Applied Mathematics Letters 23 (2010) 975–979

Contents lists available at ScienceDirect

Applied Mathematics Letters

journal homepage: www.elsevier.com/locate/aml

Logarithmic bounds on the posterior divergence time of two sequences

Radu Mihaescu [a], Mike Steel [b,*]

[a] *UC Berkeley, Department of Computer Science, United States*

[b] *University of Canterbury, Allan Wilson Center for Molecular Ecology and Evolution, New Zealand*

ARTICLE INFO

Article history:
Received 14 September 2009
Received in revised form 16 April 2010
Accepted 20 April 2010

Keywords:
Bayesian estimation
Binary symmetric channel
Sequences

ABSTRACT

When two initially identical binary sequences undergo independent site mutations at a constant rate, the proportion of site differences is often used to estimate the total time T that separates the two sequences. In this short note we study the posterior distribution of T when the prior distribution on T is exponential. We show that posterior estimates of T (for any data) cannot grow faster than the logarithm of the sequence length, and this rate is achieved for data generated at site saturation (i.e. in the limit as $T \to \infty$). The problem is motivated by information-theoretic questions arising in molecular systematic biology, in which one wishes to use DNA sequences to estimate the divergence time between present-day species.

1. Introduction

Consider the following problem: Long ago there was some (unknown) binary sequence of length k, and each position (site) in the sequence was independently subjected to two independent and identical 2-state symmetric Markov processes for a random time $T/2$ (T large) resulting in two derived sequences of length k that we observe today. Suppose we count the number M of positions where the two observed sequences are in different states (the 'Hamming' or 'mismatch' distance). On the basis of M we wish to estimate T, perhaps as a posterior given some prior distribution on this random variable.

This problem is a special case of a problem arising in molecular systematics, in which one has two present-day DNA sequences that have evolved from some common ancestral sequence that we cannot observe, and we wish to estimate how long ago the two sequences diverged [1]. Although DNA sequences involve sequences of four states (A, C, G, T), these four states are sometimes combined in pairs to form two states (purines and pyrimidines). The use of a process on just two states helps us to simplify the calculations that follow, though we expect similar results to hold for a fully symmetric 4-state process (the 'Jukes–Cantor' model).

Although the problem as stated seems to involve three sequences (the ancestral and two observed sequences) the reversibility of this particular Markov processes ensures that the problem is identical to the following problem, involving just one sequence. We have k coins, initially all heads up. Each coin is subjected independently to the same 2-state symmetric Markov process for time T and we count the number M of tails. We wish to use k to estimate the posterior distribution of T.

The maximum likelihood (ML) estimate $\hat{T}$ of T has a simple, and well-known form:

$$\hat{T} = -\frac{1}{2}\log\left(1 - 2M/k\right),$$

as can be verified by selecting the value of T to maximize the likelihood function $p_T^M(1-p_T)^{k-M}$, where $p_T = \frac{1}{2}(1 - e^{-2T})$ is the probability of observing a tail on any given coin after time T.

[*] Corresponding author.
E-mail address: mathmomike@gmail.com (M. Steel).

0893-9659/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aml.2010.04.020

Notice that it is entirely possible that $\hat{T}$ is undefined (if the term inside the logarithm is negative), but if it is defined then the largest value $\hat{T}$ can possibly take, due to the fact that M is an integer less than $k/2$, is: $\hat{T}_{\max} = \frac{1}{2}\log(k)$, which occurs when k is odd and $M = (k-1)/2$ (for related observations, see [2]). Thus if T is very large, ML will either give an undefined estimate or it will be 'small' (of order $\log(k)$). In this paper we ask whether Bayesian methods can do any better, given that they are not based on taking logarithms of quantities that can be undefined. We show that essentially the same $\frac{1}{2}\log(k)$ upper bound applies for an exponential prior on T.

More precisely we show that under an exponential prior Ψ with shape parameter $\psi > 0$, the posterior probability that T exceeds $(\frac{1}{2}+\epsilon)\log(k)$ converges to 0 as k grows for *any data*. This, in turn, implies that the expected posterior value of T grows no faster than $\frac{1}{2}\log(k)$. We then show a matching order $\log(k)$ lower bound on posterior estimates of T when the true value of T is infinity, again assuming an exponential prior on T.

The use of the exponential prior is motivated by the use of this distribution as a prior on branch lengths in phylogenetics [3]—thus we wish to consider how long sequences might need to be in order to detect long time-scales. In this short note we only consider two sequences, but in future work it would be interesting to derive more general results.

2. Bounds on posterior estimates of T for any data

Theorem 1. *For an exponential prior on T, and for all $c > 1$ we have:*

$$\mathbb{P}\left[T > \frac{c}{2}\log k | D_k\right] \to 0$$

uniformly for all binary strings D_k of length k.

Proof. Let $M = M(D)$ denote the number of mismatches in the data D (i.e. the number of positions where the two sequences take different values) and for any $\sigma > 0$ let

$$f_\sigma(m) := \mathbb{P}[\mathrm{e}^{-2T} < \sigma | M = m].$$

We first show that for smooth prior on T, $f_\sigma(m)$ is monotone increasing in m. To see this, first note that:

$$f_\sigma(m) = \frac{\int_0^\sigma (1-x)^m(1+x)^{k-m}\mathrm{d}p(x)}{\int_0^1 (1-x)^m(1+x)^{k-m}\mathrm{d}p(x)},$$

where $x = \mathrm{e}^{-2T}$, and where $p(x)$ denotes the distribution on x inherited from that on T. Now, set $\lambda(x) := \frac{1-x}{1+x}$ and $C := \left(\int_0^1 \lambda(x)^m(1+x)^k\mathrm{d}p(x)\right)^{-2}$, and observe that the conditions of the Leibniz integral rule allow us to differentiate under the integral sign in the numerator and denominator of $f_\sigma(m)$ as follows.

$$\begin{aligned}
\frac{\partial}{\partial m}f_\sigma(m) &= \frac{\partial}{\partial m}\left(\frac{\int_0^\sigma \lambda(x)^m(1+x)^k\mathrm{d}p(x)}{\int_0^1 \lambda(x)^m(1+x)^k\mathrm{d}p(x)}\right)\\
&= C\int_0^\sigma \lambda(x)^m\log(\lambda(x))(1+x)^k\mathrm{d}p(x)\int_0^1 \lambda(x)^m(1+x)^k\mathrm{d}p(x)\\
&\quad - C\int_0^\sigma \lambda(x)^m(1+x)^k\mathrm{d}p(x)\int_0^1 \lambda(x)^m\log(\lambda(x))(1+x)^k\mathrm{d}p(x)\\
&= C\int_0^\sigma \lambda(x)^m\log(\lambda(x))(1+x)^k\mathrm{d}p(x)\int_\sigma^1 \lambda(x)^m(1+x)^k\mathrm{d}p(x)\\
&\quad - C\int_0^\sigma \lambda(x)^m(1+x)^k\mathrm{d}p(x)\int_\sigma^1 \lambda(x)^m\log(\lambda(x))(1+x)^k\mathrm{d}p(x)\\
&\geq C\log(\lambda(\sigma))\int_0^\sigma \lambda(x)^m(1+x)^k\mathrm{d}p(x)\int_\sigma^1 \lambda(x)^m(1+x)^k\mathrm{d}p(x)\\
&\quad - C\log(\lambda(\sigma))\int_0^\sigma \lambda(x)^m(1+x)^k\mathrm{d}p(x)\int_\sigma^1 \lambda(x)^m(1+x)^k\mathrm{d}p(x)\\
&= 0,
\end{aligned}$$

thereby establishing the claim that $f_\sigma(m)$ is monotone increasing in m. Consequently, it suffices to prove Theorem 1 in the case when $M(D_k) = k$ (i.e. a mismatch occurs at all sites). Now, for an exponential distribution with shape parameter $\psi > 0$:

$$\mathbb{P}\left[T \geq \frac{c}{2}\log k | M = k\right] = \frac{\int_{\frac{c}{2}\log k}^\infty (1-\mathrm{e}^{-2t})^k\mathrm{e}^{-\psi t}\mathrm{d}t}{\int_0^\infty (1-\mathrm{e}^{-2t})^k\mathrm{e}^{-\psi t}\mathrm{d}t}.$$

Substituting $x = \mathrm{e}^{-2t}$ gives:

$$\begin{aligned}
\mathbb{P}\left[T \geq \frac{c}{2}\log k | M = k\right] &= \frac{\int_0^{k^{-c}} (1-x)^k x^{\psi/2-1}\mathrm{d}x}{\int_0^1 (1-x)^k x^{\psi/2-1}\mathrm{d}x} \\
&= \frac{\frac{2}{\psi} x^{\psi/2}(1-x)^k \,|_0^{k^{-c}} + \frac{2}{\psi}\int_0^{k^{-c}} x^{\psi/2} k(1-x)^{k-1}\mathrm{d}x}{\frac{2}{\psi} x^{\psi/2}(1-x)^k \,|_0^1 + \frac{2}{\psi}\int_0^1 x^{\psi/2} k(1-x)^{k-1}\mathrm{d}x} \\
&= \frac{k^{-c\psi/2}(1-k^{-c})^k + \int_0^{k^{-c}} x^{\psi/2} k(1-x)^{k-1}\mathrm{d}x}{\int_0^1 x^{\psi/2} k(1-x)^{k-1}\mathrm{d}x}.
\end{aligned}$$

Since $c > 1$, we have $(1-k^{-c})^k \to 1$ as $k \to \infty$. Breaking up the above sum, we obtain:

$$\begin{aligned}
\frac{k^{-c/2\psi}(1-k^{-c})^k}{\int_0^1 x^{\psi/2} k(1-x)^{k-1}\mathrm{d}x} &< \frac{k^{-c\psi/2}}{\int_0^1 x^{\psi/2} k(1-x)^{k-1}\mathrm{d}x} \\
&\leq \frac{k^{-c\psi/2}}{\int_{1/k}^1 x^{\psi/2} k(1-x)^{k-1}\mathrm{d}x} \\
&\leq \frac{k^{-c\psi/2}}{k^{-\psi/2}\int_{1/k}^1 k(1-x)^{k-1}\mathrm{d}x} \\
&= \frac{k^{-c\psi/2}}{k^{-\psi/2}(1-1/k)^k} \\
&\leq 4(k^{(1-c)\psi/2}) = o(1);
\end{aligned}$$

(where the last inequality uses $(1-1/k)^k \geq 1/4$ for $k > 1$) and

$$\begin{aligned}
\frac{\int_0^{k^{-c}} x^{\psi/2} k(1-x)^{k-1}\mathrm{d}x}{\int_0^1 x^{\psi/2} k(1-x)^{k-1}\mathrm{d}x} &\leq \frac{\int_0^{k^{-c}} x^{\psi/2} k(1-x)^{k-1}\mathrm{d}x}{\int_{k^{-c}}^1 x^{\psi/2} k(1-x)^{k-1}\mathrm{d}x} \\
&\leq \frac{k^{-c\psi/2}\int_0^{k^{-c}} k(1-x)^{k-1}\mathrm{d}x}{k^{-c\psi/2}\int_{k^{-c}}^1 k(1-x)^{k-1}\mathrm{d}x} \\
&= \frac{1-(1-k^{-c})^k}{1} = o(1).
\end{aligned}$$

Thus $\lim_{k\to\infty} \mathbb{P}[T \geq \frac{c}{2}\log k | D_k] = 0$, where the convergence is uniform over all possible data-sets D_k. □

Corollary 2. *Under an exponential prior with shape parameter ψ, and given $c > 1$, the following inequality holds uniformly over all possible data D_k (thus irrespective of the "true" value of T that was involved in generating the data):*

$$\mathbb{E}[T|D_k] < \frac{c}{2}\log k + o(1).$$

Proof.

$$\begin{aligned}
\mathbb{E}[T|D_k] &= \int_0^\infty \mathbb{P}[T > t|D_k]\mathrm{d}t \\
&\leq \int_0^{\frac{c}{2}\log k} \mathbb{P}[T > t|D_k]\mathrm{d}t + \int_{\frac{c}{2}\log k}^\infty \mathbb{P}[T > t|D_k]\mathrm{d}t \\
&\leq \frac{c}{2}\log k + \int_{\frac{c}{2}\log k}^\infty \mathbb{P}[T > t|D_k]\mathrm{d}t.
\end{aligned}$$

It remains to show that the last term in the above inequality converges to 0. We have:

$$\begin{aligned}
\int_{\frac{c}{2}\log k}^\infty \mathbb{P}[T > t|D_k]\mathrm{d}t &< \int_{\frac{c}{2}\log k}^\infty \mathbb{P}[T > t|M(D_k) = k]\mathrm{d}t \\
&= \frac{\log k}{2}\int_c^\infty \mathbb{P}\left[T > \frac{c'}{2}\log k | M(D_k) = k\right]\mathrm{d}c'
\end{aligned}$$

$$= \frac{\log k}{2}\int_c^\infty \left[\frac{k^{-c'\psi/2}(1-k^{-c'})^k + \int_0^{k^{-c'}} x^{\psi/2}k(1-x)^{k-1}\mathrm{d}x}{\int_0^1 x^{\psi/2}k(1-x)^{k-1}\mathrm{d}x}\right]\mathrm{d}c'$$
$$< \frac{\log k}{2}\int_c^\infty \left[4(k^{(1-c')\psi/2}) + 1 - (1-k^{-c'})^k\right]\mathrm{d}c',$$

using the same bounds as in the proof of Theorem 1. For the first term of the integral, we have:

$$\frac{\log k}{2}\int_c^\infty \left[4(k^{(1-c')\psi/2})\right]\mathrm{d}c' = \frac{4}{\psi}k^{(1-c)\psi/2} = o(1).$$

Finally, substituting $y = k^{-c'}$ and $\mathrm{d}y = -y\log(k)\mathrm{d}c'$ in the second term and applying the inequality $1-(1-y)^k < ky$ gives:

$$\begin{aligned}\frac{\log k}{2}\int_c^\infty (1-(1-k^{-c'})^k)\mathrm{d}c' &= \frac{\log k}{2}\int_0^{k^{-c}} (1-(1-y)^k)\frac{\mathrm{d}y}{y\log k}\\ &\le \frac{1}{2}\int_0^{k^{-c}} (1-(1-y)^k)\frac{\mathrm{d}y}{y}\\ &\le \frac{1}{2}\int_0^{k^{-c}} ky\frac{\mathrm{d}y}{y}\\ &= \frac{1}{2}k^{1-c} = o(1). \quad \square\end{aligned}$$

3. Lower bounds for the case where $T = \infty$ (Random sequences)

In this section we are concerned with providing bounds on the posterior distribution of T, under an exponential prior Ψ with shape parameter $\psi > 0$, in the case when the data are i.i.d. samples from $\mathbb{P}_\infty$, the probability distribution on binary sequences that arises in the limit as $T \to \infty$ (the "saturated model") in which each site is a fair coin toss.

Theorem 3. *For all $\beta < \frac{1}{4}$, there exists a sequence E_k of sets of bit strings of length k such that, as k increases:*

- $\mathbb{P}_\infty[D_k \in E_k] \to 1$, *and*
- *for all $D_k \in E_k$, $\mathbb{P}[T < \frac{\beta}{2}\log k|D_k] \to 0$, where the convergence is uniform.*

Proof. Let $E_k := \{D \in \{0,1\}^k | M(D) > m_k\}$, where

$$m_k := \frac{k}{2}\left(1 - \frac{c_k}{\sqrt{k}}\right).$$

Letting $c_k \to \infty$ yields $\mathbb{P}_\infty[E_k] \to 1$ which establishes the first claim in Theorem 3.

Now let $\beta \in (0, 1/4)$ and pick $\epsilon > 0$ sufficiently small so that $\beta(2+\epsilon) < 1/2$, and let

$$t_k = \frac{\beta}{2}\log k.$$

Note that $\mathbb{E}[M(D_k)|T = (2+\epsilon)t_k] = \frac{k}{2}(1-k^{-(2+\epsilon)\beta}) \le \frac{k}{2}(1-\frac{c_k}{\sqrt{k}}) = m_k$. We will establish the following inequality:

$$\text{For all } m \ge m_k, \mathbb{P}[T < t_k|M(D) = m] \le \frac{\mathbb{P}[M = m|T = t_k]}{\mathbb{P}[M = m|T = 2t_k]}\psi(\mathrm{e}^{-2\psi t_k} - \mathrm{e}^{-(2+\epsilon)\psi t_k})^{-1}. \quad (1)$$

To establish (1) we first show that for all $m \ge m_k$, the function $f_m(t) := \mathbb{P}[M(D_k) = m|T = t]$ is increasing in t for $t \in (0, (2+\epsilon)t_k)$. We have:

$$f_m(t) = \mathbb{P}[M(D_k) = m|T = t] = \frac{1}{2^k}(1+\mathrm{e}^{-2t})^{k-m}(1-\mathrm{e}^{-2t})^m,$$

and so

$$\frac{\partial f_m(t)}{\partial t}/f_m(t) = \frac{2m\mathrm{e}^{-2t}}{1-\mathrm{e}^{-2t}} - \frac{2(k-m)\mathrm{e}^{-2t}}{1+\mathrm{e}^{-2t}} \ge 0,$$

since

$$m \ge m_k = \frac{k}{2}(1 - c_k\mathrm{e}^{-(\log k)/2}) \ge \frac{k}{2}(1-\mathrm{e}^{-2(2+\epsilon)t_k}) \ge \frac{k}{2}(1-\mathrm{e}^{-2t}).$$

This establishes our claim regarding the increasing property of $f_m(t)$, and we now use this to establish (1). We have

$$\mathbb{P}[T \leq t_k | M = m] = \frac{\mathbb{P}[M = m | T \leq t_k]\mathbb{P}[T \leq t_k]}{\mathbb{P}[M = m]} \leq \frac{\mathbb{P}[M = m | T \leq t_k]}{\mathbb{P}[M = m]},$$

and so, since $f_m(t)$ is increasing in t for $t \in (0, (2+\epsilon)t_k)$ and $m \geq m_k$, we have:

$$\begin{aligned}\mathbb{P}[T \leq t_k | M = m] &\leq \frac{\mathbb{P}[M = m | T = t_k]}{\mathbb{P}[M = m]} \\ &\leq \frac{\mathbb{P}[M = m | T = t_k]}{\mathbb{P}[M = m | T \in [2t_k, (2+\epsilon)t_k]]\mathbb{P}[T \in [2t_k, (2+\epsilon)t_k]]} \\ &\leq \frac{\mathbb{P}[M = m | T = t_k]}{\mathbb{P}[M = m | T = 2t_k]}\mathbb{P}[T \in [2t_k, (2+\epsilon)t_k]]^{-1},\end{aligned}$$

which establishes Inequality (1). Thus, it suffices for the proof of Theorem 3 to show that:

$$\lim_{k\to\infty} \frac{\mathbb{P}[M = m | T = t_k]}{\mathbb{P}[M = m | T = 2t_k]}(\mathrm{e}^{-2\psi t_k})^{-1} = 0, \tag{2}$$

since

$$\lim_{k\to\infty} \frac{\mathrm{e}^{-2\psi t_k} - \mathrm{e}^{-(2+\epsilon)\psi t_k}}{\mathrm{e}^{-2\psi t_k}} = 1.$$

Setting $x_k := \mathrm{e}^{-2t_k} = \mathrm{e}^{-\beta \log k} = k^{-\beta}$, (2) is equivalent to:

$$\lim_{k\to\infty} x_k^{-\psi} \frac{(1 - x_k)^{m_k}(1 + x_k)^{k-m_k}}{(1 - x_k^2)^{m_k}(1 + x_k^2)^{k-m_k}} = 0. \tag{3}$$

Substituting $m_k = \frac{k}{2}(1 - \frac{c_k}{\sqrt{k}})$ condition (3) becomes:

$$\lim_{k\to\infty} k^{\psi\beta} \frac{(1 + k^{-\beta})^{c_k\sqrt{k}}}{(1 + k^{-2\beta})^{k/2 - c\sqrt{k}}} = 0.$$

Taking logarithms, and using the inequality $x > \log(1 + x) > x - x^2/2$ for $x > 0$, this reduces further to:

$$\lim_{k\to\infty} \psi\beta \log(k) + k^{-\beta}c_k\sqrt{k} - (k^{-2\beta} - k^{-4\beta}/2)(k/2 - c\sqrt{k}) = -\infty.$$

Since $0 < \beta < 1/2$, the dominant term in the above expression is $-k^{1-2\beta}/2$, and thus the limit is indeed $-\infty$, so (3) follows, as required to complete the proof. □

Acknowledgements

The second author thanks Peter Lockhart whose question prompted this note, and the Cass Field Station where this research was carried out.

References

[1] J. Felsenstein, Inferring phylogenies. Sinauer Associates, Sunderland, Mass, 2004.

[2] M. Lacey, J. Chang, A signal-to-noise analysis of phylogeny estimation by neighbor-joining: insufficiency of polynomial length sequences, Math. Biosci. 199 (2006) 188–215.

[3] Z. Yang, B. Rannala, Branch-length prior influences Bayesian posterior probability of phylogeny, Syst. Biol. 54 (3) (2005) 455–470.