Biol Philos (2011) 26:223–250
DOI 10.1007/s10539-010-9239-x

Entropy increase and information loss in Markov models of evolution

Elliott Sober · Mike Steel

Received: 13 August 2010 / Accepted: 11 November 2010 / Published online: 6 January 2011

Abstract Markov models of evolution describe changes in the probability distribution of the trait values a population might exhibit. In consequence, they also describe how entropy and conditional entropy values evolve, and how the mutual information that characterizes the relation between an earlier and a later moment in a lineage's history depends on how much time separates them. These models therefore provide an interesting perspective on questions that usually are considered in the foundations of physics—when and why does entropy increase and at what rates do changes in entropy take place? They also throw light on an important epistemological question: are there limits on what your observations of the present can tell you about the evolutionary past?

Keywords Entropy · Evolution · Markov model · Mutual information

Introduction

Entropy increase is almost always treated as a topic in the foundations of physics, but there are reasons to think that this is a mistake. The entropy of a system at a time is well-defined whenever there is a probability distribution for the system's possible states at that time, and probability distributions are described in many sciences, not just in physics. In addition, an entropy increase for states P_1, P_2,…,P_n does not entail an entropy increase for states Q_1, Q_2,…,Q_m, where the Q-properties supervene on the P-properties (Barrett and Sober 1994, 1995). Given this, the general question of when and why entropy increases must be gently removed from the exclusive grip

E. Sober (✉)
Philosophy Department, University of Wisconsin, Madison, WI, USA
e-mail: ersober@wisc.edu

M. Steel
Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand

of physicists and examined in connection with processes described in other sciences. This is a point with which physicalists should agree.

The processes considered in this paper are described in so-called Markov models of evolution. These models are widely used throughout several areas of evolutionary biology, including phylogenetics (see, for example, Semple and Steel 2003; Felsenstein 2004), DNA sequence evolution (Durrett 2002), population genetics (Crow and Kimura 1970), and modeling of speciation and extinction (Pinelis 2003). However, in these applications the topic of entropy change has rarely been explicitly considered. The abstractness of some of these models means that they can apply to many systems that change through time, not just to lineages undergoing biological evolution. Moreover, some of these models apply to systems that are not at equilibrium as well as to ones that are.

Since probability has both objective and subjective interpretations, entropy has the same dual nature. In what follows, we will investigate both. Markov models of evolution describe how trait frequencies change value in a population. Since trait frequencies are objective quantities, the probabilities that describe these quantities are objective as well, and so are the entropies. But these models also tell you how observations of the present state of a population should affect your uncertainty about the past or future state of that population. Here entropy describes something subjective.

Although "entropy" may sound like a concept from physics, it is definitionally related to another concept, that of *mutual information*, and this concept not only *sounds* like it has a wider scope of application; it does. The relationship of entropy to mutual information entails that tracing the changes in entropy that a lineage experiences has epistemological significance; it allows you to estimate how much information the present state of the lineage provides about the lineage's past. Do present-day traces unambiguously and definitively reveal what the past was like, or is the past forever lost to us, with no *petite Madeleine* in sight? The truth is to be found between these two extremes. Information often decays, but the rate at which it decays depends heavily on the specifics of the processes involved. This may seem to give rise to a paradox: to judge how much information the present provides about the past, you need information about the processes linking present to past. But there is no paradox here. Your genotype provides information about the genotypes of your parents, but only because the process of genetic inheritance has certain features. The same is true of evolution, though evolutionary descendants and ancestors are separated by far greater reaches of time.

We begin with the simplest Markov model, in which a lineage at a point in time has just two states (coded as 0 and 1), but most of our discussion extends seamlessly to general finite-state Markov processes.[1] When a probability distribution attaches to the states that the lineage might occupy at a given time, we can compute the lineage's entropy via the formula[2] $-\sum p_i(\log p_i)$. The entropy is maximal when the

[1] In this paper 'Markov process' will refer to either a discrete Markov process (a Markov chain) or a continuous-time Markov process, on a finite state space.

[2] Throughout this paper, log is with base e, rather than 2; the conclusions remain the same regardless of the base of the log, though some formulae change slightly.

two states are equiprobable and minimal when they have probabilities of 1 and 0. An evolutionary process may lead to a change in the frequency of the two states, and so the entropy may change.

Markov modeling of the evolution of a binary character in a lineage starts with the idea of "instantaneous" probabilities of change. Let u be the probability that a lineage in state 0 changes to state 1 in a brief period of time; let v be the possibly different probability that a lineage in state 1 changes to state 0 in that brief moment. By allowing the 'brief' period of time to be sufficiently short, we may assume that u and v are both less than 0.5. The next step is to use these instantaneous transition probabilities to model what happens in a lineage that has some number of these small instants of time as its duration. If X_t denotes the state of the system at time t, a Markov model describes the values of the conditional probabilities $\Pr(X_t = j|X_0 = i)$; this is the probability that a lineage will end in state j, given that it begins in state i and has a duration of t units of time. For convenience, we will describe a discrete time rather than a continuous time model.[3] There are four lineage transition probabilities to consider:

$$\begin{aligned}
\Pr(X_t = 1|X_0 = 0) &= \frac{u}{u+v} - \frac{u}{u+v}(1-u-v)^t \\
\Pr(X_t = 0|X_0 = 0) &= \frac{v}{u+v} + \frac{u}{u+v}(1-u-v)^t \\
\Pr(X_t = 0|X_0 = 1) &= \frac{v}{u+v} - \frac{v}{u+v}(1-u-v)^t \\
\Pr(X_t = 1|X_0 = 1) &= \frac{u}{u+v} + \frac{v}{u+v}(1-u-v)^t
\end{aligned}$$

In each equation, the first addend fails to mention the amount of time t between the lineage's start and finish; the second addend does, and it quickly shrinks towards zero as t increases. This means that the first addend (namely $u/(u + v)$ and $v/(u + v)$) describes the probability that obtains in the limit as the time in the lineage tends towards infinity. These are the so-called *equilibrium probabilities*. When we consider a short period of time, the values of these transition probabilities are mainly determined by the lineage's initial state; if the lineage begins in a given state, it will almost certainly end in that same state.[4] As the duration of the lineage is increased, the process plays a progressively 1 narger role in determining the probability of the final state and the initial condition of the lineage is steadily forgotten. For example, if $u = v$, the first addend in each of these four equations equals ½, to which the second addend adds or subtracts a quantity that shrinks as the duration of the lineage is increased.

This simple model can be used to describe the difference between selection and drift. Pure drift is represented by the idea that $u = v$ and selection for character state 1 by the idea that $u > v$. These constraints on the instantaneous probabilities of

[3] The formulae for a continuous-time model are similar; just replace the term $(1-u-v)^t$ by $e^{-(u+v)t}$.

[4] For example, at $t = 0$, $\Pr(X_t = 1|X_0 = 1) = \Pr(X_t = 0|X_0 = 0) = 1$, and $\Pr(X_t = 1|X_0 = 0) = \Pr(X_t = 0|X_0 = 1) = 0$.

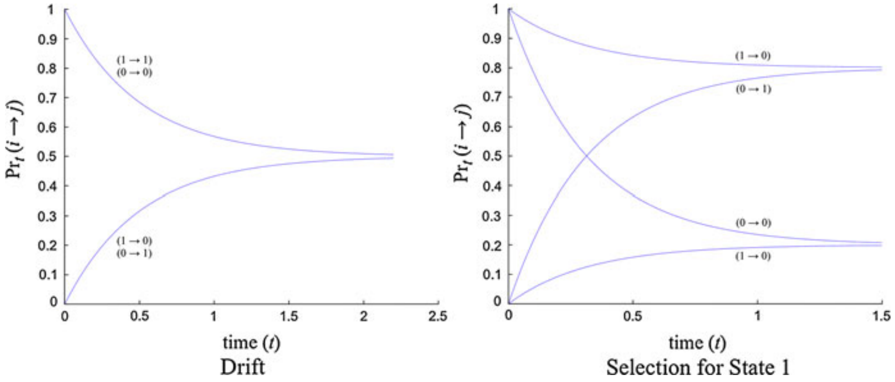


Fig. 1 The effect of drift and of selection for state 1 on transition probabilities in a continuous-time two-state model. Time is measured in units of the expected number of substitutions for the process in equilibrium

change have implications concerning how different lineage transition probabilities will be related. If the traits are subject just to drift, then $\Pr(X_t = 0|X_0 = 0) = \Pr(X_t = 1|X_0 = 1)$ and $\Pr(X_t = 1|X_0 = 0) = \Pr(X_t = 0|X_0 = 1)$; selection for state 1 means that $\Pr(X_t = 1|X_0 = 0) > \Pr(X_t = 0|X_0 = 1)$ and that $\Pr(X_t = 1|X_0 = 1) > \Pr(X_t = 0|X_0 = 0)$. The impact of selection and drift on these transition probabilities is represented in Fig. 1.

To address the question of how entropy changes in this Markov process, we begin by considering the initial state of the lineage and then trace the lineage forward in time. This initial state does not need to be the moment at which the lineage comes into existence; it is simply the time at which the investigator wants to start tracing the lineage forward. At the start of the process, the two possible states have probabilities Pr(the lineage starts in state 0) = p_0 and Pr(the lineage starts in state 1) = p_1. Then the lineage starts evolving. In the limit, as time tends to infinity, the probabilities of the lineage's state converge on the equilibrium probabilities $\pi_0 = v/(u + v)$ and $\pi_1 = u/(u + v)$. As the probabilities evolve—from their initial values to their equilibrium values—there are three different questions that can be asked about entropy increase:

(1) *Definite starting state*: If the lineage definitely starts in state 1 (or in state 0), under what circumstances will entropy increase?
(2) *Equilibrium starting state*: If the lineage's starting state is characterized by the equilibrium probabilities π_0 and π_1, under what circumstance will entropy increase?
(3) *Probabilistic starting state*: If the lineage's starting state is characterized by the probabilities p_0 and p_1, under what circumstance will entropy increase?

It should be clear that (3) is the most general question of the three; (1) and (2) are special cases. By "increasing," we mean that the entropy sometimes increases and never declines; by "strictly increasing," we mean that the entropy at time t' is always strictly greater than the entropy at time t whenever $t < t'$.[5] When entropy isn't always increasing, the exceptions of interest are cases in which entropy always declines, goes up and then down, or never changes.

Conditional entropy versus ordinary entropy

Our main task is to track the behavior of the *conditional entropy* when t units of time have elapsed after the lineage's initial state. We explain this first with reference to a two-state process, using the usual information-theoretic terminology (Cover and Thomas 1991). Conditional on the system starting in state 0, the entropy after t units of time have passed is:

$$\begin{aligned} H(X_t|X_0 = 0) = &-\Pr(X_t = 0|X_0 = 0)\log(\Pr(X_t = 0|X_0 = 0)) \\ &-\Pr(X_t = 1|X_0 = 0)\log(\Pr(X_t = 1|X_0 = 0)) \end{aligned} \tag{1}$$

Similarly, conditional on the lineage starting in state 1, the entropy after t units of time will be:

$$\begin{aligned} H(X_t|X_0 = 1) = &-\Pr(X_t = 0|X_0 = 1)\log(\Pr(X_t = 0|X_0 = 1)) \\ &-\Pr(X_t = 1|X_0 = 1)\log(\Pr(X_t = 1|X_0 = 1)) \end{aligned} \tag{2}$$

If p_0 is the lineage's probability of starting in state 0 and p_1 is its probability of starting in state 1, then the lineage's conditional entropy, denoted by $H(X_t|X_0)$, is a weighted average of the two expressions just described (i.e., p_0 times term (1) + p_1 times term (2)). In other words, the conditional entropy is a weighted average of the values $H(X_t|X_0 = 0)$ and $H(X_t|X_0 = 1)$ according to the prior probabilities p_i. Note that the p_i values are constants whereas the $\Pr(X_t = j|X_0 = i)$ terms change value with time. The conditional entropy should be contrasted with ordinary *(i.e. unconditional) entropy*, $H(X_t)$, which is:

$$H(X_t) = -\Pr(X_t = 0)\log(\Pr(X_t = 0)) - \Pr(X_t = 1)\log(\Pr(X_t = 1)).$$

Conditional entropy also should not be confused with the quantities $H(X_t|X_0 = 0)$ and $H(X_t|X_0 = 1)$, even though these expressions involve conditional probability distributions.

Markov models of evolution apply to lineages evolving through time. The probabilities deployed in these models pertain to population properties—for example, to the frequencies of traits in a population. Entropy in these models is therefore a property of the probability distribution of a population's possible states; it is not a property of the organisms in those populations. If we consider the various possible trait values that the organisms in a population might have for some trait, a monomorphic population will have a lower entropy than one that is highly polymorphic. The entropy of the population says nothing about the complexity of its

[5] A sufficient condition for a function to be strictly increasing is that it has a strictly positive slope everywhere, though this condition is not necessary. This can be seen by considering the function $f(x) = 1 + (x - 1)^3$, which is strictly increasing on the interval [0,2] but has zero slope at $x = 1$.

member organisms. This simple point is worth bearing in mind. There is a large literature on the conditions that will lead organisms and their genomes to evolve towards greater complexity (see, for example, Brooks and Wiley 1988; Weber and Depew 1988; Yockey 2005). This is not our subject.

Analysis of the two-state process

Returning to our three-case breakdown, we can easily assess Case (1), in which the lineage definitely starts in one state (equivalently: the probabilities of the two starting states are 0 and 1), by consulting Fig. 1. The starting entropy is minimal, since the starting probabilities p_0 and p_1 have values of 0 and 1. If the lineage starts in state 0, we need to track the values of $\Pr(X_t = 0|X_0 = 0)$ and $\Pr(X_t = 1|X_0 = 0)$ as the lineage duration t is increased; if the lineage starts in state 1, we need to track the values of $\Pr(X_t = 0|X_0 = 1)$ and $\Pr(X_t = 1|X_0 = 1)$. When a drift process is in place, the first two probabilities draw closer together and, in the limit, converge on 0.5; the same pattern holds for the second two. Our first observation, therefore, is:

Proposition 1a *In a two-state Markov process in which there is pure drift and the lineage starts evolving in a definite state, conditional entropy strictly increases.*

The situation is different when selection is present. The right-hand side of Fig. 1 represents the lineage transition probabilities when there is selection for state 1. Notice that $\Pr(X_t = 1|X_0 = 1)$ and $\Pr(X_t = 0|X_0 = 1)$ draw closer together, but remain well-separated, as the lineage duration increases. The entropy steadily increases. The pattern is different when the lineage begins in state 0. In this case, $\Pr(X_t = 0|X_0 = 0)$ and $\Pr(X_t = 1|X_0 = 0)$ cross at 0.5. This means that:

Proposition 1b *In a two-state Markov process in which there is selection for state 1, the quantity $H(X_t|X_0 = i)$ is everywhere increasing if the lineage begins in state $i = 1$, but it increases and then decreases if the lineage begins in state $i = 0$.*

These two trajectories are depicted in Fig. 2.

We now turn to Case (2). If the lineage's initial state is characterized by the equilibrium probabilities π_0 and π_1, the probability of being in state 0 never changes with time; the same is true of the probability of being in state 1. That is, the *expected state* of the system never changes. It follows that:

Proposition 2a *In a two-state Markov process in which the lineage's initial state is characterized by the equilibrium probabilities, the unconditional entropy never changes.*

But what happens to the conditional entropy? A special case of a later result (Proposition 5(ii)) has the following implication:

Proposition 2b *In a two-state (continuous) Markov process in which the lineage's initial state is characterized by the equilibrium probabilities, the slope of the conditional entropy is always positive.*

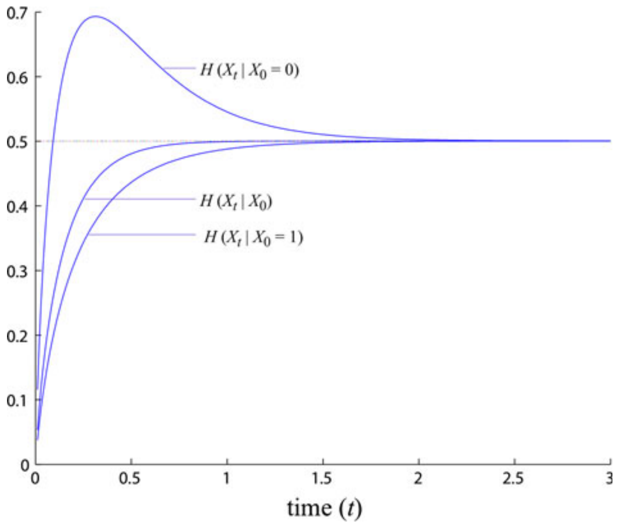


Fig. 2 Change in entropy functions with t for a continuous-time two-state model with selection for state 1 ($\pi_0 = 0.2$ and $\pi_1 = 0.8$). $H(X_t|X_0 = i)$ is everywhere increasing for $i = 1$, but for $i = 0$ it increases to a local maximum and then declines. If these two entropies are averaged according to their equilibrium distribution, the resulting conditional entropy, $H(X_t|X_0)$, is everywhere increasing. Time is measured in units of the expected number of substitutions for the process in equilibrium

An example in which there is selection for state 1 and the conditional entropy initially increases and then levels off is depicted in Fig. 2; this is a continuous-time Markov process in equilibrium where the equilibrium probabilities are $\pi_0 = 0.2$ and $\pi_1 = 0.8$.

Case (3) is the most general of the three, in that (3) subsumes both (1) and (2). One obvious fact about this general case follows from Proposition (1a). If the lineage's entropy increases if it starts in state 0 and also increases if it starts in state 1, then any weighted average of these two will increase as well:

Proposition 3a *In a two-state Markov process in which there is pure drift, the conditional entropy is strictly increasing.*

No such obvious generalization attaches to Proposition 1b where the pattern of entropy change depends on the starting state. However, Proposition 2b can be strengthened. If the lineage's starting probabilities are 'close' to the equilibrium values π_0 and π_1, the conditional entropy will be strictly increasing. The following result places Proposition 2b in a more general setting.

Proposition 3b *In a two-state Markov process, the conditional entropy always increases if the (ordinary) entropy does.*

Notice that this result obtains regardless of whether the process at work is drift or selection. Proposition 3b is a special case of Proposition 5(i), to be stated below.

General Markov processes

We turn now to the analysis of a general Markov process, X_t (either discrete or continuous),[6] on a finite state space with N states. For such a process, a probability distribution on states is an *equilibrium distribution* $\pi = (\pi_1,\ldots,\pi_N)$ if for each initial starting state i, the probability that the process is in any state j at time t converges to π_j as t grows.

The two-state Markov process has an equilibrium distribution, except in degenerate cases.[7] Markov processes on more than two states can fail to have equilibrium distributions for more generic reasons, and we describe some shortly. A sufficient (but not necessary) condition for a Markov process to possess an equilibrium distribution is that all the transition probabilities (or rates for a continuous process) are strictly positive. This is a very strong requirement, since often it is not possible to move directly from one state to another in a single jump, so we can look for a weaker sufficient condition which is called 'irreducibility': for any two states i and j, there is a sequence of states starting with i and ending with j and for which each transition has strictly positive probability (or rate). This condition must hold for each ordered pair of states i, j (including the case $i = j$ if the chain is discrete).

For a continuous-time Markov process, this irreducibility condition is sufficient for the existence of an equilibrium distribution for the process. For a discrete chain, it is 'almost' sufficient, but not quite (because, roughly speaking, it is possible for a discrete process to traverse states in a deterministic cyclic fashion). Rather, a sufficient condition for a discrete Markov chain to have an equilibrium distribution is that the chain is both irreducible and 'aperiodic,'[8] in which case the (discrete) Markov process is sometimes said to be *regular*. Regularity turns out to be mathematically equivalent to the following condition: for some fixed number t of time steps, it is possible to move from any state to any other in *exactly* t steps. Moreover, t can be taken to be $N^2 - 2N + 2$ (Seneta 1973). We will use the phrase 'regular Markov process' to refer to either a regular Markov chain or an irreducible continuous-time Markov process.

The 'Markov chain convergence theorem' (see, e.g., Häggström 2002) says that any finite-state Markov process that is regular has an equilibrium distribution. Moreover, this distribution is strictly positive on all states, and it is the unique stationary distribution[9] for the process.[10]

[6] We also assume, as usual, that the process is time-homogeneous—that is, the transition probabilities (or rates) do not change over time.

[7] The exceptional cases arise if both u, v are zero (and in the discrete chain case if u, v are both 1).

[8] See, for example, Häggström (2002) for a precise definition of aperiodicity. A sufficient condition for it to hold is that $\Pr(X_t = i|X_0 = i) > 0$ for each state i and $t = 1$. Moreover, if the chain is irreducible a sufficient condition is simply that $\Pr(X_t = i|X_0 = i) > 0$ for at least *one* state i and $t = 1$.

[9] A *stationary distribution* of a Markov process is any distribution on the states that satisfies the condition that if X_0 is chosen according to that distribution, then X_t also has this distribution for all $t > 0$. An equilibrium distribution is stationary, but not conversely; indeed, a Markov chain may have infinitely many stationary distributions but no equilibrium distribution.

[10] The Markov chain convergence theorem is a consequence of the well-known Perron-Frobenius theorem (as applied to irreducible matrices), see e.g. Grimmett and Stirzaker (2001, p. 295).

Entropy for a general Markov process

For any Markov process the conditional entropy is given by[11]:

$$H(X_t|X_0) = \sum_i p_i H(X_t|X_0 = i) \tag{3}$$

where p_i is the probability that the initial state (X_0) is i, and

$$H(X_t|X_0 = i) = -\sum_j \Pr(X_t = j|X_0 = i)\log(\Pr(X_t = j|X_0 = i))$$

is the entropy conditional on an actual initial state i (Case 1 above, the 'definite starting state'). Formula (3), which also can be written as

$$H(X_t|X_0) = -\sum_{i,j} p_i \Pr(X_t = j|X_0 = i)\log(\Pr(X_t = j|X_0 = i))$$

generalises our earlier description of conditional entropy for the two-state process.

This conditional entropy should be contrasted with the ordinary (i.e. unconditional) *entropy*, $H(X_t)$, which is defined as:

$$H(X_t) = -\sum_j \Pr(X_t = j)\log(\Pr(X_t = j)) \tag{4}$$

Notice that $\Pr(X_t = j)$ is related to the quantities used to define conditional entropy by the equation:

$$\Pr(X_t = j) = \sum_i p_i \Pr(X_t = j|X_0 = i).$$

Thus, $p_i = \Pr(X_0 = i)$, and in the special case where $p_i = \pi_i$ (the equilibrium frequency) for each state i, we have $\Pr(X_t = j) = \pi_j$ for all states j and all $t \geq 0$.

Example

To further explain the two concepts of entropy, we provide an example from population genetics. Consider a haploid population consisting of N individuals, each of which has one of two character states (A or B). Consider how the frequencies of these two character states vary through time under the usual population-genetic models that allow pure drift (no selection) and some low symmetric rate of mutation between the states (so neither state reaches fixation in the population). If we assume that the population size remains constant and there are discrete time steps (e.g., a Wright-Fisher model[12] with symmetric mutation), then the number X_t of individuals carrying character state A at time step t forms a finite-state Markov chain on the state space $\{0,1,2,\ldots,N\}$. Moreover, it is easily seen to be a regular Markov chain

[11] Summations are over all the states (or pairs of states). We use i throughout to denote the initial state and j to denote the state at time t.

[12] The Wright-Fisher model is a classical Markov chain in population genetics in which each generation is replaced by the next (see e.g. Durrett 2002).

and so has an equilibrium distribution (whose shape depends on the magnitude of the mutation rate). Assume that this process is in equilibrium; then the expected value of X_t will be $N/2$, and classical population genetics furnishes a probability distribution for each of the $N + 1$ values that is symmetric about the mean. The entropy does not change with time, as we assume the process is in equilibrium. By contrast, suppose that at a particular moment (call it time = 0) one were to count how many individuals actually carry character state A. This observation will provide information about the frequency of trait A at any future time $t > 0$, even if the observed value at time 0 happens to be the expected frequency (namely, $N/2$). Note that the entropy of the probability distribution of different states at time t given a particular present observation is *not* the conditional entropy. Rather, the former quantity corresponds to the analogue of Case (1) ('Definite starting state'), though here the state space is of size $N + 1$ rather than 2. It is the term $H(X_t|X_0 = i)$ above, if i is the present observation. By contrast, conditional entropy $H(X_t|X_0)$ is the *expected value of* $H(X_t|X_0 = I)$, if we were to observe the unknown (random) number of individuals I carrying character A at time 0. It thus is the analogue of Case (3) ('Probabilistic starting state') above.

We can modify this simple example (while retaining the assumption of small mutational input) to include frequency independent selection, as well as frequency-dependent selection in which one or both character states become increasingly favored the higher its frequency is in the population. For example, suppose both states are subject to equal positive frequency-dependent selection. Formally this could be modeled by a Moran-type model[13] in which all self and adjacent transition probabilities $p_{ii}, p_{ii-1}, p_{ii+1}$ are positive and the difference $p_{ii+1} - p_{ii-1}$ is positive and increasing as i gets closer to N, and negative and decreasing as i gets closer to 0. Intuitively, it would seem that if the initial state is larger than $N/2$, then the process will tend to get pushed towards N and stay near there; similarly, it may seem that for an initial state less than $N/2$, the population will tend to get pushed towards 0 and will stay near there. So it would seem that observing the state of the system even infinitely far into the future will provide some information about its starting state. This intuition is mistaken; the process described is regular and so the Markov Chain convergence theorem applies. Multiple adaptive peaks do not immunize a lineage from information loss if the lineage has a nonzero probability of crossing valleys. Rather, as we will see in §8, peaks merely slow the rate of information loss.

Inequalities and dynamics

Although the two quantities—entropy and conditional entropy—can (and usually do) differ in value, there is a fundamental and classic inequality from information theory that connects them: the conditional entropy of the lineage after t units of time never exceeds the entropy of the lineage at that time. That is:

[13] In the Moran model each step of the process involves the death of a single individual and its replacement by a new one (see e.g. Durrett 2002).

$$H(X_t|X_0) \leq H(X_t) \tag{5}$$

Inequality (5) is a consequence of a far more general inequality, namely that for *any* two random variables X, Y (regardless of Markov processes or time) we have $H(X|Y) \leq H(X)$. Inequality (5) is maximal at $t = 0$ (since $H(X_t|X_0) = 0$ for $t = 0$) and the non-negative difference $H(X_t) - H(X_t|X_0)$ converges to zero as t grows for any Markov process that has an equilibrium distribution.[14] Inequality (5) makes intuitive sense when entropy is interpreted as a measure of uncertainty: learning the system's state at $t = 0$ cannot on average increase your uncertainty about what its state will be later on.

There are two ways in which the weak inequality (5) can be an equality. The first is if $p_i = 1$ for some state i. In that case, observing the state at time 0 tells you nothing about the process (including its likely state at time t) that you did not know without observing it; you knew already what state it must be in at time 0. In this special case, the following equalities hold:

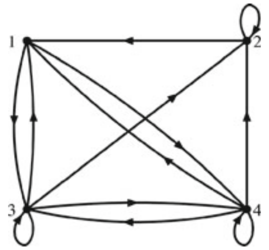
$$\begin{aligned} H(X_t) &= -\sum_j \Pr(X_t = j)\log(\Pr(X_t = j)) = -\sum_j p_{ij}\log(p_{ij}) = H(X_t|X_0 = i) \\ &= H(X_t|X_0). \end{aligned}$$

But even when the initial distribution is not this extreme, there are simple discrete-time Markov chains for which (5) is an equality for all positive values of t; i.e., $H(X_t|X_0) = H(X_t)$ for all $t > 0$, in which case the initial state tells you nothing about any future states. An example, from Mossel (1998), is the Markov chain with four states 1,2,3,4, with an arbitrary non-zero initial distribution p_1, p_2, p_3, $p_4 > 0$, and transition probabilities:

$$\begin{aligned} p_{11} &= p_{12} = p_{23} = p_{24} = 0;\ p_{13} = p_{14} = p_{21} = p_{22} = 1/2, \text{ and} p_{3j} = p_{4j} \\ &= 1/4, \text{ for all} j. \end{aligned}$$

The transition matrix P for this chain, and the transition digraph, are:

$$P = \begin{bmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$$



This Markov chain is regular. Yet after just two time steps we have no idea (from X_2) which of the four possible initial states the process was in at time 0, since the four states have equal probability after two time steps, regardless of the starting state.[15] Notice also that this particular matrix P is singular (i.e., its determinant

[14] Since both $H(X_t)$ and $H(X_t|X_0)$ converge to $-\sum_i \pi_i \log(\pi_i)$.

[15] We describe a further interesting property of this chain in the Section "Tree-like evolution".

det(P) is zero[16]). The condition det(P) = 0 is not possible for a continuous-time Markov process.[17] In particular, for the two-state process described in the Introduction, we always have det(P) > 0 (technically, this discrete process embeds within a continuous-time process). Let us call a Markov process *non-singular* if it is either a continuous-time process, or if discrete, its transition matrix has non-zero determinant. With this in hand, we have the following result (the proof of which is given in the Appendix), which states that for a non-singular Markov process the expected amount of information that the initial state provides about a future state never vanishes completely in any finite time:

Proposition 4 *For a non-singular Markov process, and with $p_i > 0$ for at least two states $i = i_1$, i_2, the weak inequality (5) is strict for all $t > 0$, i.e. $H(X_t|X_0) < H(X_t)$.*

Regarding the dynamic behavior of conditional entropy as a function of time, the non-negative difference $H(X_t) - H(X_t|X_0)$ between unconditional and conditional entropy is exactly identical to the so-called 'mutual information' I between X_0 and X_t, denoted $I(X_0; X_t)$, which measures how much observing the state of the process at time 0 tells you, on average, about which state the process will be in at time *t,* and vice versa (a formula for I is given in the Appendix). A fundamental property of Markov models is the 'data processing inequality' in information theory which states that if $Y \rightarrow W \rightarrow Z$ is a Markov chain, then $I(Y;Z) \leq I(Y;W)$.[18] Applying this to $Y = X_0,\ W = X_t,\ Z = X_{t'}$ for $t' > t$, gives a classic result:

Proposition 5a *For* any *Markov process, $I(X_0; X_t)$ is non-increasing as a function of time.*

We will see later that for non-Markovian mixtures of Markov processes, mutual information can exhibit markedly different dynamics. However, for Markov processes, Proposition 5a has the following consequences for the dynamics of conditional entropy:

Proposition 5b *In a finite-state Markov process the conditional entropy is always increasing if either: (i) the entropy is increasing; or (ii) the lineage's initial state is characterized by a non-trivial[19] equilibrium distribution; or (iii) the process has a uniform non-trivial equilibrium distribution.*

Part (i) is equivalent to Theorem 4 of Sober and Barrett (1992); both it and part (ii) follow directly from Proposition 5a, while part (iii) requires more detailed argument (see Cover and Thomas 1991, p. 35).

[16] An equivalent definition of singularity is that we can write some row of the matrix as a linear combination of the other rows. Singular matrices are thus, in some sense, 'exceptional'.

[17] Since the transition matrix of a continuous process acting for time t with intensity matrix Q can be written as exp(Qt), and Jacobi's formula assures us that det(exp(Qt) = exp(tr(Q)t) > 0.

[18] $Y \rightarrow W \rightarrow Z$ means that W screens off Y from Z. If, in addition, Z screens off Y from W, then the inequality $I(Y;Z) \leq I(Y;W)$ becomes an equality (see, e.g., Cover and Thomas 1991).

[19] Non-trivial here means that the equilibrium distribution assigns strictly positive probability to at least two states.

The rate of information loss

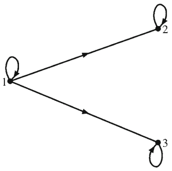
When the mutual information $I(X_0; X_t)$ between X_0 and X_t converges to zero as t grows, the present state of a system tells you progressively less about the future state the further you look into the future. Moreover, for a regular Markov process this loss of information is rapid. Not only does the future become harder to accurately predict with time; it becomes, in a sense, 'exponentially harder' with time. This is made precise by the following result (the proof of which is given in the Appendix):

Proposition 6 *For any finite-state Markov process that is regular, $I(X_0; X_t) \leq Ce^{-ct}$ for all $t > 0$, where the constants C, $c > 0$ depend on just the process. Thus, $I(X_0; X_t)$ converges to zero exponentially quickly with increasing t.*

However, for non-regular Markov processes, $I(X_0; X_t)$ need not converge to zero with increasing t. We illustrate this with a 'toy' example from Sober and Steel (2002) of a chain on three states, and with a variation of our population-genetic example of drift with symmetrical mutation.

Consider first the chain on three states, with uniform distribution for X_0 ($p_1 = p_2 = p_3 = {}^1/_3$) and transition matrix and transition digraph given by:

$$P = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$



This Markov chain is aperiodic (but not irreducible, since one cannot move from state 2 or 3 to any other state). Moreover, det(P) is non-zero. However, $I(X_0; X_t)$ converges to a non-zero value as t tends to infinity. In other words, observing the state at time 0 provides an expected amount of information about the future state that never vanishes no matter how far into the future you look.

As a second and more biologically-relevant example, consider the population genetic process described earlier, in which the frequency X_t of character state A in a population of size N at time t is subject to pure drift. Suppose there is no mutation between states, so the population eventually becomes fixed either with $X_t = 0$ or $X_t = N$ (these are the two 'absorbing states' of the Markov chain). Suppose that at time 0 each of the $N + 1$ possible states for X_0 is equally probable. Thus p_i is uniform and the population is equally likely to become fixed at all-A or at all-B. In that case, the unconditional entropy $H(X_t)$ *decreases*, from its maximal value of log(N) at time 0 to log(2) as t tends to infinity. The behaviour of the conditional entropy $H(X_t|\mathrm{X}_0)$ is more interesting, and is described in the following result, whose proof is given in the Appendix.

Proposition 7 *In the Wright-Fisher model for neutral genetic drift with no mutation in a haploid population of size N (large) where at time 0 each of the $N + 1$ possible states for X_0 is equally probable, $H(X_t|X_0)$ initially increases (as t grows) from its value of zero at time 0, but later decreases to a limit that is strictly less than the corresponding limit of $H(X_t)$.*

Moreover, as with the modification to our ‘toy’ example, the mutual information $I(X_0; X_t)$ in this process of drift without mutation does not converge to zero as t tends to infinity. In this process, any observation at time 0 (other than $N/2$ if N is even) tells you which state (0 or N) is more likely to be the absorbing state, and the closer your observation lies to one of these two absorbing states the more information you gain.

Notice that introducing any non-zero symmetrical mutation into this model, no matter how small, results in a fundamental jump in the limit of $I(X_0; X_t)$ as t tends to infinity; it jumps from a non-zero value to 0, by the Markov chain convergence theorem. Allowing a low non-zero mutation rate in the model means that Proposition 6 applies (so $I(X_0; X_t)$ converges to zero exponentially fast with t) but the constants C, c depend on the model. In particular, as the mutation rate shrinks to zero, one or both of these constants will also converge to zero.

So whether there is zero mutational input matters a lot to the trajectory of the mutual information. With zero mutation, the mutual information asymptotes to zero; with nonzero, the mutual information does not. However, the difference between these two situations disappears if we consider only finite lineage durations. For any finite time t, the mutual information is positive. What matters for practical purposes is not what happens in the infinite limit but how much information an observation at one time provides about another that is finitely far away.

Another example in which the rate of information loss is lowered involves frequency-dependent selection. Consider a population in which each individual has trait A or trait B, and selection favors the trait that is in the majority. Selection will push the lineage to either all-A or to all-B, depending on the lineage’s starting state. If we observe this lineage at one time, how much does this observation tell you about the lineage’s state earlier or later? If there is mutational input, Proposition 6 applies and you know that the mutual information decays to zero (and exponentially fast) and in the infinite limit it is zero. Without mutational input, Proposition 6 does not apply (since the process is not irreducible) and the mutual information is positive even in the limit. But for finite temporal intervals, the mutual information is positive, regardless. It is larger when there is no mutation. And in both cases, the mutual information is larger when there is frequency dependent selection for the majority trait. These differences are reflected in the magnitude of the constants C and c in Proposition 6. Once again, the actual rate of information loss at some time t in the finite future is heavily influenced by the properties of the model.

Looking into the past

One also can consider the conditional entropy of the state of a Markov process t time steps back into the past, given the present state. This case is described by the following formal identity from information theory[20]:

[20] Equation (6) follows from symmetry of the mutual information function I, so that $H(X_t) - H(X_t|X_0) = I(X_0; X_t) = I(X_t; X_0) = H(X_0) - H(X_0|X_t)$.

$$H(X_0|X_t) = H(X_t|X_0) + H(X_0) - H(X_t) \tag{6}$$

This equation applies whether or not the process is in equilibrium. If it is in equilibrium (as it is in our population-genetic example of drift), then $H(X_0) = H(X_t)$, in which case it follows that $H(X_0|X_t) = H(X_t|X_0)$. In other words, if we are going to observe the system's present state, our expected uncertainty concerning the state of the system t time steps into the future is the same as our expected uncertainty concerning the state of the system t time steps in the past. This does not mean that the Markov process would appear the same when run forward or backward in time (this is what 'time reversibility' means in Markov chain theory[21]), as the last equation applies for any Markov process in equilibrium. Note also that the rate of information loss as we move t time steps into the past is the same as that described in Proposition 6 for t time steps into the future, provided once again that the process is in equilibrium.

If the process is not in equilibrium, the comments above need to be modified. The difference $H(X_0) - H(X_t)$ tells you whether your present observation is expected to tell you more about the process t time steps in the future or about t time steps in the past (the former applies if $H(X_0) - H(X_t) > 0$, the latter if $H(X_0) - H(X_t) < 0$). This difference in the unconditional entropies is the criterion for whether conditionalizing on an observation is expected to provide more information about the future or about the past.

Non-Markov processes

A discrete time process has the Markov property precisely when, for any time t, states i and j, and history h we have:

$$\begin{aligned}&\Pr(\text{system is in state } j \text{ at time } t+1|\text{system is in state } i \text{ at time } t) =\\&\Pr(\text{system is in state } j \text{ at time } t+1|\text{system is in state } i \text{ at time } t\\&\,\&\text{ system had history } h \text{ before time } t).\end{aligned}$$

At first glance, it may seem that every sensible model of a physical process must have the Markov property; otherwise, the model will entail some sort of weird "action at a temporal distance" whereby the past can influence the future without having that influence transmitted through the present. In fact, a perfectly reasonable model of a process can violate the Markov property if the model is causally incomplete. For example, consider the fact that Mendelian inheritance is generally thought to be Markovian; grandparents influence the genotypes of their grandchildren only by way of influencing their children. There is no action at a temporal distance in genetic transmission. However, this fact about the process does not mean that there can't be models in population genetics in which the state of generation 3 depends not just on the state of generation 2 but on that of generation 1 as well. The

[21] A process is *time-reversible* precisely when it is in equilibrium and $\pi_i \Pr(X_t = j|X_0 = i) = \pi_j \Pr(X_t = i|X_0 = j)$ for all states i, j and $t > 0$. Any two-state process in equilibrium is time-reversible, whether it involves drift or selection. For any $n > 2$ there are n-state Markov processes that are not time-reversible (Häggström 2002).

theory of inbreeding provides a case in point; when all matings in an infinite population are between sibs, the heterozygosity (Z_t) in generation t depends on the heterozygosity in the previous two generations, $t - 1$ and $t - 2$ (Crow and Kimura 1970, p. 87):

$$Z_t = \left(\frac{1}{2}\right)Z_{t-1} + \left(\frac{1}{4}\right)Z_{t-2}$$

The *process* of inheritance is Markovian, but *models* of the process can fail to be when they are causally incomplete.

One formal device for constructing a non-Markov model is by taking a *mixture* of two or more Markov processes; Chang (1996, p. 213) makes this point in connection with phylogenetic inferences that use DNA substitution models. Mixture models have becomes widely used in molecular phylogenetics to model rate heterogeneity across DNA sequence sites (and, more recently, to model changing processes in a tree). To illustrate a particularly simple example of rate heterogeneity across sites, and how this leads to non-Markovian behavior, consider a simple model of DNA site substitution in which each site can either be ‘on’ (free to undergo substitution) or ‘off’ (invariable due to functional or structural constraints), and these two classes remain fixed through time. Then the state X_t of a randomly-selected site at time t (namely, A, C, G, or T) can be modeled as a mixture of two Markov processes (one a regular model of site substitution, the other a trivial Markov process in which each state remains unchanged). It is easily seen that X_t is not a Markov process, as the past is not conditionally independent of the future given a present observation. For example, we have:

$$\Pr(X_t = A|X_{t-1} = A\&X_{t-2} = A) > \Pr(X_t = A|X_{t-1} = A),$$

since knowing that a site has remained unchanged for two time steps provides some evidence that it is in the ‘off’ class and so is more likely to be found in that same state at the next time step.

Another simple non-Markov model can be constructed by using the two-state Markov model described earlier. We previously represented selection for state 1 by way of the inequality $u > v$; symmetrically, selection for state 0 can be expressed by the inequality $u < v$. Each of these models is Markovian. Now let us mix these two models together by supposing that there will be selection for state 1 if the lineage begins in state 1 and selection for state 0 is the lineage begins in state 0. The lineage transition probabilities for this new model are depicted in Fig. 3 (using the assumption that $u = x_1 > > v = x_2$ when there is selection for state 1 and $u = x_2 < < v = x_1$ when there is selection for state 0). This mixed model violates the Markov property because the probability a lineage has of being in state 1 at time $t + 1$ depends, not just on the state of the lineage at time t, but also on the values of u and v, and those values depend on the state of the lineage at time $t = 0$.

Each of the Markov models (the one with $u > v$, the other with $u < v$) is regular as long as $0 < u,\ v < 1$, and so the Markov chain convergence theorem guarantees that the mutual information that characterizes the relationship between a past and a future time slice of the process asymptotes to zero exponentially fast as the temporal

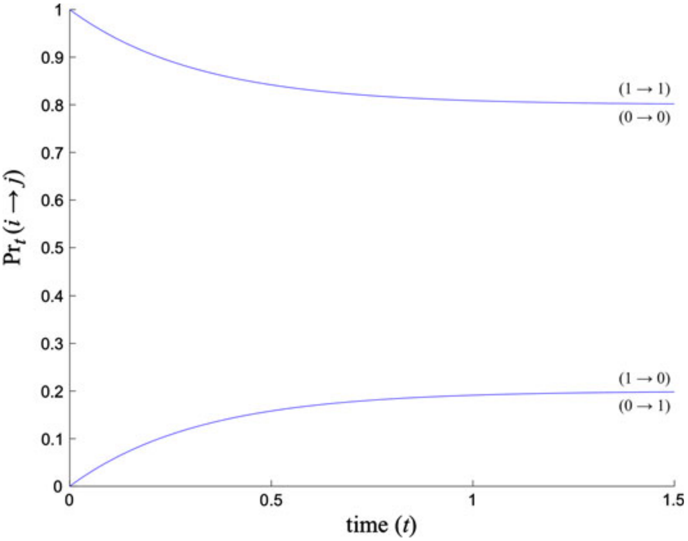


Fig. 3 A non-Markov model in which there is selection for state 1 when the lineage begins in state 1 and selection for state 0 when the lineage begins in state 0

separation of the two slices increases. However, the new mixed model is not Markovian and it has the interesting property that the mutual information between past and future does not asymptote to zero. As can be seen from Fig. 3, ascertaining the initial state of the lineage helps you predict the lineage's state in the future, even when the initial and future moments are infinitely separated. The same point holds for inferring past from present; even if the past time slice is infinitely long ago, observing the present state of the lineage helps you infer its initial state.[22]

In the example just described of a non-Markov process, the conditional entropy monotonically increases (see Fig. 4). However, it is easy to describe another example in which the conditional entropy behaves differently. Consider the mixed model that we have just described but with a twist: selection now favors the *opposite* state to the one in which the lineage begins. Thus, let us suppose the lineage starts, with equal probability, in state 0 or 1, and if the lineage starts in state 0 the process converges to an equilibrium of (x, $1 - x$), where $0 < x < ½$, while if the lineage starts in state 1 the process converges, at the same rate, to an equilibrium of ($1 - x$, x). In this case the conditional entropy $H(X_t|X_0)$ initially rises with increasing t from its initial value of 0 at $t = 0$ to a maximum value of log(2) at $t = f(x)$, where f is a function that tends to infinity as x increases towards ½. However, for values of t greater than $f(x)$, the conditional entropy $H(X_t|X_0)$ *declines* as t grows towards an asymptotic limit of the entropy of the equilibrium distribution (x, $1 - x$) of each process, namely $g(x) =_{def} -x\log(x) - (1-x)\log(1-x)$ (see the Appendix, Part B, for details). In other words, when we observe the present state of the system, our

[22] This discussion corrects some of what Sober (2008, pp. 300–306) says about a model of frequency dependent selection for the majority trait; the model is nonMarkovian.

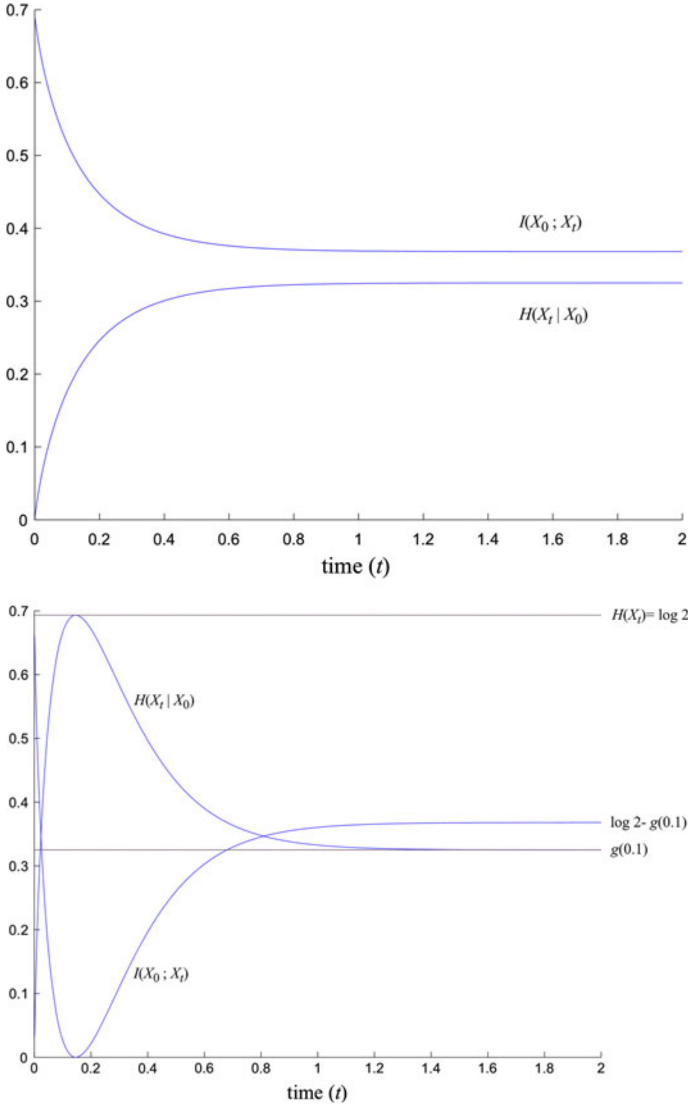


Fig. 4 The dynamics of conditional entropy and mutual information for two non-Markov processes. Each process is a mixture of a matching pair of two-state regular Markov processes; each non-Markov process models a form of selection. *Upper:* If selection favors the lineage's initial state, then mutual information and conditional entropy change monotonically. *Lower:* If selection works against the lineage's initial state, then mutual information can drop to zero and then start increasing, a behavior that is impossible for any Markov process (*cf.* Proposition 5a). See text for further details

expected uncertainty about the state of the system in the near future will be greater than our uncertainty for the distant future.

The stronger the selection, the closer to zero is the limiting value of the system's conditional entropy in the distant future (since $g(x)$ tends to 0 as x decreases to 0).

Notice that, as selection weakens and more and more resembles the case of pure drift (i.e., as x increases towards ½), the decline of $H(X_t|X_0)$ is pushed further into the future by the stated property of $f(x)$. Furthermore, in this example, the entropy $H(X_t)$ remains constant at log(2) for all values of t, and so the mutual information $I(X_0;X_t)$ initially declines from its initial value of log(2), reaching 0 at $t = f(x)$ before increasing towards its asymptotic limit of $\log(2) - g(x)$. These dynamical aspects of mutual information and entropy are illustrated in Fig. 4 for the case in which $x = 0.1$.

In the mixed models just described, the transition probabilities deterministically depend on the lineage's starting state. The more general point is that the mixed model will fail to be Markovian if the lineage transition probabilities are probabilistically dependent on the state of the lineage at any time. Non-Markovian models don't always have the mutual information remaining positive in the infinite limit, but a general sufficient condition for this can now be stated.

Suppose we have $k > 1$ regular Markov processes, $M_1, M_2, \ldots, M_k$, each on the same state space S, each with its own equilibrium distribution $\pi^1, \pi^2, \ldots, \pi^k$. Consider the following random process $(X_t : t \geq 0)$ on state space S. At time $t = 0$ we select model M_α with probability q_α and select a state according to some initial distribution (p^α) of model M_α. This is the (random) initial state X_0. We then evolve this state according to this Markov process (M_α) for time t. We refer to the process X_t as a *mixture of Markov processes*. We say that this mixture is *proper* if (i) $q_\alpha > 0$ for at least two values of α, (ii) the initial distributions p^α are not all equal, and (iii) for each state i, $p_i^\alpha > 0$ for at least one α.

Recall that for any regular Markov process the mutual information between the initial state and the state at time t always decays to zero (and exponentially fast). Yet, with mixtures of two such processes the behavior of the mutual information function with increasing time is markedly different, as the following result shows.

Proposition 8 *For any proper mixture of* any *two regular Markov processes with different equilibria:* $\lim_{t\to\infty} I(X_0;X_t) > 0$.

That is, the mutual information between X_0 and X_{t} does not decay to zero. Notice that the condition that the initial distributions are not all identical is implied by the non-identity of equilibria condition when the initial distribution for each chain is its equilibrium distribution.

Proposition 8 is an immediate corollary of the following more general result for mixtures of $k \geq 2$ regular Markov processes, whose proof is given in the Appendix. Let Π be the $k \times |S|$ matrix whose rows are the vectors $\pi^1, \pi^2, \ldots, \pi^k$.

Proposition 9 *For a proper mixture of any $k \geq 2$ regular Markov processes,* $\lim_{t\to\infty} I(X_0;X_t) > 0$ *if the rows of Π are linearly independent.*

We end this section by noting that mixtures do not provide the only way to generate a non-Markov model from Markov models. A more general device is to take a Markov model defined on some large state space, but then to suppose that the process we are able to observe is merely some property or 'shadow' of the states in that larger space. More precisely, given a Markov process Z_t on a state space Ω, and a function $f : \Omega \rightarrow S$ the process $X_t = f(Z_t)$ is sometimes referred to as a *lumped Markov process*.[23] As an example, consider the model of DNA site substitution with invariable sites described near the start of this Section. This was described as a mixture of two regular Markov processes on four states, but it can be equally well described as a lumped Markov process in which Ω is the eight-element set $\{A_{on}, A_{off}, C_{on},\ldots\}$ and f simply ignores the subscript on each state. This is no coincidence, as *any* finite mixture of Markov processes can be described as a lumped Markov process for a suitable choice of f and Ω (Appendix, part B, provides a proof). Moreover, this alternative viewpoint using a lumped process can accommodate biological phenomena that cannot easily be described by mixtures. For example, suppose that in the model of DNA site substitution with invariable sites we make the following adjustment: sites that are 'off' may turn 'on' and sites that are 'on' may turn 'off', according to a further Markovian mechanism. The resulting 8-state Markov process on Ω is a simple 'covarion drift' process (based on Walter Fitch's covarion model from the early 1970s; see Tuffley and Steel 1997b). As before, the lumped process on the (observable) states {A,C,G,T} is no longer Markovian; but the twist now is that this lumped process provably *cannot* be represented as a finite mixture of Markov processes.[24]

Tree-like evolution

We have considered changes in the conditional and unconditional entropies of a Markov process as it unfolds in a single lineage through time. However, Markov processes are also widely used in evolutionary biology to study the evolution of a discrete character on a bifurcating phylogenetic tree (Felsenstein 2004). In particular, the states observed at the tips (leaves) of an evolutionary tree (which correspond to present-day species) contain information about the state at an ancestral node. Consider, for example, the tree shown in Fig. 5a, and assume that the same Markov process applies to each of the seven edges in this tree. Under the usual Markov assumptions, the leaves D_2 and D_3 are conditionally independent given the state at C. But D_2 and D_3 are not conditionally independent given the state at the root node R. By the 'information-processing inequality' (described just prior to Proposition 5a), the pair D_2 and D_3 conveys at least as much information concerning the state at C as at R. But an interesting question arises for this tree: From which two pairs of leaves—(D_1, D_4) or (D_2, D_3)—do we learn more about the ancestral state at R? The answer turns out to depend on the type of process at work.

It might seem that the pair (D_1, D_4) is preferred as it provides two independent observations generated from R, as compared with the two observations (D_2, D_3), which are non-independent, owing to a period of 'shared history'. Indeed, the mutual information between the states at (D_1, D_4) and R is at least as large as the mutual information between the states at (D_2, D_3) and R for certain Markov

[23] A lumped Markov process is not generally a Markov process; necessary and sufficient conditions for it to be so are well known (see, for example, Kemeny and Snell 1976).

[24] This follows from results in Tuffley and Steel (1997b).

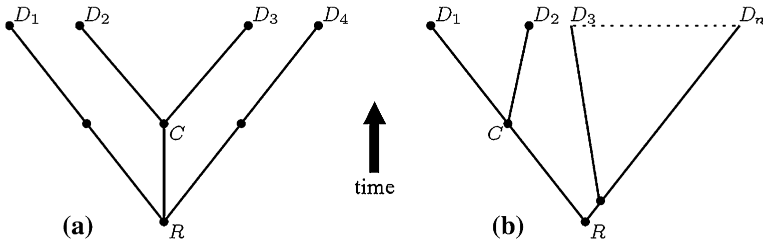


Fig. 5 **a** A tree in which the pair of leaves D_1, D_4 can tell you either more—or less—than the pair D_2, D_3 does about the state of an ancestral node R, depending on the model. **b** A tree in which the present-day leaves can tell you more about an ancient node R than about a more recent one C. See text for details

processes, such as the two-state symmetric process (pure drift) (Evans et al. 2000). Remarkably, however, for other (discrete) Markov processes exactly the opposite inequality can hold, as shown by (Mossel 1998).[25] The latter paper furnishes a particularly striking instance of this inequality based on the four-state Markov chain described in Section "Inequalities and dynamics". If we apply this transition matrix to each edge of the tree in Fig. 5a, then the mutual information between the pair (D_2, D_3) and R is strictly positive; yet for (D_1, D_4) and R, the mutual information is identically zero.[26]

For a general Markov process on the tree in Fig. 5a, the collection of states at the leaves D_1, …, D_4 may tell you more about the ancestral state at R than about C if arbitrary transition matrices on the seven edges of the tree are allowed. This last result may seem to go contrary to the information-processing inequality (which, loosely stated, says that the present provides more information about the recent past than it does about the more distant past), but this isn't so; the information-processing inequality concerns a "chain internal" comparison. The four leaves will be more informative about R than about C in a two-state drift model in which the rate of substitution is very low on the edges leading to D_1 and D_4, yet very high on the other three edges of the tree.

Allowing rates to vary among edges in this way seems to involve 'cheating' and suggests a more interesting question: is it possible for the leaves of a tree to be more informative about the state of a deep ancestral node than a more recent node, under a two-state drift model that has a constant rate of substitution? Perhaps surprisingly, the answer is *yes*, though we need to go to larger trees to find cases in which this is true. Consider the tree in Fig. 5b, in which the root node R is t years in the past, C is $t/2$, and the 'triangular' tree is a completely balanced binary tree with $n - 2 = 2^h$ leaves, with edges of equal length (and equal to the edge joining the triangular tree to R). Suppose there is a substitution rate r on all edges of the tree. Using results

[25] Sober (1989) uses a likelihood framework to show that with a two-state drift process, the pair of observations $D_1 = 1$ and $D_4 = 1$ provides stronger evidence that $R = 1$ than does the pair of observations $D_2 = 1$ and $D_3 = 1$; however, with other processes, the relationship reverses.

[26] The mutual information is positive for (D_2, D_3) and R because if we observe state 1 at D_2 and state 3 at D_3 then C must be have been either in state 3 or 4, and so R cannot have been in state 2. The mutual information is zero if we replace (D_2, D_3) by (D_1, D_4) because taking two steps in the chain produces the uniform distribution from any starting state.

from Evans et al. (2000), the following result can be established: for any $\delta > 0$ (no matter how small) one can select n sufficiently large, and an appropriate substitution rate r so that if X is the collection of states observed at the n leaves, and X_R, X_C are the states at the nodes R, C respectively, then:

$$I(X_R; X) > 0.1 \text{ and yet } I(X_C; X) < \delta.$$

In other words, present-day observations could, in certain cases, be more informative about a node that is 1 billion years old than they are about a node that is merely 500 million years old.[27]

Mutual information can also be analysed on larger trees. For example, for the two-state symmetric (drift) model, a general bound on the mutual information between the state at the root X_0 of a phylogenetic tree t years in the past and the states at the set Y_t of n present-day leaves of the tree has been described (Sober and Steel 2002). The result, based on a crucial inequality from (Evans et al. 2000), states that

$$I(X_0, Y_t) \leq ne^{-4rt}$$

where r is the rate of substitution. This inequality provides limits on the accuracy one can hope to achieve when reconstructing ancestral states deep in the past, depending on the conflicting interplay of the substitution rate and the number of present-day observations.

To illustrate the implications of this inequality, consider any phylogenetic tree in which a single ancestral species that existed $t = 9$ million years ago now has $n = 100$ current descendants, where drift is the process at work in branches. If the rate of substitution is 1 per 3 million years, then the mutual information connecting the 100 descendants to their most recent common ancestor is at most 0.0006. What does this number mean? Before you look at the states of a character across the 100 leaf species, your prior probabilities for the two possible states of their most recent common ancestor are ½, ½. If you now look at the state of the 100 descendants, how much will your observation change these prior probabilities? The answer is that the change is, in expectation, about 0.0006. Moreover, no method can estimate the root state from the leaf states with accuracy more than (about) ½ + 0.0006; this is a consequence of Fano's inequality (see Cover and Thomas 1991). In other words, your prospects in this situation are essentially no better than what you'd be able to achieve by ignoring the leaf states and just tossing a fair coin. With less time separating root and leaves, or more descendants, or a smaller substitution rate, the epistemic situation would be rosier. Further results in this direction have been developed by (Mossel 2003) and (Mossel and Steel 2005).

We end this section by noting that, for any tree, and any Markov process, we maximize the expected information about the state at an ancestral node by observing the states at *all* present-day leaves; with respect to the tree in Fig. 5a, this means using the observed states at all four leaves $D_1 - D_4$. Stated formally, if X is the collection of states at all the leaves, and Y is the collection of states at just some of the leaves, then, for any Markov process, $I(X_0; X) \geq I(X_0; Y)$, where X_0 is the state at

[27] A related example, involving maximum parsimony, was described by Fischer and Thatte (2009).

any ancestral node in the tree.[28] This result from information theory helps justify *the principle of total evidence.*

This inequality does not preclude the possibility that some methods for inferring anestral states may perform better on certain trees when the observations used are restricted to the states at just some of the leaves. A case in point is maximum parsimony (which is ordinaly equivalent with maximum-likelihood, if there is a symmetric (drift) Markov process at work on each character on each branch and branch lengths are considered as nuisance parameters (Tuffley and Steel 1997a, Theorem 6). In this case, it has recently been shown that the accuracy[29] of ancestral state estimation can sometimes be improved by using the observed states at a subset of the leaves (for details, see Fischer and Thatte 2009, Theorem 1).[30] However, for maximum likelihood estimation in the absence of nuisance parameters, accuracy is provably maximized by using all the observations available (see e.g. Berger 1985, p. 159).

Concluding comments

There is no general principle of entropy increase in Markov processes of evolution. True, if a Markov process is in equilibrium, then the conditional entropy $H(X_t|X_0)$ always increases. But if a lineage does not begin at equilibrium, the entropy can be strictly increasing, or strictly decreasing, or increasing and then decreasing. This point holds whether it is entropy or conditional entropy that one wishes to track. Nor is there any simple relationship between the entropy trajectory and the question of whether the process has a direction (as in selection) or not (as in drift). Drift without mutation leads entropy to decline, whereas drift with mutation can lead entropy to increase. And selection can destroy variation (in which case the entropy goes down), but it also can produce a stable polymorphism (in which case the entropy may go up).

Matters become more orderly when we leave the value of a single entropy behind and focus on the difference between two entropies—the unconditional and the conditional. For *any* Markov process, the mutual information $I(X_0; X_t) = H(X_t) - H(X_t|X_0)$ is non-increasing. It converges to zero if the chain is regular but for other chains it can converge to a nonzero limit. It vanishes completely in finite time only for 'singular' processes (Proposition 4). Since continuous-time Markov processes can't be singular, the bad news of information zeroing out in finite time does not arise in that case. For a regular chain, the rate at which $I(X_0; X_t)$ goes to zero is

[28] This inequality is an immediate consequence of the data-processing inequality (referred to just before Proposition 5a) since $X_0 \rightarrow X \rightarrow Y$ is a Markov chain.

[29] 'Accuracy' here refers to the expected probability of correctly reconstructing the ancestral state by the method.

[30] If the state of all the leaves on a tree provides at least as much information about the state of the root as any subset of the leaves provides, and if parsimony sometimes does better at reconstructing the state of the ancestor when it consults only a subset of the leaves, then parsimony sometimes misinterprets what the full data set is saying. Of course, the sub-optimal performance of parsimony in this context leaves open that the method might perform optimally under other models of evolution.

exponential with t, but the rate depends heavily on the details of the model (e.g., frequency-dependent selection can make the rate of information loss much slower than the rate associated with drift + mutation or with frequency independent selection).

Mixtures of Markov processes—and more generally lumped Markov processes—can fail to be Markovian. Even when a non-Markovian mixed process is regular, the mutual information between two time slices 'usually' fails to asymptote to zero as the temporal separation of the slices increases; exceptions to this rule arise if the component Markov processes have equilibria that are linearly dependent or the initial distributions of the component processes are the same.

The loss of information within a tree—as one moves from present-day observations at the leaves towards ever deeper ancestral nodes—involves additional complexities. It is possible for the leaves to tell you more about more ancient nodes than about ones that are more recent. The (exponential) rate of loss of information with time in a chain is tempered by the number of present-day observations; ancient ancestors often have more present day descendants than more recent ancestors do, and in this respect the character states of ancient ancestors may be more accessible. In addition, information about an ancestral node is maximized by using all present-day observations even though some estimation procedures can be more accurate when they use only a subset of the observations.

Acknowledgments We thank the editor and referee for useful suggestions. ES thanks the William F. Vilas Trust of the University of Wisconsin-Madison for financial support and MS thanks the Royal Society of New Zealand for funding under its James Cook Fellowship scheme.

Appendix: Technical details

Part A: Proof of propositions

First recall that the 'mutual information' I between X_0 and X_t is defined, as usual, as:

$$I(X_0;X_t) = \sum_{ij} \Pr(X_0 = i \& X_t = j) \log\left(\frac{\Pr(X_0 = i \& X_t = j)}{\Pr(X_0 = i)\Pr(X_t = j)}\right). \tag{7}$$

This can be rewritten in terms of the transition probabilities $(\Pr(X_t = j|X_0 = i))$, initial distribution (p_i) and subsequent distribution $(\Pr_\mathrm{t}(j))$ of states as:

$$I(X_0;X_t) = \sum_{ij} p_i \Pr(X_t = j|X_0 = i) \log\left(\frac{\Pr(X_t = j|X_0 = i)}{\Pr(X_t = j)}\right). \tag{8}$$

A straightforward and well-known inequality in information theory is that:

$$I(X_0;X_t) = H(X_t) - H(X_t|X_0). \tag{9}$$

To justify Proposition 4, it suffices, by (9), to show that $I(X_0;X_t) > 0$. Notice that $I(X_0;\ X_t)$ is the Kullback–Leibler distance between the probability distributions $p_i \Pr(X_t = j|X_0 = i)$ and $p_i \Pr(X_t = j)$. In particular, $I(X_0;\ X_t) = 0$ implies that these two probability distributions are identical—that is, $p_i \Pr(X_t = j|X_0 = i) =$

$p_i \Pr(X_t = j)$ for all pairs of states i, j (including $i = j$). As there are two distinct values of i for which $p_i > 0$ then the two corresponding rows of the matrix $P = [p_{ij}] = [\Pr(X_t = j|X_0 = i)]$ are identical (since $p_{ij} = \Pr(X_t = j)$ in both cases), and so $\det(P) = 0$. This contradicts our assumption in Proposition 4, and so $I(X_0; X_t) > 0$.

Next we turn to the proof of Proposition 6. The regularity assumption implies that, for all states i, j we have: $\pi_i \geq \varepsilon$ and $|\Pr(X_t = j|X_0 = i) - \pi_j| \leq Be^{-ct}$ for strictly positive constants B, c, ε (see, for example, Rozanov, 1969; Theorem 7.4). It follows that the logarithmic term in $I(X_0; X_t)$ from (7), namely, $\log\left(\frac{\Pr(X_t=j|X_0=i)}{\Pr(X_t=j)}\right)$, can be written as $\log(1 + O(e^{-ct})) = O(e^{-ct})$, where O is the usual order notation. In particular, for some $C > 0$ we have $I(X_0; X_t) \leq Ce^{-ct} \sum_{ij} p_i \Pr(X_t = j|X_0 = i) = Ce^{-ct}$, as claimed.

We turn now to the proof of Proposition 7. For the Wright-Fisher model, it is well known that:

$$\lim_{t\to\infty} \Pr(X_t = j|X_0 = i) = \left(\begin{array}{cc} i/N, & \text{if } j = N; \\ (N-i)/N, & \text{if } j = 0; \\ 0, & \text{otherwise.} \end{array}\right.$$

Thus,

$$\lim_{t\to\infty} H(X_t|X_0 = i) = -\frac{i}{N}\log\left(\frac{i}{N}\right) - \left(1 - \frac{i}{N}\right)\log\left(1 - \frac{i}{N}\right). \tag{10}$$

Now, for i selected from the uniform distribution, $\lim_{t\to\infty} H(X_t|X_0) = \frac{1}{N+1}\sum_{i=0}^{N} \lim_{t\to\infty} H(X_t|X_0 = i)$, and so, from Eq. (10), $\lim_{t\to\infty} H(X_t|X_0) = \log(N) - \frac{2}{N(N+1)}\sum_{i=1}^{N} i\log(i)$. Thus,

$$\lim_{t\to\infty} H(X_t|X_0) = \frac{2}{(N+1)}\sum_{i=1}^{N} -\frac{i}{N}\log\left(\frac{i}{N}\right) \sim 2\int_0^1 -x\log(x)\mathrm{d}x = \frac{1}{2},$$

where $\sim$ refers to asymptotic equivalence as $N \to \infty$. In summary, we have the following inequality: $\lim_{N\to\infty}\lim_{t\to\infty} H(X_t|X_0) = \frac{1}{2} < \log(2) = \lim_{N\to\infty}\lim_{t\to\infty} H(X_t)$. Now consider $H(X_1|X_0)$. Conditional on $X_0 = i$, the distribution of X_1 under the Wright-Fisher model is binomial with N trials and probability $p = \frac{i}{N}$. Thus, $H(X_1|X_0 = i)$ is the entropy of a binomial distribution with these parameters. Since i is uniformly distributed between 0 and N it can be shown that $H(X_1|X_0)$ grows at the order $\log(N)$ so, for N sufficiently large, $0 = H(X_0|X_0) < H(X_1|X_0)$, and also $H(X_1|X_0) > \lim_{t\to\infty} H(X_t|X_0)$. This justifies the claims in Proposition 7.

Finally we justify Proposition 9. Let us suppose that $\lim_{t\to\infty} I(X_0; X_t) = 0$; we will show that this implies the rows of Π are linearly dependent. Notice that the condition $\lim_{t\to\infty} I(X_0; X_t) = 0$ implies (via Pinsker's inequality) that:

$\lim_{t\to\infty}[\Pr(X_t = j\&X_0 = i) - \Pr(X_t = j)\Pr(X_0 = i)] = 0$, for all pairs of states i, j, which in turn implies that:

$$\lim_{t\to\infty}[\Pr(X_t = j|X_0 = i) - \Pr(X_t = j)] = 0, \tag{11}$$

for all i, j (note that the event $X_0 = i$ has strictly positive probability for all states i by the assumption that the mixture is proper). This further implies the following condition holding for all states i, i', j:

$$\lim_{t\to\infty}[\Pr(X_t = j|X_0 = i) - \Pr(X_t = j|X_0 = i')] = 0, \tag{12}$$

by using Eq. (11) twice (once for the conditioning event $X_0 = i$ and once for $X_0 = i'$). Let Z denote the index $\alpha \in \{1,\ldots,k\}$ of the model M_α that is selected at the start of the process; thus $Z = \alpha$ with probability q_α. We have:

$$\Pr(X_t = j|X_0 = i) = \sum_{\alpha=1}^{k}\Pr(X_t = j|X_0 = i\&Z = \alpha)\Pr(Z = \alpha|X_0 = i).$$

Now, since the Markov process M_α is regular, the Markov chain convergence theorem assures us that, $\lim_{t\to\infty}\Pr(X_t = j|X_0 = i\&Z = \alpha) = \pi_j^\alpha$, for all i, and so:

$$\lim_{t\to\infty}\Pr(X_t = j|X_0 = i\&Z = \alpha) = \sum_{\alpha=1}^{k}\pi_j^\alpha\Pr(Z = \alpha|X_0 = i).$$

Combining this with Eq. (12) gives the following constraint:

$$\sum_{\alpha=1}^{k}\pi_j^\alpha\cdot(\Pr(Z = \alpha|X_0 = i) - \Pr(Z = \alpha|X_0 = i')) = 0, \tag{13}$$

for all states i, i', j. Now, suppose that:

$$\Pr(Z = \alpha|X_0 = i) - \Pr(Z = \alpha|X_0 = i') = 0, \tag{14}$$

for all states i, i' and index α. We have $\Pr(Z = \alpha|X_0 = i) = \Pr(X_0 = i|Z = \alpha)q_\alpha/\Pr(X_0 = i)$, by Bayes' Theorem. Furthermore, since $\Pr(X_0 = i|Z = \alpha) = p_i^\alpha$ (the initial state distribution for M_α) Eq. (14) implies that $p_i^\alpha/\Pr(X_0 = i) = p_{i'}^\alpha/\Pr(X_0 = i')$, for all states i, i' and index α. It follows that we can write $p_i^\alpha = c_i d_\alpha$, for all states i, and indices α, where c_i does not depend on α and d_α does not depend on i.[31] Thus, from which the identity: $\sum_i p_i^\alpha = 1 = \sum_i p_i^\beta$ for any two distinct indices α, β we obtain $d_\alpha = d_\beta$ and thus $p_i^\alpha = p_i^\beta$ for all i; that is, the initial distribution of any two models are equal, in violation of our assumption.

Thus we may suppose that there exists states i, i' and an index α for which $\Pr(Z = \alpha|X_0 = i) - \Pr(Z = \alpha|X_0 = i') \neq 0$. Then the row vector $\Delta = [\Delta_\alpha]$ defined by

$$\Delta_\alpha := \Pr(Z = \alpha|X_0 = i) - \Pr(Z = \alpha|X_0 = i'),$$

[31] To see this, note that we can take $c_i = \frac{\Pr(X_0=i)}{\Pr(X_0=1)}$; $d_\alpha = p_1^\alpha$.

for $\alpha = 1, 2, \ldots, k$, is not equal to the zero vector 0 and yet from Eq. (13) we have $\Delta\Pi = 0$, so the rows of Π are linearly dependent. This completes the proof.

Part B: Proofs of other claims in Section "Non-Markov processes"

First we justify the analysis reported for the mixed model in which the lineage starts, with equal probability, in state 0 or 1, and if the lineage starts in state 0 the process converges to an equilibrium of (x, $1 - x$), where $0 < x < ½$, while if the lineage starts in state 1 the process converges, at the same rate, to an equilibrium of ($1 - x$, x). From (7) we have:

$$I(X_0; X_t) = \sum_{i=0}^{1}\sum_{j=0}^{1} \Pr(X_t = j \& X_0 = i) \log\left(\frac{\Pr(X_t = j \& X_0 = i))}{\Pr(X_t = j)\Pr(X_0 = i)}\right).$$

Now, $\Pr(X_t = j) = \frac{1}{2}$ for all $t \geq 0$. Also, for $i = 0{,}1$, if $\Pr^i$ refers to the Markov process that applies if the initial state is i, then $\Pr(X_t = j \& X_0 = i) = \frac{1}{2}\Pr^i(X_t = j | X_0 = i)$, and so:

$$I(X_0; X_t) = \sum_{i=0}^{1}\sum_{j=0}^{1} \frac{1}{2}\Pr^i(X_t = j | X_0 = i) \log(2\Pr^i(X_t = j | X_0 = i)). \qquad (15)$$

Setting $a(t) := x + (1 - x)e^{-t/\gamma}$; $\quad b(t) := (1 - x)(1 - e^{-t/\gamma})$, gives:

$$\Pr^0(X_t = 0 | X_0 = 0) = \Pr^1(X_t = 1 | X_0 = 1) = a(t);$$
$$\Pr^0(X_t = 1 | X_0 = 0) = \Pr^1(X_t = 0 | X_0 = 1) = b(t).$$

Consequently, by Eq. (15), we have:

$$I(X_0; X_t) = a(t)\log(2a(t)) + b(t)\log(2b(t)). \qquad (16)$$

Since $a(t) = b(t) = \frac{1}{2}$ for the value t_x of t for which $e^{-t/\gamma} = \frac{1-2x}{2(1-x)}$, we have $I(X_0; X_t) = 0$ at $t = t_x$. Routine analysis with Eq. (16), now gives $\lim_{t\to\infty} I(X_0; X_t) = \log(2) - g(x)$, as claimed.

Finally, we justify the assertion that a finite mixture of Markov processes can be described as a lumped Markov process. Suppose a process X_t is described by a finite mixture of Markov processes $M_\alpha (\alpha = 1, \ldots, k)$, each on state space S, with model M_α selected with probability q_α, and with p_i^α as the initial distribution of state i in model M_α. Let $\Omega = \{(s, \alpha) : s \in S, \alpha = 1, \ldots, k\}$ and consider the following Markov process Z_t on Ω. Select the initial state $(i, \alpha) \in \Omega$ with probability $q_\alpha \cdot p_i^\alpha$ and define transition probabilities on all ordered pairs of states from the set Ω as follows: $\Pr(Z_t = (j, \alpha) | Z_0 = (i, \alpha)) = \Pr^\alpha(X_t = j | X_0 = i)$, where $\Pr^\alpha(X_t = j | X_0 = i)$ is the transition probability within the model M_α, and $\Pr(Z_t = (j, \beta) | Z_0 = (i, \alpha)) = 0$ for all $\alpha \neq \beta$ and all states i, j. Then, under the function $f : \Omega \to S$ defined by $f((s, \alpha)) = s$, the processes (X_t; $t \geq 0$) and ($f(Z_t)$; $t \geq 0$) have the same distribution; that is, mixtures are just a special case of lumped processes, as claimed.

References

Barrett M, Sober E (1994) The second law of probability dynamics. Br J Philos Sci 45:941–953

Barrett M, Sober E (1995) When and why does entropy increase? In: Savitt S (ed) Time's arrow today. Cambridge University Press, Cambridge, UK, pp 230–258

Berger JO (1985) Statistical decision theory and Bayesian analysis, 2nd edn. Springer Series in Statistics, Springer-Verlag, Berlin

Brooks D, Wiley E (1988) Evolution as entropy. University of Chicago Press, Chicago

Chang J (1996) Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. Math Biosci 134:189–215

Cover TM, Thomas JA (1991) Elements of information theory. Wiley, New York

Crow J, Kimura M (1970) An introduction to population genetics theory. Burgess Publishing Company, Minneapolis

Durrett R (2002) Probability models for DNA sequence evolution. Springer-Verlag, New York

Evans W, Kenyon C, Peres Y, Schulman LJ (2000) Broadcasting on trees and the Ising model. Adv Appl Probab 10:403–433

Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, Sunderland, MA

Fischer M, Thatte B (2009) Maximum parsimony on subsets of taxa. J Theor Biol 260:290–293

Grimmett G, Stirzaker D (2001) Probability and random processes, 3rd edn. Oxford University Press, Oxford, UK

Häggström O (2002) Finite Markov chains and algorithmic applications. Cambridge University Press, Cambridge, UK

Kemeny JG, Snell JL (1976) Finite Markov chains. Springer-Verlag, New York

Mossel E (1998) Recursive reconstruction on periodic trees. Random Struct Algorithm 13(1):81–97

Mossel E (2003) On the impossibility of reconstructing ancestral data and phylogenies. J Comput Biol 10:669–678

Mossel E, Steel M (2005) How much can evolved characters tell us about the tree that generated them? In: Gascuel O (ed) Mathematics of evolution and phylogeny. Oxford University Press, Oxford, pp 384–412

Pinelis I (2003) Evolutionary models of phylogenetic trees. Proceedings of the Royal Society B 270(1522):1425–1431

Rozanov YA (1969) Probability theory: a concise course. Dover Publications Inc, New York

Semple C, Steel M (2003) Phylogenetics. Oxford University Press, Oxford, UK

Seneta E (1973) Non-negative matrices: an introduction to theory and applications. Wiley, New York, pp 52–54

Sober E (1989) Independent evidence about a common cause. Philos Sci 56:275–287

Sober E (2008) Evidence and evolution: the logic behind the science. Cambridge University Press, Cambridge, UK

Sober E, Barrett M (1992) Conjunctive forks and temporally asymmetric inference. Aust J Philos 70:1–23

Sober E, Steel M (2002) Testing the hypothesis of common ancestry. J Theor Biol 218:395–408

Tuffley C, Steel MA (1997a) Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull Math Biol 59(3):581–607

Tuffley C, Steel MA (1997b) Modeling the covarion hypothesis of nucleotide substitution. Math Biosci 147:63–91

Weber B, Depew D (1988) Entropy, information, and evolution: new perspectives on physical and biological evolution. MIT Press, Cambridge

Yockey H (2005) Information theory, evolution, and the origin of life. Cambridge University Press, Cambridge