J. Appl. Prob. **36,** 621–631 (1999)
Printed in Israel
© *Applied Probability Trust* 1999

A NEW STOCHASTIC MODEL OF MICROSATELLITE EVOLUTION

RICHARD DURRETT,* *Cornell University*
SEMYON KRUGLYAK,** *University of Southern California*

Abstract

We introduce a continuous-time Markov chain model for the evolution of microsatellites, simple sequence repeats in DNA. We prove the existence of a unique stationary distribution for our model, and fit the model to data from approximately 10^6 base pairs of DNA from fruit flies, mice, and humans. The slippage rates from the best fit for our model are consistent with experimental findings.

Keywords: Microsatellite evolution; Markov chain model

AMS 1991 Subject Classification: Primary 73P05
Secondary 60J27

1. Introduction

Microsatellites are repeats of short patterns (2–6 base pairs) in DNA. These repeating patterns are unstable with respect to mutations in length, resulting in a high degree of variability. This high degree of variability and the fact that microsatellites occur with high frequency in the DNA of many organisms has made them popular for use as markers in constructing whole genome maps of humans and other organisms [5, 32], and in phylogenetic studies [10, 28]. Another reason for interest is that microsatellite loci play an important role in human genetic diseases such as fragile X syndrome, myotonic dystrophy and Huntington's disease [2]. In these cases the disorders result when the number of repeats exceeds a certain threshold. For example, Huntington's disease appears when the number of repeated CAGs at a certain locus is 36 or more.

The first thing to notice is that the occurrence of microsatellites is much more frequent than one would predict on the basis of chance. Data on primate repeat sequences from 6 994 799 nucleotides drawn from GenBank 84.0 reported on p. 415 of Bell and Jurka [3] shows 30 dinucleotide repeats consisting of 20 repeat units or more, including one of length 40. If we assume that the 4 possible bases A, C, G, and T of DNA occur with equal probability then the probability of a dinucleotide repeat of length 20 that begins with a specified pair of nucleotides is $4^{-38} = 1.32 \times 10^{-23}$. Since the sample consists of $\leq 7 \times 10^6$ nucleotides, the probability of a single occurrence of a dinucleotide repeat of length 20 by random chance is less than 10^{-16}. Thus, it is clear that these repeat sequences do not arise from random chance events.

There are two primary mechanisms that may lead to a change in the length (number of repeat units) of a microsatellite. The first is 'polymerase slippage'. Replicating strands of DNA may dissociate, and reassociate in a misaligned manner leading to an alteration of the number of

Received 12 February 1998; revision received 6 May 1998.

* Postal address: Department of Mathematics, 528 Mallott Hall, Cornell University, Ithaca, NY 14853, USA.
Email address: rtd1@cornell.edu
** Postal address: 293 Denney Research Building, University of Southern California, Los Angeles, CA 90089, USA.

repeat units [13]. In the microsatellites associated with diseases, unstable transmission can result in increases of ≥ 10 repeats in one generation [15]. Here, we will concentrate on the evolution of the presumably neutral markers used in genetic mapping. In these the slippage usually leads to changes of ± 1 repeat units, but greater changes are possible. For example, a study by Weber and Wong [31] found that 82% of mutation events observed in humans involved a gain or loss of a single repeat unit.

The stepwise mutation model of Ohta and Kimura [17] and Moran [16], originally formulated to model the charge state of proteins as inferred from electrophoretic mobility, has recently been much used as a model of the change in size of microsatellites [6–10, 12, 20, 24, 25, 30]. According to the basic stepwise mutation model, microsatellite length follows a simple random walk with a reflecting barrier at one repeat unit. One problem with the use of this model is that such a reflecting random walk on the line is null recurrent, so there is no stationary distribution for microsatellite lengths. However, this problem is easily solved by looking at the difference of the lengths in a sample of n individuals. The differences will have a well-defined equilibrium distribution since the genealogies of the sampled individuals coalesce to a single common ancestor following the usual rules for neutral genealogies [11, 29]. The technique of looking at differences can be used to determine the variance in lengths of a sample from a population. A genetic distance measure can then be derived by comparing variances in different species or populations [20]. An alternate solution to avoiding some of the problems of the stepwise mutation model is to introduce mutation bias into the evolution process [8]. If short alleles have a preponderance of length-increasing mutations, whereas long alleles tend to decrease in size, then an equilibrium length will be reached. Zhivotovsky *et al.* analyse this model and use it to derive a genetic distance measure [35].

Most users of the stepwise mutation model ignore a second important mechanism; a point mutation (insertion, deletion, or nucleotide substitution) will destroy the perfect nature of the repeat and cut a microsatellite into two smaller pieces. A notable exception is Bell and Jurka [3], who let their repeats of length n break into one of length $j-1$ and one of length $n-j$ at rate a. At the lower end of their state space they have an absorbing barrier at 1, but allow new repeats to be born at length 2 at a positive rate, c. At the upper end they impose an artificial cutoff at 30 by declaring that microsatellites were killed when they reached that length. At first one might worry that even in a bounded interval the branching might lead to an exponentially growing population of repeats. This fear can be quieted by noting that the total of the lengths of all the repeats is decreased by 1 when a split occurs, so an equilibrium will become established. Unfortunately, it seems to be very difficult to compute anything for this model except by simulation.

In this paper we introduce a simple continuous-time Markov chain model for the evolution of microsatellites in the DNA of various organisms, which shows that, as Schug *et al.* suggested [22], the equilibrium length of repeats is 'determined by a balance between the rate at which DNA slippage introduces new repeat units, and the rate at which point mutations or insertion/deletion mutations occur within the repeat itself'. To motivate the definition of our model, we fix our attention on a pair of nucleotides, and let N_t be the number of times that the pair is repeated in the sequence beginning with the chosen pair and reading to the right. When $N_t = \ell$ there are three types of transition that can occur:

- *polymerase slippage*: $\ell \to \ell + k$ at rate $r_{\ell,k}$
- *point mutations that destroy the perfect repeat*: for $1 \leq j < \ell$, $\ell \to j$ at rate a
- *substitutions that start a new repeat*: $1 \to 2$ at rate c.

In the stepwise mutation model $r_{l,1} = r_{l+1,-1} = r$ for all $l \geq 1$. However, experimental evidence clearly indicates that microsatellite mutation rates depend on their length [13, 21, 34]. We suppose instead that $r_{\ell,k}$ is of the form $(\ell-1)b_k$ (with minor exceptions to this rule when ℓ is small). In writing this formula, we imagine that slippage events of size k occur at a constant rate b_k in any of the $\ell-1$ positions between repeated units. The simplest choice, which we call the nearest neighbor case, has $b_1 = b_{-1} = b$ with the other $b_k = 0$. However, as mentioned above, one cannot ignore the possibility of slippage by more than 1, so we will consider general finite range models, i.e. $b_k = 0$ when $|k| > K$.

The choice of $r_{\ell,k}$ described in the previous paragraph leads to an absorbing state at 1. However, the third transition fixes this problem. Here we are thinking, for example, of a single CA becoming CACA through one or two substitutions. Note that since any substitution and any single insertion or deletion will cut a repeat, while only a judicious substitution or two will extend it, c will typically be 5–10 times smaller than a. While on the subject of sizes of parameters, we should note that estimates of b_1 given below are usually several hundred times as large as a.

Logically, the first thing we have to do is to show that our model has a stationary distribution. By Markov chain theory, it is sufficient to check that the expected time to return to 1 starting from 1 is finite. To do this, we can obviously throw away slippage down and consider the following example.

Example 1. The upper bound model: $r_{\ell,k} = (\ell-1)b_k$ for $1 \leq k \leq K$, and $r_{\ell,k} = 0$ otherwise.

Theorem 1. *Let $u = \sum_{k=1}^{K} kb_k$. In the upper bound model, and hence in any finite range model, there is a unique stationary distribution π with*

$$\sum_{\ell=1}^{\infty}(\ell-1)\pi(\ell) \leq \frac{2(c+u)}{a}.$$

Note that c is much smaller than u, so the first term can be ignored. If we do this in the case when $b_1 = b$ and $b_k = 0$ for $k > 1$, the upper bound reduces to $2b/a$.

2. Comparison with experimental results

Since the stationary distribution of lengths exists, it is natural to ask if its shape is similar to the distributions observed in nature. To test this, approximately 10^6 nucleotides of DNA from fruit flies, mice, and humans were collected for each organism from World Wide Web sites maintained by the Whitehead Institute (http://www-seq.wi.mit.edu) and the Lawrence Berkeley Laboratory (http://www.lbl.gov). We processed the data by examining each pair of nucleotides and then counting the number of times the pair was tandemly repeated, starting from the chosen nucleotides, and scanning to the right in the sequence. A microsatellite was defined as a sequence consisting of five or more repeat units. This definition is used for several reasons. We will be comparing the predictions of our model to experimental studies, and experimental studies tend to restrict their attention to repeats of length five or more repeat units. Another reason for this definition is that short repeats will often occur by random chance, and not due to the slippage/point mutation interaction. Finally, trying to fit the model to the frequently occurring short repeats makes the fit insensitive to the tail of the distribution.

Our method of processing the data, i.e. examining every pair and then counting to the right, has the property that a repeat of $n > 5$ units will generate repeats of lengths $n-1, \ldots, 5$. Readers who want to recover the usual laboratory viewpoint should note that if $p(n)$ is the frequency of microsatellites of the length n when each repeat is counted only once, and $\pi(n)$ is our stationary distribution then

$$\pi(n) = \sum_{m \geq n} p(m),$$

since each repeat of length exactly $m \geq n$ generates exactly one repeat of length n in our counting scheme. In words, the equilibrium distribution of our chain is the tail of the distribution function for the more natural counting scheme in which each repeat is counted only once. Once we estimate the distribution function $\pi(n)$ by fitting the model, the density function $p(m)$ can be obtained by taking $p(m) = \pi(m) - \pi(m+1)$, and rescaling.

Our first step in fitting our model to the data is to reduce the number of parameters. To eliminate c we note that it follows easily from the equations for the stationary distribution (see e.g. (1) below for the nearest neighbor case) that the value of c only controls the value of $\pi(1)$ relative to $(\pi(2), \pi(3), \dots)$, so if we restrict our attention, as we have, to repeats of length 5 or more, then the parameter c is not relevant. In other words, the actual value of c will cancel when we condition the stationary distribution on $n \geq 5$. To eliminate a we set $a = 2 \times 10^{-8}$ [14], a rough guess for the rate at which point mutations hit one of the two nucleotides in the repeat unit. We then vary $b_1 = b_{-1} = b$ and a multiplicative constant M to fit $M\pi(n)$, $n \geq 5$ to the data. We note here that the one true parameter of the model is the ratio, b/a. If we were to vary the value of a, the value of b would change proportionally.

Our estimate of b was chosen to be the slippage rate that minimized the sum of the absolute differences between the observed and expected length distributions. Using absolute differences rather than squared differences seemed to work better since the squared error criteria assigned too much weight to the first few values of $M\pi(i)$ and was insensitive to the tail. Figure 1 shows our stationary distribution fit to the data for humans, which is typical of the other two fits. There are two main reasons for only fitting the stationary distribution to microsatellites consisting of five or more repeat units. First, shorter repeats are typically not classified as microsatellites in the experimental setting. Since we are interested in comparing our results to experimental findings, we only count length 5 or greater microsatellites in the data, and condition the stationary distribution accordingly. Second, the number of times that a pair of nucleotides is repeated 1–4 times in a long sequence is very large. Attempting to fit these short repeats would obscure the fitting of the tail of the distribution. Table 1 shows the best fit slippage rates for dinucleotide repeats for organisms that we considered and compares with experimental estimates. The first column is the value of $2b$ (slippage up plus slippage down, assuming $a = 2 \times 10^{-8}$) that gave the best fit to the data. To compare our per nucleotide (per generation) slippage rates to the per locus rates quoted in the literature, we have to multiply the rate in the first column by $(l - 1)$, where l is the average microsatellite length in the experimental study. The result is given in the second column for comparison with the experimental results in the third.

It is comforting to note that the slippage parameter that led to the best fit was highest in mice, followed by humans and then fruit flies. There are several reasons for differences between the fitted values and the experimental results. Most importantly, our model cannot include all of the biological complexities (e.g. it ignores unequal crossing over [27]), and it can only be expected to provide rough estimates. A second important factor is that there is still considerable uncertainty in the experimental results because they are based on having found few mutations. For example the estimate in Schug *et al.* [23] is based on finding 3 mutations

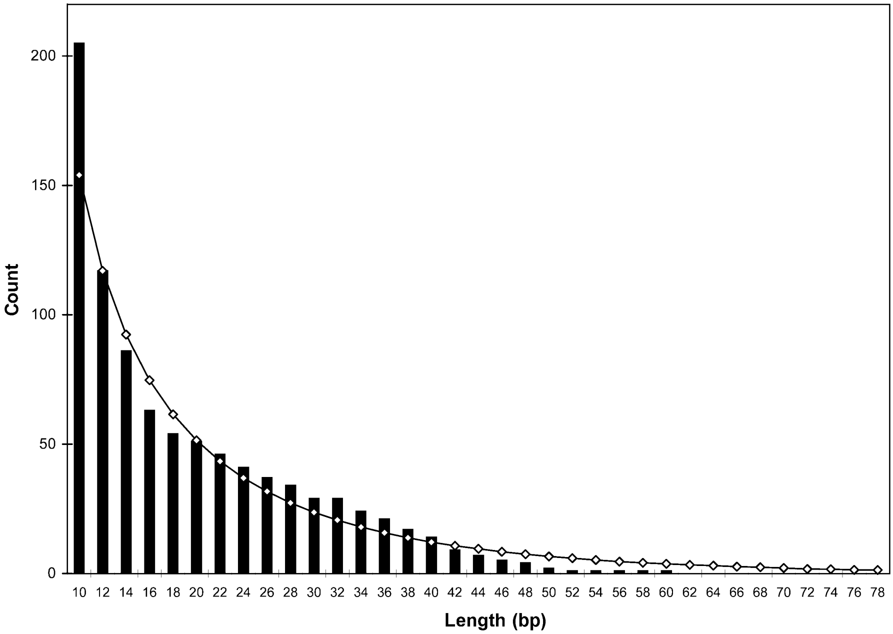

FIGURE 1: The best fit results of the model to human dinucleotide microsatellite data. Observed counts are given by the histogram, and expected counts are given by the line graph. The counts are presented according to the counting scheme described in the text, and the lengths are given in number of nucleotides.

after screening multiple markers in 6570 allele generations at 39 loci, so it should not be surprising that their 95% confidence interval includes our estimate.

Two important conclusions follow from the fit of the model:

- The fact that in each organism the stationary distribution is able to explain the shape of the empirical distribution shows that a simple difference in slippage rate can cause the different microsatellite length distributions observed various organisms.

- Furthermore, the agreement between the value of experimentally determined slippage rates, and the slippage rate estimates determined by the model, indicates that the stationary distribution can be used to estimate slippage rates in organisms where sufficient sequence data is available.

TABLE 1: Results from fitting the model.

	Rate per unit	Rate per locus	Experiment	Reference
Human	4.8×10^{-6}	2.3×10^{-4}	1.3×10^{-4}	[18]
Mouse	1.0×10^{-5}	1.9×10^{-4}	4.7×10^{-4}	[4]
			4.5×10^{-5}	[5]
Fruit fly	2.3×10^{-7}	2.8×10^{-6}	9.3×10^{-6}	[23]
			1.9×10^{-7}	[33]

3. A better upper bound

For our fit of the human data, the parameter $b = 2.4 \times 10^{-6}$ per repeat unit. Recalling that we have set $a = 2 \times 10^{-8}$, this means that $b/a = 120$ and the bound on the mean microsatellite length in equilibrium given in Theorem 1 is $1 + 2b/a = 241$, which drastically overestimates the lengths shown in Figure 1. In this section we turn our attention to finding better bounds, or more generally to the question: how does the equilibrium distribution depend on the parameters of the model? Some insight can be gained from exact solutions. Letting d stand for down and e for expand, we can consider the following example.

Example 2. A more general nearest neighbor model. Let

$$r_{\ell,-1} = d_\ell \text{ for } \ell \geq 2, \quad r_{\ell,1} = e_\ell \text{ for } \ell \geq 1$$

and note that we have incorporated c into $r_{1,1}$.

Because of the large jumps that come from point mutations, Example 2 is not a birth and death chain. However, by considering the flow out of and into the set $[1, i]$ we can conclude that the stationary distribution π satisfies

$$e_i \pi(i) = ai \sum_{j=i+1}^{\infty} \pi(j) + d_{i+1}\pi(i+1). \tag{1}$$

Introducing $\sigma(i) = \sum_{j=i+1}^{\infty} \pi(j)$ then leads to the coupled recursion

$$\pi(i+1) = \frac{e_i}{d_{i+1}}\pi(i) - \frac{ai}{d_{i+1}}\sigma(i), \qquad \sigma(i+1) = \sigma(i) - \pi(i+1). \tag{2}$$

In our basic model $d_{i+1} = ib_{-1}$, for $i \geq 1$, $e_1 = c$, and $e_j = (j-1)b_1$ for $j \geq 2$. If, instead, we take $e_j = je$ for all $j \geq 1$, and change variables $b_{-1} = d$ then the coefficients in (2) are constant and we have an exponential solution with $\pi(i) = \beta\lambda^i$ and $\sigma(i) = \beta\lambda^{i+1}/(1-\lambda)$, with $\beta = (1-\lambda)/\lambda$. It follows from (2) that,

$$\lambda^{i+1} = \frac{e}{d}\lambda^i - \frac{a}{d}\frac{\lambda^{i+1}}{1-\lambda} \quad \text{or} \quad 0 = d\lambda^2 - (a+d+e)\lambda + e.$$

If $d = 0$, i.e. slippage down is impossible, then $\lambda = e/(a+e)$. If $d > 0$ then solving the quadratic equation and noticing that we must have $\lambda < 1$ gives

$$\lambda = \frac{1}{2}\left\{\left(1 + \frac{a+e}{d}\right) - \sqrt{\left(1 + \frac{a+e}{d}\right)^2 - \frac{4e}{d}}\right\}. \tag{3}$$

Some authors have suggested [1, 19] that microsatellites exhibit a preponderance of expansions over contractions. However, one can also argue for a symmetric model by noting that a slippage event which results in an addition of k units to the strand being replicated would (in most cases) also result in a decrease of k units if it happened on the template strand. If in Example 2, $d = e = b$, then the expression under the square root in (3) becomes

$$\left(2 + \frac{a}{b}\right)^2 - 4 = \frac{4a}{b} + \left(\frac{a}{b}\right)^2.$$

If a/b is small then the square root is roughly $2(a/b)^{1/2}$ and we have

$$\lambda \approx 1 - (a/b)^{1/2}. \tag{4}$$

Note that in the above calculation β is a constant needed to make the π's sum to 1 and make $\pi(i),\ i \geq 1$ into a density function. For example, if $d = 0$ and $\lambda = e/(a+e)$ then

$$\sum_{i=1}^{\infty} \lambda^i = e/a,$$

and if we let N denote mean microsatellite length at equilibrium, then $\beta = a/e$ and the density function is

$$\begin{aligned} P(N = j) &= \frac{a}{e}\left(\frac{e}{a+e}\right)^j \\ &= \frac{a}{a+e}\left(1 - \frac{a}{a+e}\right)^{j-1}, \quad j = 1, 2, \ldots \end{aligned}$$

which is geometric with mean $(a+e)/a$.

If $\lambda = 1 - (a/b)^{1/2}$, then

$$\begin{aligned} \mathrm{E}N &= \left(\sum_{i=1}^{\infty} \lambda^i\right)^{-1} \sum_{i=1}^{\infty} i\lambda^i \\ &= \frac{1-\lambda}{\lambda} \frac{\lambda}{(1-\lambda)^2} \\ &= (1-\lambda)^{-1} = (b/a)^{1/2}. \end{aligned}$$

The last calculation is exact for the model with $d_{i+1} = bi$ and $e_i = bi$ for $i \geq 1$, but a little thought reveals that it also gives the asymptotic exponential decay of $\pi(i)$ in the model with $d_{i+1} = bi$ for $i \geq 1$, $e_1 = c$, and $e_j = b(j-1)$ for $j \geq 2$. This decay rate for the tail of the distribution suggests that if slippage is symmetric then microsatellites should typically be of size $\sqrt{b/a}$.

Our final result gives a bound which extends the above rule of thumb for the size of microsatellites to the non-nearest neighbor case.

Example 3. The symmetric model. Suppose $b_k = b_{-k}$ for $1 \leq k \leq K$, and $b_k = 0$ otherwise.

$$r_{\ell,k} = (\ell - 1)b_k, \quad -(\ell - 1) \leq k \leq (\ell - 1).$$

Here, the lower limit on k is needed to avoid transitions to 0, while the upper one is forced by symmetry.

Theorem 2. *Let $v = \sum_k k^2 b_k$. The stationary distribution for the symmetric model has*

$$\sum_{\ell=1}^{\infty} (\ell - 1)\pi(\ell) \leq \left(\frac{3v}{2a}\right)^{1/2} + \left(\frac{3c}{2a}\right)^{1/3}.$$

Note that for the parameters of interest to us, the second term on the right is always smaller than 1 and can be ignored. If we do this in the nearest neighbor model in which $b_1 = b_{-1} = b$ then $v = 2b$ and the upper bound becomes $(3b/a)^{1/2}$. Inserting our fit for the human data which has $b/a = 240$ yields an upper bound on the mean microsatellite length of $1 + (720)^{1/2} = 27.83$. This still overestimates the mean (as it must by the proof of Theorem 2 given in the next section) but is substantially closer than the crude estimate of Theorem 1.

4. Proofs of the main results

Let N_t denote the length of our microsatellite at time t. We begin with the proof of Theorem 1.

Proof of Theorem 1. If N_t were null recurrent or transient then starting from $N_0 = 1$ we would have $N_t \to \infty$ in probability and by Fatou's lemma that

$$\liminf_{t\to\infty} \mathrm{E}_1 N_t = \infty.$$

Thus we can prove the existence of a stationary distribution by showing that $\mathrm{E}_1 N_t$ stays bounded as $t \to \infty$. To do this we begin by recalling a standard fact about Markov chains. The polynomial bound here on f is more than enough to justify the use of the dominated convergence theorem in the proof of Lemma 1 (which is left to the reader).

Lemma 1. *Let $q(x, y)$ denote the transition rate from state x to state y in our Markov chain. If $|f(x)| \le C(1 + |x|^n)$ then*

$$\frac{\mathrm{d}}{\mathrm{d}t} E_x f(N_t)\bigg|_{t=0} = \sum_y q(x, y)[f(y) - f(x)].$$

Our first step is to apply Lemma 1 to $f(x) = x$. To facilitate computation we will write q as a sum of q_a for the transition events in parts 1 and 3 of the definition, and q_b for the base pair substitutions in part 2. Recalling the various definitions shows

$$\sum_y q_a(x, y)(y - x) \le c + \sum_k (x - 1)b_k \cdot k = c + (x - 1)u,$$

and with a little more arithmetic we get

$$\sum_y q_b(x, y)(y - x) = \sum_{1\le y<x} a(y - x) = -a\sum_{m=1}^{x-1} m = -a\left(\frac{x(x-1)}{2}\right).$$

Using Lemma 1 with $\mathrm{E}N_t^2 \ge (\mathrm{E}N_t)^2$ and the trivial $N_t \ge 1$, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}_1 N_t &\le c + u\mathrm{E}_1(N_t - 1) - \frac{a}{2}\mathrm{E}_1(N_t^2 - N_t),\\ &\le \mathrm{E}_1 N_t \left(c + u - \frac{a}{2}\mathrm{E}_1(N_t - 1)\right) < 0,\end{aligned}$$

if $\mathrm{E}_1 N_t > 1 + 2(c + u)/a$. Since $\mathrm{E}_1 N_0 = 1$, an easy argument by contradiction shows that we must have $\mathrm{E}_1 N_t \le 1 + 2(c + u)/a$ for all t.

The last conclusion shows the existence of a stationary distribution. Uniqueness of the stationary distribution and convergence to it starting from state 1, now follow from the fact that our continuous time Markov chain is irreducible. Another use of Fatou's lemma then shows that the limit distribution N_∞ has

$$\mathrm{E}_1 N_\infty \le \liminf_{t\to\infty} \mathrm{E}_1 N_t \le 1 + 2(c+u)/a.$$

Proof of Theorem 2. We follow the approach of the previous proof and use its notation, but this time take $f(x) = (x-1)^2$. Noting that

$$f(x+k) - f(x) = 2k(x-1) + k^2,$$

and that c is the rate of jumps from 1 to 2, we have

$$\sum_y q_a(x,y)(f(y)-f(x)) \le c + \sum_k (x-1)b_k \cdot (2k(x-1)+k^2) = c + (x-1)v,$$

since symmetric slippage implies $\sum_k kb_k = 0$. For the other term we begin with

$$\sum_y q_b(x,y)(f(y)-f(x)) = \sum_{1\le y<x} a((y-1)^2 - (x-1)^2) = a\sum_{m=1}^{x-2} m^2 - a(x-1)^3$$

and use approximating sums for the integral to get

$$\sum_{m=1}^{x-2} m^2 \le \int_1^{x-1} y^2\,\mathrm{d}y \le \frac{(x-1)^3}{3}.$$

Combining the last three displays with Lemma 1 we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}_1(N_t-1)^2 \le c + v\mathrm{E}_1(N_t-1) - \frac{2a}{3}\mathrm{E}_1\{(N_t-1)^3\}.$$

If we start with the stationary distribution, then the left-hand side is 0, so introducing $\mu = \mathrm{E}_1(N_t-1)$ and using Jensen's inequality gives

$$0 \le c + v\mu - \frac{2a\mu^3}{3}.$$

The cubic $c + vx - 2ax^3/3$ on the right-hand side goes to $-\infty$ as $x \to \infty$, so μ must be smaller than the largest root, ρ. To estimate this we note that if $x > (3v/2a)^{1/2} + (3c/2a)^{1/3}$ then we have

$$\begin{aligned}\frac{2a}{3}x^3 &> \frac{2a}{3}x\left(\frac{3v}{2a} + \left(\frac{3c}{2a}\right)^{2/3}\right)\\ &> vx + \frac{2a}{3}\cdot\left(\frac{3c}{2a}\right)^{1/3}\cdot\left(\frac{3c}{2a}\right)^{2/3} = vx + c.\end{aligned}$$

From this we have $\mu \le \rho \le (3v/2a)^{1/2} + (3c/2a)^{1/3}$, and the proof of Theorem 2 is complete.

Acknowledgements

We thank Charles Aquadro and Malcolm Schug for providing the biological motivation for this work. RTD was supported by a grant from the NSF, and SK was supported by the Department of Defense National Graduate Fellowship.

References

[1] AMOS, W., SAUCER, S., FEAKES, R. AND RUBINSZTEIN, D. (1996). Microsatellites show mutational bias and heterozygote instability. *Nature Genet.* **13**, 390–391.

[2] ASHLEY, C. T. AND WARREN, S. T. (1995). Trinucleotide repeat expansion and human disease. *Ann. Rev. Genet.* **29,** 703–728.

[3] BELL, G. AND JURKA, J. (1997). The length distribution of perfect dimer repetitive DNA is consistent with its evolution by an unbiased single step mutation process. *J. Mol. Evol.* **44,** 414–421.

[4] DALLAS, J. (1992). Estimation of microsatellite mutation rates in recombinant inbred strains of mouse. *Mammalian Genome* **3,** 452–456.

[5] DIETRICH, W., KATZ, H., LINCOLN, S. E., SHIN, H. S., FRIEDMAN, J., DRACOPOLI, N. C. AND LANDER, E. S. (1992). A genetic map of the mouse suitable for typing interspecific crosses. *Genetics* **131,** 423–447.

[6] DIRIENZO, A., PETERSON, A. C., GARZA, J. C., VALDES, A. M., SLATKIN, M. AND FREIMER, N. B. (1994). Mutational processes of simple sequence repeat loci in human populations. *Proc. Nat. Acad. Sci. USA* **91,** 3166–3170.

[7] FELDMAN, M. W., BERGMAN, A., POLLOCK, D. D. AND GOLDSTEIN, D. B. (1997). Microsatellite genetic distances with range constraints: Analytic description and problems of estimation. *Genetics* **145,** 207–216.

[8] GARZA, J. C., SLATKIN, M. AND FREIMER, N. B. (1995). Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. *Mol. Biol. Evol.* **12,** 594–603.

[9] GOLDSTEIN, D. B., RUIZ-LINARES, A., CAVALLI-SFORZA, L. L. AND FELDMAN, M. W. (1995). An evaluation of genetic distances for use with microsatellite loci. *Genetics* **139,** 463–471.

[10] GOLDSTEIN, D. B., RUIZ-LINARES, A., CAVALLI-SFORZA, L. L. AND FELDMAN, M. W. (1995). Genetic absolute dating based on microsatellites and modern human origins. *Proc. Nat. Acad. Sci.* **92,** 6723–6727.

[11] HUDSON, R. R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, Vol. 7, ed. D. J. Futuyama and J. Antonovics. OUP, Oxford, pp. 1–44.

[12] KIMMEL, M. AND CHAKRABORTY, R. (1996). Measures of variation at DNA repeat loci under a general stepwise mutation model. *Theoret. Popul. Biol.* **39,** 30–48.

[13] LEVINSON, G. AND GUTMAN, G. A. (1987). Slipped–strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* **4**, 203–221.

[14] LI, W. H. (1997). *Molecular Evolution.* Sinauer Associates, Massachusetts, pp. 177–236.

[15] MCMURRAY, C. T. (1995). Mechanisms of DNA expansion. *Chromosoma* **104,** 2–13.

[16] MORAN, P. A. P (1975). Wandering distributions and the electrophoretic profile. *Theoret. Pop. Biol.* **8,** 318–330.

[17] OHTA, T. AND KIMURA, M. (1973). The model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a genetic population. *Genet. Res.* **22,** 201–204.

[18] PETRUKHIN, K. E., SPEER, M. C., CAYANIS, E., DEFATIMA BONALDO, M., TANTRAVAHI, U., SOARES, M. B., FISCHER, S. G., WARBURTON, D., GILLIAM, T. C. AND OTT, J. (1993). A microsatellite genetic linkage map of human chromosome 13. *Genomics* **15,** 76–85.

[19] PRIMMER, C. R., ELLEGREN, H., SAINO, N. AND MOLLER, A. P. (1996). Directional evolution in germline microsatellite mutations. *Nature Genet.* **13**, 391–393.

[20] PRITCHARD, J. K. AND FELDMAN, M. W. (1996). Statistics for microsatellite variation based on coalescence. *Theoret. Popul. Biol.* **50**, 325–344.

[21] SCHLOTTERER, C. AND TAUTZ, D. (1992). Slippage synthesis of simple sequence DNA. *Nucleic Acids Res.* **20,** 211–216.

[22] SCHUG, M. D., WETTERSTRAND, K., GAUDETTE, M., LIM, R., HUTTER, C. AND AQUADRO, C. F. (1997). The distribution and frequency of microsatellite loci in *Drosophila melanogaster*. Submitted to *Nucleic Acids research* 1/29/97.

[23] SCHUG, M. D., HUTTER, C. M., WETTERSTRAND, K. A., GAUDETTE, M. S., MACKAY, T. F. C. AND AQUADRO, C. F. (1998). The mutation rate of di- tri– and tetranucleotide repeats in Drosophila melanogaster. *Mol. Biol. Evol.* **15**, 1751–1760.

[24] SHRIVER, M. D., JIN, L., CHAKRABORTY, R. AND BOERWINKLE, E. (1993). VNTR allele frequency distributions under the stepwise mutation model: A computer simulation approach. *Genetics* **134,** 983–993.

[25] SLATKIN, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139,** 457–462.

[26] SLATKIN, M. (1995) Hitchhiking and associative overdominance at a microsatellite locus. *Mol. Biol. Evol.* **12,** 473–480.

[27] SMITH, G. P. (1973). Unequal crossover and the evolution of multigene families. *Cold Spring Harbor Symp. Quant. Biol.* **38**, 507–513.

[28] TAKEZAKI, N. AND NEI, M. (1996). Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics* **144,** 389–399.

[29] TAVARÉ, S. (1984). Line-of-descent and genealogical processes and their application in population genetics models. *Theoret. Pop. Biol.* **26,** 119–164.

[30] VALDES, A., SLATKIN, M. AND FREIMER, N. B. (1993). Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* **133,** 737–749.

[31] WEBER, J. L. AND WONG, C. (1993). Mutation of human short tandem repeats. *Hum. Mol. Genet.* **2,** 1123–1128.

[32] WEISSENBACH, J., GYAPAY, G., DIB, C., VIGNAL, A., MORISSETTE, J., MILLASSEAU, P., VAYSSEIX, G. AND LATHROP, M. (1992). A second-generation linkage map of the human genome. *Nature,* **359,** 794–801.

[33] WETTERSTRAND, K. S. (1997). Microsatellite polymorphism and divergence in worldwide populations of *Drosophila Melanogaster* and *D. simulans*. M.Sc. Thesis, Cornell University, Ithaca, NY.

[34] WIERDL, M., DOMINISKA, M. AND PETES, T. D. (1996). Microsatellite instability in yeast: dependence on the length of the microsatellite. *Genetics,* **146**, 769–779.

[35] ZHIVOTOVSKY, L. A., FELDMAN, M. W. AND GRISHECHKIN, S. A. (1997). Biased mutations and microsatellite variation. *Mol. Biol. Evol.* **14**, 926–933.