Time and Knowability in Evolutionary Processes

Elliott Sober and Mike Steel*†

Historical sciences like evolutionary biology reconstruct past events by using the traces that the past has bequeathed to the present. Markov chain theory entails that the passage of time reduces the amount of information that the present provides about the past. Here we use a Moran process framework to show that some evolutionary processes destroy information faster than others. Our results connect with Darwin's principle that adaptive similarities provide scant evidence of common ancestry whereas neutral and deleterious similarities do better. We also describe how the branching in phylogenetic trees affects the information that the present supplies about the past.

1. Introduction. What is the epistemic relation of present to past? Absent a time machine, we are trapped in the present and must rely on present traces to learn about the past. There are memory traces inside the skull, but outside there are tree rings, fossils, and traces of other kinds. People use these traces to reconstruct the past. Sometimes they simply assume that the traces provide unerring information about the past, but often they realize that the jump from present to past is subject to error. A bevy of epistemic concepts can be pressed into service to investigate the relation of present traces to past events, ranging from strong concepts like knowledge and certainty to more modest ones like justified belief and evidence.

Received February 2014; revised April 2014.

*To contact the authors, please write to: Elliott Sober, Philosophy Department, University of Wisconsin, Madison, WI, USA; e-mail: ersober@wisc.edu. Mike Steel, Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand.

†Elliott Sober presented this paper in 2012 at the University of Bordeaux, the London School of Economics, and the Institute for Mathematical Philosophy at the Ludwig Maximilian University in Munich, and received valuable comments. We are grateful for these and also to Elchanan Mossel for helpful comments. Elliott Sober thanks the William F. Vilas Trust of the University of Wisconsin–Madison, and Mike Steel thanks the Allan Wilson Centre (New Zealand).

Philosophy of Science, 81 (October 2014) pp. 558–579. 0031-8248/2014/8104-0009$10.00

We are interested in how the natural processes connecting past to present constrain our ability to know about the past by looking at the traces found in the present. An optimistic view of these processes is that the past is potentially an open book; all we need do is understand the connecting processes correctly and look around for the right traces. If the relation of the past state of a system to its present state were deterministic and one to one, this optimistic view would be correct. If only we could know the present state with sufficient precision, and if only we could grasp the true mapping function that connects present and past, we would be home free. This optimism is something that Laplace (1814, 4) affirmed when he discussed *une intelligence* (now referred to as "a demon"):

> We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes.

It is worth noting that determinism is not sufficient for the optimistic view to be true; without the one-to-one assumption, distinct states of the past may map onto the same state of the present, with the consequence that the exact state of the past cannot be retrieved from even a perfectly precise grasp of the present.

Is determinism necessary for the optimistic view to be true? It is not, provided that we set to one side strong concepts like knowledge and certainty and take up an epistemic evaluation that is more modest. Consider, for example, a process in which the system is, at each moment, in one of two states (coded 0 and 1). Suppose Past = 0 makes Present = 0 extremely probable (say, 0.96) and that Past = 1 makes Present = 1 extremely probable as well (say, 0.98). This means that when we observe the system's present state, we gain strong evidence that discriminates between the two hypotheses Past = 0 and Past = 1. We cannot infer from Present = 0 that the past state was certainly 0; in fact, we cannot even infer that the past state was probably 0. But we can conclude that the observation favors the hypothesis that Past = 0 over the hypothesis that Past = 1. This conclusion is licensed by what Hacking (1965, 59–62) calls the Law of Likelihood:

> Observation O favors hypothesis H_1 over hypothesis H_2 if and only if $\Pr(O|H_1) > \Pr(O|H_2)$.

Royall (1997, 9–11) suggests that this qualitative principle should be supplemented by a quantitative measure of favoring:

> The degree to which O favours H_1 over H_2 is given by the likelihood ratio $\Pr(O|H_1)/\Pr(O|H_2)$.

Royall further suggests that a reasonable convention for separating strong evidence from weak is a ratio of 8. Royall's suggestion entails that the probabilistic process just described has the consequence that Present = 0 provides strong evidence favoring Past = 0 over Past = 1, since the likelihood ratio is $0.96/0.02 = 48$.

This simple example should not be overinterpreted. If there were a process connecting past to present in the way described, the present would provide strong evidence about the past. Do not forget the *if*. Perhaps there are such processes, especially when the past under discussion is the recent past. But what if we consider not just the recent past, but past times that are more and more ancient? How does increasing the temporal separation between present and past affect the amount of information that the present provides about the past?

2. Two Theorems. A simple theorem provides an answer to the question just posed (Cover and Thomas 2006). Consider a system that at any time is in one of n possible states $(s_1, s_2, \ldots, s_n)$. For simplicity we shall think of the system as evolving in discrete time steps. We stipulate that the system has the following two properties:

> **The Markov property**. For any two times $t_1 < t_2$, the state of the system at time t_1 screens-off the system's history prior to t_1 from the state at t_2. That is, for all states x and y:
>
> $\Pr(\text{system is in state } y \text{ at } t_2 | \text{system is in state } x \text{ at } t_1) =$
> $\Pr(\text{system is in state } y \text{ at } t_2 | \text{system is in state } x \text{ at } t_1 \ \& \ \text{system's history prior to } t_1)$.

Note that the Markov property does not require that the transition probabilities are constant with time (often called 'time-homogeneous' chains); rather, they may vary from step to step.

The second property we stipulate is sometimes referred to, in the time-homogeneous, finite-state setting, as the Markov chain's being "regular"; we will use the same term in the broader setting we are considering in which time homogeneity is relaxed:

> **Regularity**. For some positive integer n, and some strictly positive real value ε, the following holds for all ordered pairs of states (i, j): given that the system at any given time t is in state i, the probability that the system is in state j at time $t + n$ is at least ε.

This condition asserts that it is possible to move from any given state to any other state (including the original state) in a fixed finite number of steps with a probability that remains bounded above zero. Regularity for finite-state time-homogeneous Markov processes is equivalent to the condition that the Markov chain is 'aperiodic' and 'irreducible' (for details, see Häggström [2002, corollary 4.1]).

For any system of this sort, the following result holds (a precise statement and proof is provided in app. A):

> **Exponential information loss theorem**. If a finite-state system satisfies the Markov property and regularity, then I(Past; Present) is less than or equal to a term that approaches zero exponentially fast as the time between the present and past increases.

Here $I(X; Y)$ is the "mutual information" linking the two variables. If the variables are discrete, the formula for this quantity is

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right),$$

where $p(x, y)$ is the joint probability that $X = x$ and $Y = y$, and $p(x)$, $p(y)$ are the (marginal) probabilities that $X = x$, and that $Y = y$, respectively. Mutual information measures how much (on average) you learn about the state of one of the variables by observing the state of the other. Its value is zero when X and Y are independent; otherwise it is positive. Mutual information is symmetrical: $I(X; Y) = I(Y; X)$.

The exponential information loss theorem generalizes the special case where transition probabilities are constant with time (Sober and Steel 2011, proposition 6). This generalization is important, since many processes (including the biological examples we will discuss) often change their rates from one period of time to another.

Note that the theorem does not ensure a monotonic decline in information as the temporal separation of past and present is increased. That extra element is provided by a different result, the so-called Data Processing Inequality (DPI; Cover and Thomas 2006, 32):

> *The Data Processing Inequality*: In a causal chain from a distal cause D to a proximate cause P to an effect E, if P screens-off D from E, then $I(E; D)$ is less than or equal to both $I(E; P)$ and $I(P; D)$.

For a discrete-state process, these two inequalities are strict whenever P is neither perfectly correlated with D or with E, nor is P independent of them (see app. B). The Data Processing Inequality does not require that the process linking D to P is the same as the process linking P to E.

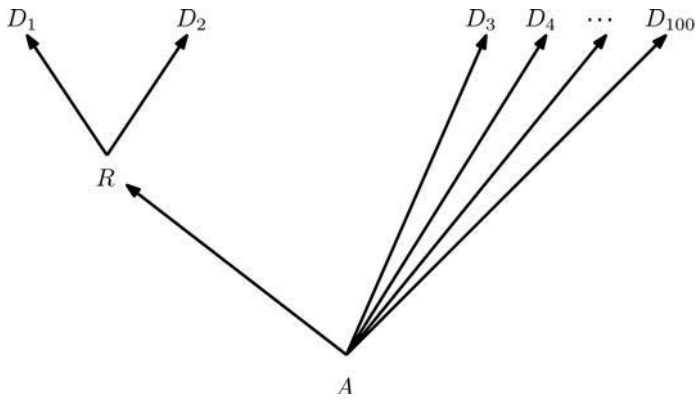

Figure 1. A case in which it is possible for the present to provide more information about an ancient event (A) than about a more recent event (R).

The information-processing inequality is "chain internal." It does not say that the present always provides more information about recent events than about ones that are older. Consider figure 1. Suppose that R screens-off A from D_1 & D_2 Then the information-processing inequality says that $I(D_1 \,\&\, D_2; R) \geq I(D_1 \,\&\, D_2; A)$. It does not say that $I(D_1 \,\&\, D_2 \,\&\, \ldots \,\&\, D_{100}; R) \geq I(D_1 \,\&\, D_2 \,\&\, \ldots \,\&\, D_{100}; A)$. It is perfectly possible that the hundred descendants that now exist provide more information about A than they do about R. Note that R does not screen-off A from D_1 & D_2 & . . . & D_{100}. The heightened information that the leaves provide about A is not due to the simple fact that A has a larger number of outgoing lineages than R does; it is possible to modify this tree so that each nonleaf node (including the root node A) has just two descending lineages (Sober and Steel 2011, fig. 5b, 243–34), and still the leaves provide more information about A than they do about R, owing to the values of the transition probabilities attaching to branches.

Both the exponential information loss theorem and the Data Processing Inequality are very general. They characterize any system whose laws of motion have the requisite probabilistic features. The system might be a chamber of gas, but it also might be an evolving population of organisms. Indeed, if there were disembodied spirits that changed probabilistically, the results would apply to them. Both results are more general than physics—they cover the systems and properties that are discussed in the laws of physics, but they also apply to systems and properties that are not. In addition, both results are a priori mathematical truths, although of course it is an empirical matter whether a given system satisfies the antecedent of the conditional that each result expresses (Sober 2011a).

3. Five Evolutionary Processes. Since our interest here is in how evolutionary processes affect the amount of information that the present provides about the past, it is worth making clear how the exponential information loss theorem applies to models of biological evolution. In phylogenetics, the rapid loss of information in models of nucleotide substitution with time has been highlighted as a significant problem for using DNA sequence data to accurately resolve deep divergences of species lineages (for a recent review, see Salichos and Rokas [2013]) and for inferring ancestral states deep within a given tree (Mossel 2003; Gascuel and Steel 2014).

We will return to phylogenies in a later section, but for now we consider information loss in population genetics. The application of information theory to population genetics has been investigated a bit. For example, Frieden, Plastino, and Soffer (2001) explore a variational principle ("extreme physical information") based on Fisher information to study genotype frequency changes. The questions we consider here are different, as our primary interest is in the relative ranking of likelihood ratios across different evolutionary processes, including both drift and different types of selection. The Moran (1962) models of evolution will be our workhorse in what follows. We consider a population containing N individuals. The population evolves through a sequence of discrete temporal "moments" (how long a moment is will not matter). At each moment, one of those N individuals produces a copy of itself and one of those N individuals dies. We consider two traits A and B; each individual has one of them or the other. At any moment, the population is in one of $N + 1$ states (ranging from 0% A to 100% A). This Moran framework can be articulated in different ways to represent different evolutionary processes. For example, if individuals are chosen at random to reproduce and die, then we have a drift process. Selection processes of different kinds can be represented by letting A individuals have chances of dying or reproducing that differ from those possessed by B individuals. A population undergoing a Moran process forms a Markov chain, with its recent past screening-off its more remote past from the present.

Suppose we observe the population in the present and see that all N individuals are in state A. How much information does that observation provide about the state of the population at some earlier time? If all states are accessible to each other (which requires that mutations can prevent the population from getting "stuck" at 100% A or 100% B), then the exponential information loss theorem applies and so the mutual information declines asymptotically to zero with time. However, if there is no mutation, then the population will evolve to either 100% A or 100% B and will stay there. In this case, the present state of the population provides information about its past even if the two are infinitely separated. For example, if we observe that the population is now 100% A and the population has been evolving by drift, this

observation favors the hypothesis that the population was at more than 95% A at some earlier time over the hypothesis that it was 5% A or less, and this is true regardless of the time separation between the past and the present.

John Maynard Keynes (1924, chap. 3) once said that "in the long run, we are all dead." His point was to pooh-pooh the relevance of claims about the infinite long run. What should matter to us mortals is finite time. This point applies to the bearing of the exponential information loss theorem on our knowledge of the evolutionary past. Who cares if mutual information goes to zero as the time separating present from past goes to infinity? Life on earth is a mere 3.8 billion years old. What is relevant is that information decays monotonically in Markov chains. But in addition, we know that there are different kinds of evolutionary process. Which of these speeds the loss of information and which slows it?

The five processes we want to investigate are represented in figure 2; each of them can result in our present observation that all N of the individuals in the population now have trait A. In panel i, trait A was favored by selection. In panel iii, there was selection against trait A. In panel ii, the traits are equal in fitness, and so the traits evolved by pure drift. In panel iv, selection favored the majority trait. And in panel v, selection favored the minority trait. We are not asking which of these process hypotheses is most plausible, given the observed state of the population at present. Rather, we want to explore what happens to information loss under each of these five scenarios, in each case thinking of the process in the context of the Moran framework of a finite population of fixed size N.

To compare the five processes in this respect, we calculate the following likelihood ratio for each of them:

$$R_{ij} = \frac{\Pr(\text{Present} = N|\text{Past} = i)}{\Pr(\text{Present} = N|\text{Past} = j)}, \text{ for } i > j.$$

Although our main interest is to compare cases in which i is close to N and j is close to 0, the Moran framework makes it easy to derive results for the more general case in which $i > j$.

So the observation is that all N of the individuals now in the population have trait A. Does this observation favor the hypothesis that there were exactly i individuals at some past time who had trait A over the hypothesis that there were exactly j, where $i > j$? It does; $R_{ij} > 1$ for each of the five processes we are considering (see app. C). Our question is how the magnitude of R_{ij} depends on the underlying evolutionary process. It is worth noting that although the likelihood ratio R_{ij} is "past-directed" (in that it describes the degree to which a present observation discriminates between two hypotheses about the past), evaluating this ratio requires one to consider

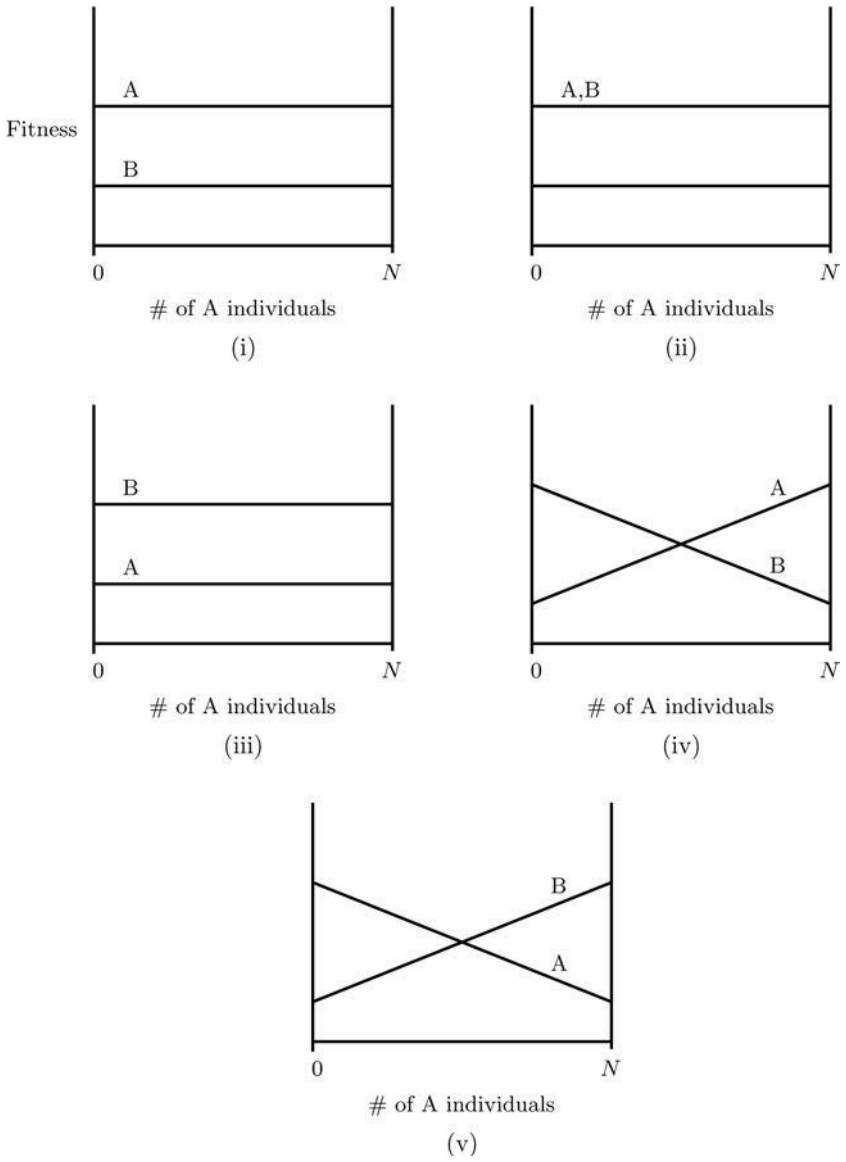

Figure 2. Five processes that can result in all N of the individuals now in the population having trait A.

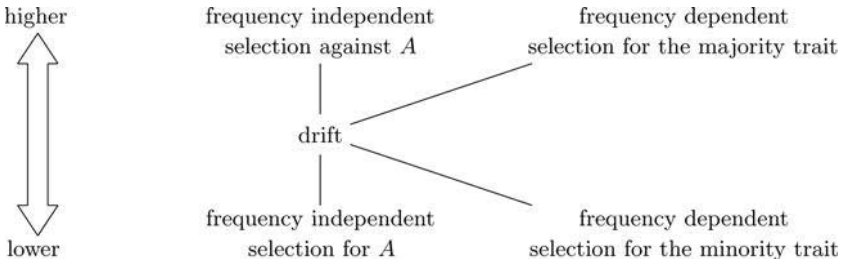

Figure 3. Comparing R_{ij} values for five processes, assuming infinite temporal separation of Past from Present and zero mutation. The relation of the two frequency-dependent processes to drift is derived using the assumption that $j < N/2$.

two "future-directed" probabilities—the probability of reaching Present $= N$ if the system begins at Past $= i$ and the probability of reaching Present $= N$ if the system begins at Past $= j$.

We begin by adopting an assumption that we treated above with Keynesian disdain. Let us assume that the temporal separation of past and present is infinite and that there is zero mutation. We will relax this idealization in due course. For each of the processes we are considering, we now can describe what the value of R_{ij} is for each pair of values for i and j such that $i > j$. For example, under neutral evolution the value of R_{ij} is i/j. The values for the other four processes are given in appendix C. The ordering of the R_{ij} values for the five processes is depicted in figure 3.

Let us first consider three of the cases described in figure 3—drift and the two cases of frequency-independent selection. The ordering of R_{ij} values for these three processes means that the observation that all the individuals in the population now have trait A provides more information about the past state of the population the less probable it was that A would evolve to fixation. Selection for A is at the bottom of the pile, with neutrality next, and selection against A at the top.

The ordering of these three processes has an intuitive interpretation. Suppose we observe that trait A is fixed in a population and we wish to estimate whether A was common (at frequency i) or rare (at frequency $j < i$) at some time in the past. Selection for A makes it easy for A to go to 100% in the population, regardless of whether it starts off common or starts off rare, which is why selection for A comes in last in the part of figure 3 that describes frequency-independent processes. The evidence in favor of the former hypothesis relative to the latter, as measured by the likelihood ratio, is stronger under selection against A than under a drift model. This is because the drift model provides more opportunity for a rare allele A to fix at 100% in the population than the model in which A is selected against. For a

formal proof concerning how the R_{ij} values for different processes compare, see appendix C.

The three results depicted in figure 3 that describe frequency-independent processes echo an insight that Darwin expresses in the *Origin*: "Adaptive characters, although of the utmost importance to the welfare of the being, are almost valueless to the systematist. For animals belonging to two most distinct lines of descent, may readily become adapted to similar conditions, and thus assume a close external resemblance; but such resemblances will not reveal—will rather tend to conceal their blood-relationship to their proper lines of descent" (Darwin 1859, 427). Darwin illustrates this idea by giving an example: whales and fish both have fins, but this is not strong evidence for their common ancestry, since the trait is an adaptation for swimming through water. Far stronger evidence for common ancestry is provided by similarities that are useless or deleterious. One of us has called this idea Darwin's Principle, in view of its centrality to Darwin's framework (Sober 2008, 2011b).

Darwin's topic in the passage quoted is inferring common ancestry, not inferring the past state of a lineage from its present state, but the epistemologies are similar. Here is a simplified argument that illustrates why. Suppose trait A has probability p of being fixed in a recent species. If two species x and y diverged from their most recent common ancestor very recently, the probability that they both have trait A is approximately p. On the other hand, if they have no common ancestor, then the probability of them both having A is p^2. This means that the likelihood ratio of the hypotheses "x and y have a recent common ancestor" and "x and y have no common ancestor" is approximately $p/p^2 = 1/p$ under Markovian trait evolution. Thus, the smaller p is, the greater this likelihood ratio will be. And the value for p if there is selection for A is larger than the value for p if there is drift, which in turn is larger than the value for p if there is selection against A.

There are two cases of frequency-dependent selection represented in figure 3. One of them (frequency-dependent selection for the majority trait) shows that it would be an overstatement to say that the current state of the population (all individuals having trait A) always provides scant evidence concerning the population's past state if trait A evolved because of natural selection. It matters a great deal what sort of selection process we are talking about. Frequency-dependent selection for the majority trait is better than drift in terms of how much information the present state of a lineage provides about its past. Figure 3 also locates the evidential meaning of frequency-dependent selection for the minority trait; it has an informational yield that is worse than that provided by drift. As noted in appendix C, our results for both cases of frequency-dependent selection make the assumption that $j < N/2$. This assumption ensures that the ordering of the frequency-

dependent selection cases is always maintained as shown in figure 3; it is an innocent assumption since, as noted above, we are mainly interested in the case where i and j are majority and minority values, respectively.[1]

Figure 3 provides only a partial ordering of the five cases depicted. The reason for this is that a comparison of, say, frequency-independent selection against A with frequency-dependent selection for the majority trait would depend on the values of specific parameters.

We now can remove the idealization of infinite time and zero mutation. The ordering of the R_{ij} ratios for the five processes, when time is infinite and mutation is zero, is the same as the ordering of those processes for the following slightly different likelihood ratio, when time is finite (and sufficiently large) and there is a sufficiently small mutational input:

$$\frac{\Pr(\text{Present} = N|\text{Past} \approx N)}{\Pr(\text{Present} = N|\text{Past} \approx 0)}.$$

See appendix D for a proof of this ordinal equivalence result. Here "Past ≈ N" and "Past ≈ 0" just mean any states close to N and to 0 (respectively), which are then held fixed across the five models. With mutational input, the exponential information loss theorem applies to all five processes. The mutual information between present and past declines monotonically as their temporal separation increases, but the decline is faster under some processes than it is under others.

Our results are derived within the setting of the Moran model of population genetics. This is a finite-state Markov chain that forms a "continuant process" (Ewens 2010). That is, at each step the number of individuals carrying a particular allele in the population of fixed size N either goes up by 1, or goes down by 1, or it stays the same. We expect similar conclusions concerning the partial ordering of the R_{ij} ratios for other continuant processes, provided the transition probabilities faithfully reflect the various types of selection being compared. Our results also apply to a slightly different problem: how does the kind of evolutionary process at work in lineages affect the strength of the evidence that similarity provides for common ancestry (Sober 2008)?

4. Mutual Information and the Likelihood Ratio. We began our discussion of information loss by using the concept of mutual information but then shifted to considering the likelihood ratio. Have we illicitly changed

1. The situation is messier if you want to consider frequency-dependent processes for the full range of all i, j values such that $i > j$. If $i > j > N/2$, then frequency-dependent selection for the majority trait is rather like frequency-independent selection for trait A, and we know that R_{ij} for frequency-independent selection for trait A is less than R_{ij} for drift.

horses in midstream? We think not. Consider the mutual information between the binary random variable E that takes the value 1 if allele A is fixed in the present population ($E = 0$ otherwise), and the frequency X of allele A in the past population, under some strictly positive prior distribution. In this case, mutual information and the likelihood ratio are linked as follows:

$$I(X; E) = 0 \text{ precisely when } R_{ij} = 1 \text{ for all past states } i, j.$$

If we now observe $E = 1$, then instead of comparing $I(X; E)$ across the different processes, it is more relevant to compare the values of R_{ij}. That is, we are thinking of a single observation (Present $= N$), so we are not considering all the possible observations that we might make of the present state. This is why we used the likelihood ratio R_{ij} rather than mutual information to carry out the cross-process comparison.

5. The Impact of Branching on Information. We have emphasized that loss of information within a lineage is a fact of life. However, the branching that takes place in evolution is a force that pushes in the opposite direction, since it creates new lineages. As illustrated in figure 1, it is possible for an ancient ancestor to have more present-day descendants than a more recent ancestor has, and this means that the information lost to the passage of time can be offset by the proliferation of descendants that each bear witness to the ancient ancestor's state. If the process is a symmetric one on two states, it is possible to describe precisely how often branching must occur if information loss is to be offset in this way (Evans et al. 2000); for more general processes, one must usually be content with upper and lower bounds.

It might be asked why one needs to worry about observing descendants to infer the characteristics of ancestors. Doesn't observing fossils provide a simpler and more definitive solution? Our answer has three parts. First, one cannot assume that a fossil comes from an ancestor of extant species; it may just be an ancient relative. Second, fossils provide evidence about the morphological hard parts of ancient organisms; molecular characters, not to mention phenotypic features of physiology and behavior, typically do not fossilize. And finally, fossil traces degrade and are subject to the exponential information loss theorem if fossils change state in conformity with the regularity assumption and have the Markov property.

Just as the process at work in lineages has an impact on information loss, so too does the topology of the branching process itself. It might seem intuitive to conjecture that the star phylogeny shown in figure 4i is "better" than the bifurcating topology shown in figure 4ii in the sense that the former topology allows the observations to provide more information about the root than the latter topology does. In order to hold other factors fixed, we

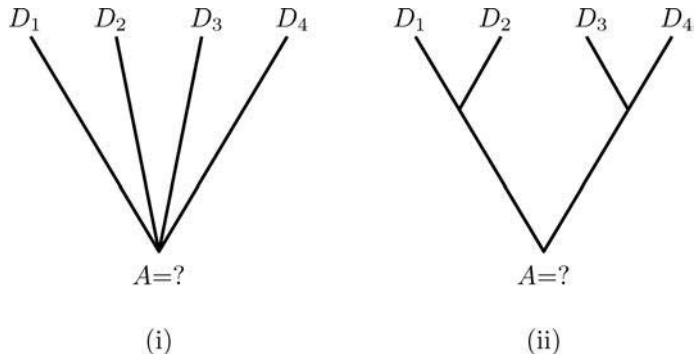

Figure 4. Does observing the four leaves of the star phylogeny (i) provide more information about the state of the root than observing the four leaves of the bifurcating phylogeny (ii)?

assume that the two topologies have the same number of leaves and that the process at work in branches is the same in the two topologies (in particular, that the expected number of substitutions—the branch lengths—between the root and each leaf match up for the two trees). The conjecture just stated seems reasonable, since in figure 4i the observations are independent of each other, conditional on the state of the root, whereas in 4ii, the observations are not conditionally independent. The guess is correct for a two-state symmetrical process when we compare the two trees in figure 4 (Sober 1989, 280). More generally, this guess holds whenever we compare any binary tree with a matching star phylogeny on the same number of leaves when a two-state symmetric process is at work (Evans et al. 2000). But, surprisingly, the guess is not always true for symmetric Markov processes with more than two states.

More precisely, consider a completely balanced binary tree with 2^n leaves, on which a constant symmetric process on five (or more) states operates with a constant substitution probability on each edge of the tree. Then if this substitution probability lies in a certain region, and n is large enough, the mutual information between the leaf states and the root state can be higher for the binary tree than for a comparable star tree on the same number of leaves (Sly 2011, theorem 1.2). Here "comparable" means that the expected number of substitutions from the root to any tip is the same in both trees. In other words, the root state of a tree can sometimes be more accurately predicted from the state of its leaves when the tree is binary (and the leaves are correlated) than when the tree is a star (and the leaves are independent), where the two trees have the same marginal distribution. A similar result holds for strongly asymmetric two-state models (Mossel 2001, theorem 1). So the intuitive maxim that testimony from independent witnesses of an

event provides more information about the event than testimony from otherwise similar dependent witnesses is sometimes false.

6. Conclusion. In summary, the view of evolution as an "information-destroying process" is basically right, but it overlooks some interesting details. For some special processes (e.g., zero mutation in simple population genetic models), information never completely disappears, even after infinite time. For example, under a drift model with zero mutation, some information concerning whether allele A was initially in the minority or the majority is always detectable at any time in the present frequency of A (which eventually fixes at 0 or N) in the population. At the other extreme there are certain (discrete time) processes for which the information can collapse completely to zero in finite time (Mossel 1998; Sober and Steel 2011, 233).

The more usual situation lies between these two extremes; for processes that are Markovian and regular, the information between past and present decays at an exponential rate and vanishes only in the limit. Here we can still compare the relative support such models provide in estimating an ancestral state from an observation today. For the five models considered, this support varies in a predictable way depending on the type of model assumed.

We are well aware that the Moran framework contains various simplifications (e.g., constant population size), but the model's simplicity allows for explicit calculations and results and does not raise questions as to whether the ordering might depend on the complexities that might be introduced in a more intricate model. We hope that our results will provide a foundation for exploring how adding complexities affects information loss. In philosophy as well as in science, it is reasonable to walk before one runs.

The estimation of an ancestral state from the leaves of a phylogenetic tree exhibits further subtleties, along with a surprise: the independent estimates obtained from a star phylogeny may or may not be more informative than the correlated estimates obtained from a binary tree, depending on the number of states, the size of the tree, and the substitution rate. The epistemological principle that "the testimony of independent witnesses always provides more evidence than the testimony of otherwise similar dependent witnesses" is wrong.

Appendix A

Formal Statement and Proof of the Exponential Loss Theorem

Proposition 1. Suppose X_t; $t \geq 0$ is any discrete, finite-state Markov process that satisfies the following condition. For some $\varepsilon > 0$, and integer $N > 0$ the following inequality holds for all $t \geq 0$ and states i, j:

$$\Pr(X_{t+N} = j | X_t = i) \geq \varepsilon. \tag{A1}$$

Then $I(X_0; X_t) \leq C \exp(-ct)$ for constants $C, c > 0$.

Proof. Let Y_n be the 'N-step' Markov chain defined by $Y_n = X_{nN}$ for all $n \geq 0$. Note that equation (A1) implies that Y_n satisfies the following inequality for all n, i, j:

$$\Pr(Y_{n+1} = j | Y_n = i) \geq \varepsilon. \tag{A2}$$

We will show that

$$I(Y_0; Y_n) \leq B \exp(-bn) \tag{A3}$$

for constants $B, b > 0$, from which proposition 1 follows because if $t = nN + r$ for $0 \leq r < N$ then

$$I(X_0; X_t) = I(Y_0; X_t) \leq I(Y_0; Y_n) \leq B \exp(-bn) < Be^b \exp(-(b/N)t),$$

where the first inequality is from the Data Processing Inequality, and the second inequality is from (A3). Thus we can take $C = Be^b$, and $c = b/N$ to obtain (A1) from (A2).

To establish inequality (A3) we apply a standard type of coupling argument. For $k \geq 0$ let $\pi_k(i) := \Pr(Y_k = i)$ for each state i. Consider the following Markov process Y'_k defined as follows: $Y'_0 = Y_0 (= X_0)$, and for each $k \geq 0$, the state of Y'_{k+1} is determined as follows: At each step of this chain, we toss a biased coin, which returns a head (H) with probability ε (independently of the chain) or a tail ($\overline{H}$) with probability $1 - \varepsilon$. If a head H is returned, Y'_{k+1} is assigned a random state according to the distribution π_{k+1}. If the coin toss results in a tail $\overline{H}$ outcome, then Y'_{k+1} selects a state that depends on Y'_k as follows: if $Y'_k = i$, then Y'_{k+1} is assigned state j with probability

$$\frac{\Pr(Y_{k+1} = j | Y_k = i) - \varepsilon \pi_{k+1}(j)}{(1 - \varepsilon)}. \tag{A4}$$

Note that this expression is greater than or equal to zero (by 2), and is less than or equal to 1; moreover,

$$\sum_j \frac{\Pr(Y_{k+1} = j | Y_k = i) - \varepsilon \pi_{k+1}(j)}{(1 - \varepsilon)} = 1,$$

so (A4) describes a legitimate probability distribution conditional on $\overline{H}$.

By the law of total probability we can express $\Pr(Y'_{k+1} = j | Y'_k = i)$ as follows:

$$\begin{aligned}\Pr(Y'_{k+1} = j | H \,\&\, Y'_k = i)\Pr(H | Y'_k = i) + \Pr(Y'_{k+1} = j | \overline{H} \,\&\, Y'_k = i)\Pr(\overline{H} | Y'_k = i) \\ = \pi_{k+1}(j)\varepsilon + [\Pr(Y_{k+1} = j | Y_k = i) - \varepsilon\pi_{k+1}(j)] = \Pr(Y_{k+1} = j | Y_k = i).\end{aligned}$$

Summarizing, we have: $Y'_0 = Y_0$, and $\Pr(Y'_{k+1} = j | Y'_k = i) = \Pr(Y_{k+1} = j | Y_k = i)$, and so Y_k and Y'_k describe the same Markov chain. In particular, $I(Y_0; Y_n) = I(Y'_0; Y'_n)$, and $\pi_n(j) = \Pr(Y'_n = j)$. Let

$$p_n(i, j) := \Pr(Y_0 = i \,\&\, Y_n = j) = \Pr(Y'_0 = i \,\&\, Y'_n = j),$$

and let $\mathcal{A}_n$ be the event that H occurs at least once in the first n biased coin tosses that are performed in the construction of the chain $Y'_1, Y'_2, \ldots$. Then $p_n(i, j)$ can be written as the sum of two terms:

$$\begin{aligned}&\Pr(Y'_n = j | \mathcal{A}_n \,\&\, Y'_0 = i)\Pr(\mathcal{A}_n \,\&\, Y'_0 = i) \\ &\quad + \Pr(Y'_n = j | \overline{\mathcal{A}_n} \,\&\, Y'_0 = i)\Pr(\overline{\mathcal{A}_n} \,\&\, Y'_0 = i). \qquad \text{(A5)}\end{aligned}$$

Now, the first term in (A5) is exactly $\pi_n(j)(1 - (1 - \varepsilon)^n)\pi_0(i)$ since

- conditional on $\mathcal{A}_n$, Y'_n is independent of Y'_0 and, in addition, Y'_n is distributed as π_n; and
- $\mathcal{A}_n$ and Y'_0 are independent, and so

$$\Pr(\mathcal{A}_n \,\&\, Y'_0 = i) = \Pr(\mathcal{A}_n)\Pr(Y'_0 = i) = [1 - (1 - \varepsilon)^n]\pi_0(i).$$

The second term in (A5) can be written as $\pi_0(i)$ multiplied by a term that lies between 0 and $(1 - \varepsilon)^n$, since

$$0 \le \Pr(\overline{\mathcal{A}_n} \,\&\, Y'_0 = i) \le \Pr(\overline{\mathcal{A}_n}) = (1 - \varepsilon)^n.$$

Thus, selecting $b > 0$ so that $e^{-b} = (1 - \varepsilon)$ we can write:

$$p_n(i, j) = \pi_0(i)[\pi_n(j) + O(e^{-bn})], \qquad \text{(A6)}$$

where, as usual, '$f(n) = O(g(n))$' is shorthand for the statement that $f(n)$ is at most some constant times $g(n)$. Finally, observe that, from (A2), $\pi_n(j) \ge \varepsilon > 0$ and so, from (A6):

$$I(Y'_0; Y'_n) = \sum_{i,j} p_n(i, j)\log\left(\frac{p_n(i, j)}{\pi_0(i)\pi_n(j)}\right) = \sum_{i,j} p_n(i, j)\log(1 + O(e^{-bn})) \le Be^{-bn},$$

for a constant $B > 0$. Since $I(Y_0; Y_n) = I(Y'_0; Y'_n)$, this establishes (A3), as required. QED

Appendix B

Proposition 2. Suppose that $X \rightarrow Y \rightarrow Z$ forms a Markov chain, where the state spaces for X, Y, and Z are discrete, and $\Pr(X = x \,\&\, Y = y) > 0$ and $\Pr(Y = y \,\&\, Z = z) > 0$ for all choices of states x, y, z for X, Y, Z, respectively. Then the DPI is an equality if and only if X, Y, and Z are mutually independent.

Proof. First, observe that if X, Y, Z are independent then they are pairwise independent and so $I(X; Y) = I(X; Z) = I(Y; Z) = 0$ and thus equality holds trivially. Next, suppose that $\Pr(X = x \,\&\, Y = y) > 0$ and $\Pr(Y = y \,\&\, Z = z) > 0$ for all choices of states x, y, z for X, Y, Z, respectively. Then $\Pr(X = x \,\&\, Y = y \,\&\, Z = z) > 0$ holds also (since $X \rightarrow Y \rightarrow Z$ is a Markov chain). Suppose further that the DPI is an equality; we will show that X, Y, and Z are independent. Since the DPI is an equality, $X \rightarrow Y \rightarrow Z$ and $X \rightarrow Z \rightarrow Y$ are both Markov chains (Cover and Thomas 2006). We write $p(xyz)$ as shorthand for the probability $\Pr(X = x \,\&\, Y = y \,\&\, Z = z)$ and similarly for conditional and marginal probabilities (thus, e.g., $p(x|z) = \Pr(X = x|Z = z)$). First observe that the positivity condition $p(xyz) > 0$ for all (x, y, z) implies that $p(xy), p(xz), p(yz), p(x), p(y), p(z)$ are also strictly positive. Since $X \rightarrow Y \rightarrow Z$ is a Markov chain, and $p(xy) > 0$:

$$p(xyz) = p(z|xy)p(xy) = p(z|y)p(xy), \tag{B1}$$

and since $X \rightarrow Z \rightarrow Y$ is also a Markov chain, and $p(xz) > 0$, we have $p(xyz) = p(y|xz)\, p(xz) = p(y|z)\, p(xz)$. Applying Bayes's theorem, the last term can be written as

$$\frac{p(z|y)p(y)}{p(z)} p(xz)$$

(note that $p(z) > 0$) and so, combining this with equation (B1) gives

$$p(xyz) = p(z|y)p(xy) = \frac{p(z|y)p(y)}{p(z)} p(xz),$$

and so

$$p(z|y)p(xy)p(z) = p(z|y)p(y)p(xz). \tag{B2}$$

Since $p(z|y) > 0$ (because $p(yz) > 0$) we can cancel this term on the left and right of equation (B2) to obtain:

$$p(xy)p(z) = p(y)p(xz). \tag{B3}$$

Now, we can further write $p(xy) = p(x|y)\ p(y)$ and $p(xz) = p(x|z)\ p(z)$ which, upon substitution into equation (B3) gives:

$$p(x|y)p(y)p(z) = p(y)p(x|z)p(z),$$

in other words, $p(x|y) = p(x|z)$ (noting that $p(y), p(z) > 0$). Now, this equation must hold for all choices of x, y, and z so $p(x|y)$ must be constant as y varies—which implies that X and Y are independent. Similarly X and Z are independent. Finally, reversing the two Markov chains gives that Z is independent of Y. Thus X, Y and Z are pairwise independent.

Moreover they are independent as a triple since $X \to Y \to Z$ is a Markov chain and so

$$p(xyz) = p(z|xy)p(xy) = p(z|y)p(x)p(y) = p(yz)p(x) = p(x)p(y)p(z).$$

QED

Appendix C

R_{ij} Ratios with Zero Mutation at the Infinite Time Limit

Consider the Moran model in population genetics, with two trait values A and B and population size N. Let $X_t \in \{0, 1, \ldots, N\}$ be the number of copies of A in the population at time t. In this section we assume zero mutation, and we consider neutral evolution, selection for A, selection against A, and frequency-dependent selection (for the majority state and against the majority state). Since each of these Markov processes has absorbing states 0 and N (because of zero mutation) eventually one allele will be fixed and the other lost. Let $E \in \{0, N\}$ be this end state, and $S \in \{0, 1, \ldots, N\}$ the starting state (thus $S = X_0$ and $E = \lim_{t\to\infty} X_t$).

We are interested in comparing the ratio of conditional probabilities:

$$R_{ij} := \frac{\Pr(E = N|S = i)}{\Pr(E = N|S = j)},$$

for $i > j$ under the various models.

Proposition 3.

(i) Under neutral evolution $R_{ij} = (i/j)$, for all $0 < i, j \leq N$.
(ii) Under frequency-independent selection $R_{ij} = [(1 - c^i)/(1 - c^j)]$, for all $0 < i, j \leq N$, where c is a positive constant with $c < 1$ when there is selection for A and $c > 1$ when there is selection against A.

(iii) For any two values $i, j \in \{1, 2, \ldots, N\}$ with $i > j$ the R_{ij} value for selection against A exceeds the R_{ij} value for neutral evolution, which in turn exceeds the R_{ij} value for selection for A.

(iv) For frequency-dependent selection, where the fitness of trait A is proportional to its frequency, the associated R_{ij} value exceeds that for neutral evolution for all $i, j \in \{1, 2, \ldots, N\}$ with $i > j$ provided that $j < N/2$.

(v) For frequency-dependent selection, where the fitness of trait A is proportional to the frequency of the alternative trait B, the associated R_{ij} value is lower than that for neutral evolution for all $i, j \in \{1, 2, \ldots, N\}$ with $i > j$, provided that $j < N/2$.

(vi) In all the cases considered above we have $R_{ij} > 1$ for all $i > j$.

Proof: *Part i*. For any integer $x : 0 \leq x \leq N$ it is a well-known result (for many neutral models) that

$$\Pr(E = N|S = x) = x/N$$

(see, e.g., eq. 3.49 of Ewens 2010), from which i immediately follows.

Part ii. From Ewens (2010, eq. 3.66) we have, for any $x \in \{1, \ldots, N\}$:

$$\Pr(E = N|S = x) = \frac{1 - c^x}{1 - c^N},$$

for a positive constant c which is greater than 1 for selection against A and less than 1 for selection for A. Part ii now follows immediately.

Part iii. By parts i and ii, part iii is equivalent to the assertions that for $i > j$, if $c \in (0, 1)$ then:

$$\frac{1 - c^i}{1 - c^j} < \frac{i}{j},$$

while if $i > j$ and $c > 1$ then:

$$\frac{1 - c^i}{1 - c^j} > \frac{i}{j}.$$

Now, using the identity $1 - c^k = (1 - c)\,(1 + c + c^2 + \ldots + c^{k-1})$ we have

$$\frac{1 - c^i}{1 - c^j} = \frac{1 + \ldots + c^{i-1}}{1 + \ldots + c^{j-1}}, \tag{C1}$$

and so we wish to compare

$$\frac{1 + \ldots + c^{i-1}}{1 + \ldots + c^{j-1}} \quad \text{and} \quad \frac{i}{j},$$

which is equivalent to comparing

$$\frac{1 + \ldots + c^{i-1}}{i} \quad \text{and} \quad \frac{1 + \ldots + c^{j-1}}{j}.$$

Now, the left-hand side is merely the average of the terms c^k from $k = 0$ up to $k = i - 1$ while the right-hand side is the average of these terms up to $j - 1$ and since $i > j$ the left-hand side is smaller than the right when $c < 1$ and greater when $c > 1$. This completes the proof.

Part iv. When selection is frequency dependent, we need to use the expression:

$$\Pr(E = N|S = i) = \left(1 + \sum_{j=1}^{i-1} \prod_{k=1}^{j} \frac{g_k}{f_k}\right) \Big/ \left(1 + \sum_{j=1}^{N-1} \prod_{k=1}^{j} \frac{g_k}{f_k}\right), \tag{C2}$$

where f_k and g_k denote the fitnesses of alleles A and B, respectively, when k individuals have allele type A (see, e.g., Huang and Traulsen 2010, eq. 6). If we now take the fitness of trait A to be proportional to its frequency, that is $f_k = a \cdot k$ for a constant $a > 0$, and take the fitness of trait B to also be proportional to its frequency (with the same coefficient)—that is, $g_k = a \cdot (N - k)$—then equation (C2) gives

$$R_{ij} = \sum_{m=0}^{i-1} \binom{N-1}{m} \Big/ \sum_{m=0}^{j-1} \binom{N-1}{m}. \tag{C3}$$

As before, this ratio exceeds i/j precisely when the average of the first i terms

$$\binom{N-1}{m}$$

(for $m = 0, 1, 2 \ldots$) exceeds the average of the first j terms

$$\binom{N-1}{m};$$

this holds for all $i > j$ with $j < N/2$.

Part v. If we take the fitness of trait A to be proportional to the frequency of B, that is $f_k = a \cdot (N - k)$ for a constant $a > 0$, and take the fitness of trait

B to be proportional to the frequency of A (with the same coefficient)—that is, $g_k = a \cdot k$—then equation (C2) gives

$$R_{ij} = \sum_{m=0}^{i-1} \binom{N-1}{m}^{-1} \Big/ \sum_{m=0}^{j-1} \binom{N-1}{m}^{-1}, \tag{C4}$$

and this ratio is lower than i/j precisely when the average of the first i terms

$$\binom{N-1}{m}^{-1}$$

(for $m = 0, 1, 2 \ldots$) is lower than the average of the first j terms

$$\binom{N-1}{m}^{-1};$$

this holds for all $i > j$ with $j < N/2$.

Part vi. The inequality $R_{ij} > 1$ for $i > j$ is trivial for neutral evolution, while for the other processes the inequality follows from equations (C1), (C3), and (C4), noting that when $i > j$ the i terms in the numerator include the j terms in the denominator along with some additional positive terms.

Note that N plays no role in the expression of the ratio R_{ij} in cases i–iii, but it does for iv and v.

Appendix D

R_{ij} Ratios with Finite Time and Nonzero Mutation Rate

The results in the previous section assume zero mutation and consider the infinite time limit. However, they also have some bearing on what happens at finite time and for nonzero mutation. First assume zero mutation and consider the ratio R_{ij} at finite times t. Since these ratios are continuous functions of t and converge to values that satisfy the partial order described in figure 3, this ordering also holds for t sufficiently large (but finite). Select any such sufficiently large value t_0 of t, and consider this ratio R_{ij} at t_0 as a function of the mutation rates. Again, R_{ij} is a continuous function of these mutation rates (with t fixed to t_0), and so, for sufficiently small (but nonzero) mutation rates, the five R_{ij} values will still be ordered as in figure 3 at t_0. In summary, for a sufficiently large value of time, we can take small but strictly positive mutation rates that preserve the order of the R_{ij} ratios shown in figure 3. This ordering may change if the mutation rates are then held fixed and time increased. In other words, the order of quantifiers here is important. We are merely asserting that for any sufficiently large value of time, there exist positive mutation rates that preserve the orderings of the

ratios—if we select a larger time, the mutation rates may need to be reduced (but will still be strictly positive).

REFERENCES

Cover, Thomas M., and Joy A. Thomas. 2006. *Elements of Information Theory*. 2nd ed. New York: Wiley.

Darwin, Charles R. 1859. *On the Origin of Species by Means of Natural Selection*. London: Murray.

Evans, William, Claire Kenyon, Yuval Peres, and Leonard J. Schulman. 2000. "Broadcasting on Trees and the Ising Model." *Advances in Applied Probability* 10:410–33.

Ewens, Warren J. 2010. *Mathematical Population Genetics*. Vol. 1, *Theoretical Introduction*. New York: Springer.

Frieden, B. Roy, Angelo Plastino, and B. H. Soffer. 2001. "Population Genetics from an Information Perspective." *Journal of Theoretical Biology* 208 (1): 49–64.

Gascuel, Olivier, and Mike Steel. 2014. "Predicting the Ancestral Character Changes in a Tree Is Typically Easier than Predicting the Root State." *Systematic Biology* 63 (3): 421–35.

Hacking, Ian. 1965. *The Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Häggström, Olle. 2002. *Finite Markov Chains and Algorithmic Applications*. Cambridge: Cambridge University Press.

Huang, Weini, and Arne Traulsen. 2010. "Fixation Probabilities of Random Mutants under Frequency Dependent Selection." *Journal of Theoretical Biology* 263 (2): 262–68.

Keynes, John Maynard. 1924. *A Tract on Monetary Reform*. London: Macmillan.

Laplace, Pierre-Simon. 1814. *A Philosophical Essay on Probabilities*. Trans. from the French 6th ed. by F. Truscott and F. Emory. New York: Dover, 1951.

Moran, Patrick. 1962. *The Statistical Processes of Evolutionary Theory*. Oxford: Oxford University Press.

Mossel, Elchanan. 1998. "Recursive Reconstruction on Periodic Trees." *Random Structures and Algorithms* 13 (1): 81–97.

———. 2001. "Reconstruction on Trees: Beating the Second Eigenvalue." *Annals of Applied Probability* 11:285–300.

———. 2003. "On the Impossibility of Reconstructing Ancestral Data and Phylogenies." *Journal of Computational Biology* 10:669–78.

Royall, Richard. 1997. *Statistical Evidence—a Likelihood Paradigm*. Boca Raton, FL: Chapman & Hall.

Salichos, Leonidas, and Antonis Rokas. 2013. "Inferring Ancient Divergences Requires Genes with Strong Phylogenetic Signals." *Nature* 497:327–31.

Sly, Allan. 2011. "Reconstruction for the Potts Model." *Annals of Probability* 39:1365–1406.

Sober, Elliott. 1989. "Independent Evidence about a Common Cause." *Philosophy of Science* 56: 275–87.

———. 2008. *Evidence and Evolution: The Logic behind the Science*. Cambridge: Cambridge University Press.

———. 2011a. "*A Priori* Causal Models of Natural Selection." *Australasian Journal of Philosophy* 89:571–89.

———. 2011b. *Did Darwin Write the Origin Backwards?* Amherst, NY: Prometheus Books.

Sober, Elliott, and Mike Steel. 2011. "Entropy Increase and Information Loss in Markov Models of Evolution." *Biology and Philosophy* 26:223–50.