Microsatellite Evolution: Markov Transition Functions for a Suite of Models

Joseph C. Watkins

Department of Mathematics, University of Arizona, Tucson, Arizona 85721 USA

Abstract

This paper takes from the collection of models considered by Whittaker et. al. (2003) derived from direct observation of microsatellite mutation in parent-child pairs and provides analytical expressions for the probability distributions for the change in number of repeats over any given number of generations. The mathematical framework for this analysis is the theory of Markov processes. We find these expressions using two approaches, approximating by circulant matrices and solving a partial differential equation satisfied by the generating function. The impact of the differing choice of models is examined using likelihood estimates for time to most recent common ancestor. The analysis presented here may play a role in elucidating the connections between these two approaches and shows promise in reconciling differences between estimates for mutation rates based on Whittaker's approach and methods based on phylogenetic analyses.

Key words and phrases: microsatellites, Markov process, generating functions

1. Introduction

Microsatellites are portions of the genome consisting of a sequence of repeats of a given string of nucleotides. These strings generally have lengths from one to six bases. Such a structure has suggested the alternate names *short* or *simple tandem repeats* (STRs), *simple sequence repeats* (SSRs), *simple sequence length polymorphisms* (SSLPs) and *variable number tandem repeats* (VNTRs). The typical microsatellite mutations, and the only ones we shall consider here, are those which result in a change in the number of repeats from parent to offspring. Because of their abundance in the genome and their rapid mutation rates, microsatellites have received much attention. (For an overview, see the book edited by Goldstein and Schotterer, 1999. For a popular account, see Moxon and Wills, 1999.)

For the purposes of population genetics, many microsatellites offer what are considered to be independent neutrally mutating segments of DNA. Moreover, in cases in which the nucleotide sequences in the flanking regions of the microsatellite are known, specific primers can be designed to amplify the microsatellite by the polymerase chain reaction (PCR). Thus, data are relatively easy to collect and the evolution of the microsatellites provides several modeling advantages. The models we develop will not take into account any mutations that

might occur during the collection of data. Such issues are considered, e.g., by Lai and Sun (2004).

The mathematical modeling of microsatellites has a long history. The classical 1973 symmetric single step mutation model of Ohta and Kimura was applied to microsatellites soon after their discovery in the late 1980's. In 1994, DiRenzo et. al. contemplated multiple step mutations through their "two phase model" prescribing a geometric random variable for the distribution of multirepeat mutations. Fu and Chakraborty (1998) consider inference for multiple step models. The notion that longer microsatellites mutated more rapidly and that microsatellites could undergo point mutation (Ellegren, 2000) led to the proportional slippage model of Kruglyak et. al. (1998) with extensions in 2001 by Calabrese et. al. Asymmetry in mutation rates was modeled by Walsh in 1987 as a linear birth and death process. Calabrese and Durrett (2003) generalize this model by having a positive minimum microsatellite repeat number and by considering quadratic mutation rates. Garza et. al. (1995) and Zhivotovsky et. al. (1997) consider models in which mutational bias depends on repeat number.

A review of microsatellite models can be found in Calabrese and Sainudiin (2004). A detailed analysis of these and other models was undertaken by Sainudiin et. al. (2004) in the context of the split time for humans and chimpanzees.

Whittaker et. al. (2003) note that the "most straightforward and conclusive method by which to study mutation is direct observation of allele transmissions in parent-child pairs, ..." They carry through with this method by analyzing 118,866 parent-offspring transmissions of AC microsatellites, finding 53 length mutations. Their statistical methods are solid - using log-likelihood tests for nested models augmented by Akaike's information criteria to assist in comparisons for non-nested models.

The Whittaker et. al. approach to model building is an ideal basis for the next steps taken in this paper. If time is measured in units that are equal to the per generation probability of a mutation, then a single generation is a small discrete time step. Thus, even though mutation probabilities for microsatellites are much higher than for other mutational types, they are sufficiently small (4.5×10^{-4} in the study above) so that the use of continuous time stochastic models provides very accurate answers. The current understanding of microsatellite evolution places it under the circumstances in which the tools of a particular class of stochastic models, namely, time homogeneous Markov processes, apply. Specifically, the future chances of mutational events depend on the past history only through the current length of the microsatellite. The Whittaker et. al. data on parent-child pairs provide the short term or "infinitesimal" criterion needed to characterize the Markov process under consideration. The theory of Markov processes is thus available to determine the long term probabilities of microsatellite length changes. This, in turn, gives us the building blocks for the likelihood functions that form the bases for the next stage of analysis.

In the next section, we describe in some detail a subset of the suite of parameterized models that were considered by Whittaker et. al. leading up to the case in which mutation

rate increases geometrically with microsatellite length.

The analysis in subsequent sections will lead to explicit formulas for Markov transition functions, which, naturally, have more complex expressions with increasingly complex mutation models. The choice of model for any particular inferential question necessarily involves assessing the increased value of that choice weighed against the incurred computational overhead. To begin to elucidate the impact of the model choice on estimation values, we focus our examples on the simple question of the time to most recent common ancestor for a pair of individuals.

Any of the Markov transition functions can be incorporated, at appropriate points, to improve on sophisticated (e.g., coalescent based) inferential frameworks using full or approximate likelihood approaches and Bayesian techniques. For example, the Markov chain Monte Carlo approach introduced by Wilson and Balding (1998) can be adapted to include the models developed here by replacing the transitions function in their equation (3) with the transitions functions developed in this work. The transition functions can also be incorporated in the extension to the Bayesian approach undertaken by Wilson, Weale, and Balding (2003). The importance sampling strategy introduced by Stephens and Donnelly (2000) and extended to microsatellites by DeIorio et. al. (2005) do not directly use Markov transitions functions. Nevertheless, the techniques developed in this paper lead to insights in building the importance sampler for the models presented here.

Finally, in the discussion, we explore the extent of the applicability of the techniques developed here and note some of the interesting biophysical properties of the models favored by the Whittaker et. al. analysis.

2. The Suite of Models

In setting the mutation models, the first questions concern the step size - either only single steps are considered or multiple steps are possible. The multiple step model considered by Whittaker et. al. has microsatellite repeat numbers that change according to a geometric random variable. Call α the parameter for this geometric distribution.

The second set of questions concerns the dependence on microsatellite length for the chance of a mutational event from a parent to an offspring. The Whittaker et. al. analysis selects a model in which the intensity of mutations increases geometrically with repeat number. Let β denote the parameter for this intensity for mutation.

Whittaker et. al. do analyze some additional possibilities for microsatellite evolution. For example, they incorporate asymmetric mutation direction by adding a parameter p, the probability that a mutation increases length. Thus, the value $p = 1/2$ takes us back to the symmetric case. Whittaker et. al. consider addition models beyond those described above. We shall postpone discussion of these models until the final section.

This provides, for each model under consideration, a straightforward aim: Find the Markov transition function $P_{j,j+n}(\tau)$ - given an individual who has a microsatellite with

repeat number j, find, for progeny τ time units later, the probability that the number of repeats changes by an amount n. Note that n can be either positive or negative. The paper develops these notions by considering the models according to their increasing level of complexity. We use this as an opportunity to introduce the additional mathematical tools - approximation by circulant matrices and use of eigenvalues and eigenvectors, or generating functions, complex variables and the residue calculus - as needed. The approach of using generating functions in single step mutation models back at least Wehrbahn in 1975. For those familiar with such tools, the paper can be read by skimming immediately to the most complex model and retrieving all of the previous results as special cases of this comprehensive model.

Nearly every model has limits to applicability and this suite of models is no exception. For example, the models developed in this paper fail to give reliable results and become computational infeasible for times that are too long. From the biological perspective, at sufficiently long time scales, mutational events not considered in this work begin to claim a more prominent role. (See Kruglyak et. al., 1998.) From the mathematical perspective, the models considered will allow for microsatellite repeat numbers to take any integer value. This is similar to using a normal approximation for a manifestly positive quantity. This is used because the normal approximation in that case and the repeat number in the cases under consideraton here yield exceedingly small probabilities to negative values. If one insists on models having values only in the range seen in reality, then this is rectified by conditioning the lengths to be biologically meaningful.

3. Single Step Mutation Models

The computation of $P_{j,j+n}(\tau)$ for the *symmetric* single step mutation model has appeared in a variety of places in the genetics literature (See e.g. Walsh, 2001.) and in textbooks in probability theory. (See e.g. Fristedt and Gray, 1997, page 292.) Their results involve I_n, the n-th modified Bessel function of the first kind. From the many identities of these well studied special functions (See Watson, 1944.), four will be particularly useful in our analysis. (See Abramowitz and Stegun, 1972, Section 9.6, pages 374-377.)

- The generating function:

$$G_I(s,z) = \exp(\frac{s}{2}(\frac{1}{z} + z)) = \sum_{n=-\infty}^{\infty} I_n(s) z^n. \tag{1}$$

- A derivative identity:

$$I_n'(s) = \frac{n}{s} I_n(s) + I_{n+1}(s). \tag{2}$$

- An integral identity:

$$I_n(s) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{in\theta} \exp(s\cos\theta)\ d\theta = \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(n\theta) \exp(s\cos\theta)\ d\theta. \tag{3}$$

- A multiplication identity:

$$I_n(\rho s) = \frac{1}{\rho^n} \sum_{k=0}^{\infty} \frac{(\rho^2-1)^k}{k!} (\frac{1}{2}s)^k I_{n-k}(s), \quad |\rho^2 - 1| < 1. \tag{4}$$

The derivative identity follows from either the generating function or the integral identity. Additionally, we can use either identity to see that $I_n(s) = I_{-n}(s)$. The Cauchy integral formula states that the integral identity follows form the generating function by evaluating the contour integral $\frac{1}{2\pi i} \int_C z^{-n} G_I(s,z) \, \frac{dz}{z}$ where C is the unit circle centered at the origin.

3.1 Symmetric Case The Markov process for the single step mutation model has a mutation rate that does not depend on the length of the microsatellite. Such processes having spatially homogeneous changes are examples of *compound Poisson processes*. The incremental change during the time τ can be represented by

$$X_\tau = \sum_{k=1}^{N_\tau} Y_k. \tag{5}$$

- $\{Y_k; k \geq 1\}$ are the independent and identically distributed sizes of the mutation event. In the symmetric single step model, the values are equally likely to be $+1$ and -1.
- N_τ, a Poisson process independent of $\{Y_k; k \geq 1\}$, gives the mutation event times. Normalizing time according to the mutation rate sets the parameter for the process to 1.

To compute $P_{j,j+n}(\tau) = P\{X_\tau = n\}$, begin with the generating function

$$G_0(\tau, z) = Ez^{X_\tau} = \sum_{n=-\infty}^{\infty} P\{X_\tau = n\} z^n. \tag{6}$$

For a single mutation: $Ez^{Y_k} = z^{-1}P\{Y_k = -1\} + z^1 P\{Y_k = 1\} = \frac{1}{2}(\frac{1}{z} + z) = g_0(z)$.
For m mutations: $Ez^{Y_1 + \cdots + Y_m} = Ez^{Y_1} \cdots Ez^{Y_m} = (\frac{1}{2}(\frac{1}{z} + z))^m = g_0(z)^m$.
For N_τ mutations:

$$\begin{aligned} G_0(\tau, z) &= \sum_{m=0}^{\infty} E[z^{X_\tau} | N_\tau = m] P\{N_\tau = m\} \\ &= \sum_{m=0}^{\infty} (\frac{1}{2}(\frac{1}{z} + z))^m e^{-\tau} \frac{\tau^m}{m!} = e^{-\tau} \exp(\frac{\tau}{2}(\frac{1}{z} + z)) = e^{-\tau} G_I(\tau, z). \end{aligned} \tag{7}$$

By equating the coefficients of z^n in the expressions (1) and (6), we find that

$$P_{j,j+n}(\tau) = P\{X_\tau = n\} = e^{-\tau} I_n(\tau).$$

To illustrate how this might be used, we begin with data consisting of the repeat numbers for m homologous microsatellites collected from the nonrecombining region of two human Y chromosomes. The aim is to estimate the time to their most recent common ancestor. Assume that each of these microsatellites evolve independently according to the single step mutation model with a common mutation rate. As Eckert et. al. (2002) have shown, differing microsatellites are likely to have differing mutation rates. Such difference in rates, if known, can easily be accomodated in the likelihood function. Set:

- m_n - the number of sites that differ by n, $n = 0, \ldots, n_{max}$. Thus, $m = \sum_{n=0}^{n_{max}} m_n$.
- $d = \sum_{n=0}^{n_{max}} n m_n$ - the *Manhattan distance*.

For a maximum likelihood estimation, the sufficient statistics for τ are $\mathbf{m} = (m_0, \ldots, m_{n_{max}})$. The *likelihood function*,

$$L(\tau|\mathbf{m}) = \frac{m!}{m_0! \cdots m_{n_{max}}!} \prod_{n=0}^{n_{max}} (e^{-2\tau} I_n(2\tau))^{m_n}.$$

Denoting the multinomial factor by c, we have the log-likelihood,

$$\log L(\tau|\mathbf{m}) = \log c + \sum_{n=0}^{n_{max}} m_n(-2\tau + \log I_n(2\tau)) = \log c - 2m\tau + \sum_{n=0}^{n_{max}} m_n \log I_n(2\tau).$$

Take a derivative and use the Bessel function derivative identity (2) to obtain

$$\frac{d}{d\tau} \log L(\tau|\mathbf{m}) = -2m + \frac{1}{\tau} d + 2 \sum_{n=0}^{n_{max}} m_n \frac{I_{n+1}(2\tau)}{I_n(2\tau)}. \tag{8}$$

For the maximum likelihood, find $\hat{\tau}$, the value of time that makes the expression (8) equal to zero. We apply this to the 3 person data set in Table 1, the repeat numbers of 8 microsatellites. The sufficient statistics and the estimate for the time to most recent common ancestor under this single step model is given in Table 2.

Table 1

Data on 8 microsatellite lengths from the nonrecombining region of human Y chromosome for three individuals, taken from the GATC database at the University of Arizona.

sample	DYS19	DYS90	DYS391	DYS393	DY426	DYS607	H4	DYS442
1	15	23	10	13	11	14	11	13
2	13	25	10	13	12	15	10	12
3	17	18	11	13	12	12	11	12

Table 2

Using the symmetric singlse step mutation model, estimated time to most recent common ancestor from data in Table 1.

pair	m_0	m_1	m_2	m_3	m_4	m_5	m_6	m_7	$\hat{\tau}$
1-2	2	4	2	0	0	0	0	0	0.870
1-3	2	3	2	0	0	1	0	0	2.112
2-3	3	2	0	1	1	0	0	1	4.632

3.2 Asymmetric Case In this case, the step size in the compound Poisson process X_τ in equation (5) is

$$Y_k = \begin{cases} +1, & \text{with probability} \quad p, \\ -1, & \text{with probability} \quad 1-p. \end{cases}$$

Then the generating function for the single step, Y_k, is $Ez^{Y_k} = (1-p)/z + pz = g_{0,p}(z)$, and the generating function for X_τ is

$$\begin{aligned} Ez^{X_\tau} &= \exp \tau(g_{0,p}(z) - 1) = e^{-\tau} e^{(2p-1)z\tau} \exp\left(2(1-p)\tau \frac{1}{2}(\frac{1}{z} + z)\right) \\ &= e^{-\tau} \sum_{\ell=0}^{\infty} \frac{(2p-1)^\ell}{\ell!} \tau^\ell z^\ell \cdot \sum_{k=-\infty}^{\infty} I_k(2(1-p)\tau) z^k \\ &= \sum_{n=-\infty}^{\infty} \left(e^{-\tau} \sum_{\ell=0}^{\infty} \frac{(2p-1)^\ell}{\ell!} \tau^\ell I_{n-\ell}(2(1-p)\tau) \right) z^n, \quad n = k + \ell. \end{aligned}$$

Now use the multiplication identity (4) with $s = 2(1-p)\tau$ and $\rho = \sqrt{p/(1-p)}$ to obtain

$$P\{X_\tau = n\} = e^{-\tau} \sum_{\ell=0}^{\infty} \frac{(2p-1)^\ell}{\ell!} \tau^\ell I_{n-\ell}(2(1-p)\tau) = e^{-\tau} \left(\frac{p}{1-p}\right)^{n/2} I_n(2\sqrt{p(1-p)}\ \tau). \quad (9)$$

Formula (9) and general likelihood methods were used by Cooper et. al., (1999) to suggest a mutational bias of human Y chromosome microsatellites towards increasing length ($p > 1/2$). Interestingly, by using (9) and following the steps in (8), we can see that the likelihood function takes its maximum at a time that is independent of p.

The restriction on ρ in (4) limits this formula to cases with $p < 2/3$. However, if we write $Ez^{X_\tau} = \exp \tau(g_{0,p}(z) - 1) = \exp(-\tau) \exp((1-2p)\tau/z) \exp(2p\tau(1/z + z)/2)$ and proceed as before, we include all values of $p \in (0, 1)$.

If $p > 1/2$, then $p/(1-p) > 1$ and the formula shows how over time mutations are more likely to increase the length of the microsatellite. This is demonstrated in Figure 1.

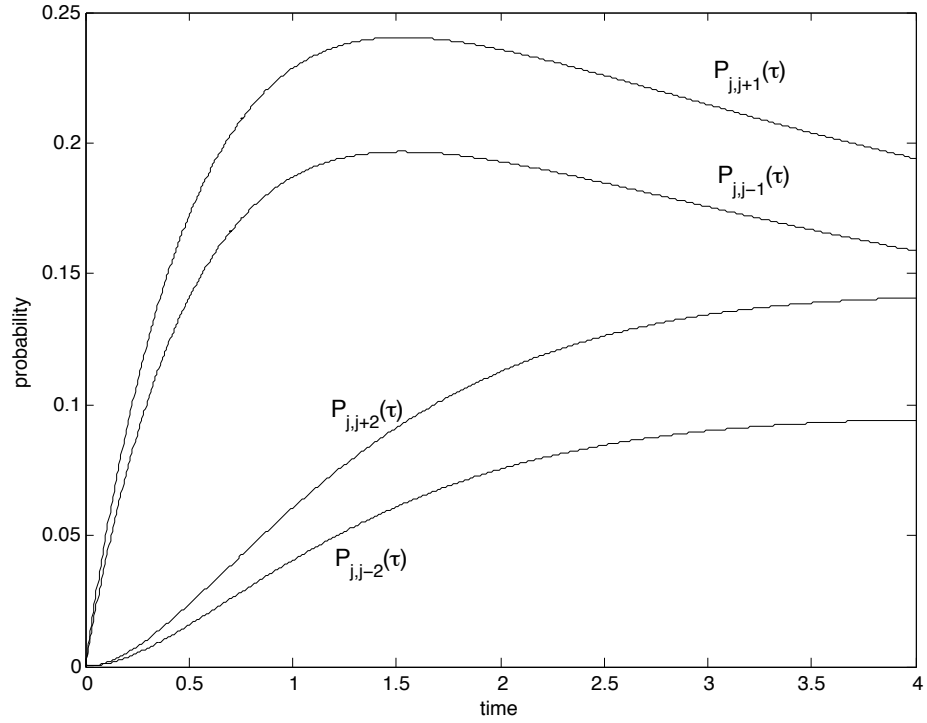

Figure 1: Reading from top to bottom, the plot of the probability $P_{j,j+n}(\tau)$, for $n = 1, -1, 2, -2$ for the asymmetric ($p = 0.55$) single step mutation model. The time τ is measured in mutation rate times generations.

To give some additional insight, fix the number of mutations to be n_0. The probability that mutations result in an increase of n repeats is given by the binomial probability

$$\binom{n_0}{(n_0+n)/2} p^{(n_0+n)/2}(1-p)^{(n_0-n)/2}.$$

The probability that mutations result in a decrease in its length by n repeats is equal to

$$\binom{n_0}{(n_0-n)/2} p^{(n_0-n)/2}(1-p)^{(n_0+n)/2}.$$

Because the two binomial coefficients are equal, the quotient of the two expressions above,

$$\left(\frac{p}{1-p}\right)^n,$$

is exactly the same as the ratio of $P\{X_\tau = n\}$ to $P\{X_\tau = -n\}$.

In addition, the time parameter in the argument of the Bessel function $2\sqrt{p(1-p)}\ \tau$ matches the reduction in the standard deviation of a step from the symmetric case.

4. Geometric Step Mutation Model

This model now allows multiple step mutation. As suggested by Whittaker et. al., we now consider models in which the step size follows a geometric distribution. We start with the case in which the step direction is symmetric.

4.1 Symmetric Case This mutation model is also a compound Poisson process as described in (5). For this case, the step sizes Y_k have common distribution

$$P\{Y_k = n\} = \left(\frac{1-\alpha}{2}\right)\alpha^{|n|-1},\ n \neq 0.$$

We shall now develop an expression for $P_{j,j+n}(\tau)$ for this model. Written in matrix form $P_{j,j+n}(\tau)$ is the solution to the *Kolmogorov forward equation*,

$$P'(\tau) = P(\tau)Q, \quad P(0) = I, \text{ the identity matrix.} \tag{10}$$

Here, Q is the *infinitesimal generator*, i.e., the matrix of infinitesimal transitions.

Because time is scaled so that mutations occur at rate 1, the diagonal elements of Q, $Q_{j,j} = -1$. The structure theorem for pure jump Markov chains (Breiman, page 329) states that the off-diagonal element $Q_{j,j+n}$ is given by $P\{Y_k = n\}$, the probability of making a jump from j to $j+n$, multiplied by 1, the jump rate. Thus, for $n \neq 0$, $Q_{j,j+n} = P\{Y_1 = n\}$.

We will not attempt to find the solution $P(\tau) = \exp \tau Q$ to the forward equation (10) directly. Rather, in the appendix, we choose a particularly convenient sequence $\{Q_m; m \geq 1\}$ of finite dimensional matrices and take a limit. To relate this to previous modeling strategies, the Q_m is some variant of the $2m-1$-alleles model.

We find in equation $(A.1)$ that

$$\exp(\tau Q_m))_{\ell,j} = \frac{1}{2m}\sum_{k=-(m-1)}^{m} e^{ik\pi(\ell-j)/m}\exp(p_m(e^{ik\pi/m})\tau). \tag{11}$$

One key to success in choosing this strategy is to select the polynomials p_m so that as $m \to \infty$, we have $p_m(e^{i\theta}) \to \lambda(\theta)$, for some continuous function λ and for $\theta = \pi k/m$. If this holds, the sum (11) is a Riemann sum for the integral

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i(\ell-j)\theta}\exp(\lambda(\theta)\tau)\ d\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi}\cos((\ell-j)\theta)\exp(\lambda(\theta)\tau)\ d\theta. \tag{12}$$

We carry out this plan in the appendix deriving the expression $(A.2)$,

$$\lambda(\theta) = -1 + \frac{(1-\alpha)(\cos\theta - \alpha)}{1 - 2\alpha\cos\theta + \alpha^2}. \tag{13}$$

Consequently, $P_{j,j+n}(\tau) = e^{-\tau} I_n^{\alpha}(\tau)$ where

$$I_n^{\alpha}(\tau) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(n\theta) \exp\left(\frac{(1-\alpha)(\cos\theta - \alpha)\tau}{1 - 2\alpha\cos\theta + \alpha^2} \right) \, d\theta. \tag{14}$$

Because $I_n^0(\tau) = I_n(\tau)$, the case $\alpha = 0$ returns the single step symmetric mutation model.

The impact of these choice of models is demonstrated in Figure 2 and Table 3. The infinite alleles model, i. e., a model in which no back mutations take place, has the most rapidly decreasing probability for identity in state; the single step model has the least rapidly decreasing probability. As we increase α from 0 to 1, the probabilities drop from the single step model to the infinite alleles model.

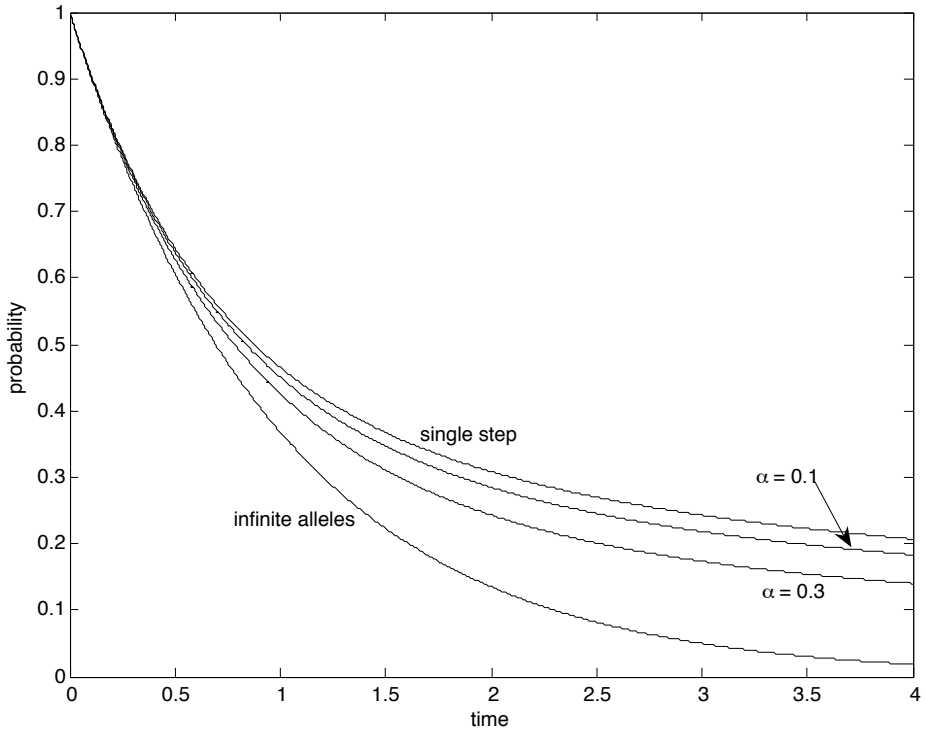

Figure 2: Plot of the probability of identity in state as a function of time, measured in mutation rate times generations. Reading from top to bottom, the graphs are for the single step model, geometric step size model $\alpha = 0.1$, geometric step size model $\alpha = 0.3$, and the infinite alleles model.

Although we would not choose to limit our estimates to the time to most recent common ancestor to the fraction of microsatellites that are identical in state, we show in Table 3 the

impact of that estimator. If half of the microsatellites are identical in state, the difference in estimated times differs by at most 21% for the 4 models considered. However, if only 10% of the microsatellites are identical in state, then the estimates range over a set of value that differ up to a factor of seven. From the case $\alpha = 0$ to the case $\alpha = 0.3$ note that estimate time drops by only 11% when the fraction identical in state is 0.50, but drops by 58% when the fraction is 0.10.

Table 3

Time to most recent common ancestor as a function of probability of identity in state, for four mutation models.

	single	multiple step model		infinite
probability	step model	$\alpha = 0.1$	$\alpha = 0.3$	alleles
0.50	0.877	0.837	0.781	0.693
0.20	4.261	3.458	2.519	1.609
0.10	16.172	12.180	6.863	2.303

Using (14) and using the same sufficient statistics as in the single step mutation model, we can redo the maximum likelihood analysis to estimate the time to the most recent common ancestor. The results of our computation are collected in Table 4.

Table 4

Estimated time to most recent common ancestor from data in Table 1 applied to models having multiple repeat number mutations.

	single	multiple step model	
pair	step model	$\alpha = 0.1$	$\alpha = 0.3$
1-2	0.870	0.757	0.678
1-3	2.112	1.483	0.871
2-3	4.632	3.263	1.195

4.2 Asymmetric Case The strategy for obtaining $P_{j,j+n}(\tau)$ in the *asymmetric* geometric step model is found in combining the approaches taken in the asymmetric single step model and the symmetric geometric step model. This mutation model remains in the class of compound Poisson processes as introduced in (5). Letting n denote a positive integer, we see that the step size

$$Y_k = \begin{cases} +n & \text{with probability} \quad p(1-\alpha)\alpha^{n-1}, \\ -n. & \text{with probability} \quad (1-p)(1-\alpha)\alpha^{n-1}. \end{cases}$$

The details of the computation for this model are found in the appendix equation $(A.3)$ where we see that the function $\lambda(\theta)$ defined by equation (13) is generalized to

$$\lambda_p(\theta) = -1 + \frac{(1-\alpha)(\cos\theta - \alpha + i(2p-1)\sin\theta)}{1 - \alpha\cos\theta + \alpha^2}.$$

Thus, $P_{j,j+n}(\tau) = e^{-\tau} I_n^{\alpha,p}(\tau)$ with

$$I_n^{\alpha,p}(\tau) \;=\; \frac{1}{2\pi}\int_{-\pi}^{\pi} \cos\left(n\theta + \frac{(1-\alpha)(2p-1)\tau\sin\theta}{1-\alpha\cos\theta+\alpha^2}\right)\exp\left(\frac{(1-\alpha)\tau\cos\theta}{1-\alpha\cos\theta+\alpha^2}\right)\, d\theta.$$

Note that $I_n^{\alpha,1/2}(\tau) = I_n^{\alpha}(\tau)$ as it should.

5. Modeling Increasing Instability with Length

The data considered in Whittaker et. al. suggest that longer microsatellites are more likely to mutate with a mutation rate that increases geometrically with the repeat number of the microsatellite. Call this rate parameter β. When a mutation occurs, we shall maintain the repeat change distributions as described in the previously developed models.

For example, in the asymmetric geometric mutation model having this given propensity to mutate, we have for the generator Q the entries

$$Q_{j+n,j} = \beta^{j+n}(1-p)(1-\alpha)\alpha^{n-1}, \quad Q_{j,j} = -\beta^j, \quad Q_{j-n,j} = \beta^{j-n}p(1-\alpha)\alpha^{n-1},\ n>0. \tag{15}$$

The scaling of the model yields the identity $P_{j,j+n}(\tau) = P_{0,n}(\beta^j\tau)$. Thus, we need only find solutions in the case $j = 0$. Alternatively, we can call some fixed length of the microsatellite the $j = 0$ state and look for the difference in the number of repeats from this length.

In the appendix, we see in equation $(A.4)$ that the generating function

$$G_{p,\alpha,\beta}(\tau, z) = \sum_{n=-\infty}^{\infty} P_{0,n}(\tau)z^n, \tag{16}$$

satisfies partial differential equation

$$\frac{\partial}{\partial\tau}G_{p,\alpha,\beta}(\tau, z) = (g_{p,\alpha}(z) - 1)G_{p,\alpha,\beta}(\tau, \beta z) \tag{17}$$

where $g_{p,\alpha}$ is the generating function for a length of a mutation. Because $P_{0,0}(0) = 1$ and $P_{0,n}(0) = 0$ if $n \neq 0$, we have the initial condition $G_{p,\alpha,\beta}(0, z) = 1$.

For a power series solution for (16) in the time variable τ,

$$G_{p,\alpha,\beta}(\tau, z) = \sum_{k=0}^{\infty} a_k(z)\tau^k, \tag{18}$$

the differential equation (17) becomes

$$\sum_{k=0}^{\infty} k a_k(z)\tau^{k-1} = (g_{p,\alpha}(z) - 1)\sum_{k=0}^{\infty} a_k(\beta z)\tau^k.$$

Equating coefficients of powers of τ gives $a_k(z) = (g_{p,\alpha}(z) - 1)a_{k-1}(\beta z)/k$. The initial condition gives $a_0(z) = 1$. We recursively solve for the coefficients and obtain

$$a_k(z) = \frac{1}{k!}\prod_{\ell=1}^{k}(g_{p,\alpha}(\beta^{\ell-1}z) - 1).$$

Note that the case $\beta = 1$ returns the solution $G_{p,\alpha,1}(\tau, z) = \exp \tau(g_{p,\alpha}(z) - 1)$.

To complete the analysis, we note that $P_{0,n}(\tau)$ is the coefficient of z^n in the power series expansion of $G_{p,\alpha,\beta}(\tau, z)$. This can be obtained by expanding the products $a_k(z)$. Alternatively, from the Cauchy integral formula, we see that this coefficient can be determined by evaluating the integral $\frac{1}{2\pi i}\int_C z^{-n}G_{p,\alpha,\beta}(\tau, z)\ \frac{dz}{z}$ where C is any positively oriented simple closed contour that contains the origin and avoids the singularities of $G_{p,\alpha,\beta}$. For example, if we take the unit circle parameterized by $z = e^{i\theta}, -\pi < \theta \leq \pi$, we obtain

$$P_{0,n}(\tau) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-in\theta}G_{p,\alpha,\beta}(\tau, e^{i\theta})\ d\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-in\theta}\sum_{k=0}^{\infty}\frac{\tau^k}{k!}\prod_{\ell=1}^{k}(g_{p,\alpha}(\beta^{\ell-1}e^{i\theta}) - 1)\ d\theta.$$

Note that in the case $\beta = 1$, $P_{0,n}(\tau) = e^{-\tau}\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-in\theta}\exp \tau g_{p,\alpha}(e^{i\theta})\ d\theta$ yielding results for the asymmetric geometric step model.

Both Whittaker, et. al. and Eckert et. al. conclude that $\beta > 1$. Figure 3 examines at a symmetric single step mutation model with $\beta = 1.05$. Even though the distribution of change in repeat number for a mutation is symmetric from any given repeat number, the slightly higher probability for mutations decreasing the length of a microsatellite is a general property of the symmetrical single step mutation model with increased instability for longer microsatellites.

To continue this point, we have, for $\alpha = 0$, the quadratic approximations in τ:

$$P_{0,1}(\tau) \approx p\tau - \frac{1}{2}p(1+\beta)\tau^2 \text{ and } P_{0,-1}(\tau) \approx (1-p)\tau - \frac{1}{2}(1-p)(1+\frac{1}{\beta})\tau^2. \qquad (19)$$

Thus, for $p = 1/2$, $P_{0,1}(\tau) < P_{0,-1}(\tau)$ for small values of τ, but the inequality reverses for $p > 1/2$.

Note that if either $p \neq 1/2$ or $\beta > 1$, the model distinguishes between forward and backward in time and the maximum likelihood estimate for time to most recent common ancestor must take this in account. Clearly, inclusion of $\beta > 1$ in the model reduces the impact of the long microsatellites, e.g., DYS90 in Table 1, in the estimation using these model parameters. In contrast to the case $\alpha = 0$, $\beta = 1$, the estimates for coalescent times do depend on p.

The impact of β on these estimates for the three individuals is illustrated in Table 5.

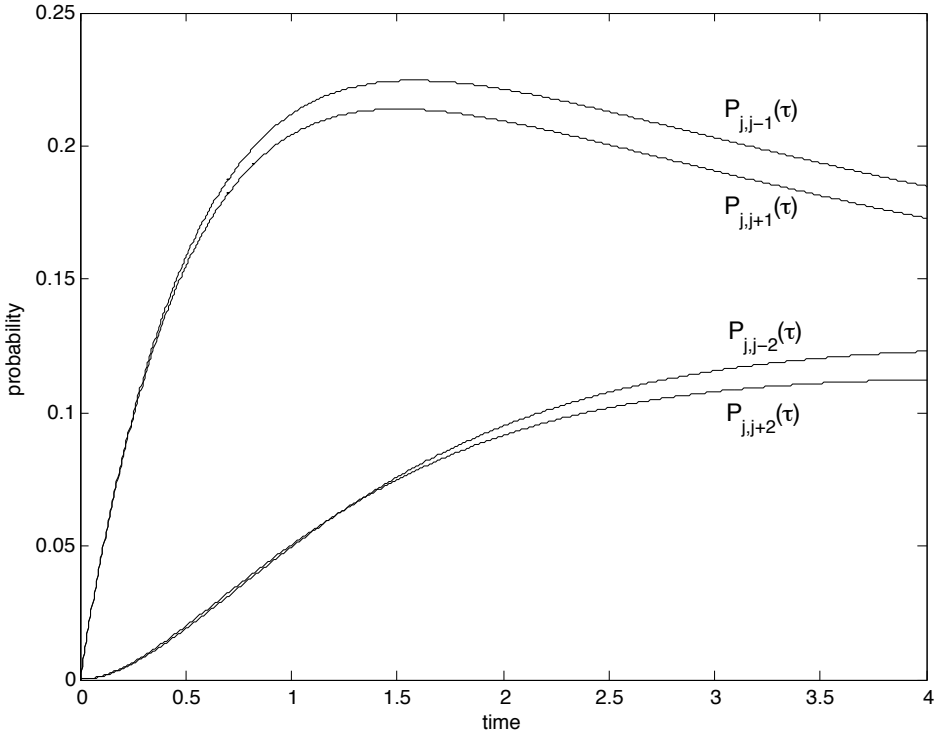

Figure 3: Reading from top to bottom, a plot the probability of $P_{j,j+k}(\tau), k = -1, 1, -2, 2$, for a single step symmetric mutation model with β, the parameter for the geometric increase in intensity for mutation, set to 1.05. Time is measured in units equal to the per generation probability of mutation for a microsatellite having repeat number j.

Table 5

Estimated time to most recent common ancestor from data in Table 2 applied to symmetric single step geometric intensity rate model. Time is measured in units that are equal to the per generation probability of a mutation for a 10 repeat microsatellite.

	single	geometric intensity model	
pair	step model	$\beta = 1.03$	$\beta = 1.06$
1-2	0.870	0.711	0.611
1-3	2.112	1.545	1.195
2-3	4.632	3.285	2.445

In addition, we can find the sensitivity of the parameters by differentiating the generating function, evaluating at the parameter values for the symmetric single step mutation model,

and examining the coefficients of z^n. We find that

$$\begin{aligned}
\frac{\partial}{\partial p}P_{0,n}^{1/2,0,1}(\tau) &= \tau e^{-\tau}\left(I_{n-1}(\tau) - I_{n+1}(\tau)\right), \\
\frac{\partial}{\partial \alpha}P_{0,n}^{1/2,0,1}(\tau) &= \frac{1}{2}\tau e^{-\tau}\left(I_{n-2}(\tau) - I_{n-1}(\tau) - I_{n+1}(\tau) + I_{n+2}(\tau)\right), \\
\frac{\partial}{\partial \beta}P_{0,n}^{1/2,0,1}(\tau) &= \frac{1}{8}\tau^2 e^{-\tau}\left(I_{n-2}(\tau) - 2I_{n-1}(\tau) + 2I_{n+1}(\tau) - I_{n+2}(\tau)\right).
\end{aligned}$$

We have exact results for biased single step models in (9). Figures 4 and 5 gives the graphs of the partial derivatives with respect to the parameters α and β for $n = 0, 1, 2, 3$.

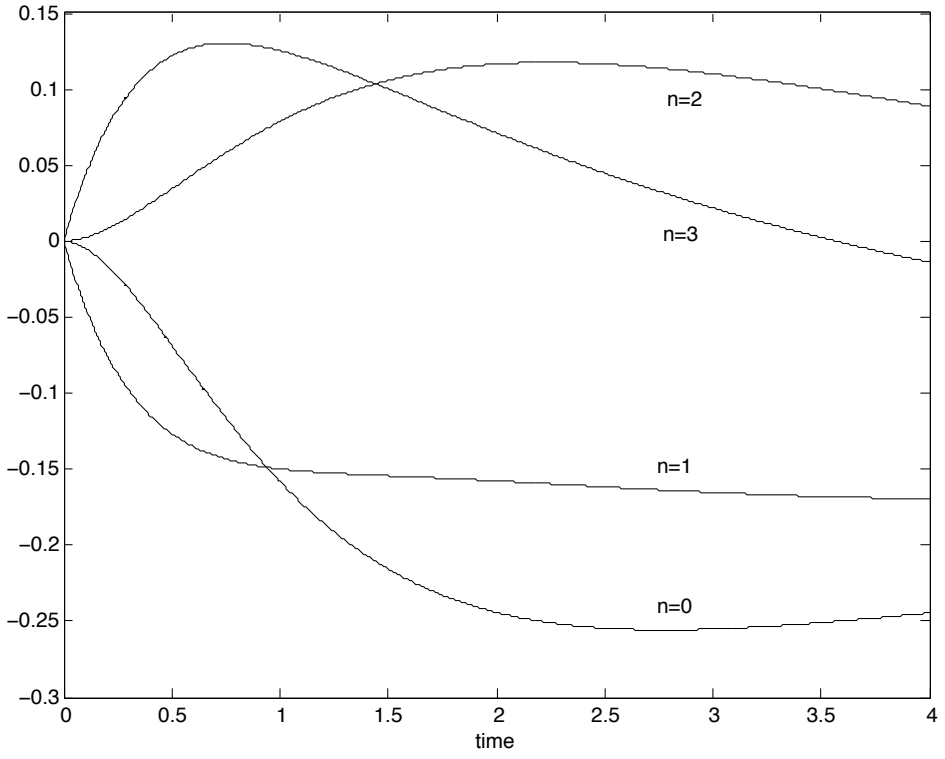

Figure 4: Sensitivity of parameter α for the mutation repeat length distribution as given by $\partial P_{0,n}^{1/2,0,1}(\tau)/\partial\alpha$ for $n = 0, 1, 2, 3$.

Figures 6 and 7 displays the impact of the choice of α and β in the maximum likelihood estimate for time to the most recent common ancestor for the three individuals given in

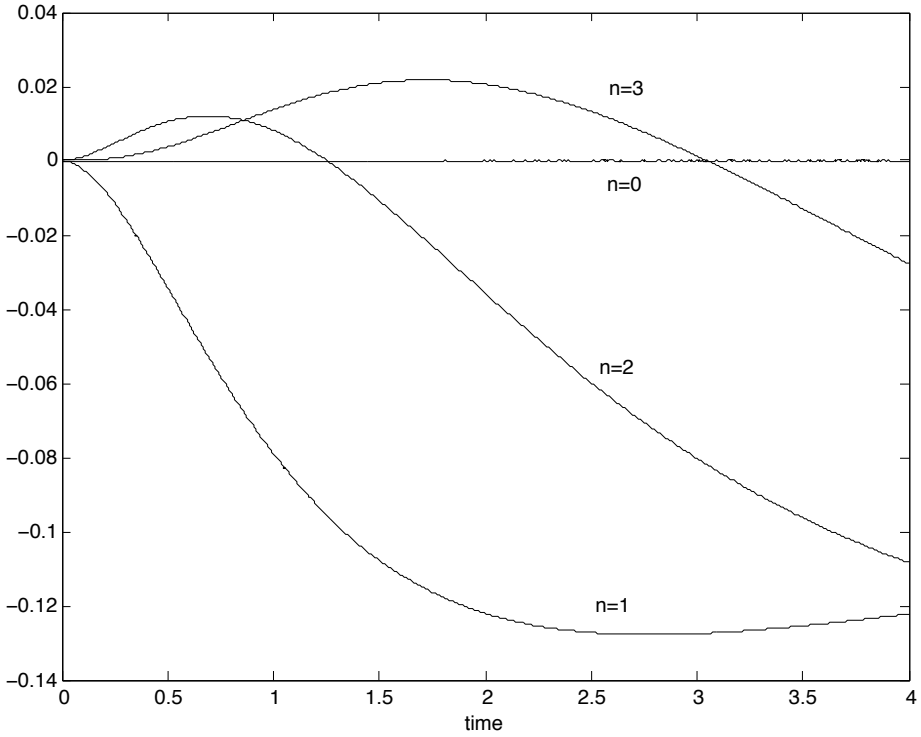

Figure 5: Sensitivity of parameter β for the mutation intensity as given by $\partial P_{0,n}^{1/2,0.1}(\tau)/\partial\beta$ for $n = 0, 1, 2, 3$.

Table 1. In Figure 6, we show the range of these times $\hat{\tau}_{12}$ for the most closely related pair, 1 and 2, and $\hat{\tau}_{23}$ the most distantly related pair, 2 and 3. The comparison of these two surfaces shows that the impact of the choice of parameter values is much greater for longer coalescent times. The estimates $\hat{\tau}_{12}$ ranges from 0.87 for the single step mutation model down to 0.48 for $\alpha = 0.3, \beta = 1.06$ whereas the estimates $\hat{\tau}_{23}$ ranges from 4.63 to 0.80 for the same set of parameter values. As coalescent times increase much above 5, the computation of the likelihood becomes much more complex and unless numerical methods are sophisticated the concerns of numerical underflow become more acute.

In the case $\beta \neq 1$, the mutation rate depends of the repeat number. Thus, in setting the time scale, we must fix a repeat number from which to measure time in units of the per generation probability of mutation. For this example, we have set repeat number 10 to have this property. Because the ratio of the times $\hat{\tau}_{23}/\tau_{12}$ is independent of this choice, it gives a better sense of the consequences of the parameter choice. These ratios, graphed in Figure 7, vary from 5.32 to 1.65.

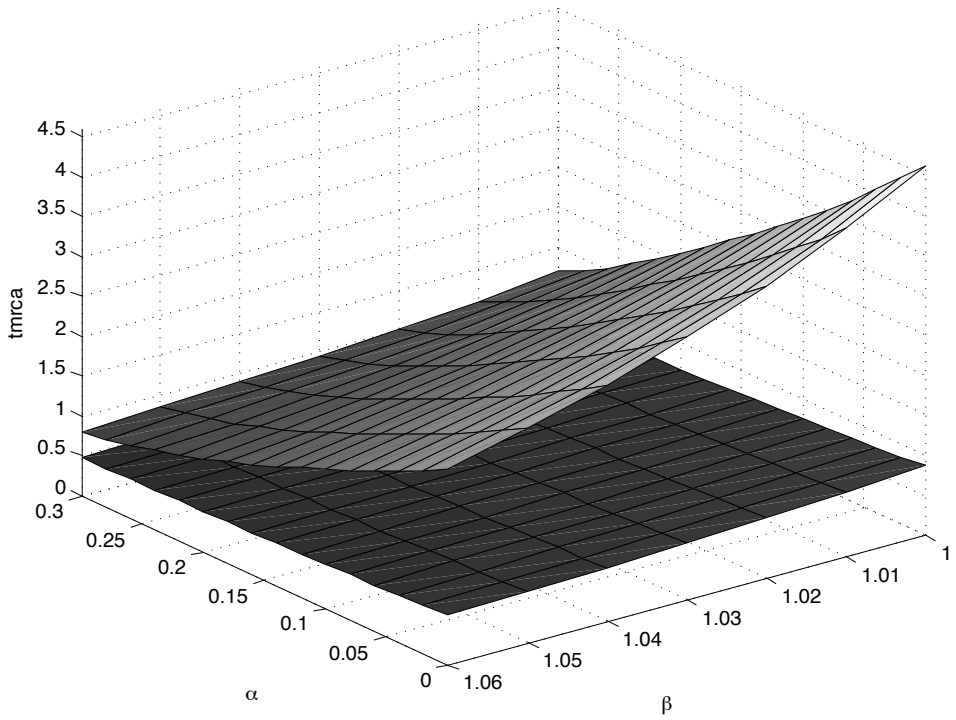

Figure 6: Estimated coalescent time under model parameters $0.0 \leq \alpha \leq 0.3$, $1.00 \leq \beta \leq 1.06$. The times $\hat{\tau}_{12}$ for individuals 1 and 2 is shown on the lower surface, and $\hat{\tau}_{23}$ for individuals 2 and 3 on upper surface.

6. Discussion

The mutations one finds in the genome have been extensively employed as a tool for dating genealogical events. For these purposes, one scans for polymorphic pieces of DNA and then develops mathematical models for their evolution. Questions that can be addressed using only neutral mutations over non-recombining regions of the DNA simplify both the model building and the ensuing statistical analysis.

Several types of mutations have been particularly fruitful in applying this approach to the population genetics of humans - single nucleotide polymorphisms (SNPs), and microsatellites on the Y chromosome along the patriline and the mitochondrial hypervariable region for inference along the matriline.

For the purposes of this discussion, we focus on the human patriline questions. In examining a parent-child pair, estimates for the probability of a point mutation are approximately one in a billion to one in 10 billion whereas the probabilities of a microsatellite mutation are about one in a thousand to one in 10 thousand.

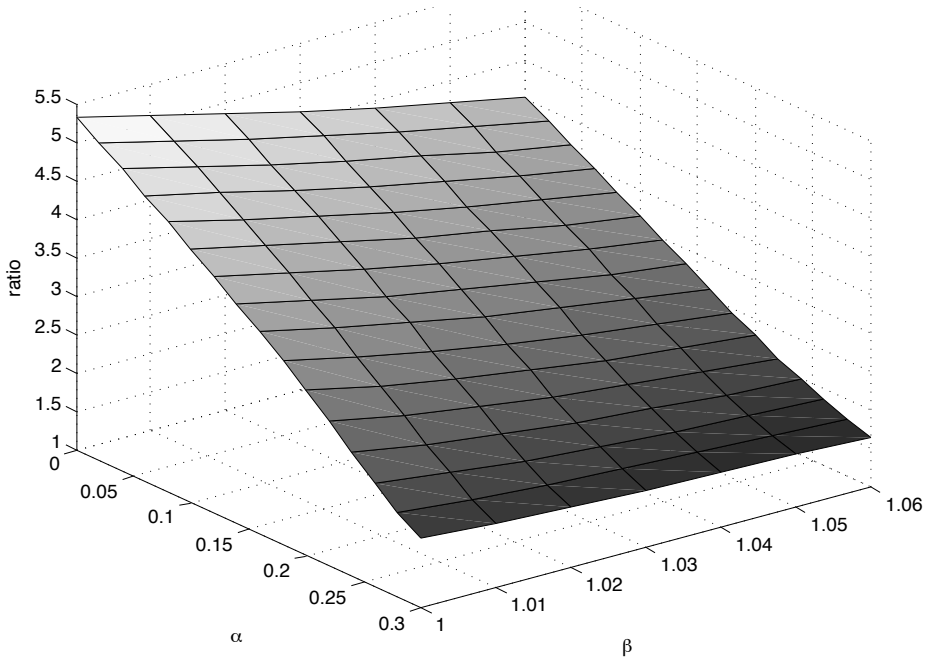

Figure 7: Ratio $\hat{\tau}_{23}/\hat{\tau}_{12}$ of coalescent time under model parameters $0.0 \leq \alpha \leq 0.3$, $1.00 \leq \beta \leq 1.06$.

In terms of mathematical modeling using Markov processes, this creates two quite distinct time scales. For those time scales in which one expects few SNP mutations, the microsatellites have undergone many mutations. Moreover, relatively rare mutational types, e. g., point mutations, play a more important role in microsatellite evolution. Analysis on this time scale can sometimes be fairly based on the assumption that microsatellites repeat number vary according to a stationary distribution. (See Kruglyak et. al. (1998) and Calabrese et. al. (2001).)

On the shorter of the two time scales, the microsatellite modeling must be refined so that the best understood assumptions feed into the model and the analysis applies well for comparisons of microsatellite lengths that differ by a small amount. The first of these needs is supplied by Whittaker and his colleagues, the second is supplied here.

The approach taken in this paper is to consider models step by step in increasing complexity. However, the final method captures all of the Whittaker models and many others provided that the change in the number of repeats experienced by a mutation does not depend on microsatellite length and intensity of mutation depends geometrically on microsatellite length. Moreover, the method is sufficiently robust to accommodate any distribution of muta-

tion size. In particular, let $\ell\{n\} = Prob\{\text{mutation of } n \text{ repeats}\}$ and $g(z) = \sum_{n=-\infty}^{\infty} \ell\{n\} z^n$ denote the generating function for the size of a mutation and let β denote the parameter for the geometric dependence of mutation intensity on microsatellite length, then

$$P_{j,j+n}(\tau) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-in\theta} \sum_{k=0}^{\infty} \frac{(\beta^j \tau)^k}{k!} \prod_{\ell=1}^{k} (g(\beta^{\ell-1} e^{i\theta}) - 1) \; d\theta.$$

Whittaker et. al. assayed mutation rates for AC microsatellites only. The structure of mutation types are anticipted to be similar for other microsatellites. However, the parameters of the model are expected to depend on the particular nucleotides that constitute any given microsatellite. (See Ellegren, 2000.) As we have witnessed in the examples presented, the estimates can vary considerably based on the choices of parameter values. Indeed, to minimize the effect of variations in mutation rates, we have chosen in our examples only microsatellites having repeats of length 4 nucleotides.

Whittaker et. al. also discuss the issue of length-dependent mutational bias. This would result, for example, in a value of p in equation (15) that depends on repeat number. In their study, up mutations exceed down mutations for short microsatellites with the reverse holding for long microsatellites. The transition in behavior appears to occur at 20 repeats. Such models are not included among those whose transition functions are developed here. However their study should be amenable to analysis as perturbations of the models presented here. (See Kato, 1982.) Such an analysis may be needed to distinguish between the competing hypotheses of length-dependent mutational bias as discussed by, e.g., Whittaker et. al. or Sainudiin et. al. (2004).

Estimates for mutation probabilities have been obtained using both information on parent-child pairs and from inference from genealogical models. The evolutionary modeling pursued here is based on the first approach. The second approach is often based on the assumption of a stationary distribution of microsatellite length. The model parameters are then estimated using a collection of DNA sequences. (See Silby et. al., 2001). With the analytical expressions established for the models above, the assumption of stationarity is no longer necessary for short time genealogies. This holds the promise of resolving the difference in estimations for mutation rates based on these two approaches.

Finally, one of the most intriguing aspects of this endeavor are those members of this class of models that satisfy the thermodynamic constraint of *detailed balance*. Mathematically, this means that given a generator, Q, we can find a function, m, of the microsatellite repeat number that satisfies $m\{n+j\}Q_{n+j,n} = m\{n\}Q_{n,n+j}$. In other words,

$$\frac{m\{n+j\}}{m\{n\}} = \frac{Q_{n,n+j}}{Q_{n+j,n}} = \frac{\ell\{j\}\beta^n}{\ell\{-j\}\beta^{n+j}}.$$

If the distribution of repeat changes for a mutation is symmetric, then we obtain a model satisfying detailed balance by choosing $m\{n\} = \beta^{-n}$. For the single step ($\alpha = 0$) models in

(15), we can satisfy detailed balance with

$$\frac{m\{n+1\}}{m\{n\}} = \frac{Q_{n,n+1}}{Q_{n+1,n}} = \frac{p}{(1-p)\beta}.$$

This leads to a choice of $m\{n\} \propto (p/((1-p)\beta))^n$.

The physics perspective is derived from consideration of the energetics of the enzymatic cycle. (See Hill, 1989.) Microsatellites, like all DNA, are replicated using DNA polymerase. The primary mutational mechanism is most often attributed to replication slippage. (See Ellegren, 2000.) If the length k of the microsatellite corresponds to an energy level E_k, can we determine a force F so that a change of length n in the microsatellite results in a change of energy by an amount Fn? Such a force F must necessarily be proportional to $\log \beta$. (See Goel, Astumian, and Herschback, 2003 for a mechanistic model of DNA polymerase.)

Another approach, based on *in vitro* experiments to estimate DNA polymerase error frequencies for microsatellite sequences, has appeared in the literature. (See Kunkel (1990) and Eckert, et. al., (2002).) They also found a geometric relationship between mutation intensities and the length of the microsatellite. In addition, they found that this intensity depended on the unit length of the microsatellite and on the nucleotides on the template strand. The mathematical models and their analysis presented in this work may provide key insights to a biophysical model that provides a deeper understanding of the mechanism for microsatellite mutation.

Appendix

A.1 Symmetric Geometric Step Model Our choice in approximation will be *circulant matrices*. Such matrices have the advantage of having eigenvalues and eigenvectors that are easy to determine. This will facilitate finding the matrix exponential as explained below. (See Davies, 1979.)

Consider the $2m \times 2m$ matrix R_m, having 1's along the superdiagonal and an additional 1 in the lower left corner. Notice that $R_m^{2m} = I$. Because the $2m$ solutions to the *minimal polynomial* equation $\lambda^{2m} - 1 = 0$ are distinct, they are the eigenvalues for R_m. The solutions are the $2m$-th roots of unity, $\epsilon_{k,m} = \exp(\frac{ik\pi}{m})$, $k = -(m-1), \ldots, m$. Denote the standard basis vectors e_k, $k = -(m-1), \ldots, m$. Then, the normalized eigenvector for $\epsilon_{k,m}$ is

$$u_k = \frac{1}{\sqrt{2m}} \sum_{j=-(m-1)}^{m} \epsilon_{k,m}^{-j} e_j \text{ and } e_\ell = \frac{1}{\sqrt{2m}} \sum_{k=-(m-1)}^{m} \epsilon_{l,m}^{k} u_k.$$

Check that the inner product $\langle e_\ell, u_k \rangle = \epsilon_{\ell,m}^k$.

Circulant matrices are polynomials in the matrix R_m.

$$\begin{aligned} p_m(R_m) &= p_{0,m}I + p_{1,m}R_m + \cdots + p_{2m-1,m}R_m^{2m-1} \\ &= \begin{pmatrix} p_{0,m} & p_{1,m} & \cdots & p_{2m-2,m} & p_{2m-1,m} \\ p_{2m-1,m} & p_{0,m} & \cdots & p_{2m-3,m} & p_{2m-2,m} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{1,m} & p_{2,m} & \cdots & p_{2m-1,m} & p_{0,m} \end{pmatrix}. \end{aligned}$$

The matrix $p_m(R_m)$ has eigenvalues $\lambda_{k,m} = p_m(\epsilon_{k,m})$ with associated eigenvectors u_k. Thus, finding the value of the exponential applied to the eigenvectors has a particularly simple form, namely, $\exp(\tau p_m(R_m))u_k = \exp(\lambda_{k,m}\tau)u_k$.

Consequently, for the standard basis vectors,

$$\begin{aligned} \exp(\tau p_m(R_m))e_\ell &= \frac{1}{\sqrt{2m}} \sum_{k=-(m-1)}^{m} \langle e_\ell, u_k \rangle \exp(\tau p_m(R_m))u_k \\ &= \frac{1}{\sqrt{2m}} \sum_{k=-(m-1)}^{m} \epsilon_{\ell,m}^{k} e^{\lambda_{k,m}\tau} \left(\frac{1}{\sqrt{2m}} \sum_{j=-(m-1)}^{m} \epsilon_{k,m}^{-j} e_j \right) \\ &= \sum_{j=-(m-1)}^{m} \left(\frac{1}{2m} \sum_{k=-(m-1)}^{m} \epsilon_{k,m}^{\ell} \epsilon_{k,m}^{-j} e^{\lambda_{k,m}\tau} \right) e_j \\ &= \sum_{j=-(m-1)}^{m} \left(\frac{1}{2m} \sum_{k=-(m-1)}^{m} e^{ik\pi(\ell-j)/m} \exp(p_m(e^{ik\pi/m})\tau) \right) e_j. \end{aligned}$$

Therefore,

$$\exp(\tau p_m(R_m)))_{\ell,j} = \frac{1}{2m} \sum_{k=-(m-1)}^{m} e^{ik\pi(\ell-j)/m} \exp(p_m(e^{ik\pi/m})\tau). \qquad (A.1)$$

To complete the argument for the symmetric geometric step size mutation model, it suffices to choose polynomials p_m so that as $m \to \infty$,

1. $p_m(R_m)$ generates a Markov process,
2. $p_m(R_m)e_\ell \to Qe_\ell$, and
3. $p_m(e^{i\theta}) \to \lambda(\theta)$, for $\theta = i\pi k/m$.

Then, as $m \to \infty$ $\exp(p_m(R_m)\tau)e_\ell \to P(\tau)e_\ell$ and $P_{\ell,j}(t)$ has the given integral representation (12). (See Ethier and Kurtz, 1986, pages 28, 29.)

For the polynomial associated with the m-th circulant matrix, choose the coefficients

$$p_{0,m} = -1 \quad \text{and} \quad p_{k,m} = c_m \frac{1-\alpha}{2}(\alpha^{k-1} + \alpha^{2m-k-1}),\ k = 1, \ldots 2m-1$$

where $c_m = 1 - \alpha^{2m-1}$.

Check that the off diagonal entries are non-negative and the row sum is zero. Thus, $Q_m = p_m(R_m)$ is the generator of a Markov process and 1 is satisfied.

As $m \to \infty$, $c_m \to 1$, and $\alpha^{2m-|n|-1} \to 0$, and 2 is satisfied.

Finally, we verify 3 and identify the function λ. For $\theta = \pi k/m$, use the fact that $e^{i2m\theta} = 1$ to see that

$$\begin{aligned}
p_m(e^{i\theta}) &= -1 + c_m \frac{1-\alpha}{2}\left((\alpha^0 + \alpha^{2m-2})e^{i\theta} + (\alpha^1 + \alpha^{2m-3})e^{i2\theta} + \cdots + (\alpha^{m-2} + \alpha^0)e^{i(2m-1)\theta}\right) \\
&= -1 + c_m \frac{1-\alpha}{2}\left(\frac{e^{i\theta} - \alpha^{2m-1}e^{i2m\theta}}{1 - e^{i\theta}\alpha} + \frac{e^{i(2m-1)\theta} - \alpha^{2m-1}}{1 - e^{-i\theta}\alpha}\right) \\
&= -1 + c_m \frac{1-\alpha}{2}\left(\frac{e^{i\theta} - \alpha^{2m-1} - \alpha + e^{-i\theta}\alpha^{2m} + e^{-i\theta} - \alpha^{2m-1} - \alpha + e^{i\theta}\alpha^{2m}}{1 - e^{-i\theta}\alpha - e^{-i\theta}\alpha + \alpha^2}\right) \\
&= -1 + c_m \frac{1-\alpha}{2}\left(\frac{2\cos\theta(1 + \alpha^{2m}) - 2\alpha(1 + \alpha^{2m-2})}{1 - 2\alpha\cos\theta + \alpha^2}\right) \to \lambda(\theta)
\end{aligned}$$

as $m \to \infty$ where

$$\lambda(\theta) = -1 + \frac{(1-\alpha)(\cos\theta - \alpha)}{1 - 2\alpha\cos\theta + \alpha^2}. \tag{A.2}$$

A.2 Asymmetric Geometric Step Model In the *asymmetric* geometric step size model, a suitable polynomial for the m-th circulant matrix has coefficients $p_{0,m} = -1$ and $p_{k,m} = c_m(1-\alpha)(p\alpha^{k-1} + (1-p)\alpha^{2m-k-1})$. Thus, 1 and 2 are satisfied, let's check 3.

$$\begin{aligned}
p_m(e^{i\theta}) &= -1 + c_m(1-\alpha)\left(p\frac{e^{i\theta} - \alpha^{2m-1}e^{i2m\theta}}{1 - e^{i\theta}\alpha} + (1-p)\frac{e^{i(2m-1)\theta} - \alpha^{2m-1}}{1 - e^{-i\theta}\alpha}\right) \\
&= -1 + c_m(1-\alpha)\left(\frac{p(e^{i\theta} - \alpha^{2m-1} - \alpha + e^{-i\theta}\alpha^{2m}) + (1-p)(e^{-i\theta} - \alpha^{2m-1} - \alpha + e^{i\theta}\alpha^{2m})}{1 - \alpha\cos\theta + \alpha^2}\right) \\
&= -1 + c_m(1-\alpha)\left(\frac{(pe^{i\theta} + (1-p)e^{-i\theta}) - \alpha^{2m}(pe^{-i\theta} + (1-p)e^{i\theta}) - \alpha(1 - \alpha^{2m-2})}{1 - \alpha\cos\theta + \alpha^2}\right) \to \lambda_p(\theta)
\end{aligned}$$

where

$$\lambda_p(\theta) = -1 + \frac{(1-\alpha)(\cos\theta - \alpha + i(2p-1)\sin\theta)}{1 - \alpha\cos\theta + \alpha^2}. \tag{A.3}$$

A.3 Modeling Increasing Instability with Length We derive a partial differential equation for the generating function for the model showing increasing instability with the length of the microsatellite.

Beginning with the generator Q given in (15), we have, for the forward equation, $P'(\tau) = P(\tau)Q$, the expansion

$$\begin{aligned} P'_{0,n}(\tau) &= \sum_{k=-\infty}^{\infty} P_{0,n+k}(\tau) Q_{n+k,n} \\ &= \sum_{k=1}^{\infty} P_{0,n-k}(\tau)\beta^{n-k}p(1-\alpha)\alpha^{k-1} - \beta^n P_{0,n}(\tau) + \sum_{k=1}^{\infty} P_{0,n+k}(\tau)\beta^{n+k}(1-p)(1-\alpha)\alpha^{k-1}. \end{aligned}$$

Thus,

$$\begin{aligned} \frac{\partial}{\partial t} G_{p,\alpha,\beta}(t,z) &= \sum_{n=-\infty}^{\infty} P'_{0,n}(\tau) z^n \\ &= \sum_{n=-\infty}^{\infty}\sum_{k=1}^{\infty} P_{0,n-k}(\tau)\beta^{n-k}p(1-\alpha)\alpha^{k-1}z^n - \sum_{n=-\infty}^{\infty}\beta^n P_{0,n}(\tau)z^n \\ &\quad + \sum_{n=-\infty}^{\infty}\sum_{k=1}^{\infty} P_{0,n+k}(\tau)\beta^{n+k}(1-p)(1-\alpha)\alpha^{k-1}z^n \\ &= \sum_{n=-\infty}^{\infty}\sum_{k=1}^{\infty} P_{0,n}(\tau)\beta^{n}p(1-\alpha)\alpha^{k-1}z^{n+k} - \sum_{n=-\infty}^{\infty}\beta^n P_{0,n}(\tau)z^n \\ &\quad + \sum_{n=-\infty}^{\infty}\sum_{k=1}^{\infty} P_{0,n}(\tau)\beta^{n}(1-p)(1-\alpha)\alpha^{k-1}z^{n-k} \\ &= \left(p(1-\alpha)\frac{z}{1-\alpha z} - 1 + (1-p)(1-\alpha)\frac{z^{-1}}{1-\alpha z^{-1}}\right)\sum_{n=-\infty}^{\infty} P_{0,n}(\tau)(\beta z)^n. \end{aligned}$$

The third equality follows by shifting the index on the summation on n. The last follows by summing on k. Consequently,

$$\frac{\partial}{\partial \tau} G_{p,\alpha,\beta}(\tau,z) = (g_{p,\alpha}(z) - 1) G_{p,\alpha,\beta}(\tau,\beta z) \tag{A.4}$$

where $g_{p,\alpha}$ is the generating function for a length of a mutation.

Acknowledgements

Thanks to Bruce Walsh for introducing me to microsatellite evolution and to Tatiana Karafet and the Genomic Analysis and Technology Core, University of Arizona, for the microsatellite sequence data. The author wishes to acknowledge his appreciation for the referee's many helpful comments. This work was supported by National Science Foundation grant BCS-0432262

References

Abramowitz, M. and Stegun, I. A. (editors), 1972 Handbook of Mathematical Functions and Formulas, Graphs, and Mathematical Tables, 9th printing. New York: Dover.

Breiman, L., 1968 Probability, Addison-Wesley, New York.

Calabrese, P. P. and Durrett, R. T., 2003 Dinucleotide repeats in the Drosophila and human genomes have complex length-dependent mutation processes. Mol. Biol. Evol. **20**: 715-725.

Calabrese, P. P., Durrett, R. T., and Aquadro, C. T., 2001 Dynamics of microsatellite divergence. Genetics **159**: 839-852.

Calabrese, P. P, and Sainudiin, R., 2004 Models of Microsatellite Evolution, in R. Nielsen (Ed.), Statistical Methods in Molecular Evolution, Series: Statistics for Biology and Health, Springer, New York.

Cooper, G., Burroughs, N. L., Rand, D. A., Rubensztein, D. C., and Amos, W., 1999 Markov chain Monte Carlo analysis of human Y-chromosome microsatellites provides evidence of biased mutation. Proc. Natl. Acad. Sci. USA **96**: 11916-11921.

Davies, P. J., 1979 Circulant Matrices, John Wiley & Sons, New York.

DeIorio, Maria, Griffiths, Robert C., Leblois, Raphael, and Rousset, François, 2005 Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models, Theoretical Population Biology **68**: 41-53.

DiRenzo, A., Peterson, A. C., Garza, J. C., Valdes, A. M., Slatkin, M., et. al., 1994 Mutational processes of simple-sequence repeat loci in human populations. Proc. Natl. Acad. Sci. USA **91**: 3166-3170.

Eckert, K. A., Mowery, A., and Hile, S. E., 2002 Misalignment-Mediated DNA Polymerase$^{\beta}$ Mutations: Comparison of Microsatellite and Frame-Shift Error Rates Using a Forward Mutation Assay. Biochemistry, **41**(33): 10490-10498.

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

Ethier, S. N. and Kurtz, T. G., 1986 Markov Processes: Characterization and Convergence, John Wiley & Sons, New York.

Fristendt, B. and Gray, L. F., 1997 A Modern Approach to Probability Theory. Birkhäuser, Boston.

Fu, Y. and Chakraborty, R., 1998 Simultation estimation of all the parameters of a step-wise mutation model. Genetics **150**: 487-497.

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

Goel, A., Astumian, R. D., and Herschbach, D., 2003 Tuning and switching a DNA polymerase motor with mechanical tension. Proc. Natl. Acad. Sci. USA, **100**:9699-9704.

Goldstein, D. B., and Schlotterer, C., 1999 Microsatellites: Evolution and Applications, Oxford University Press, Oxford.

Hill, T. L., 1989 Free Energy Transduction and Biochemical Cycle Kinetics. Springer-Verlag, New York.

Kato, T. 1982 A Short Introduction to Perturbation Theory of Linear Operators. Springer, New York.

Kruglyak, S., Durrett, R. T., Malcolm, S., and Aquadro, C.F., 1998 Equilibrium distribution of microsatellite repeat lengths resulting from a balance between slipping events and point mutations. Proc. Natl. Acad. Sci. USA **95**: 10774-10778.

Kunkel, T. A., 1990 Misalignment-mediated DNA synthesis errors, Biochemistry, **29**(35): 8003-8011.

Lai, Y, and Sun, F., 2004 Sampling distribution for microsatellites amplified by PCR: mean field approximation and its applications to genotyping. Journal of Theoretical Biology **228**: 185-194.

Moxon, E. R. and Wills, C., 1999 DNA microsatellites: Agents of evolution? Scientific American, **280-1**: 94-99.

Ohta, T. and Kimura, M., 1973 The model of mutation appropriate to calculate the number of electrophoretically detectable alleles in a genetic population. Genet. Res. **22**: 201-204.

Sainudiin, R., Durrett, R. T., Aquadro C.F., and Nielsen, R. 2004 Microsatellite Mutation Models: Insights From a Comparison of Humans and Chimpanzees. Genetics **168**: 383-395.

Silby, R. M., Whittaker, J. C., and Tolbot, M., 2001 A maximum-likelihood approach to fitting equilibrium models of microsatellite evolution. Molecular Biology and Evolution **18**: 413-417.

Stephens, M., and Donnelly P., 2000 Inference in moledular population genetics. J. Roy. Statist. Soc. B **62**: 605-655.

Walsh, J. B., 1987 Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. Genetics **115**: 471-478.

Walsh, J. B., 2001 Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals. Genetics **158**: 897-912.

Watson, G. N., 1944 A Treatise on the Theory of Bessel Functions, Cambridge University Press, Cambridge.

Wehrbahn, C. F., 1975 The evolution of selectively similar electrophoretically detectable alleles in finite populations. Genetics **80**: 375-394.

Whittaker, J. C., Harbord, R. M., Boxall, N., Mackay, I., Dawson, G., and Silby, R. M., 2003 Likelihood-based estimation of microsatellite mutation rates. Genetics **164**: 781-787.

Wilson, Ian. J. and Balding, David J., 1998 Genealogical Inference from microsatellite data, Genetics **150**: 499-510.

Wilson, Ian. J., Weale, Michael E., and Balding, David J., 2003 Inference from DNA data: populations histories, evolutionary processes and forensic match probabilities. J. Roy. Statist. Soc. A **166**: 155-201.

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