J. Appl. Prob. **46,** 601–607 (2009)
Printed in England
© *Applied Probability Trust* 2009

A STOCHASTIC MODEL FOR PHYLOGENETIC TREES

THOMAS M. LIGGETT,* *University of California, Los Angeles*
RINALDO B. SCHINAZI,** *University of Colorado*

Abstract

We propose the following simple stochastic model for phylogenetic trees. New types are born and die according to a birth and death chain. At each birth we associate a fitness to the new type sampled from a fixed distribution. At each death the type with the smallest fitness is killed. We show that if the birth (i.e. mutation) rate is subcritical, we obtain a phylogenetic tree consistent with an influenza tree (few types at any given time and one dominating type lasting a long time). When the birth rate is supercritical, we obtain a phylogenetic tree consistent with an HIV tree (many types at any given time, none lasting very long).

Keywords: Phylogenetic tree; influenza; HIV; stochastic model

2000 Mathematics Subject Classification: Primary 60K35

1. Introduction

The influenza phylogenetic tree is peculiar in that it is very skinny: one type dominates for a long time and any other type that arises quickly dies out. Then the dominating type suddenly dies out and is immediately replaced by a new dominating type. The models proposed so far are very complex and make many assumptions. See, for instance, Koelle *et al.* (2006) and van Nimwegen (2006). We use a simple stochastic model for such a tree. The other motivation for this work comes from the comparison between influenza and HIV phylogenetic trees. An HIV tree is characterized by a radial spread outward from an ancestral node, in sharp contrast with an influenza tree. Moreover, Korber *et al.* (2001) noted that the influenza virus is less diverse worldwide than the HIV virus is in Amsterdam alone. However, both types of tree are supposed to be produced by the same basic mechanism: mutations. Can the same mathematical model produce two trees that are so different? Our simple stochastic model will show a striking difference in behavior depending on the mutation rate.

Our model has a birth-and-death component and a fitness component. For the birth and death components, we do the following. If there are $n \geq 1$ types at a certain time t then there is birth of a new type (by mutation) at rate $n\lambda$. We think of a birth as the appearance of one new type, not the replacement of one type by two new types. If there are $n \geq 2$ types then there is death of one type at rate n. If only one type is left, it cannot die. That is,

$$n \to n+1 \quad \text{at rate } n\lambda, \qquad n \to n-1 \quad \text{at rate } n \text{ if } n \geq 2.$$

Received 9 October 2008; revision received 6 February 2009.
* Postal address: Department of Mathematics, University of California, Los Angeles, CA 90095-1555, USA.
Partially supported by NSF grant DMS-0301795.
** Postal address: Department of Mathematics, University of Colorado, Colorado Springs, CO 80933-7150, USA.
Email address: schinazi@math.uccs.edu
Partially supported by NSF grant DMS-0701396.

https://doi.org/10.1239/jap/1245676110 Published online by Cambridge University Press

Moreover, each new individual is assigned a fitness value chosen from a fixed distribution, independently each time. Every time there is a death event then the type that is killed is the one with the smallest fitness. Since all that matters are the ranks of the fitness, we might as well take their distribution to be uniform on [0, 1]. For simplicity, the process is started with a single type.

We give no specific rule on how to attach a new type after a birth to existing types (in order to construct a tree). Our results do not depend on such a rule. Two natural possibilities are to either attach the new type to the type which has the maximum fitness or to a type taken at random.

Theorem 1. *Take* $\alpha \in (0, 1)$. *If* $\lambda \leq 1$ *then*

$$\lim_{t\to\infty} \mathrm{P}(\textit{maximal types at times } \alpha t \textit{ and } t \textit{ are the same}) = \alpha,$$

while if $\lambda > 1$ *then this limit is* 0.

We see that if $\lambda < 1$, the dominating type (i.e. the fittest type) at time t has likely been present for a time of order t and at any given time there will not be many types. This is consistent with the observed structure of an influenza tree. On the other hand, if $\lambda > 1$ then the dominating type at time t has likely been present for a time of order shorter than t and at any given time there will be many types. This is consistent with an HIV tree.

2. Proof of Theorem 1

The proof is divided into three cases, depending on whether the birth-and-death chain is positive recurrent, null recurrent, or transient. We present them in order of difficulty.

2.1. $\lambda < 1$ case

Let $\tau_1, \tau_2, \ldots$ be the (continuous) times between successive visits of the chain to 1, let $T_n = \tau_1 + \cdots + \tau_n$, let $\sigma_1, \sigma_2, \ldots$ be the number of new types introduced in cycles between successive visits to 1, and let $S_n = 1 + \sigma_1 + \cdots + \sigma_n$. Note that the τs and σs are not independent of each other, but the sequence $(\tau_1, \sigma_1), (\tau_2, \sigma_2), \ldots$ is independent and identically distributed (i.i.d.) and independent of the fitness sequence. Define the usual renewal process $N(t)$ corresponding to the τs by $\{N(t) = n\} = \{T_n \leq t < T_{n+1}\}$.

For $0 < s < t$, recalling that $T_{N(t)} \leq t < T_{N(t)+1}$, and noting that the maximal type is increasing in time, we see that

$$\mathrm{P}(\text{maximal types at times } s \text{ and } t \text{ are the same}, N(s) < N(t)) \tag{1}$$

lies between

$$\mathrm{P}(\text{maximal types at times } T_{N(s)} \text{ and } T_{N(t)+1} \text{ are the same}, N(s) < N(t))$$

and

$$\mathrm{P}(\text{maximal types at times } T_{N(s)+1} \text{ and } T_{N(t)} \text{ are the same}, N(s) < N(t)).$$

Let $\mathcal{F}$ be the σ-algebra generated by $(\tau_1, \sigma_1), (\tau_2, \sigma_2), \ldots$. Then, for $k \leq l$, since the fitness sequence is i.i.d. and independent of $\mathcal{F}$,

$$\mathrm{P}(\text{maximal types at times } T_k \text{ and } T_l \text{ are the same} \mid \mathcal{F}) = \frac{S_k}{S_l}.$$

https://doi.org/10.1239/jap/1245676110 Published online by Cambridge University Press

More precisely, conditional on $\mathcal{F}$, there are S_l fitnesses observed by time T_l and S_k of them are observed by time T_k. We claim that in n i.i.d. observations the probability that the largest occurs among the first m is m/n, since any one of the n is equally likely to be the largest. Since $N(s)$ and $N(t)$ are $\mathcal{F}$ measurable, it follows that (1) lies between

$$\mathrm{E}\left[\frac{S_{N(s)}}{S_{N(t)+1}},\ N(s) < N(t)\right] \quad \text{and} \quad \mathrm{E}\left[\frac{S_{N(s)+1}}{S_{N(t)}},\ N(s) < N(t)\right]. \tag{2}$$

Since $\lambda < 1$, $\mathrm{E}[\tau] < \infty$, and the renewal theorem gives

$$\frac{N(s)}{s} \to \frac{1}{\mathrm{E}[\tau]} \quad \text{almost surely (a.s.)},$$

while the strong law of large numbers gives $S_{N(s)}/N(s) \to \mathrm{E}[\sigma]$ a.s., so that $S_{N(s)}/s \to \mathrm{E}[\sigma]/\mathrm{E}[\tau]$ a.s. It follows by the bounded convergence theorem that

$$\lim_{t\to\infty} \mathrm{P}(\text{maximal types at times } \alpha t \text{ and } t \text{ are the same}) = \alpha. \tag{3}$$

This completes the proof of Theorem 1 in the subcritical case.

2.2. $\lambda > 1$ case

Define the τs and σs as above, except that now the cycles used are between the successive times the chain reaches a new high. In other words, T_n is the hitting time of $n+1$, σ_n is the number of new types born during a first passage cycle from n to $n+1$, and S_n is the number of new types seen up to time T_n. Of course, the σs and τs are no longer identically distributed. However, $(\tau_1, \sigma_1), (\tau_2, \sigma_2), \ldots$ are independent. The key to the proof is the following lemma.

Lemma 1. *Assume that $\lambda > 1$. Then $\mathrm{e}^{-(\lambda-1)t} N(t)$ is a.s. bounded.*

In order to prove Lemma 1, we will need the following recursion formula, which is easy to prove by induction.

Lemma 2. *Let a_n and b_n be two sequences of real numbers such that $a_1 = \lambda^{-1} b_1$ and, for $n \geq 2$, $\lambda a_n = b_n + a_{n-1}$. Then,*

$$a_n = \sum_{j=1}^{n} \lambda^{-j} b_{n+1-j}.$$

Proof of Lemma 1. Our first step in this proof is to estimate the first two moments of τ_n. Following Keilson (1979, Equation (5.1.2)) we note that τ_n has the same distribution as

$$\frac{X}{(1+\lambda)n} + Y(\tau_{n-1} + \tau_n') \quad \text{for } n \geq 2, \tag{4}$$

where X has a mean 1 exponential distribution, τ_n' has the same distribution as τ_n, Y is a Bernoulli random with $\mathrm{P}(Y = 1) = 1/(\lambda+1)$, and X, Y, τ_{n-1}, and τ_n' are independent. Letting $\mu_n = \mathrm{E}[\tau_n]$, it follows from (4) that

$$\lambda\mu_n = \frac{1}{n} + \mu_{n-1} \quad \text{for } n \geq 2 \quad \text{and} \quad \mu_1 = \frac{1}{\lambda}. \tag{5}$$

https://doi.org/10.1239/jap/1245676110 Published online by Cambridge University Press

Applying Lemma 2 to (5) we obtain

$$\mu_n = \sum_{j=1}^{n} \frac{\lambda^{-j}}{n+1-j}.$$

Writing $1/(\lambda - 1)$ as a geometric series we have

$$\mu_n - \frac{1}{n(\lambda-1)} = \sum_{j=1}^{n} \lambda^{-j}\left(\frac{1}{n+1-j} - \frac{1}{n}\right) - \frac{1}{n}\sum_{j=n+1}^{\infty} \lambda^{-j}.$$

Changing the order of summation gives

$$\sum_{n=1}^{\infty}\sum_{j=1}^{n} \lambda^{-j}\left(\frac{1}{n+1-j} - \frac{1}{n}\right) = \sum_{j=1}^{\infty} \lambda^{-j} \sum_{n=j}^{\infty}\left(\frac{1}{n+1-j} - \frac{1}{n}\right).$$

Note that, for $j \geq 2$,

$$\sum_{n=j}^{\infty}\left(\frac{1}{n+1-j} - \frac{1}{n}\right) = \sum_{k=1}^{j-1} \frac{1}{k},$$

and this term is 0 for $j = 1$. Hence,

$$\sum_{n=1}^{\infty} \frac{1}{n} \sum_{j=n+1}^{\infty} \lambda^{-j} = \sum_{j=2}^{\infty} \lambda^{-j} \sum_{k=1}^{j-1} \frac{1}{k}.$$

We conclude that

$$\sum_{n=1}^{\infty} \left|\mu_n - \frac{1}{n(\lambda-1)}\right| < \infty$$

and

$$\sum_{n=1}^{\infty} \left(\mu_n - \frac{1}{n(\lambda-1)}\right) = 0.$$

Therefore,

$$\mathrm{E}[T_n] - \frac{1}{\lambda-1}\sum_{k=1}^{n} \frac{1}{k} \quad \text{converges to 0.} \tag{6}$$

We also need an almost-sure result for T_n, and for this, we will estimate the second moment of τ_n. Let $v_n = \mathrm{var}(\tau_n)$. It is easy to check that if Y is a Bernoulli random variable and is independent of a random variable Z, then

$$\mathrm{var}(ZY) = \mathrm{E}[Y]\,\mathrm{var}(Z) + \mathrm{var}(Y)(\mathrm{E}[Z])^2.$$

Using this remark and (4), we have, for $n \geq 2$,

$$v_n = \frac{1}{(1+\lambda)^2 n^2} + \frac{1}{1+\lambda}(v_n + v_{n-1}) + \frac{\lambda}{(1+\lambda)^2}(\mu_n + \mu_{n-1})^2.$$

Therefore, for $n \geq 2$,

$$\lambda v_n = b_n + v_{n-1}, \tag{7}$$

https://doi.org/10.1239/jap/1245676110 Published online by Cambridge University Press

where

$$b_n = \frac{1}{(1+\lambda)n^2} + \frac{\lambda}{1+\lambda}(\mu_n + \mu_{n-1})^2.$$

Set $\mu_0 = 0$. Then $\lambda v_1 = b_1$. Hence, Lemma 2 applies to (7), giving

$$v_n = \sum_{j=1}^{n} \lambda^{-j} b_{n+1-j}. \tag{8}$$

Since $\mu_n \sim 1/n(\lambda - 1)$ (that is, the ratio converges to 1), $b_n \sim C/n^2$, where C depends on λ only. From (8) we obtain

$$v_n \sim C' b_n \sim \frac{CC'}{n^2},$$

where C' depends on λ only. This implies the almost-sure convergence of the random series $\sum(\tau_n - \mathrm{E}[\tau_n])$ (see, for instance, Corollary 47.3 of Port (1994)). Therefore, the partial sums converge a.s. and

$$T_n - \mathrm{E}[T_n] \quad \text{converges a.s.}$$

Using (6), we find that $T_n - \log n/(\lambda - 1)$ converges a.s. and is therefore a.s. bounded. Now use the fact that $\{N(t) \geq n\} = \{T_n \leq t\}$ to conclude that $N(t)\exp(-(\lambda - 1)t)$ is a.s. bounded. This concludes the proof of Lemma 1.

We are now ready to complete the proof of Theorem 1 in the supercritical case. Let $(Z_i)_{i\geq 1}$ be a discrete-time random walk starting at 0 that goes to the right with probability $\lambda/(\lambda + 1)$ and to the left with probability $1/(\lambda + 1)$. For every $n \geq 1$, let $Z_{i,n}$ be a discrete-time random walk starting at 0 with the same rules of evolution as Z_i except that the random walk $Z_{i,n}$ has a reflecting barrier at $-n + 1$. For every $n \geq 1$, the two random walks Z_i and $Z_{i,n}$ are coupled so that they move together until (if ever) they hit $-n + 1$, and thereafter we still couple them so that $Z_i \leq Z_{i,n}$ for every $i \geq 0$. Let U and U_n be the hitting times of 1 for the random walks Z_i and $Z_{i,n}$, respectively.

First note that a new type appears every time there is a birth. Therefore, σ_n is the number of steps to the right of the random walk $Z_{i,n}$ stopped at 1. That is, σ_n is $(1 + U_n)/2$. We now show that U_n converges a.s. to U. Let $\delta > 0$. We have

$$\mathrm{P}(|U_n - U| > \delta) \leq \mathrm{P}(U > U_n) \leq \mathrm{P}(Z_i = -n + 1 \text{ for some } i \geq 1).$$

The last probability decays exponentially with n. Therefore,

$$\sum_{n\geq 1} \mathrm{P}(|U_n - U| > \delta) < \infty.$$

An easy application of the Borel–Cantelli lemma implies that U_n converges a.s. to U. Since $U_n \leq U$, the dominated convergence theorem implies that, for every $k \geq 1$, the kth moment of σ_n converges to the kth moment of $(1 + U)/2$. In particular, $\mathrm{var}(\sigma_n)$ is a bounded sequence. This is enough to prove that

$$\frac{1}{n}\sum_{i=1}^{n}(\sigma_i - \mathrm{E}[\sigma_i]) \quad \text{converges a.s. to 0;}$$

see, for instance, Proposition 47.10 of Port (1994). Since $\mathrm{E}[\sigma_n]$ is a convergent sequence, we find that S_n/n converges a.s. to the limit of $\mathrm{E}[\sigma_n]$.

https://doi.org/10.1239/jap/1245676110 Published online by Cambridge University Press

Since $N(t) \to \infty$ a.s., this strong law of large numbers shows that $S_{N(t)}/N(t)$ converges to the limiting expectation of σ_n. This together with Lemma 1 shows that the two terms in (2) converge to 0 when we let $s = \alpha t$ and t go to ∞. The proof of Theorem 1 in the supercritical case is complete.

2.3. $\lambda = 1$ case

In this subsection we go back to the notation of Subsection 2.1 where T_n is the time of the nth visit of the chain to 1.

Lemma 3. *Let* $\lambda = 1$. *Then,*

$$\frac{T_n}{n \log n} \to 1 \quad \textit{in probability.}$$

Proof. When the chain hits 1, it waits a mean 1 exponential time and then jumps to 2. Hence,

$$T_n = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} H_i,$$

where the X_i are independent mean 1 exponential times and H_i are the hitting times of 1 starting at 2. The H_i are i.i.d. with distribution function F. From the backward Kolmogorov equation,

$$\int_0^{F(t)} \frac{\mathrm{d}s}{1 + s^2 - 2s} = t,$$

we obtain

$$F(t) = \frac{t}{1+t}.$$

We now use a weak law of large numbers; see Theorem 2 of Feller (1971, Section VII.7). It is easier to redo the short proof rather than check the hypotheses of the theorem. The key is the following consequence of Chebyshev's inequality applied to the truncated random variables:

$$\mathrm{P}\left(\left|\frac{1}{nm_n}\sum_{i=1}^{n} H_i - 1\right| > \varepsilon\right) \le \frac{1}{n\varepsilon^2 m_n^2} s_n + n(1 - F(\rho_n)),$$

where

$$m_n = \int_0^{\rho_n} t F'(t)\,\mathrm{d}t \quad \text{and} \quad s_n = \int_0^{\rho_n} t^2 F'(t)\,\mathrm{d}t;$$

see (7.13) of Feller (1971, Section VII.7). We will take $\rho_n = n\sqrt{\log n}$. A little calculus shows that

$$m_n \sim \log \rho_n \sim \log n \quad \text{and} \quad s_n \sim \rho_n.$$

With our choice of ρ_n, $n(1 - F(\rho_n))$ converges to 0 and

$$\frac{1}{nm_n}\sum_{i=1}^{n} H_i$$

converges to 1 in probability. This completes the proof of Lemma 3.

https://doi.org/10.1239/jap/1245676110 Published online by Cambridge University Press

Since the events $N(t) \geq n$ and $T_n \leq t$ are the same, it follows that

$$N(t)\frac{\log t}{t} \to 1 \tag{9}$$

in probability as $t \uparrow \infty$.

Now, S_n/n^2 converges in distribution to a one-sided stable law of index $\frac{1}{2}$ (see Theorem (7.7) of Durrett (2004)). By (9), it follows that $S_{N(t)}/N(t)^2$ also has this distributional limit. In fact,

$$\frac{S_{N(\alpha t)}}{N^2(t)} \quad \text{converges to } Y_\alpha$$

in the sense of convergence of finite-dimensional distributions, where Y_α is a stable subordinator (increasing stable process) of index $\frac{1}{2}$. (Note that independence between the σs and τs is not required here, which is good since they are highly dependent. All that is needed is that the limit in (9) is constant and that both S_n and $N(t)$ are monotone.) So, the limit in (3) is

$$\lim_{t\to\infty} \mathrm{E}\left[\frac{S_{N(\alpha t)}}{S_{N(t)}}\right] = \mathrm{E}\left[\frac{Y_\alpha}{Y_1}\right] = \alpha.$$

To check the final equality, it is enough to verify by monotonicity for rational α. If $\alpha = m/n$, this boils down to the simple fact that if $V_1, \ldots, V_n$ are i.i.d. and positive, then

$$\mathrm{E}\left[\frac{V_i}{V_1 + \cdots + V_n}\right] = \frac{1}{n}.$$

Acknowledgement

We thank an anonymous referee whose remarks helped improve the presentation of the paper and simplify several of our proofs.

References

Durrett, R. (2004). *Probability: Theory and Examples*, 3rd edn. Duxbury press, Belmont, CA.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, Vol. 2, 2nd edn. John Wiley, New York.

Keilson, J. (1979). *Markov Chain Models—Rarity and Exponentiality* (Appl. Math. Sci. **28**). Springer, New York.

Koelle, K., Cobey, S., Grenfell, B. and Pascual, M. (2006). Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. *Science* **314,** 1898–1903.

Korber, B. *et al.* (2001). Evolutionary and immunological implications of contemporary HIV-1 variation. *British Med. Bull.* **58,** 19–42.

Port, S. C. (1994). *Theoretical Probability for Applications*. John Wiley, New York.

Van Nimwegen, E. (2006). Influenza escapes immunity along neutral networks. *Science* **314,** 1884–1886.

https://doi.org/10.1239/jap/1245676110 Published online by Cambridge University Press