process models that one could consider (far more than the eight examples described above), perhaps the problem could be simplified by taking a single, realistic process model and discarding the rest. The most realistic process model will be maximally complex. This is because the assumptions A, B, and C are *idealizations*—we know that they are false. On the other hand the model –A & -B & -C is entirely open-minded. For example, this model does not assert that different traits follow different rules, but merely assigns each trait its own suite of parameters; these may be equal in value, or unequal. This realistic model allows the data to decide what the best settings of these parameters are; it does not stipulate in advance that parameters for different traits must have exactly the same values.

Although realism and open-mindedness are often scientific virtues, they spell disaster in this and many other inference problems (Sober, 2001). The reason is that the most general/realistic model (M_r) leads all tree topologies to have the same likelihood, namely the maximum value of 1 — when T_i & M_r and T_j & M_r are each fitted to the data, L(T_i & M_r) and L(T_j & M_r) will have the same maximal likelihood. As Lewis (1998) notes, increasing the complexity of the process model leads the likelihoods of all topologies to increase, but also to draw closer to each other.

We now describe a mathematical result (Theorem 2) that more precisely captures the problem at hand. Suppose that G_0 is the true underlying graph describing the evolution of a collection of species, and suppose that a process model M_0 with its parameters set at P_0 governs character evolution. Then, for any other graph G_i, there exists an associated process model M_i whose parameters are set at P_i such that G_i & M_i & P_i fits the data at hand as well as T_0 & M_0 & P_0 does. Here either the true graph G_0 or the alternative G_i may involve the CA-*i* for any values of *i*. In other words, a false topology can mimic the fit-to-data achieved by the true topology, once that false topology is equipped with a suitably chosen process model.

Before describing this result formally, we provide a simple example, which is germane to our project of testing CA-1 against competing hypotheses of the form CA-*i* (for $i > 1$). This example is illustrated in Fig. 6. The tip species A and B either (6a) trace back to a common ancestor C, or (6b) are the products of separate origination events. Suppose that (6a) is true and that dichotomous characters evolve on the branches of this tree by following a symmetric model of character evolution, wherein

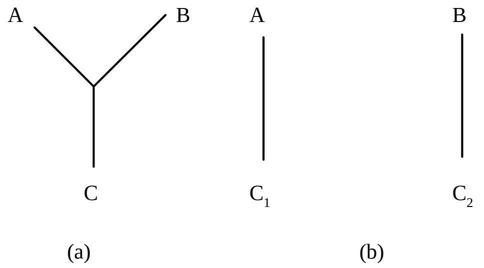


FIG. 6.

$P(0 \rightarrow 1) = P(1 \rightarrow 0) = 0.25$ on each branch. Suppose that $C = 0$ is the ancestral condition for each character. Then the joint distribution of possible character states for A and B is

$$P(A = 1 \,\&\, B = 1) = 3/16,$$

$$P(A = 1 \,\&\, B = 0) = 3/16,$$

$$P(A = 0 \,\&\, B = 1) = 3/16,$$

$$P(A = 0 \,\&\, B = 0) = 7/16.$$

We can also calculate the marginal probabilities $P(A = 1) = P(B = 1) = 3/8$. Notice that when A and B evolve from their common ancestor C according to the process model just stated, their character states will be correlated, in that

$$P(A = 1 \,\&\, B = 1) > P(A = 1)P(B = 1).$$

If A and B have correlated character states in virtue of their common ancestry as shown in Fig. 6(a), how can the hypothesis of separate ancestry depicted in Fig. 6(b) mimic this result? The answer is simple: we merely adopt a different, and more complex, process model, wherein different traits evolve according to different rules. More precisely, suppose, that (6b) is true and that traits of two types evolve on this topology. Traits of type-1 all begin with C_1 and C_2 is state 1. Traits of type-2 all begin with C_1 and C_2 in state 0. Let 25% of the traits be of the first type and 75% of the second. If the probability of change is 1/4 on each branch, the hypothesis of separate origination (6b) will mimic the predictions of the hypothesis of common ancestry (6a). Symmetrically, if the data are generated according to model (6b), it is easy to mimic the resulting data by suitable parameter settings for the model described by (6a).

We generalize this idea as follows. Suppose we have a fixed, finite set of character states (e.g. A, C, G, T), and let $p(a_1,\ldots, a_n)$ denote the joint probability of generating the character state a_1 for species 1, a_2 for species 2, …, a_n for species n. Under the extreme hypothesis CA-n, of independent origin of species, and allowing for a mixture of different processes, we can write

$$p(a_1,...,a_n) = \sum_{j=1}^{N} \pi_j \prod_{i=1}^{n} f_{ij}(a_i) \qquad (1)$$

for suitable probability distributions f_{ij} and π and some number N of categories.

On the other hand, if we have a "tree-model" but also allow mixed processes (e.g. for DNA sites to evolve at different rates), and for each category j of classes we have a Markov model M on a binary tree, with non-degenerate (positive, but finite) parameters (e.g. branch lengths) θ_j, we can write

$$p(a_1,...,a_n) = \sum_{j=1}^{N'} \pi'_j P(a_1,...,a_n|T, M, \theta_j), \quad (2)$$

where $P(a_1,\ldots,a_n|T, M, \theta_j)$ is the probability of generating the character $a_1,\ldots,a_n$ by Markov process M on tree T with parameters θ_j.

Let $p = [p(a_1,\ldots, a_n)]$ denote the vector of all these r^n joint probabilities, where r is the size of the underlying state space. Whether we can distinguish between these two models depends on what restrictions we place on the model M, and on N and N' (the number of categories, in relation to n) and the values θ_j can take, which may depend on constraints that apply across categories. Certainly, there are cases where we can distinguish between these models, for example, if $N' = N = 1$. However, if we impose no restrictions then the two models confer the same probabilities on any possible observation of the character states at the tip species. More precisely, we have the following result whose proof is given in the Appendix:

Theorem 2. *For any probability distribution p on C^n, we can represent p exactly by a probability distribution of the type of eqn* (1), *and we can represent p arbitrarily closely* (*and using any tree T*) *by a probability distribution of the type of eqn* (2). *In particular it is impossible to distinguish between models described by eqns* (1) *and* (2) *in terms of their fit to data without imposing additional constraints or assumptions.*

However, once some assumptions are in place it becomes possible to distinguish between the models. For example, suppose we require that M is a continuous-time symmetric, stationary Poisson-process model (e.g. Jukes–Cantor, or Kimura 3ST), but we do not impose any further restrictions (i.e. the θ_j and π'_j are completely unconstrained). Then it can be shown [from eqn (2)] that $p(a_1,\ldots, a_n)$ is maximized when $a_1 = a_2 = \cdots = a_n$; however, this is not necessarily the case for the "non-tree" model. A more interesting example, is the following: if M is any stationary reversible model, and, for each j, the branch lengths θ_j satisfy a molecular clock (but are otherwise unconstrained) then this places considerable constraints on p, even without knowing anything about N', π' or the further details of M. Indeed p suffices to reconstruct T (and from just the induced pairwise marginal distributions, see Theorem 4 of Steel & Penny, 2001). Another restriction that might prove useful in distinguishing between the models is to limit the size of N relative to n, since the proof of Theorem 2 requires N to grow quickly with n.

6. A Way Forward

Standard frequentist statistics provides a methodology for coping with the problem we are considering, but its scope of application is limited. If two models are nested, one can use a likelihood ratio test to decide whether the simpler model should be rejected. Thus, it is perfectly possible in this framework to compare A & B & C & G$_i$ and A & B &-C & G$_i$ and to compare A & B & C & G$_j$ and A & B &-C & G$_j$. If the simpler model cannot be rejected in either topology, one then can compare A & B & C & G$_i$ and A & B & C & G$_j$. But suppose that the data permit one to reject A & B & C & G$_i$, but do not permit one to reject A & B & C & G$_j$. One then needs to compare A & B &-C & G_i and A & B & C & G_j. The likelihood ratio test does not apply in this instance—one cannot compare different topologies that have different process models attached to them.

A solution to this problem is furnished by the Akaike information criterion (AIC). AIC is based on a theorem (Akaike, 1973; Sakamoto *et al.*, 1986) that describes how the predictive accuracy of a model M containing adjustable parameters can be estimated:

An unbiased estimate of the predictive accuracy of M $\approx$ log-likelihood[L(M)]$-k$.

L(M) is the hypothesis obtained from M by assigning values to adjustable parameters that maximize the probability of the data. Good fit-to-data enhances a model's estimated predictive accuracy, but there also is in Akaike's theorem a penalty for complexity, represented by k, the number of adjustable parameters in M. One does not automatically embrace the more complex model—its gain in likelihood must be sufficient to compensate for its loss in simplicity. AIC applies to nested and non-nested models alike, and it was recently applied in the context of phylogenetic analysis of molecular data by Posada & Crandall (1998, 2001a, b).

With this methodology in hand, we can return, finally, to the problem with which we began. To test CA-1 against various CA-i ($i>1$), one should consider many specific evolutionary graphs (G_1, G_2, ..., G_n) and a variety of different process models (M_1, M_2, ..., M_m). One then should calculate the AIC score for each conjunction (G_i & M_j). This will involve taking account of the likelihood of L(G_i & M_j), relative to the data set one is using, but also one will have to attend to the number of parameters in G_i & M_j.§§ We do not advocate accepting the hypothesis G_i & M_j that has the highest score and rejecting all the rest. Rather, one should look down the list of hypotheses, from best to worst, and note which of these entails CA-1, which entails CA-2, and so on. Suppose that the first n hypotheses on this list entail CA-1 and that a hypothesis entailing CA-i (for $i>1$) appears only further down. The larger n is, the stronger the evidence is for the hypothesis of common ancestry. In addition, one needs to attend to the magnitude of the AIC scores, not just to their ordering (Burnham & Anderson 1998). It would be desirable to have a method for quantifying the weight of evidence that each CA-i hypothesis receives, given the range of models and topologies considered, but we are unsure how best to do this. Nonetheless, we suggest that an AIC-based approach will help illuminate the problem of testing the hypothesis of common ancestry. Although it is important in this procedure that several process models be considered, it may not be necessary for the set of terminal taxa to be large. For example, one could address the question of whether human beings and a species of yeast have a common ancestor—here there are just two topologies, the ones depicted in Fig. 6.

7. Conclusion

The hypothesis of common ancestry is central to contemporary evolutionary theory. However, a valid methodology for testing that hypothesis that allows one to look at suites of characters has not, until now, been available. We hope that biologists will use the protocol we have described for different sets of species. It may turn out that all is well with the conventional wisdom on this subject. What Crick defended by considering a single characteristic—the genetic code—may be vindicated when one considers sets of characteristics whose functional significance is less well understood. But perhaps it will emerge that for some groups of species, the data do not provide unambiguous support for the idea that there is a single common ancestor. If there is a *temps perdu*—if genealogical connections are unretrievable for events that are sufficiently far in the past—this is something that biology needs to ascertain. And if one of the CA-i ($i>1$) hypotheses fares *better* than CA-1, this too would be a result of considerable interest.

§§The number of parameters in (G_i & M_j) is not a function just of M_j. For example, the simple (A&-B&C) model discussed in Section 4 has three branch parameters when applied to the topology depicted in Fig. 6(a), but only two when applied to the topology in Fig. 6(b).

We are grateful to Christopher Hitchcock for pointing out to us how Simpson's Paradox can apply to phylogenies and allow hypotheses of separate origination to predict correlations among tip species, and for other helpful comments. We also thank James Crow, and a referee report by David Penny, Michael Hendy and Anthony Poole for comments on an earlier version of this manuscript. E.S. thanks the National Science Foundation (grant SES-9906997) and M.A.S. thanks the *New Zealand Marsden Fund* for supporting this research.

REFERENCES

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In: *Second International Symposium on Information Theory* (Petrov, B. & Csaki, F., eds), pp. 267–281. Budapest: Akademiai Kiado.

Burnham, K. P. & Anderson, D. R. (1998). *Model Selection and Inference. A Practical Information-theoretic Approach*. New York: Springer-Verlag.

Cover, T. M. & Thomas, J. A. (1991). *Elements of Information Theory*. New York: John Wiley and Sons, Inc.

Crick, F. (1968). The origin of the genetic code. *J. Mol. Biol.* **38,** 367–379.

Darwin, F. (1887). *The Life and Letters of Charles Darwin*, Vol. 1–3. London: Murray. Reprinted by Johnson Reprint Corp., New York, 1969.

Evans, W., Kenyon, C., Peres, Y. & Schulman, L. J. (2000). Broadcasting on trees and the Ising model. *Adv. Appl. Prob.* **10,** 410–433.

Freeland, S., Knight, R., Landweber, L. & Hurst, L. (2000). Early fixatioin of an optimal genetic code. *Mol. Biol. Evol.* **17,** 511–518.

Huelsenbeck, J. P. (2002). Testing a covariotide model of DNA sequence substitution. *Mol. Biol. Evol.* **19,** 698–707.

Jukes, T. & Cantor, C. (1969). Evolution of protein molecules. In: *Mammalian Protein Metabolism* (Munro, H., ed.) pp. 21–132. New York: Academic Press.

Lewis, P. (1998). Maximum likelihood as an alternative to parsimony for inferring phylogeny using nucleotide sequence data. In: *Molecular Systematics of Plants II.* (Soltis, D., Soltis, P. & Doyle, J., eds), pp. 132–163. Boston: Kluwer.

Mossel, E. (2001). Reconstruction on trees: beating the second eigenvalue. *Ann. Appl. Prob.* **11,** 285–300.

Oparin, A. (1953). *The Origin of Life*. New York: Dover Books.

Penny, D., McComish, B. J., Charleston, M. A. & Hendy, M. D. (2001). Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *J. Mol. Evol.* **53,** 711–723.

Penny, D., Foulds, L. R. & Hendy, M. D. (1982). Testing the theory of evolution by Comparing phylogenetic trees constructed from five different protein sequences. *Nature* **297,** 197–200.

Posada, D. & Crandall, K. (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14,** 817–818.

Posada, D. & Crandall, K. (2001a). Selecting models of nucleotide substitution: an application to Human Immunodeficiency virus 1 (HIV-1). *Mol. Biol. Evol.* **18,** 897–906.

Posada, D. & Crandall, K. (2001b). Selecting the best-fit model of nucleotide substitution. *Syst. Biol.* **50,** 580–601.

Sakamoto, Y., Ishiguro, M. & Kitagawa, G. (1986). *Akaike Information Criterion Statistics*. New York: Springer-Verlag.

Salisbury, B. A. & Kim, J. (2001). Ancestral state estimation and taxon sampling density. *Syst. Biol.* **50,** 557–564.

Sober, E. (1999). Modus Darwin. *Biol. Philos.* **14,** 253–278.

Sober, E. (2002). Instrumentalism, parsimony and the Akaike framework. *Philos. Sci.* (in press).

Steel, M. & Penny, D. (2001). Parsimony, likelihood, and the role of models in molecular phylogenetics. *Mol. Biol. Evol.* **17,** 839–850.

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

Woese, C. R. (2000). Interpreting the universal phylogenetic tree. *Proc. Natl Acad. Sci. U.S.A.* **97,** 8392–8396.

APPENDIX

Proof of Theorem 1. By Theorem 1.3 of Evans *et al.* (2000),

$$I(X, Y) \leqslant \sum_{v} \theta_v^2,$$

where the summation is over all leaves of the tree, and

$$\theta_v = \prod_{e \in path(root,\, v)} (1 - 2p(\mathrm{e})),$$

where $p(\mathrm{e})$ is the probability of a net substitution across e. Now if substitutions occur at constant rate r throughout the tree, then

$$p(\mathrm{e}) = \tfrac{1}{2}(1 - \mathrm{e}^{-2rt(e)})$$

where $t(\mathrm{e})$ is the temporal duration of edge e. Consequently,

$$\theta_v = \mathrm{e}^{-2rt},$$

where t is the total time from the root of the tree to the present. The result now follows from the equation at the beginning of the proof. □

Proof of Theorem 2. Denote the finite state space (of size r) by the set C. Suppose that p is any probability distribution on C^n. We will show that p can be represented by eqn (1). Let us order the

$N = r^n$ elements of C^n as $a^{(1)}, \ldots, a^{(N)}$. For $j = 1, \ldots, N$, consider the following (degenerate) probability distribution f_{ij}, defined on the state space C by setting:

$$f_{ij}(b) = \begin{cases} 1 & \text{if}(a^{(j)})_i = b, \\ 0 & \text{if}(a^{(j)})_i \neq b. \end{cases}$$

Then, $p(a_1, \ldots, a_n) = \sum_{j=1}^{N} \pi_j \prod_{i=1}^{n} f_{ij}(a_i)$, where $\pi_j = p(a^{(j)})$. This completes the proof of the first claim.

Conversely, suppose p is any probability distribution on C^n. By the first part of the theorem, p can be represented by eqn (1). We will show that p can be represented arbitrarily closely by eqn (2) for any (rooted) tree T and with $N' = N$. Let us assign a uniform distribution of states from C at the root vertex of T. For each category value j, and edge, e of T, consider the non-stationary, continuous-time Markov process described by an intensity matrix Q_j(e) operating at rate λ(e) on edge e. The associated transition matrix for edge e of T is given by the familiar relationship:

$$M_j(\mathrm{e}) = \mathrm{e}^{\lambda(e)Q_j(e)}.$$

It remains to specific λ(e) and Q_j(e). These can be completely arbitrary if e is an interior edge of T. If e_i is the pendant edge of T incident say with leaf i, s*et al*l the off-diagonal entries in the i-th column of $Q_j(\mathrm{e}_i)$ equal to $f_{ij}(l)$. Using standard Markov process arguments, as $\lambda(\mathrm{e}_i)$ tends to infinity, the matrix $M_j(\mathrm{e}_i)$ converges to the matrix that has all entries in the i-th column equal to $f_{ij}(l)$. Consequently, for any $\varepsilon > 0$, one may take λ(e) sufficiently large so that $|M_j(\mathrm{e}_i)_{kl} — f_{ij}(l)| < \varepsilon$ holds for all values of k, l in C and each value of i and j. For the model as described, the probability of generating the character $(a_1, \ldots, a_n)$ is

$$\sum_{(k_1, \ldots, k_n) \in C^n} Pr(\cap_{1 \leqslant i \leqslant n} \{v(i) = k_i\})$$

$$\prod_{i=1}^{n} (M(e_i))_{k_i a_i} = \prod_{i=1}^{n} f_{ij}(a_i) + O(\varepsilon),$$

where $v(i)$ is the ancestral vertex to vertex i. If we now take a mixture of such non-stationary processes by setting $\pi'_j = \pi_j$ for $j = 1, \ldots, N$, we obtain the required representation, up to terms involving ε, via eqn (2). This completes the proof. □