Dynamics of Microsatellite Divergence Under Stepwise Mutation and Proportional Slippage/Point Mutation Models

Peter P. Calabrese,* Richard T. Durrett† and Charles F. Aquadro‡

*Department of Applied Mathematics, †Department of Mathematics and ‡Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853

Manuscript received June 15, 2000
Accepted for publication July 6, 2001

ABSTRACT

Recently Kruglyak, Durrett, Schug, and Aquadro showed that microsatellite equilibrium distributions can result from a balance between polymerase slippage and point mutations. Here, we introduce an elaboration of their model that keeps track of all parts of a perfect repeat and a simplification that ignores point mutations. We develop a detailed mathematical theory for these models that exhibits properties of microsatellite distributions, such as positive skewness of allele lengths, that are consistent with data but are inconsistent with the predictions of the stepwise mutation model. We use our theoretical results to analyze the successes and failures of the genetic distances $(\delta\mu)^2$ and D_{SW} when used to date four divergences: African *vs.* non-African human populations, humans *vs.* chimpanzees, *Drosophila melanogaster vs. D. simulans*, and sheep *vs.* cattle. The influence of point mutations explains some of the problems with the last two examples, as does the fact that these genetic distances have large stochastic variance. However, we find that these two features are not enough to explain the problems of dating the human-chimpanzee split. One possible explanation of this phenomenon is that long microsatellites have a mutational bias that favors contractions over expansions.

MICROSATELLITES are simple sequence repeats in DNA that typically have a high level of variability due to a high rate of mutations that alter their length. For this reason they have been useful for studying population structure on the time scale of thousands of generations (see Bowcock *et al.* 1994; Roy *et al.* 1994; Goldstein *et al.* 1995b; Underhill *et al.* 1996; Goldstein and Pollock 1997; Harr *et al.* 1998; Irwin *et al.* 1998; Reich and Goldstein 1998; Goldstein *et al.* 1999; Pritchard *et al.* 1999; Ruiz-Linares *et al.* 1999). To make inferences from observed patterns, one needs a statistic to measure differentiation between populations and a model to give the distribution of that statistic. Here, we consider two genetic distances: $(\delta\mu)^2$ of Goldstein *et al.* (1995a,b) and D_{SW} of Shriver *et al.* (1995).

We examine the behavior of two genetic distances $(\delta\mu)^2$ and D_{SW} in four increasingly divergent examples: (i) African *vs.* non-African human populations, (ii) human *vs.* chimpanzee, (iii) *Drosophila melanogaster vs. D. simulans*, and (iv) cattle *vs.* sheep. If one assumes the stepwise mutation model (SMM) of Ohta and Kimura (1973), then the expected value of $(\delta\mu)^2$ grows linearly in time. When used on example (i), the statistic $(\delta\mu)^2$ gives good estimates (see Goldstein *et al.* 1995b), but when applied to examples (ii) and (iii), it gives answers that are roughly one-seventh and one-thirtieth of the commonly accepted values. The nonlinear distance D_{SW} does not do as well as $(\delta\mu)^2$ at dating the human population split but has a slightly better performance for examples (ii) and (iii), yielding estimates that are about one-third and one-eighth of the commonly accepted values.

Finally, in example (iv), the two species are too far diverged for microsatellites to be useful molecular clocks. Results of Ellegren *et al.* (1997) show that roughly one-half of the microsatellites they isolated in one species were monomorphic in the other and have presumably lost their ability to mutate. This observation suggests that in the long run point mutations break up perfect repeats and reduce the mutation rates of microsatellite loci. It is natural to ask if this mechanism can explain the underestimates that arise in examples (ii) and (iii). To investigate this possibility, we introduced two new models. The first is a slight generalization of the model of Kruglyak *et al.* (1998), which we call the proportional slippage/point mutation (PS/PM) model. In this model point mutations spoil perfect repeats; the slippage rate is zero for microsatellites with fewer than κ repeat units and then increases linearly. The PS/PM model can be used to estimate slippage rates from DNA sequence data, but to address the divergence question we need a second model, called the PCR model, that keeps track of the lengths of all perfect repeats that make up an imperfect repeat.

The PCR model is complicated, but it is possible to obtain a simple formula for the variance of a repeat L_t as a function of time t in generations (see *Theorem 2*).

Corresponding author: Rick Durrett, Department of Mathematics, 523 Malott Hall, Cornell University, Ithaca, NY 14853.
E-mail: rtd1@cornell.edu

Using $a = 2 \times 10^{-8}$ as an estimate for the point mutation rate per repeat unit and a threshold of four repeat units for slippage events to be possible, this formula shows that the variance of the repeat length begins to depart from linearity when $t/(10{,}000{,}000)$ is not small relative to one. This result explains some of the problems with the use of $(\delta\mu)^2$ in the comparison of *D. melanogaster* and *D. simulans*, which diverged ~25,000,000 generations ago, but makes the failure of $(\delta\mu)^2$ in the human *vs.* chimpanzee split even more mysterious, since, as our calculations have shown, point mutations will not have had a significant effect in 250,000 generations.

To further investigate the problems in dating the human *vs.* chimpanzee split, we investigated the behavior of the PS/PM model when there are no mutations. This special case, called the PS/0M model, and denoted A_t^0, is equivalent to the binary branching process of probability theory, so it is possible to do many exact calculations. *Theorem 3* gives expressions for the first four moments of A_t^0. Expressions for the third moment show that the distribution of A_t^0 has a positive skewness, which contrasts with the symmetric distributions of the SMM, but is consistent with the skewness observed in microsatellite data.

Calculations for the fourth moment show that if β is the per locus slippage rate, and α is the initial activity of a microsatellite, *i.e.*, the length minus the threshold κ for slippage to occur, then the kurtosis becomes large when $\beta t/\alpha^2$ is large relative to one. In general fourth moments of the microsatellite lengths are larger under the PS/0M model than under the SMM. Consequently, microsatellite statistics that use these moments, such as those of Reich and Goldstein (1998) and Gonser *et al.* (2000), will have much different distribution under PS/0M than under SMM. In the case of our four examples the kurtoses are (i) 3.02, (ii) 3.93, (iii) 6.75, and (iv) 10.7, compared to 3 for the normal distribution. In the case of the human-chimpanzee split, (ii), this implies that confidence intervals are 1.21 times as large as they would be under the SMM. However, this again does not explain the magnitude of the failures of $(\delta\mu)^2$ and D_{SW} in dating the human-chimpanzee split. The last observation and the fact that the simulated microsatellite distributions given in Figures 1 and 3 have many more large microsatellites than are typically observed lead us to conclude that there are forces that constrain the growth not yet incorporated into our models. We return to this point in the discussion.

GENETIC DISTANCES

Our first step is to define the two genetic distances $(\delta\mu)^2$ and D_{SW} and to compute their values for the four examples. We then introduce our two new models, state the theoretical results we have obtained, and use them to study the four examples. To define $(\delta\mu)^2$, let μ_A and μ_B be the mean length of alleles at a microsatellite locus in populations A and B, and define genetic distance between the two populations [see (1) of Goldstein *et al.* 1995b] as

$$(\delta\mu)^2 = (\mu_A - \mu_B)^2.$$

Given data, the distance is estimated by the corresponding statistic

$$\widehat{(\delta\mu)^2} = (\overline{X}_A - \overline{Y}_B)^2,$$

where $\overline{X}_A$ and $\overline{Y}_B$ are the average lengths observed in samples from populations A and B.

To motivate the definition of our second distance, we recall (see, *e.g.*, p. 6723 of Goldstein *et al.* 1995b) that, if X and X' are the lengths of the microsatellite locus in a sample of size two from population A, and Y and Y' are a similar random sample of size two from population B, then

$$(\delta\mu)^2 = E(X - Y)^2 - \frac{E(X - X')^2 + E(Y - Y')^2}{2},$$

where in each case we take the square before computing expected value. Replacing the squares in the last formula by absolute values, we can follow Shriver *et al.* (1995) and define the genetic distance

$$D_{SW} = E|X - Y| - \frac{E|X - X'| + E|Y - Y'|}{2}.$$

Given microsatellite lengths $X_1, \ldots X_m$ from population A, and $Y_1, \ldots Y_n$ from population B, D_{SW} is estimated by

$$\hat{D}_{SW} = \frac{1}{mn}\sum_{i=1}^{n}\sum_{i=1}^{m}|X_i - Y_j| - \frac{1}{m(m-1)}\sum_{1\le i<j\le m}|X_i - X_j| - \frac{1}{n(n-1)}\sum_{1\le i<j\le n}|Y_i - Y_j|.$$

Suppose that microsatellites follow the SMM of Ohta and Kimura (1973) in which microsatellites change by ± 1 unit at a rate β independent of their length. In this case Goldstein *et al.* (1995a) have shown that if one assumes the divergence of the two populations occurred τ generations ago, then

$$(\delta\mu)^2 = 2\beta\tau.$$

Shriver *et al.* (1995) simulated the behavior of D_{SW} under the SMM and concluded that over short timescales D_{SW} was linear. In appendix a we prove the following result about the nonlinear behavior of D_{SW} under the SMM when two populations of N_e diploid individuals diverged τ generation ago.

Theorem 1: *If* $2\beta\tau$ *is large and* $\tau \ge N_e$ *then*

$$D_{SW} \approx \sqrt{\frac{2}{\pi}} \cdot \sqrt{2\beta\tau + 4\beta N_e} - \frac{4\beta N_e}{\sqrt{8\beta N_e + 1}}. \qquad (2)$$

When $\tau \gg N_e$, the terms involving N_e can be dropped and

$$D_{SW} \approx 2(\beta\tau/\pi)^{1/2},$$

so in the long run D_{SW} grows like a constant times $\tau^{1/2}$.

FOUR EXAMPLES

To test the behavior of the statistics $(\delta\mu)^2$ and D_{SW} we consider four increasingly divergent examples.

Divergence of human populations: Goldstein *et al.* (1995b) investigated 30 microsatellite loci and estimated that the value of $(\delta\mu)^2$ between African and non-African populations was 6.47. Using this in (1) with their mutation rate estimate of 5.6×10^{-4} gives a prediction of 5776 generations for the divergence time. Assuming a human generation time of 27 years, they then arrived at the estimate of 156,000 years, a figure that they argued was in agreement with previous genetic estimates and with archaeological data.

Rubinsztein *et al.* (1995) studied 24 microsatellite loci in East Anglians and Sub-Saharan Africans and obtained an estimate of 1.45 for D_{SW}. Assuming $\beta = 5.6 \times 10^{-4}$ and taking $N_e = 5000$ as the size of one of the two subpopulations, we can use (2) to give a prediction of 9880 generations for the divergence time. Multiplying by 27 years leads to an estimate of 267,000 years, which is much larger than the estimate of Goldstein *et al.* (1995b). One possible explanation is that we have chosen the wrong effective population size for our estimate. If instead we use $N_e = 750$ then an estimate of 5630 generations results, which is similar to the value estimated by Goldstein *et al.* (1995b).

Humans *vs.* chimpanzees: Rubinsztein *et al.* (1995) also studied 24 microsatellite loci in chimpanzees. Combining this with their human data, they obtained an estimate of 5.475 for D_{SW} for the human-chimpanzee comparison. They commented that the ratio of this estimate to the East Anglian *vs.* African comparison, 5.475/1.45 = 3.78, was surprising since the ratio of the divergence times for the two splits is at least 50. The nonlinearity of D_{SW} shown in *Theorem 1* helps explain this discrepancy. If we use the slippage rate of $\beta = 5.6 \times 10^{-4}$ from the previous example for both humans and chimpanzees and assume an effective population size of $N_e = 10^4$ for each population, then using Theorem 1 we arrive at an estimate of $\tau = 88{,}200$ generations for their divergence time. If we use an average lifetime of 20 years for humans and chimpanzees this translates into 1.76 million years, about one-third the accepted estimate of 5–6 million years (see, *e.g.*, Goodman *et al.* 1998 or Kumar and Hedges 1998).

Since Rubinsztein *et al.* (1995) report only the genetic distances D_{SW} for their loci, we need to turn to other sources for data we can use to calculate $(\delta\mu)^2$. Bowcock *et al.* (1994), Deka *et al.* (1994), and Garza *et al.* (1995) studied 10, 7, and 8 microsatellite loci, respectively, in these two species. The data are given in Table 1. From this, we can compute $(\delta\mu)^2$ values of 7.56, 86.19, and 40.19, respectively. Even though the second estimate is >11 times the first, we can use all 25 loci in Table 1 together to get $(\delta\mu)^2 \approx 40$. Using (1) now with the slippage rate estimate $\beta = 5.6 \times 10^{-4}$ gives 35,700 generations, or ~700,000 years, which is less than one-seventh the accepted age.

Assuming the SMM and that the above parameters remain constant, coalescent simulations show that the $(\delta\mu)^2$ and D_{SW} estimates are significantly smaller than those expected under the SMM. Specifically, for two samples of 20 individuals with 25 unlinked microsatellites in two separate random-mating populations of size 10^4, which were separated until 275,000 generations ago and with mutations following the SMM with $\beta = 5.6 \times 10^{-4}$, we expect a 95% confidence interval for $(\delta\mu)^2$ of 179–465, whereas the data were only 40, and a 95% confidence interval for D_{SW} of 7.97–14.6, whereas the data were 5.475.

Drosophila species: The divergence time between *D. melanogaster* and *D. simulans* is estimated to have occurred ~2.5 million years ago (see Hey and Kliman 1993). Wetterstrand (1997) used eight di-, four tri-, and four tetranucleotide repeats and estimated $(\delta\mu)^2 = 19.393$ between these species. Using the mutation estimate of 6.3×10^{-6} from Schug *et al.* (1997), she then used (1) to estimate that the divergence time occurred 1.52 million generations ago. Assuming 10 generations per year, she computed a divergence time of 152,000 years, which is about one-sixteenth of the estimate of Hey and Kliman (1993).

One of the problems with this estimation is that tri- and tetranucleotide repeats have considerably smaller slippage rates than dinucleotide repeats in Drosophila (see Schug *et al.* 1998). With this in mind, we applied Wetterstrand's analysis to data on 31 dinucleotide repeat loci from Hutter *et al.* (1998) given in Table 2. The average value of $(\delta\mu)^2$ for these loci is 16.09. Using the estimate $\beta = 9.3 \times 10^{-6}$ from Schug *et al.* (1998) in (1) we estimate the divergence time to be ~865,000 generations. Using the previous estimate of 10 generations per year, this translates into 86,500 years, which is about one-thirtieth of the estimate of Hey and Kliman (1993).

Independently, Harr *et al.* (1998) also used $(\delta\mu)^2$ to estimate the divergence times in the phylogeny of *D. melanogaster*, *D. simulans*, *D. sechelia*, and *D. mauritiana*. From the possible choices of the mutation rate β they list, we choose 10^{-5}, which is the closest to that of Schug *et al.* (1998). In this case their estimates differ from those of Hey and Kliman (1993) by factors of 10–30.

Our second statistic D_{SW} does much better on the data set of Hutter *et al.* (1998). The estimate of D_{SW} from their data is 3.64, so assuming an effective population size of $N = 10^6$ and using (2) with $\beta = 9.3 \times 10^{-6}$, we obtain an estimate of $\tau = 3{,}330{,}000$ generations. With

TABLE 1

Human/chimpanzee data

Locus	Human repeat motif	$\delta\mu$[a]	$(\delta\mu)^2$	Variance	
				Human	Chimp
		Bowcock *et al.* (1994)			
ACTC		2.55	6.50	117.22	5.45
084XC5		0.95	0.90	5.83	2.55
D13S118	$(CA)_{15}$	0.8	0.64	3.01	3.80
D13S119		1.75	3.06	8.67	10.61
D13S125		3.0	9.0	12.27	30.14
D13S133		5.15	26.52	105.69	33.12
D13S137		4.95	24.50	12.13	23.40
D13S193	$(AC)_{14}$	1.15	1.32	11.84	30.85
D13S227		0	0	5.35	13.59
Utsw1523		1.65	2.72	1.31	4.65
				Heterozygosity (%)	
Locus	Human repeat motif	$\delta\mu$[a]	$(\delta\mu)^2$	Human	Chimp
		Deka *et al.* (1994)			
D13S71	$(CA)_{17}$	3.06	9.36	74.2	18.6
D13S118	$(CA)_{15}$	7.06	49.29	72.2	73.4
D13S121	$(AC)_{18}$	0.94	0.89	76.7	87.7
D13S122	$(GT)_{20}$	16.8	283	83.3	62.4
D13S124		2.26	5.14	66.9	81.9
D13S193	$(AC)_{14}$	6.45	41.66	74.0	71.9
D13S197	$(AC)_6(GC)_8(AC)_{12}$	14.6	214	87.4	1.8
		Garza *et al.* (1995)			
Mfd3	$(CA)_{20}$	0.3	0.09	74.5	83.4
Mfd32	$(CA)_{12}$	−2.4	5.76	70.5	79.9
Mfd38	$(CA)_{27}$	2.3	5.29	84.0	89.9
Mfd59[b]	$(CA)_5(TN)_{12}(CA)_{21}$	8.3	68.89	87.5	86.8
Mfd75	$(AC)_{20}(TC)_{12}$	7.8	60.84	87.9	69.7
Mfd104	$(CA)_{25}$	−10.6	112.36	83.2	89.9
Mfd139	$(CA)_{15}$	7.7	59.29	88.3	65.6
Mfd142	$(CA)_{20}$	3.0	9.00	75.4	68.6

[a] $(\delta\mu)$ in repeat units.
[b] In Mfd59 *N* indicates *A* or *T*.

10 generations a year this becomes 330,000 years, which is about one-eighth of the estimate of Hey and Kliman (1993).

Again coalescent simulations with the above parameters show that these estimates are significantly smaller than those expected under the SMM. Assuming the two populations are separated until 25 million generations ago we expect a 95% confidence interval of 315–728 for $(\delta\mu)^2$ while the data are <20 and a 95% confidence interval of 10.5–18.0 for D_{SW} while the data are 3.64.

Cattle *vs.* sheep: These two species diverged ~16 million years ago, which, assuming a generation of 2 years, translates into 8 million generations. Ellegren *et al.* (1997) examined 13 loci of bovine origin and 14 of ovine origin. Discarding 3 loci of bovine origin for which there was not reliable information about their length in sheep, the data are given in Table 3.

Two of these loci studied by Ellegren *et al.* (1997) show clear signs of mutations other than microsatellite slippage events. At RM103 allele sizes are 115–151 bp in cattle *vs.* 73 bp in sheep, but the example sequence given for the repeat in cattle is $(CA)_{16}$. Thus at least part of the average 61.6 bp difference must be due to a major deletion in the sequence flanking the microsatellite in sheep or to an insertion in cattle. At RME11 we have the surprising result that this locus is much longer in sheep than in cattle but that this longer and hence presumably more mutable microsatellite is monomorphic in sheep. Note also that this is the only locus of bovine origin with a large negative $\delta\mu$. This suggests that again much of this difference in length is due to mutations involving the flanking sequence.

If we remove these two loci, which have an average $(\delta\mu)^2$ of 653, the remaining 22 loci have an average $(\delta\mu)^2$ of 74.4 per locus. If we use an average generation time of 2 years for cattle and sheep, then using (1) we

TABLE 2

Drosophila data, ($\delta\mu$) in repeat units

Locus	Focal species repeat motif	$\delta\mu$	$(\delta\mu)^2$	D_{SW}	Heterozygosity (%)	
					D. mel	*D. sim*
		D. melanogaster derived				
DM28	$(GT)_{10}$	1.60	2.55	1.08	43	48
DM30	$(TG)_6$	0.97	0.95	1.01	36	27
DM40	$(TC)_9$	1.52	2.33	0.67	82	75
DM55	$(AC)_{10}$	0.58	0.34	0.24	59	73
DM58	$(GT)_{10}$	2.10	4.41	1.82	88	34
DM73	$(AC)_{23}$	11.73	137.51	16.70	90	76
DM75	$(GT)_{10}$	−2.84	8.09	2.89	71	69
DM85	$(CA)_5$	10.47	109.59	17.69	50	19
DM88	$(GA)_{13}$	−2.21	4.88	2.58	67	59
DM92	$(CA)_{12}$	3.70	13.69	4.43	68	66
DM94	$(TG)_7$	−5.72	32.71	4.76	82	83
DM97	$(GT)_5$	−2.69	7.24	3.47	64	48
DM100	$(CT)_{12}$	−1.62	2.61	1.65	65	46
DM114	$(AC)_7$	4.05	16.40	6.02	74	44
DM122	$(AC)_{12}$	2.02	4.08	1.51	78	68
		D. simulans derived				
DSIM3	$(CA)_{11}$	2.06	4.23	3.35	0	29
DSIM6a	$(GT)_9$	5.61	31.51	7.67	39	73
DSIM6b	$(CT)_6$	−1.85	3.42	4.20	0	54
DSIM10	$(AC)_9$	−0.75	0.56	1.09	41	75
DSIM18	$(GT)_8$	−0.05	0	0.59	64	0
DSIM25	$(TG)_{12}$	−1.78	3.15	0.20	69	66
DSIM28a	$(GT)_5$	0.65	0.42	0.22	53	54
DSIM29	$(GT)_{14}$	−4.76	22.66	3.06	77	53
DSIM30	$(GT)_{11}$	2.80	7.84	3.67	33	69
DSIM36	$(GT)_{10}$	−1.10	1.21	1.55	0	77
DSIM42b	$(GT)_{13}$	−3.04	9.22	4.43	67	71
DSIM45	$(TG)_8$	5.46	29.82	6.46	75	83
DSIM50	$(GT)_{11}$	0.09	0.01	0.26	62	33
DSIM86	$(GT)_8$	1.00	1.00	0.94	31	67
DSIM103	$(GT)_{11}$	3.84	14.78	3.08	74	80
DSIM119	$(GT)_8$	4.65	21.62	5.59	71	41

can estimate that the average slippage rate must be $\beta = 4.65 \times 10^{-6}$. We could find no information about slippage rates in cattle or sheep, but this is about one-thirteenth the rate of 6×10^{-5} that Ellegren (1995) observed for microsatellites in pigs.

TWO MODELS WITH POINT MUTATIONS

In all but the first example of the African *vs.* non-African split in the human population, if we use the SMM with either of our statistics $(\delta\mu)^2$ or D_{SW}, then we underestimate divergence times. In view of this, it is natural to ask if there is some mechanism that interferes with the normal rate of growth of these divergence statistics. One possibility is that point mutations spoiling perfect repeats reduce microsatellite mutation rates over time. To investigate this we introduce a new model called the PS/PM model that is a modest generalization of the one proposed by Kruglyak *et al.* (1998).

PS/PM model: There are three types of changes that can occur:

Proportional slippage: A microsatellite of length $\ell > \kappa$ becomes length $\ell \pm 1$ at rate $b(\ell - \kappa)$ each. Microsatellites of length $\ell \leq \kappa$ do not experience slippage events.

Point mutations: For $1 \leq j < \ell$, a microsatellite of length ℓ becomes length j at rate a.

Birth of microsatellites: $\kappa \rightarrow \kappa + 1$ at rate c.

For later purposes, it is convenient to write the new proportional slippage rule succinctly as $b(\ell - \kappa)^+$, where

$$(\ell - \kappa)^+ = \begin{cases} \ell - \kappa & \text{if } \ell > \kappa \\ 0 & \text{if } \ell \leq \kappa \end{cases}$$

denotes the *positive part* of $\ell - \kappa$; *i.e.*, $\ell - \kappa$ if the difference is positive and 0 otherwise.

TABLE 3

Cattle/sheep data

Locus	Focal species repeat[a]	$\delta\mu$[b]	$(\delta\mu)^2$	Heterozygosity (%)	
				Cattle	Sheep
		Bovine origin			
RM103	$(CA)_{16}$	61.6	3749	70	0
RM088	$(CA)_{14}$	23.3	542.9	74	0
RM024	$(CA)_{11}$	9.0	81.00	29	0
RM041	$(CA)_{18}$	8.8	77.44	85	0
RM012	$(CA)_{10}$	4.6	21.16	48	0
RME11	$(GTT)_9$	−38.4	1474	78	0
RME23	$(GT)_{11}$	29.4	864.4	83	29
CSSM31	$(CA)_{25}$	12	144.0	28	67
RM011	$(CA)_{15}$	9.8	96.04	76	52
RM044	$(CA)_{20}$	−0.4	1.6	85	89
		Ovine origin			
McM136	$(CA)_{18}$	−34.6	1197	0	85
McM373	$(GT)_{25}$	−23.7	561.7	0	86
CSRD2105	$(CA)_{19}$	−18.5	342.3	0	80
OARJMP8	[c]	−9.2	84.64	0	63
CSRD240	[c]	−8.1	65.61	0	80
OARCP34	$(AC)_{17}$	−5.6	31.36	0	70
Maf209	$(TG)_{24}$	4.5	20.25	0	54
McM147	$(TG)_{20}$	−19.2	368.6	41	85
McM058	$(AC)_{25}$	−17.6	309.8	70	89
CSRD2171	$(CA)_{15}$	−16.6	275.6	46	65
Maf70	$(AC)_{39}$	−7.4	54.76	79	82
McM064	$(AC)_{21}$	−6.7	44.89	80	46
OARFCB128	$(GT)_{22}$	−3.2	10.24	81	76
OARFCB5	$(GT)_{14}$	4.9	24.01	34	75

[a] References can be found on page 855 of Ellegren *et al.* (1997).
[b] $\delta\mu$ in base pairs.
[c] Not in cited reference.

When $\kappa = 1$ the PS/PM model reduces to the original model of Kruglyak *et al.* (1998). The motivation for the change from $\kappa = 1$ to a general κ comes from several studies. Goldstein and Clark (1995) studied 17 microsatellite loci in Drosophila, plotted variance of repeat count *vs.* maximum repeat count, and found (see p. 3884) a straight line that hit zero at seven repeat units. Brinkman *et al.* (1998) studied 10,844 parent/child allelic transfers at nine short tandem repeat loci, finding 23 mutations. There were no mutations at loci with fewer than nine repeats and an approximate linear growth of mutations after that point (see Figure 3 on p. 1412). Finally, Rose and Falush (1998) studied dinucleotide repeats in the yeast genome and compared their frequency with what would be expected on the basis of random chance. The ratio was close to 1 for one to four repeat units and then the logarithm of the ratio increased linearly (see the middle figure on their p. 614).

In formulating the PS/PM model introduced above, our thought experiment consists of picking two nucleotides at random and seeing how many times they are repeated as we scan to the right, so we only need to keep track of the left one-half of a newly imperfect repeat that has been hit by a mutation. This viewpoint, along with appropriate bookkeeping, can be used to fit the model to data and estimate mutation rates (see Kruglyak *et al.* 1998). However, if we are going to look at microsatellites through the eyes of an experimentalist who only tracks the length of PCR-amplified fragments of DNA, we need to define a new process that keeps track of the lengths of all the perfect repeats in an interrupted repeat as a vector $(X_t^1, X_t^2, \ldots X_t^n)$. To explain what we have in mind, consider the concrete sequence

ca ca ca ca ca **CT**|*ca ca* **GA**|*ca ca ca ca ca ca ca* **CG**|.

Here we used lower case letters for the perfect repeat segments to make them more clearly visible. In dividing this imperfect repeat into segments it is convenient to include in each piece the final pair of nucleotides that spoil the pattern. Thus the vertical bars mark the ends of the perfect repeat segments, and we record the state as (6, 3, 8). The reason for this convention will become clear as we develop properties of the model.

In words, in our PCR fragment size model, each of

the lengths of the perfect repeat units X_t^i evolves according to the rules of the PS/PM model. In using this model we are concerned only with the life and death of existing microsatellites, so we ignore the birth of new ones.

PCR model: If the state at time t is $(X_t^1, \ldots X_t^n)$, then there are two types of changes for any of the lengths X_t^i with $1 \le i \le n$.

Proportional slippage: $X_t^i \to X_t^i \pm 1$ at rate $b(X_t^i - \bar{\kappa})^+$.
Point mutation: $(X_t^1, \ldots, X_t^i, \ldots X_t^n) \to (X_t^1, \ldots, X_t^i - y, y, \ldots X_t^n)$ at rate a if $1 \le y \le X_t^i - 1$.

Note that because we include the final imperfect repeat unit in each block, the lengths of the two new pieces created by a point mutation add up to the original length. One final minor point is that since our new bookkeeping system includes the final imperfect repeat, the $\bar{\kappa}$ here should be equal to $\kappa + 1$, where κ is the parameter of the PS/PM model.

Let $L_t = \Sigma_i X_t^i$ be the *total length* of the microsatellite and let

$$A_t = \sum_i (X_t^i - \bar{\kappa})^+$$

be its *activity*; *i.e.*, $2bA_t$ is the rate at which slippage events occur at time t. Since point mutations do not change the total length, and under proportional slippage the microsatellite is equally likely to gain or lose a repeat unit, $EL_t = L_0$. That is, the average value of the length stays constant in time. It is somewhat remarkable that there is a simple formula for the variance of L_t despite the complexity of the PCR model.

Theorem 2: *If the initial activity of the microsatellite is A_0 then at any time $t \ge 0$*

$$\mathrm{var}(L_t) = 2bA_0 t \cdot \left(\frac{1 - e^{-a\bar{\kappa}t}}{a\bar{\kappa}t}\right). \qquad (3)$$

This result is derived in appendix b. *Theorem 2* concerns the variance of the process, not the population samples. The relationship between this quantity and $(\delta\mu)^2$ is that if each sample is of size one then $2\,\mathrm{var}(L_t) = (\delta\mu)^2$. And when the samples are larger than size one and the time to the most recent common ancestor of each sample is much less than the time to the most recent common ancestor of these ancestors, then $2\,\mathrm{var}(L_t) \approx (\delta\mu)^2$. Note that if we let $\beta = 2bA_0$, the initial per locus slippage rate, then the first factor is simply βt, the answer for the SMM. We call the second term in parentheses the *correction factor*, since it indicates how much the variance has been reduced from the prediction of the SMM due to the effect of point mutations. Using the series expansion $e^{-x} = 1 - x + x^2/2 - \cdots$ we see that when $a\bar{\kappa}t$ is small, the correction factor is ≈ 1. In the other direction if $a\bar{\kappa}t = 1$ then the correction factor is $1 - e^{-1} = 0.632$ and a significant reduction has occurred. From this computation, we see that point mutations begin to make a difference when the number of generations $t \approx 1/a\bar{\kappa}$.

To understand the implications of *Theorem 2* we return to our four examples. Thinking of dinucleotide repeats, we assume a point mutation rate of $a = 2 \times 10^{-8}$ per repeat unit (see Drake *et al.* 1998). Based on the work of Rose and Falush (1998) we choose $\bar{\kappa} = 5$, so in all cases $a\bar{\kappa} = 10^{-7}$, and we expect point mutations to have a significant effect after $\sim$10 million generations. In the African *vs.* non-African comparison of human populations, $t = 6000$ generations, so $a\bar{\kappa}t = 6 \times 10^{-4}$ and the correction factor is 0.9997. For humans *vs.* chimpanzees, $t = 250{,}000$, so $a\bar{\kappa}t = 2.5 \times 10^{-2}$ and the correction factor is 0.9876, which is again $\sim$1. Coalescent simulations show that the 95% confidence intervals for $(\delta\mu)^2$ and D_{SW} for the PCR model have changed by <10% from those for the SMM for this example, so the data are not consistent with the PCR model either. (For the simulations we assumed that for all microsatellites their most recent common ancestor was a perfect repeat of length 19 and the per repeat slippage rate was $b = 1.9 \times 10^{-5}$. This assumption corresponds to a per locus slippage rate of the most recent common ancestor microsatellite being $\beta = 5.6 \times 10^{-4}$ as in the SMM.) For cattle *vs.* sheep, $t = 8{,}000{,}000$ generations, so $a\bar{\kappa}t = 0.8$ and the correction factor is 0.688. For *D. melanogaster vs. D. simulans*, $t = 25{,}000{,}000$, so $a\bar{\kappa}t = 2.5$ and the correction factor is 0.367. Coalescent simulations for this example show the 95% confidence intervals for $(\delta\mu)^2$ and D_{SW} are 79.2–342 and 6.34–11.8, whereas the observed statistics were <20 and 3.64, respectively. (For the simulations we assumed that for all microsatellites their most recent common ancestor was a perfect repeat of length 15 and the per repeat slippage rate was $b = 5.0 \times 10^{-7}$.) As predicted by *Theorem 2*, the mean $(\delta\mu)^2$ for the simulations was 184. Figure 1 shows the results of simulating the PCR model to obtain the probability density of the length of a single microsatellite that has evolved for 25 million generations with the parameters of this example. Note that 23% of the microsatellites are longer than 18 repeat units, while only 2 of 186 dinucleotide microsatellites in the original 1-Mb sample of *D. melanogaster* DNA in Kruglyak *et al.* (1998) were this long. The last two examples show that the PCR model leads to significant reductions in predicted values of $(\delta\mu)^2$, but not enough to account for the 13- and 30-fold underestimation observed.

A MODEL WITHOUT POINT MUTATIONS

Our discussion of *Theorem 2* suggests that when $a\bar{\kappa}t$ is small, as is the case for comparisons between human populations or between humans and chimpanzees, we can ignore the effects of point mutations. If we set the point mutation rate $a = 0$ in the PS/PM model and add a superscript 0 to remind ourselves that we have

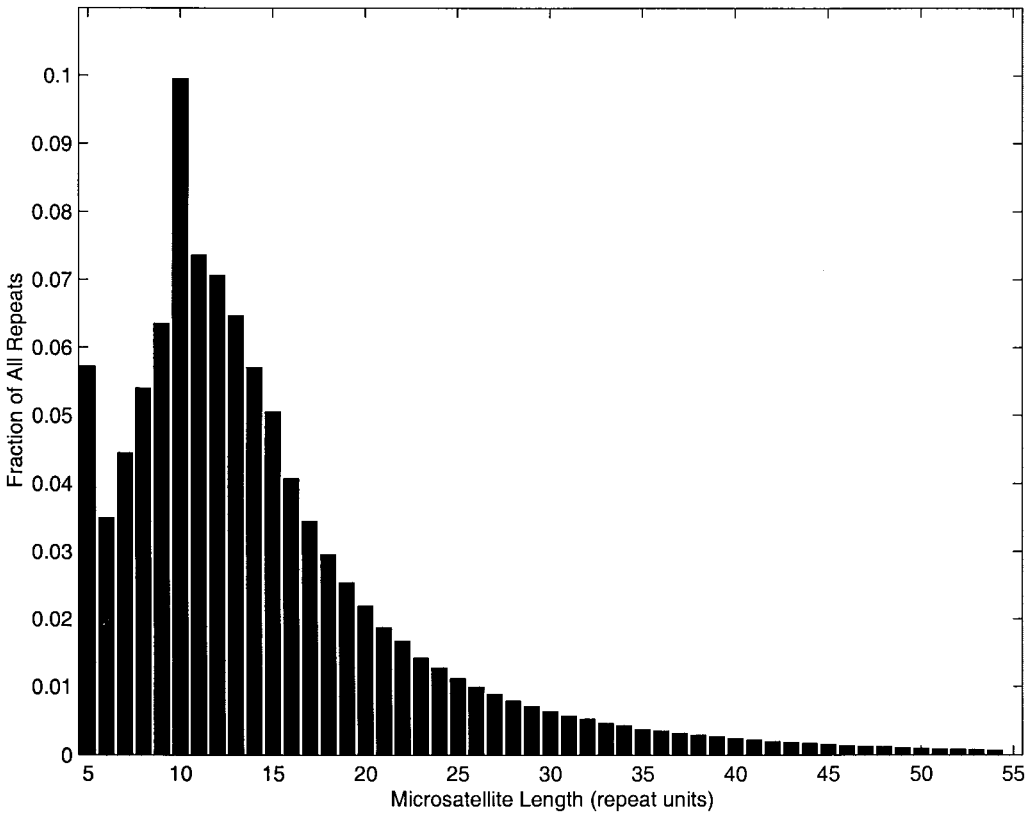

FIGURE 1.—PCR model simulation. Probability density of the length in repeat units of a single microsatellite after 25,000,000 generations is shown. Initially the microsatellite was a perfect repeat with length 15, $\bar{\kappa} = 5$, $a = 2 \times 10^{-8}$, and per repeat slippage rate $b = 5.0 \times 10^{-7}$. These are the parameters we used to study the *D. melanogaster vs. D. simulans* split.

done this, then the activity $A_t^0 = \Sigma_i (X_t^i - \bar{\kappa})^+$ follows a very simple dynamic, which we call the proportional slippage/zero mutations (PS/0M) model.

PS/0M model: *If $A_t^0 = k$ then it changes to $k \pm 1$ at rate bk.* The process A_t^0 jumps from k to $k + 1$ at rate bk, and from k to $k - 1$ at rate bk, and is thus identical to the *binary branching process* Z_t of probability theory, in which Z_t is the number of particles at time t and each particle splits into two or dies at rate b each (see, *e.g.*, ATHREYA and NEY 1972). The following is shown in APPENDIX C:

THEOREM 3: *If we use E_α to denote the expected value for the process starting from $A_0^0 = \alpha$, then*

$$E_\alpha(A_t^0 - \alpha) = 0 \tag{4}$$

$$E_\alpha(A_t^0 - \alpha)^2 = 2\alpha bt \tag{5}$$

$$E_\alpha(A_t^0 - \alpha)^3 = 6\alpha(bt)^2 \tag{6}$$

$$E_\alpha(A_t^0 - \alpha)^4 = 24\alpha(bt)^3 + 12\alpha^2(bt)^2 + 2\alpha bt. \tag{7}$$

In words, proportional slippage is equally likely to increase or decrease the average activity by one, so the average activity does not change in time. The second equation, which can be derived by setting $a = 0$ in *Theorem 2*, says that even though the slippage rate varies in time in the PS/0M model, the variance of a_t^0 is linear in time, just as in the SMM, which has constant slippage rates.

Substantial differences between the PS/0M model and the SMM appear when we look at third and higher moments. The SMM is symmetric so $E(X_t - \ell)^3 = 0$, but as (6) shows, the proportional slippage model has positive skewness. FARRALL and WEEKS (1998) performed an analysis of 4558 *AC* dinucleotide repeat loci assayed in the CEPH pedigrees and found positive skewness in the distribution of microsatellite allele lengths. RUBINSZTEIN *et al.* (1994) had earlier observed this skewness and suggested that it was evidence for "a bias in favor of gains" (see p. 1096 of RUBINSZTEIN *et al.* 1999). However, our results show such a skewed distribution can result from the PS/0M model that has no mutational bias.

Computing the fourth moment reveals another difference between our proportional slippage model and stepwise mutation. In the SMM the difference in microsatellite length, $X_t - Y_t$, between two individuals with a most recent common ancestor t generations ago is the sum of independent random variables. Thus, if t is large, $(X_t - Y_t)/\sqrt{2\beta t}$ has approximately a normal distribution and the kurtosis

$$\mathcal{K} = \frac{E(X_t - Y_t)^4}{[E(X_t - Y_t)^2]^2} \approx 3.$$

In contrast, (7) and (5) show that when $2\ell bt$ is large the kurtosis in the proportional slippage model is

$$\mathcal{K} = \frac{3bt}{\alpha} + 3 + \frac{1}{4\alpha bt} \approx 3\left(1 + \frac{\beta t}{2\alpha^2}\right), \tag{8}$$

where $\beta = 2\alpha b$ is the initial per locus slippage rate.

If the kurtosis is large then the distribution of $X_t - Y_t$ will have a heavy tail and estimation of quantities such as $(\delta\mu)^2$ will be difficult. To see when the kurtosis $\mathcal{K}$ will become large, we note that (8) implies this will occur when $\beta t/2\alpha^2$ is large. To see that this answer is reasonable, note that the expected number of slippage events in t generations is βt and recall that in n steps a random walk typically moves about $\sqrt{n}$ steps. Thus the kurtosis becomes large when the "typical amount of change" in the microsatellite, $(\beta t)^{1/2}$, exceeds its initial activity α and hence there is significant probability of microsatellite death.

In the African/non-African split if we assume $t = 6000$ generations, use an average activity $\alpha = 15$, which corresponds to an average size of 20 repeat units, and set $\beta = 5.6 \times 10^{-4}$ then $\beta t/2\alpha^2 = 0.0075$ so the kurtosis is 3.02. For the human-chimpanzee split, $t = 250{,}000$, $\beta = 5.6 \times 10^{-4}$, and $\alpha = 15$, so $\beta t/2\alpha^2 = 0.311$ and $\mathcal{K} = 3.93$. For *D. melanogaster vs. D. simulans*, $t = 25{,}000{,}000$, $\beta = 10^{-5}$, and $\alpha = 10$, so $\beta t/2\alpha^2 = 1.25$ and $\mathcal{K} = 6.75$. Finally, for cattle *vs.* sheep, we take $t = 8{,}000{,}000$ and $\alpha = 10$, so if we use the estimate $\beta = 6 \times 10^{-5}$ from pig microsatellites, $\beta t/2\alpha^2 = 2.4$ and $\mathcal{K} = 10.2$. One should note, however, that the values for *D. melanogaster vs. D. simulans* and cattle *vs.* sheep are overestimates of the kurtosis since they are based on the proportional slippage model, and our earlier calculations showed that in these cases point mutations had a significant effect on the variance.

To interpret the numerical values of the kurtosis, we observe that if a random variable V has kurtosis $\mathcal{K}$ then

$$\frac{\text{var}\ (V^2)}{(EV^2)^2} = \frac{EV^4 - (EV^2)^2}{(EV^2)^2} = \mathcal{K} - 1$$

and hence the standard deviation of V^2/EV^2 is $\sqrt{\mathcal{K} - 1}$. This shows that if the kurtosis is 3.93 as it is in the human *vs.* chimpanzee comparison, then, instead of the 3 for the normal distribution, the width of confidence intervals will be $\sqrt{(3.93 - 1)/(3 - 1)} = 1.21$ times as large or, equivalently, $1.21^2 = 1.46$ times as much data will be needed to obtain the same accuracy of estimation.

The last conclusion shows that estimates of $(\delta\mu)^2$ under the proportional slippage model are not very much more variable than under the SMM. However, the fluctuations under the SMM in this case are huge. Figure 2 gives a simulation of $(\delta\mu)^2$ under the parameters of the human-chimpanzee split. We used two populations of size $N_e = 10{,}000$ individuals, a divergence time of 250,000 generations, and a mutation rate of 5.6×10^{-4} per locus per generation. It is interesting to compare the simulations where 51% of the $(\delta\mu)^2$ values are >120 with the data in Table 1 where the largest $(\delta\mu)^2$ among 25 loci is 112. Indeed, as (1) predicts, the average value of $(\delta\mu)^2$ in the simulation is $2\beta\tau = 280$.

Further, coalescent simulations of the PS/0M model show that for the human-chimpanzee and the *D. melanogaster-D. simulans* splits the observed $(\delta\mu)^2$ and D_{SW} statistics are not within the expected 95% confidence intervals. These observations suggest that there may be some additional mechanism(s) preventing microsatellites from getting too long.

DEATH OF MICROSATELLITES

Our final topic is to compute the probability of microstellite death in the PS/0M model, *i.e.*, the probability a microsatellite will reach 0 activity in t generations. Since, as noted above, the PS/0M model is equivalent to the binary branching process of probability theory, we can compute not only all of the moments of A_t^0 but also the exact distribution of A_t^0. It follows from results on page 109 of Athreya and Ney (1972) that

Theorem 4: *Letting P_α denote the probability law for the PS/0M model starting from $A_0^0 = \alpha$,*

$$P_\alpha(A_t^0 = 0) = \left(\frac{bt}{1 + bt}\right)^\alpha \tag{9}$$

while for $k \geq 1$,

$$P_\alpha(A_t^0 = k) = \sum_{j=1}^{k\wedge\alpha} \binom{\alpha}{j}\binom{k-1}{j-1}\left(\frac{1}{1+bt}\right)^{2j}\left(\frac{bt}{1+bt}\right)^{\alpha+k-2j}. \tag{10}$$

To apply *Theorem 4* to our four examples, we begin by recalling that $b = \beta/(2\alpha)$, where β is the per locus slippage rate and α is the activity, that is, the length minus $\kappa = 4$. In the African *vs.* non-African human comparison, $t = 6000$, $\beta = 5.6 \times 10^{-4}$, and $\alpha = 15$ (*i.e.*, an average length of 19 repeat units), so (9) shows that the probability of having no activity after $t = 4 \times 10^3$ generations is $(0.11/1.11)^{15} < 10^{-15}$. In the human *vs.* chimpanzee comparison, $t = 250{,}000$, $\beta = 5.6 \times 10^{-4}$, and $\alpha = 15$, so the probability of having no activity after t generations is 0.054.

Figure 3 shows the distribution of the lengths in this case as computed from (10). Note the positive skewness in the distribution as predicted by Theorem 3. Note also that our numerical solution has 17% of the microsatellites having >30 repeat units while only 1 of 205 dinucleotide microsatellites in the original 1-Mb sample of human DNA in Kruglyak *et al.* (1998) has this length. This again suggests that there may be some additional mechanism(s) preventing microsatellites from getting too long.

For the *D. melanogaster vs. D. simulans* and cattle *vs.* sheep comparisons, the PS/0M model overestimates the number of microsatellites with no activity. But this is to be expected since our earlier results show that point mutations have slowed down microsatellite mutation processes over this amount of time.

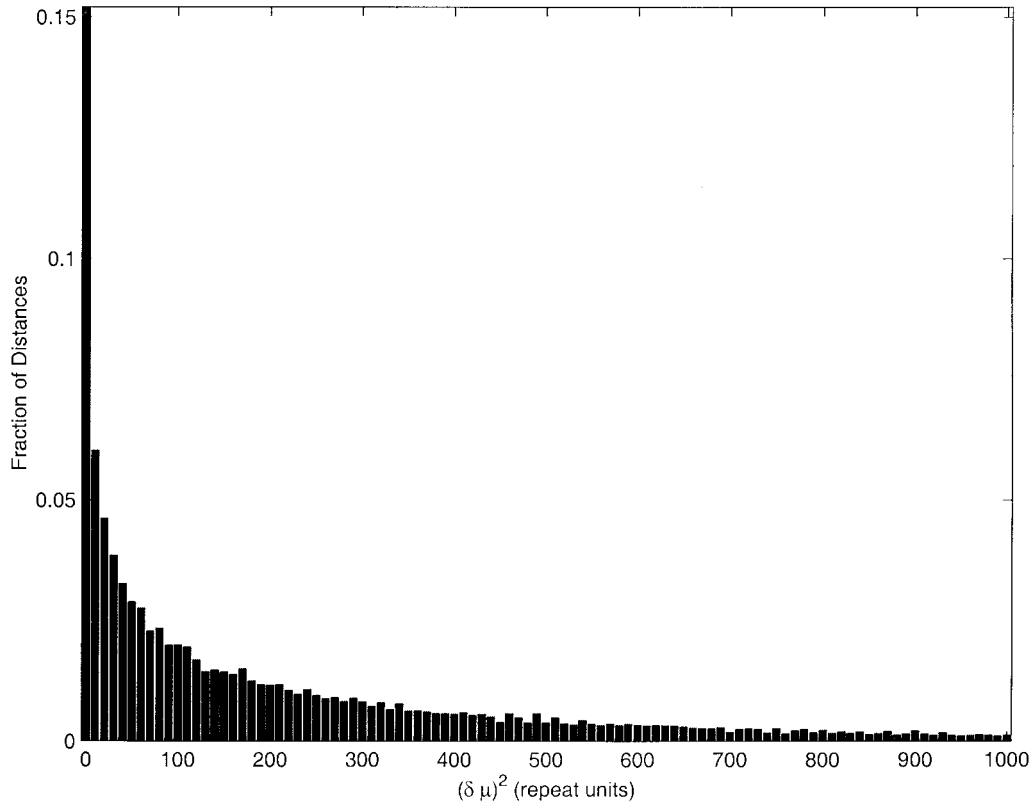

FIGURE 2.—SMM simulation. Probability density of $(\delta\mu)^2$ in repeat units for samples of size 20 from two populations with N_e = 10,000 that diverged τ = 250,000 generations ago is shown; per locus slippage rate $\beta = 5.6 \times 10^{-4}$. These are the parameters we used to study the human-chimpanzee split. In contrast, in Table 1 only 2 of 25 estimates of $(\delta\mu)^2$ are >100.

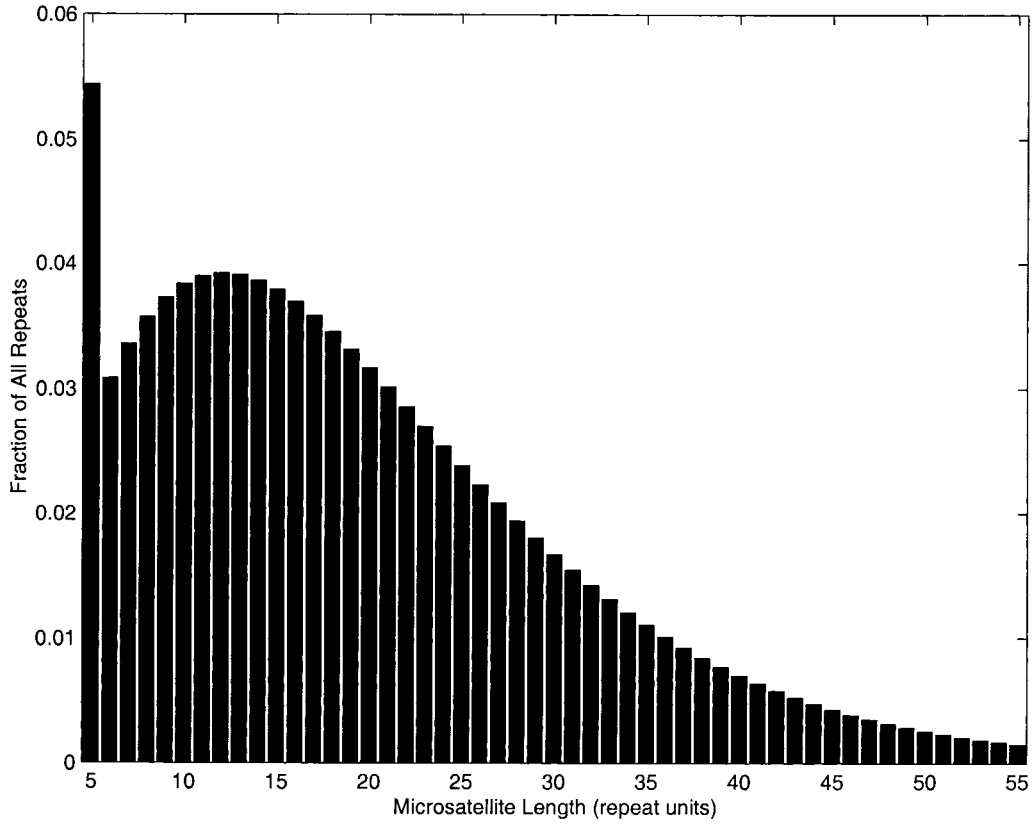

FIGURE 3.—Exact PS/0M calculation. Probability density of the length in repeat units of a single microsatellite after 250,000 generations. Initially the microsatellite has length 19, $\kappa = 4$, and per repeat slippage rate $b = 1.9 \times 10^{-5}$. These are the parameters we used to study the human-chimpanzee split. Note the positive skewness in the distribution.

DISCUSSION

In summary, microsatellite mutation models that incorporate point mutations and proportional slippage events fit the data better than the SMM. However, these two features are not enough to explain, for example, the observation that the genetic distance statistics $(\delta\mu)^2$ and D_{SW} tend to underestimate divergence times and have more difficulty with more distant comparisons. This and other evidence we presented suggests that long microsatellites are more likely to become shorter rather than longer when a mutation occurs.

One possibility is that there is selection against longer alleles. This effect is clearly noticeable in microbial genomes where selection for small genome size appears to cause microsatellites to be much shorter than they would be by chance alone (see Field and Wills 1998). The low frequency of di- and tetranucleotide repeats and the enhanced frequency of trinucleotide repeats in coding sequence in yeast (see Young *et al.* 2000) are another sign of the effects of selection.

Upper limits on allele sizes are a severe form of selection that has been incorporated in some models (*e.g.*, Nauta and Weissing 1996; Feldman *et al.* 1997; Pollock *et al.* 1998; Stefanini and Feldman 2000). This simple approach allows one to develop a detailed theory. However, it is not clear what biological mechanism sets an absolute upper bound on the length of all *CA* repeats. A second approach (see Garza *et al.* 1995; Zhivotovsky *et al.* 1997; Zhivotovsky 1999) is that there is a mutational bias such that alleles of large size mutate preferentially to alleles of smaller size. Wierdl *et al.* (1997) observed this where they inserted *GT* repeats of various sizes into the coding sequence of a yeast gene. In *D. melanogaster* mutation-accumulation lines Harr and Schlötterer (2000) observed that for long microsatellites, although the number of upward and downward mutations was identical, the size of the downward mutations was larger than the size of the upward ones. Two recent studies of mutations observed in human pedigrees also support this notion. Ellegren (2000) found a weak but statistically significant negative relationship between the magnitude of mutation and standardized allele size. Xu *et al.* (2000) examined 236 mutations at 122 tetranucleotide repeat loci and found that the rate of expansion mutations is roughly constant but contraction mutations increase with length. It will be of interest to explore the dependence of the mutational bias on length and to incorporate these effects into our proportional slippage/point mutation model to develop a more accurate model of microsatellite evolution. If mutational bias appears only when microsatellites are fairly long this bias should have only a limited impact on our previous model-based estimates of slippage rates (see Kruglyak *et al.* 1998, 2000).

We thank two anonymous reviewers, Tessa Bauer DuMont, Jennifer Calkins, Semyon Kruglyak, Willie Swanson, and Todd Vision for their many helpful comments. This work was partially supported by National Institutes of Health (NIH) grant GM36431 to C.F.A., NIH grant GM36431-14S1 to C.F.A. and R.T.D., and National Science Foundation grant DMS9877066 to R.T.D.

LITERATURE CITED

Athreya, K. B., and P. E. Ney, 1972 *Branching Processes.* Springer, New York.

Bowcock, A. M., R. A. Linares, J. Tomfohrde, E. Munch, J. R. Kidd *et al.*, 1994 High resolution of human evolutionary trees with polymorphic microsatellites. Nature **368:** 455–457.

Brinkman, B., M. Klintschar, F. Neuhuber, J. Hühne and B. Rolf, 1998 Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am. J. Hum. Genet. **62:** 1408–1413.

Deka, R., M. D. Shriver, L. M. Yu, L. Jin, C. E. Aston *et al.*, 1994 Conservation of human chromosome 13 polymorphic microsatellite $(CA)_n$ repeats in chimpanzees. Genomics **22:** 226–230.

Drake, J. W., B. Charlesworth, D. Charlesworth and J. M. Crow, 1998 Rates of spontaneous mutation. Genetics **148:** 1667–1686.

Ellegren, H., 1995 Mutation rates at porcine microsatellite loci. Mamm. Genome **6:** 376–377.

Ellegren, H., 2000 Heterogeneous mutation processes in human microsatellite DNA sequences. Nat. Genet. **24:** 400–402.

Ellegren, H., S. Moore, N. Robinson, K. Byrne, W. Ward *et al.*, 1997 Microsatellite evolution—a reciprocal study of repeat lengths at homologous loci in cattle and sheep. Mol. Biol. Evol. **14:** 854–860.

Farrall, M., and D. M. Weeks, 1998 Mutational mechanisms for generating microsatellite allele-frequency distributions: an analysis of 4,558 markers. Am. J. Hum. Genet. **62:** 1260–1262.

Feldman, M. W., A. Bergman, D. D. Pollock and D. B. Goldstein, 1997 Microsatellite genetic distances with range constraints: analytic description and problems of estimation. Genetics **145:** 207–216.

Field, D., and C. Wills, 1998 Abundant microsatellite polymorphism in *Saccharomyces cerevisiae*, and the different distribution in eight prokaryotes and *S. cerevisiae*, result from strong mutation pressures and a variety of selective forces. Proc. Natl. Acad. Sci. USA **95:** 1647–1652.

Garza, J. C., M. Slatkin and N. B. Freimer, 1995 Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. Mol. Biol. Evol. **12:** 594–603.

Goldstein, D. B., and A. G. Clark, 1995 Microsatellite variation in North American populations of *Drosophila melanogaster*. Nucleic Acids Res. **23:** 3882–3886.

Goldstein, D. B., and D. D. Pollock, 1997 Launching microsatellites: a review of mutation processes and methods of phylogenetic inference. J. Hered. **88:** 335–342.

Goldstein, D. B., A. Ruiz-Linares, L. L. Cavalli-Sforza and M. W. Feldman, 1995a An evaluation of genetic distances for use with microsatellite loci. Genetics **139:** 463–471.

Goldstein, D. B., A. Ruiz-Linares, L. L. Cavalli-Sforza and M. W. Feldman, 1995b Genetic absolute dating based on microsatellites and modern human origins. Proc. Natl. Acad. Sci. USA **92:** 6723–6727.

Goldstein, D. B., G. W. Roemer, D. A. Smith, D. A. Reich, A. Bergman *et al.*, 1999 The use of microsatellite variation to infer population structure in a natural model system. Genetics **151:** 797–801.

Gonser, R., P. Donnelly, G. Nicholson and A. Di Rienzo, 2000 Microsatellite mutations and inferences about human demography. Genetics **154:** 1793–1807.

Goodman, M., C. A. Porter, J. Czelusinak, S. L. Page, H. Schneider *et al.*, 1998 Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. Mol. Phylogenet. Evol. **9:** 585–598.

Harr, B., and C. Schlötterer, 2000 Long microsatellite alleles in *Drosophila melanogaster* have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. Genetics **155:** 1213–1220.

Harr, B., S. Weiss, J. R. David, G. Brem and C. Schlötterer, 1998

A microsatellite based multi-locus phylogeny of the *Drosophila melanogaster* species complex. Curr. Biol. **8:** 1183–1186.

Hey, J., and R. M. Kliman, 1993 Population genetics and phylogenetics of DNA sequence variation at multiple loci within the *Drosophila melanogaster* species complex. Mol. Biol. Evol. **10:** 804–822.

Hutter, C. M., M. D. Schug and C. F. Aquadro, 1998 Molecular variation in *Drosophila melanogaster* and *Drosophila simulans*: a reciprocal test of the ascertainment bias hypothesis. Mol. Biol. Evol. **15:** 1620–1636.

Irwin, S. D., K. Wetterstrand, C. M. Hutter and C. F. Aquadro, 1998 Genetic variation and differentiation at microsatellite loci in *Drosophila simulans*: evidence for founder effects in new world populations. Genetics **150:** 777–790.

Kruglyak, S., R. Durrett, M. D. Schug and C. F. Aquadro, 1998 Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc. Natl. Acad. Sci. USA **95:** 10774–10778.

Kruglyak, S., R. Durrett, M. D. Schug and C. F. Aquadro, 2000 Distribution and abundance of microsatellites in the yeast genome can be explained by a balance between slippage events and point mutations. Mol. Biol. Evol. **17:** 1210–1219.

Kumar, S., and S. Blair Hedges, 1998 A molecular timescale for vertebrate evolution. Nature **392:** 917–920.

Nauta, M. J., and F. J. Weissing, 1996 Constraints on allele size at microsatellite loci: implications for genetic differentiation. Genetics **143:** 1021–1032.

Ohta, T., and M. Kimura, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22:** 201–204.

Pollock, D. D., A. Bergman, M. W. Feldman and D. B. Goldstein, 1998 Microsatellite behavior with range constraints: parameter estimation and improved distances for use in phylogenetic reconstruction. Theor. Popul. Biol. **53:** 256–271.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun and M. W. Feldman, 1999 Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol. Biol. Evol. **16:** 1791–1798.

Reich, D. E., and D. B. Goldstein, 1998 Genetic evidence for a Paleolithic human population expansion in Africa. Proc. Natl. Acad. Sci. USA **95:** 8119–8123.

Rose, O., and D. Falush, 1998 A threshold size for microsatellite expansion. Mol. Biol. Evol. **15:** 613–615.

Roy, M. S., E. Geffen, D. Smith, E. Ostrander and R. K. Wayne, 1994 Patterns of differentiation and hybridization in North American wolflike Canids. Mol. Biol. Evol. **11:** 553–570.

Rubinsztein, D. C., W. Amos, J. Leggo, S. Goodburn, R. S. Ramesar *et al.*, 1994 Mutational bias provides a model for the evolution of Huntington's disease and predicts a general increase in disease prevalence. Nat. Genet. **7:** 525–530.

Rubinsztein, D. C., J. Leggo and W. Amos, 1995 Microsatellites evolve more rapidly in humans than in chimpanzees. Genomics **30:** 610–612.

Rubinsztein, D. C., W. Amos and G. Cooper, 1999 Microsatellite and trinucleotide repeat evolution: evidence for mutational bias and different rates of evolution in different lineages. Philos. Trans. R. Soc. Ser. B **354:** 1095–1099.

Ruiz-Linares, A., D. Ortiz-Barrientos, M. Rigueroa, N. Mesa, J. G. Munera *et al.*, 1999 Microsatellites provide evidence for Y chromosome diversity among founders of the New World. Proc. Natl. Acad. Sci. USA **96:** 6312–6317.

Schug, M. D., T. F. Mackay and C. F. Aquadro, 1997 Low mutation rates of microsatellite loci in Drosophila melanogaster. Nat. Genet. **15:** 99–102.

Schug, M. D., C. M. Hutter, K. A. Wetterstrand, M. S. Gaudette, T. F. C. Mackay *et al.*, 1998 The mutation rates of di-, tri-, and tetranucleotide repeats in *Drosophila melanogaster*. Mol. Biol. Evol. **15:** 1751–1760.

Shriver, M. D., L. Jin, E. Boerwinkle, R. Deka and R. Chakraborty, 1995 A novel measure of genetic distance for highly polymorphic tandem repeat loci. Mol. Biol. Evol. **12:** 914–920.

Stefanini, F. M., and M. W. Feldman, 2000 Bayesian estimation of range for microsatellite loci. Genet. Res. **75:** 167–177.

Underhill, P. A., L. Jin, R. Zemans, P. J. Ofener and L. L. Cavalli-Sforza, 1996 A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. Proc. Natl. Acad. Sci. USA **93:** 196–200.

Wetterstrand, K. A., 1997 Microsatellite polymorphism and divergence in worldwide populations of D. melanogaster and D. simulans. M.S. Thesis, Cornell University, Ithaca, NY.

Wierdl, M., M. Dominska and T. D. Petes, 1997 Microsatellite instability in yeast: dependence on the length of the microsatellite. Genetics **146:** 769–779.

Xu, X., M. Peng, Z. Fang and X. Xu, 2000 The direction of microsatellite mutation is dependent upon allele length. Nat. Genet. **24:** 396–399.

Young, E. T., J. S. Sloan and K. Van Riper, 2000 Trinucleotide repeats are clustered in regulatory genes in *Saccharomyces cerevisiae*. Genetics **154:** 1053–1068.

Zhivotovsky, L. A., 1999 A new genetic distance with application to constrained variation at microsatellite loci. Mol. Biol. Evol. **16:** 467–471.

Zhivotovsky, L. A., M. W. Feldman and S. A. Grischekin, 1997 Biased mutations and microsatellite variation. Mol. Biol. Evol. **14:** 926–933.

Communicating editor: S. Tavaré

APPENDIX A

To compute D_{SW} and $(\delta\mu)^2$, let $\xi_1, \xi_2, \ldots$ be independent and ± 1 with probability 1/2 each and let $S_n = \xi_1 + \cdots + \xi_n$. The ξ_i are the results of the various slippage events and S_n is the total change after n events. If each population consists of N diploid individuals then the number of slippage events, U, before the ancestors of X and X' (or of Y and Y') coalesce has a shifted geometric distribution with success probability $p = 1/(4\beta N + 1)$; that is, we have

$$P(U \geq n) = (1 - p)^n \quad \text{for } n = 0, 1, 2, \ldots \quad \text{(A1)}$$

and hence $EU = 1/p - 1 = 4\beta N$. If the two populations diverged τ generations ago then the number of slippage events before the ancestors of X and Y coalesce has the same distribution as $U + V$ where V gives the number that occurs during the first τ generations. V has a Poisson with mean $2\beta\tau$; that is, $P(V = m) = e^{-2\beta\tau}(2\beta\tau)^m/m!$ for $m = 0, 1, 2, \ldots$.

In the case of $(\delta\mu)^2$ breaking things down according to the values of U and V and using the fact that $ES_n^2 = n$ we have

$$E(X - Y)^2 - (E|X - X'|^2 + E|Y - Y'|^2)/2$$
$$= E(S_{U+V})^2 - E(S_U)^2 = E(U + V) - EU = EV = 2\beta\tau.$$

The computation for D_{SW} starts out the same,

$$E|X - Y| - \frac{E|X - X'| + E|Y - Y'|}{2} = E|S_{U+V}| - E|S_U|, \quad \text{(A2)}$$

but the computation of $E|S_U|$ and $E|S_{U+V}|$ is more complicated. To begin, we note that since $P(\xi_i = 1) = P(\xi_i = -1) = \frac{1}{2}$, considering two cases $S_{n-1} = 0$ and $S_{n-1} \neq 0$ we have

$$E|S_n| - E|S_{n-1}| = P(S_{n-1} = 0). \quad \text{(A3)}$$

Since S_n alternates between even and odd values, it can

be 0 only after an even number of steps, and simple path counting gives

$$P(S_{2k} = 0) = \binom{2k}{k}\frac{1}{2^{2k}} = \frac{(2k)!}{k!k!}\cdot\frac{1}{2^{2k}},$$

where $\binom{n}{m}$ is the usual binomial coefficient, which gives the number of ways of choosing m things out of a set of n and $k! = 1\cdot 2\cdots k$.

Let T be a random time, *e.g.*, U or $U + V$. Writing $1_{(T\geq n)}$ for the function that is 1 if $T \geq n$ and 0 otherwise, we have $|S_T| = \Sigma_{n=1}^{\infty}(|S_n| - |S_{n-1}|)\cdot 1_{(T\geq n)}$, so taking expected values and using the independent of T and S_n with (A3) we have

$$E|S_T| = \sum_{n=1}^{\infty} P(S_{n-1} = 0)\cdot P(T \geq n). \qquad \text{(A4)}$$

Changing variables $n = 2k + 1$ and using (A1) shows that in the case $T = U$ we have

$$E|S_U| = \sum_{k=0}^{\infty}\frac{(2k-1)(2k-3)\cdots 3\cdot 1}{2^k\, k!}(1-p)^{2k+1}.$$

Differentiating the function $f(x) = (1 - x)^{-1/2}$ we find its kth derivative is

$$f^{(k)}(x) = \frac{(2k-1)(2k-3)\cdots 3\cdot 1}{2^k}\cdot(1-x)^{-(2k+1)/2}.$$

Recalling the formula for the Taylor series of a function f,

$$f(x) - f(0) = \sum_{k=1}^{\infty} f^{(k)}(0)\cdot\frac{x^k}{k!},$$

and comparing with the formula for $E|S_U|$ we have that

$$E|S_U| = \frac{1}{(1-(1-p)^2)^{1/2}} = \frac{4\beta N}{\sqrt{8\beta N + 1}}, \qquad \text{(A5)}$$

the last equality following from $1 - p = 4\beta N/(4\beta N + 1)$.

In the case $T = U + V$, $P(U + V \geq 2k + 1)$ is given by

$$P(V \geq 2k+1) + \sum_{j=0}^{2k} e^{-2\beta\tau}\frac{(2\beta\tau)^j}{j!}\cdot(1-p)^{2k+1-j}. \qquad \text{(A6)}$$

Together with (A2), (A4), and (A5) this can be used to compute $E|S_{U+V}|$ numerically, but it does not seem possible to sum the series to get an exact solution. To begin to derive an approximation for $E|S_{U+V}|$, we note that if n is large, $S_n/\sqrt{n} \approx \chi$, where χ has a normal distribution, so $E|S_n| \approx n^{1/2}E|\chi| = (2n/\pi)^{1/2}$. If we let $g(n) = E|S_n|$ then $E|S_{U+V}| = Eg(U + V)$. Writing $W = U + V$ to simplify formulas and expanding in Taylor series,

$$g(W) = g(EW) + g'(EW)\cdot(W - EW) + g''(EW)\cdot\frac{(W-EW)^2}{2} + \cdots.$$

Taking expected value of each side,

$$Eg(W) \approx g(EW) + g''(EW)\cdot\frac{\mathrm{var}(W)}{2}. \qquad \text{(A7)}$$

Our next goal is to show that if $2\beta\tau$ is large and $\tau \geq N$ we can drop the second term from (A7) to end up with

$$E|S_{U+V}| \approx g(EU + EV) \approx \sqrt{\frac{2}{\pi}}\cdot(2\beta\tau + 4\beta N)^{1/2},$$

which with (A5) gives (2). To do this we note that for large x, $g(x) \approx Cx^{1/2}$ so $g''(x) \approx (C/4)x^{-3/2}$ and the ratio of the two terms is

$$\frac{g''(EW)\cdot\mathrm{var}(W)/2}{g(EW)} \approx \frac{1}{8}\cdot\frac{\mathrm{var}(W)}{(EW)^2}.$$

To see when this will be small we use formulas for the mean and variance of the Poisson and geometric distributions to conclude

$$EW = 2\beta\tau + (1-p)/p = 2\beta\tau + 4\beta N$$

$$\mathrm{var}(W) = 2\beta\tau + (1-p)/p^2 = 2\beta\tau + 4\beta N(1 + 4\beta N)$$

since $p = 1/(1 + 4\beta N)$. From this we see that the ratio of interest is

$$\frac{1}{8}\cdot\frac{2\beta\tau + 4\beta N + (4\beta N)^2}{(2\beta\tau)^2 + (4\beta\tau + 4\beta N)(4\beta N)}. \qquad \text{(A8)}$$

If $2\beta\tau$ is large and $\tau \geq N$ we can drop the $2\beta\tau + 4\beta N$ from the numerator and then divide top and bottom by $4\beta^2$ to see that the last expression is

$$\leq\frac{1}{8}\cdot\frac{(2N)^2}{(\tau + 2N)^2} \leq \frac{1}{18} \qquad \text{(A9)}$$

when $\tau \geq N$. In words the error we make by neglecting the second term in (A7) is at most 5.5%, and as τ/N increases, the error will become smaller.

APPENDIX B

Writing $\boldsymbol{X}_t = (X_t^1, \ldots X_t^n)$ and e^i for the vector that has one in the ith place and zero otherwise, it follows from the definition of the PCR model and the Kolmogorov differential equations for the associated Markov chain that

$$\frac{d}{dt}Ef(\boldsymbol{X}_t) = E[\mathcal{A}_1 f(\boldsymbol{X}_t) + \mathcal{A}_2 f(\boldsymbol{X}_t)], \qquad \text{(B1)}$$

where the two parts of the right-hand side correspond to proportional slippage and point mutation events:

$$\mathcal{A}_1 f(\boldsymbol{X}) = b\sum_i (X^i - \bar{\kappa})^+[f(\boldsymbol{X} + e^i) - 2f(\boldsymbol{X}) + f(\boldsymbol{X} - e^i)]$$

$$\mathcal{A}_2 f(\boldsymbol{X}) = a\sum_i \sum_{y=1}^{X^i-1}[f(X^1, \ldots, X^i - y, y, \ldots, X^n) - f(\boldsymbol{X})].$$

If we let $g_1(\boldsymbol{X}_t) = \Sigma_i X_t^i$ be the total length then $\mathcal{A}_2 g_1(\boldsymbol{X}_t) =$

0 since point mutations do not change the length and $\mathcal{A}_1 g_1(\boldsymbol{X}_t) = 0$ by computation so

$$Eg_1(\boldsymbol{X}_t) = g_1(\boldsymbol{X}_0). \tag{B2}$$

To prepare for the computation of the variance, let $h(\boldsymbol{X}_t, j) = \Sigma_i (X_t^i - j)^+$, where $j \le \bar{\kappa}$. Since proportional slippage is a fair game and no slippage occurs for pieces of length j, $\mathcal{A}_1 h = 0$. To compute the other term, we note that

$$\mathcal{A}_2 h(\boldsymbol{X}, j) = a \sum_i \sum_{y=1}^{X_t^i - 1} [(X_t^i - y - j)^+ + (y - j)^+ - (X_t^i - j)^+]$$

$$= a \sum_i [-(X_t^i - 1)(X_t^i - j)^+ + \sum_{z=1}^{(X_t^i - 1 - j)^+} 2z].$$

To evaluate the sum we use the identity $\Sigma_{z=1}^k 2z = k(k + 1)$ to conclude

$$\sum_{z=1}^{(X_t^i - 1 - j)^+} 2z = (X_t^i - 1 - j)^+ [(X_t^i - 1 - j)^+ + 1]$$

$$= (X_t^i - 1 - j)(X_t^i - j)^+.$$

To check the second equality note that if $X_t^i \le j + 1$ it says $0 = 0$, while for $X_t^i > j + 1$ the positive parts are irrelevant. Combining our computations,

$$\mathcal{A}_2 h(\boldsymbol{X}_t, j) = a \sum_i - j(X_t^i - j)^+ = - aj\, h(\boldsymbol{X}, j).$$

Using this with (B1) and solving the differential equation we have

$$Eh(\boldsymbol{X}_t, j) = h(\boldsymbol{X}_0, j) e^{-ajt}. \tag{B3}$$

Turning now to $g_2(\boldsymbol{X}_t) = (\Sigma_i X_t^i)^2$, we have $\mathcal{A}_2 g_2(\boldsymbol{X}_t) = 0$ since point mutations do not change the total length. For the other term we note that if $L_t = \Sigma_i X_t^i$ then $(L_t + 1)^2 - 2L_t^2 + (L_t - 1)^2 = 2$ so $\mathcal{A}_1 g_2(\boldsymbol{X}_t) = 2bh(\boldsymbol{X}_t, \bar{\kappa})$. Using (B3) we have $d/dt\, Eg_2(\boldsymbol{X}_t) = 2bh(\boldsymbol{X}_0, \bar{\kappa}) e^{-a\bar{\kappa}t}$. Integrating gives

$$\operatorname{var}(L_t) = Eg_2(\boldsymbol{X}_t) - g_2(\boldsymbol{X}_0) = \frac{2bh(\boldsymbol{X}_0, \bar{\kappa})}{a\bar{\kappa}}(1 - e^{-a\bar{\kappa}t}). \tag{B4}$$

APPENDIX C

From the definition of the PS/0M model and the Kolmogorov differential equations, it follows that

$$\frac{d}{dt} E_\ell f(Z_t) = E\mathcal{A} f(Z_t), \tag{C1}$$

where $\mathcal{A}f(x) = bx[f(x + 1) - 2f(x) + f(x - 1)]$. If we let $f_1(x) = x$ then $\mathcal{A}f_1(k) = 0$ so $E_\ell Z_t$ is constant and

$$E_\ell Z_t = \ell. \tag{C2}$$

Letting $f_2(k) = (k - \ell)^2$ we have $\mathcal{A}f_2(k) = bk \cdot 2$ so using (C2) and integrating

$$E_\ell (Z_t - \ell)^2 = 2b \int_0^t E_\ell Z_s\, ds = 2b\ell t. \tag{C3}$$

Letting $f_3(k) = (k - \ell)^3$ we have $\mathcal{A}f_3(k) = bk \cdot 6\,(k - \ell)$ so

$$E_\ell (Z_t - \ell)^3 = 6b \int_0^t E_\ell (Z_s (Z_s - \ell))\, ds.$$

To compute the right-hand side we note that $E_\ell(Z_s(Z_s - \ell)) = E_\ell(Z_s - \ell)^2 + \ell E_\ell(Z_s - \ell) = 2b\ell s$ by (C3) and (C2), so we have

$$E_\ell (Z_t - \ell)^3 = 6b^2 \ell t^2. \tag{C4}$$

Letting $f_4(k) = (k - \ell)^4$ we have $\mathcal{A}f_4(k) = bk \cdot [12(k - \ell)^2 + 2]$ so

$$E_\ell (Z_t - \ell)^4 = b \int_0^t E_\ell [12 Z_s (Z_s - \ell)^2 + 2Z_s]\, ds.$$

To compute the right-hand side we note that $E_\ell(Z_s(Z_s - \ell)^2) = E_\ell(Z_s - \ell)^3 + \ell E_\ell(Z_s - \ell)^2 = 6b^2\ell s^2 + 2b\ell^2 s$ by (C4) and (C3), so we have

$$E_\ell (Z_t - \ell)^4 = 24 b^3 \ell t^3 + 12 b^2 \ell^2 t^2 + 2b\ell t. \tag{C5}$$