Dinucleotide Repeats in the Drosophila and Human Genomes Have Complex, Length-Dependent Mutation Processes

Peter Calabrese and Rick Durrett†*

*University of Southern California; and †Cornell University

We use methods of maximum likelihood estimation to fit several microsatellite mutation models to the observed length distribution of dinucletoide repeats in the Drosophila and human genomes. All simple models are rejected by this procedure. Two new models, one with quadratic and another with piecewise linear slippage rates, have the best fits and agree with recent experimental studies by predicting that long microsatellites have a bias toward contractions.

Introduction

Microsatellites are tandem repeats of motifs that consist of two to six nucleotides. The oldest and most commonly used model of their evolution is the stepwise mutation model, in which the number of repeat units increases or decreases by 1 at a constant rate subject to the constraint that the number of repeat units is not allowed to become smaller than 1. This model is unrealistic in that it predicts microsatellite lengths that do not reach an equilibrium distribution but will grow over time. To address this problem, Kruglyak et al. (1998) proposed the proportional slippage (PS) model, in which (1) microsatellites have an equal probability of expansions and contractions, (2) slippage (expansions and contractions) occurs at a rate proportional to the length of the repeat, and (3) point mutations break up long repeats. The PS model has a stationary distribution which gave a good fit to the small samples of microsatellites in humans, mice, fruit fly, and yeast available to Kruglyak et al. (1998), and to all of the microsatellites in the yeast genome (Kruglyak et al. 2000).

In a study of the use of microsatellites to date divergences between species, Calabrese, Durrett, and Aquadro (2001) have shown that the simple PS model predicts larger length differences in interspecies comparisons than are observed. One possible explanation is that long microsatellites have a mutational bias that favors contractions over expansions. This explanation is supported by empirical work both in *Drosophila melanogaster* (Harr and Schlötterer 2000) and in yeast (Wierdl, Dominska, and Petes 1997). In the other direction, pedigree studies both in humans (Amos et al. 1996) and in barn swallows (Primmer et al. 1996), have suggested that microsatellites have an upward bias; *i.e.*, expansions occur more frequently than contractions.

One possible resolution of these conflicting views suggested by Garza, Slatkin, and Freimer (1995) is that microsatellites have a target length, and that microsatellites shorter than this target have a mutational bias up, whereas longer microsatellites have a bias down. This idea is supported by the work of both Xu et al. (2000) and Huang et al. (2002). In two large human pedigree studies, Xu et al. observed that the rate of expansions is independent of microsatellite ''length'' but that the rate of contractions increases exponentially as a function of microsatellite ''length.'' Huang et al. (2002) found a statistically significant negative relationship between the magnitude and direction of mutation and ''length.'' In the two preceding sentences we have put the word *length* in quotation marks, because neither group of researchers measured the actual length of a microsatellite but rather the total length of the PCR product that consists of the microsatellite and a variable amount of flanking sequence. They then applied the inverse of the distribution function of the observed lengths to obtain a number in [0,1], which they called the ''length.''

In this article, we investigate a number of different mutation models, including most of those that have previously been considered in the literature, to see which ones can best explain the observed dinucleotide microsatellite distributions in the genome sequences of both Drosophila and humans. We define a dinucleotide repeat to be a microsatellite if it consists of five or more adjacent pairs of the same dinucleotide motif. In an attempt to have microsatellites that evolve independently, we only keep track of microsatellites that are at least 50 bases from the closest dinucleotide microsatellite. We call such microsatellites *isolated*. Further we call a microsatellite *perfect* if on both sides of the microsatellite there are four or more bases that do not intersect a segment of DNA with two or more adjacent pairs of the same repeat motif that is contained in the microsatellite. Otherwise, we call the microsatellite *imperfect* or *interrupted*. Because interrupted microsatellites can have multiple interruptions, their state space is large, and consequently we choose to model only perfect microsatellites. Thus the two microsatellites in,

ca ca ca ca ca ca CT *ca ca ca ca ca ca ca ca ca*

will not be counted because they are interrupted.

The first of the models that we consider, the multinomial model, is not really a model at all. There is a parameter for each possible microsatellite length—the probability a microsatellite has that length. This model offers no insight into the mutational mechanism, but it provides a benchmark against which to judge the fits of the other models. For all models, we assume microsatellites are struck by point mutations at a rate proportional to their length, and at a location chosen uniformly along it.

Because of the assumed point mutations, the collection of microsatellite lengths will have an equilibrium distribution. We use this distribution to solve for the likelihood of the genome data for a given model and

Key words: microsatellites, genome, human, Drosophila.

E-mail: rtd1@cornell.edu.

Mol. Biol. Evol. 20(5):715–725. 2003
DOI: 10.1093/molbev/msg084

Table 1
Model Descriptions

Model	$X \to X+1$ at Rate	$X \to X-1$ at Rate
Multinomial[a]		
SMM	β	β
PS	bX	bX
PSwK	$b(X-\kappa)^+$	$b(X-\kappa)^+$
ConExp	β	$\delta e^{c(X-lo)}$
AsyLin	$\beta + b(X-lo)$	$\delta + d(X-lo)$
AsyQuad	$\beta + b_1(X-lo) + b_2(X-lo)^2$	$\delta + d_1(X-lo) + d_2(X-lo)^2$
PLBias[a]		

[a] These models have a different form, and are explained in the text.

Table 2
Drosophila Data

Motif	All[a]	Isolated[b]	Perfect[c]	Isolated and Perfect
AC/TG	11,410 (56.0%)	10,119	9,036	7,997 (56.0%)
AT/TA	6,223 (30.5%)	5,702	4,710	4,361 (30.5%)
AG/TC	2,650 (13.0%)	2,329	2,107	1,840 (12.9%)
GC/CG	105 (0.52%)	85	101	81 (0.57%)
Total	20,388	18,235 (89.4%)	15,954 (78.3%)	14,279 (70.0%)

[a] All includes isolated and nonisolated, perfect and interrupted.
[b] Isolated includes perfect and interrupted.
[c] Perfect includes isolated and nonisolated.

parameters. The assumption that the microsatellites are separated allows us to assume that their evolutions are independent and the joint likelihood is the product of the individual likelihoods. We then optimize over the space of model parameters to find the maximum likelihood of the data for that model. We do this for each model, and in order to compare models, we calculate the Akaike information criterion (Akaike 1974),

$$\text{AIC} = -2\log(\text{maximum likelihood}) + 2(\text{number of parameters}).$$

The models with the lowest AIC scores are deemed optimal. This approach is similar to hypothesis testing, in that models are rewarded for high likelihoods but penalized for the number of parameters. In addition, this approach computes the maximum likelihood parameters for each model. By repeatedly using the bootstrap to simulate new distributions, we can then calculate an approximate confidence interval for the parameters.

The seven dynamic models we consider differ in terms of their rates for slippage up and down, which are given in table 1. The first three models are symmetric: the stepwise mutation model (SMM), the PS model of Kruglyak et al. (1998), and the generalization due to Calabrese, Durrett, and Aquadro (2001), in which slippage occurs only when the length exceeds some threshold κ (PSwK). The constant, exponential model (ConExp) is motivated by the work of Xu et al. (2000), who asserted that in terms of ''length,'' slippage up was constant but slippage down grew exponentially. The final three models are new and asymmetric. In the asymmetric linear model (AsyLin) and the asymmetric quadratic model (AsyQuad) the expansion and contraction rates are two different linear and quadratic functions. The final model is the piecewise linear bias model (PLBias model), in which the mutation rate is assumed to be constant, but the bias up or down is a piecewise linear function of microsatellite length. This model has two additional parameters which we fix: the length of the linear segments, and the number of linear segments. To the right of the last segment, the bias remains constant.

For each model, and for a choice of parameters, we solve for the stationary distribution of isolated, perfect microsatellites with lengths $lo, lo+1, \ldots, hi$. The lower limit lo is chosen larger than the minimum length at which microsatellites evolve by polymerase slippage. The upper limit hi is imposed so that there are a finite rather than infinite number of equations to solve for the stationary distribution. The value of hi is chosen large enough so that the truncation has very little effect on the stationary distribution.

We assume that microsatellites originate (i.e., achieve length lo) at a constant rate and mutate independently. The number of possible states for imperfect microsatellites is very large, so if a microsatellite becomes imperfect, the microsatellite then exits the model. This framework is equivalent to a network of queues in which microsatellites correspond to customers, and microsatellite lengths correspond to stations. In the *Appendix*, we show how we can use results from queueing theory to solve for the stationary distribution.

Results

Drosophila

We downloaded the Drosophila genome from the Berkeley Drosophila Genome Project: http://www.fruitfly.org/sequence/dlMfasta.shtml, release 2, December 11, 2000 version. Separated by motif, the numbers are listed in table 2. The motifs are not equally represented. Fewer than 1% of dinucleotide microsatellites have the motif GC/CG, while over 50% have the motif AC/TG. This dearth of GC/CG microsatellites is also found in the genomes of humans, mice, and yeast (see e.g., Kruglyak et al. 1998). Restricting ourselves to both perfect and isolated dinucleotide microsatellites, for the three most popular motifs, figure 1 shows the natural logarithm of one plus the number of microsatellites as a function of microsatellite length. We have added one before taking the logarithm in order to differentiate between lengths with zero and one microsatellite.

The AIC model scores are listed in table 3. To make the scores easier to compare, for each column we have subtracted the number listed in the last row. Recall the models with the lowest AIC scores are deemed optimal. We solved for the stationary distribution of microsatellite lengths $lo = 5, \ldots, hi = 30$.

For all three motifs, the PS model scores much lower than the SMM; however, only for the AT/TA motif does the PSwK model represent a substantial improvement over the PS model. None of these symmetric models have a lower AIC score than the multinomial benchmark.

For all three motifs, the asymmetric models have lower AIC scores than the symmetric ones. Furthermore

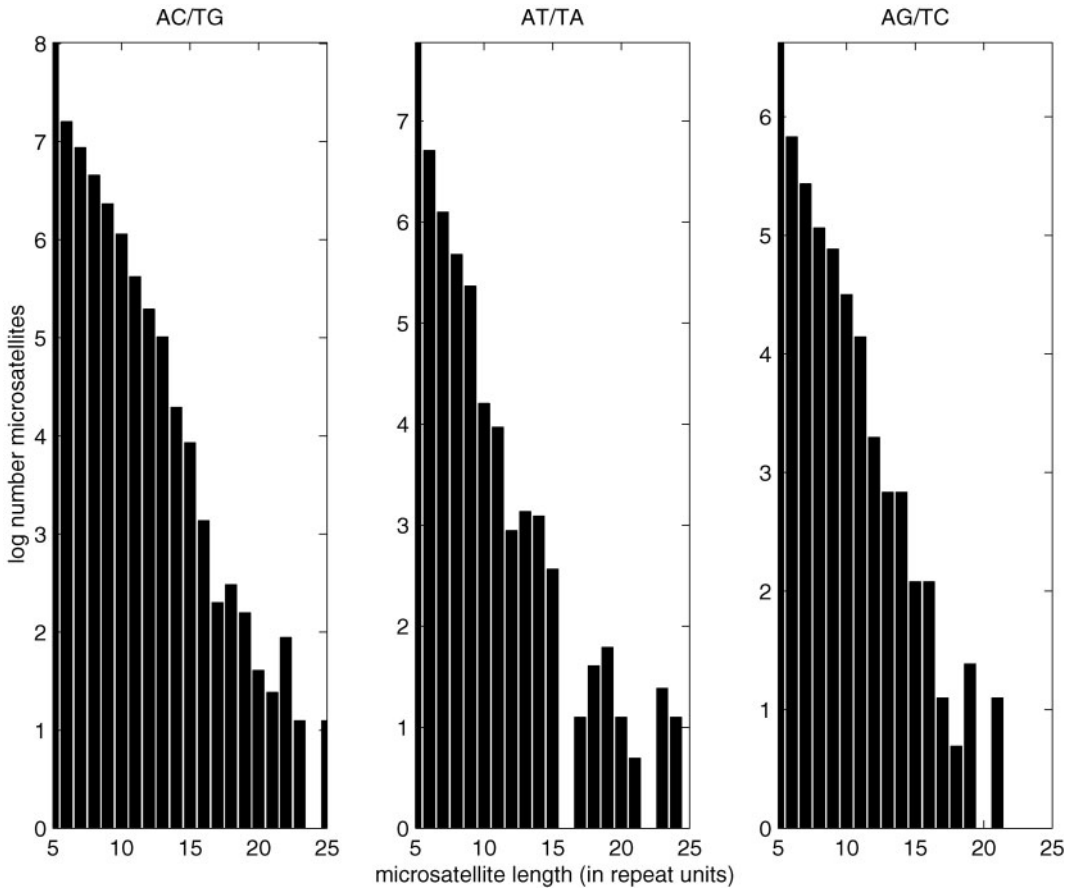

FIG. 1.—Drosophila microsatellite distribution: Natural logarithm of one plus the number of isolated, perfect microsatellites of different lengths.

the optimal parameters imply that long microsatellites have a downward bias. The AsyLin model predicts a bias down for all microsatellites longer than 7 repeat units (and a bias up for shorter microsatellites). For the AsyQuad model, microsatellites with lengths longer than 10 have a bias down; and for the PLBias model, microsatellites of all lengths have a bias down. For the two best models, AsyQuad and PLBias, figure 2 shows the probability of contraction as a function of microsatellite length. As a further test, for all three motifs we performed a parametric bootstrap test (see e.g., Huelsenbeck and Rannala 1997) with the PSwK model as the null and the AsyQuad model as the alternate. In each case, the PSwK model was significantly rejected ($P < .01$). Only for the AG/TC motif do the AsyQuad and PLBias models have a lower AIC score than the multinomial benchmark.

One problem with our approach is that for the asymmetric models, although we can solve for the contraction bias we cannot solve for the absolute mutation rates. As an example, consider one of the simplest possible asymmetric models: the asymmetric stepwise mutation model, where $X \rightarrow X+1$ at rate β and $X \rightarrow X-1$ at rate δ. We fix the per repeat point mutation rate at 2×10^{-8} (Drake et al. 1998), all inferred slippage rates are scaled relative to this estimate. In Drosophila, for the motif AC/GT, for the two parameter sets $\beta = 1.0598 \times 10^{-4}, \delta = 1.5628 \times 10^{-4}$ and $\beta = 1.0576 \times 10^{-2}, \delta = 1.5661 \times 10^{-2}$, the relative difference in log-likelihood scores is 3.73×10^{-5}. Depending on the starting point, the numerical optimization routine used would determine either of these parameter estimates to be a local maximum. But the problem is not really one of multiple local maxima, since all points on the line between these two points in parameter space have nearly equal log-likelihood scores. Because these parameters span two orders of magnitude, we cannot determine mutation rates. Nonetheless for all points along this line, the mutation bias is almost the same. This pattern of subsets of

Table 3
Drosophila AIC Model Scores

Model	No. of parameters	AIC		
		AC/TG	AT/TA	AG/TC
Multinomial	25	0	0	0
SMM	1	375	526	71
PS	1	191	216	12
PSwK	2	193	52	10
ConExp	3	217	272	27
AsyLin	4	148	48	3
AsyQuad	6	9	37	−13
PLBias	5	24	33	−19
		+30,804	+12,663	+6,724

NOTE.—To make the scores easier to compare, for each column we have subtracted the number listed in the last row.

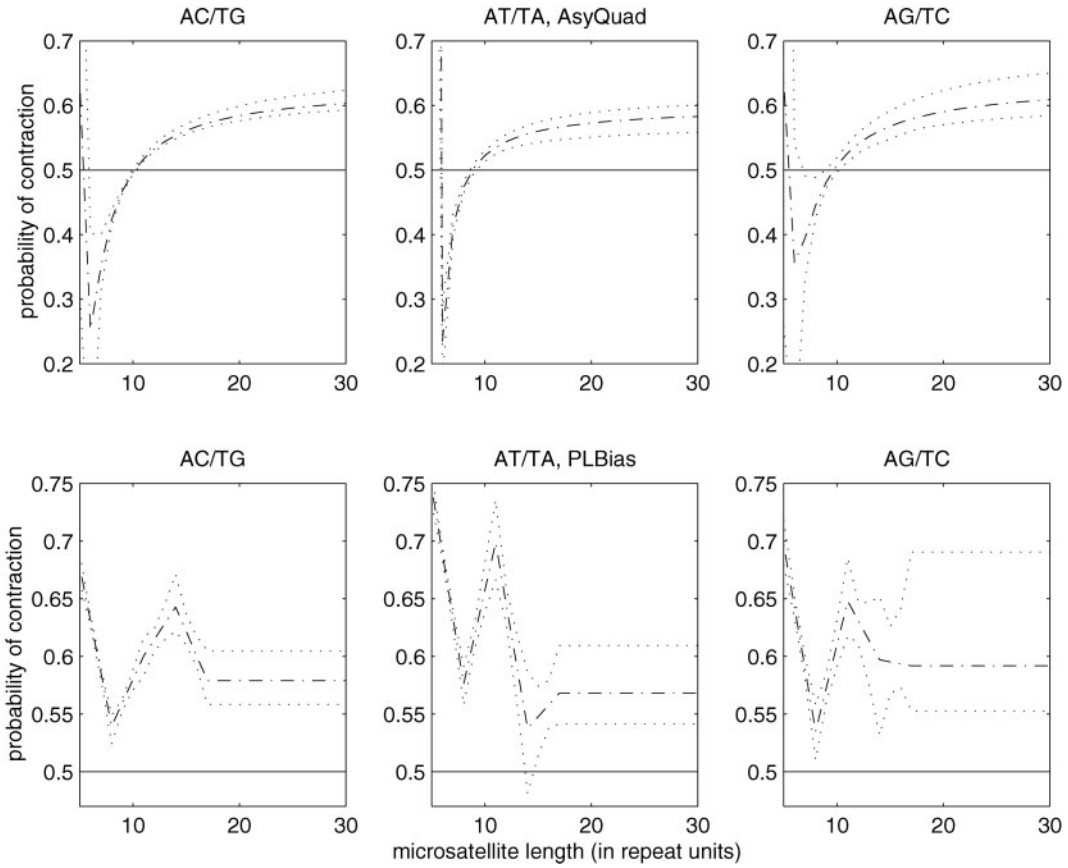

FIG. 2.—Drosophila: Probability of contraction as a function of microsatellite length in AsyQuad and PLBias model fits. The first row is the AsyQuad model, and the second row is the PLBias model. The dotted and dashed line is the model fit; the dotted lines form the 95% confidence envelope.

parameter space with nearly equal log-likelihood scores, but very different mutation estimates, and similar bias predictions is present in all asymmetric models tested. The explanation for this phenomenon is simple: the point mutation rate is the only quantity that establishes the size of the mutation rates, but in asymmetric models it plays only a very weak role in shaping the distribution. This is not a problem for symmetric models.

Humans

We downloaded the human genome from the University of California at Santa Cruz assembly of the International Human Genome Project: http://genome.ucsc.edu, December 12, 2000 version. The numbers are listed in table 4, and the natural logarithm of the distribution is shown in figure 3. As in the Drosophila genome, in humans the motifs are not equally represented. If we consider all dinucleotide microsatellites, then only 0.52% have the motif GC/CG, whereas 45.5% have the motif AC/TG. Furthermore if we only consider isolated, perfect microsatellites with length greater than or equal to 10, then an even larger fraction, 72.7%, have the motif AC/TG. In contrast to the Drosophila genome, the three most abundant dinucleotide motifs in humans have distributions with strikingly different shapes.

Table 4
Human Data

Motif	All[a]	Isolated[b]	Perfect[c]	Isolated, Perfect	Isolated, Perfect, and Length $\geq$ 10
AC/TG	176,885 (45.5%)	129,830	140,804	103,267	33,306 (72.7%)
AT/TA	110,062 (28.3%)	64,794	86,324	64,794	8,399 (18.3%)
AG/TC	99,770 (25.7%)	80,174	83,998	80,174	4,107 (8.96%)
GC/CG	2,051 (0.53%)	782	1,765	727	5 (0.01%)
Total	388,768	275,580 (70.9%)	312,891 (80.5%)	248,962 (64.0%)	45,817 (11.8%)

[a] All includes isolated and nonisolated, perfect and interrupted.
[b] Isolated includes perfect and interrupted.
[c] Perfect includes isolated and nonisolated.

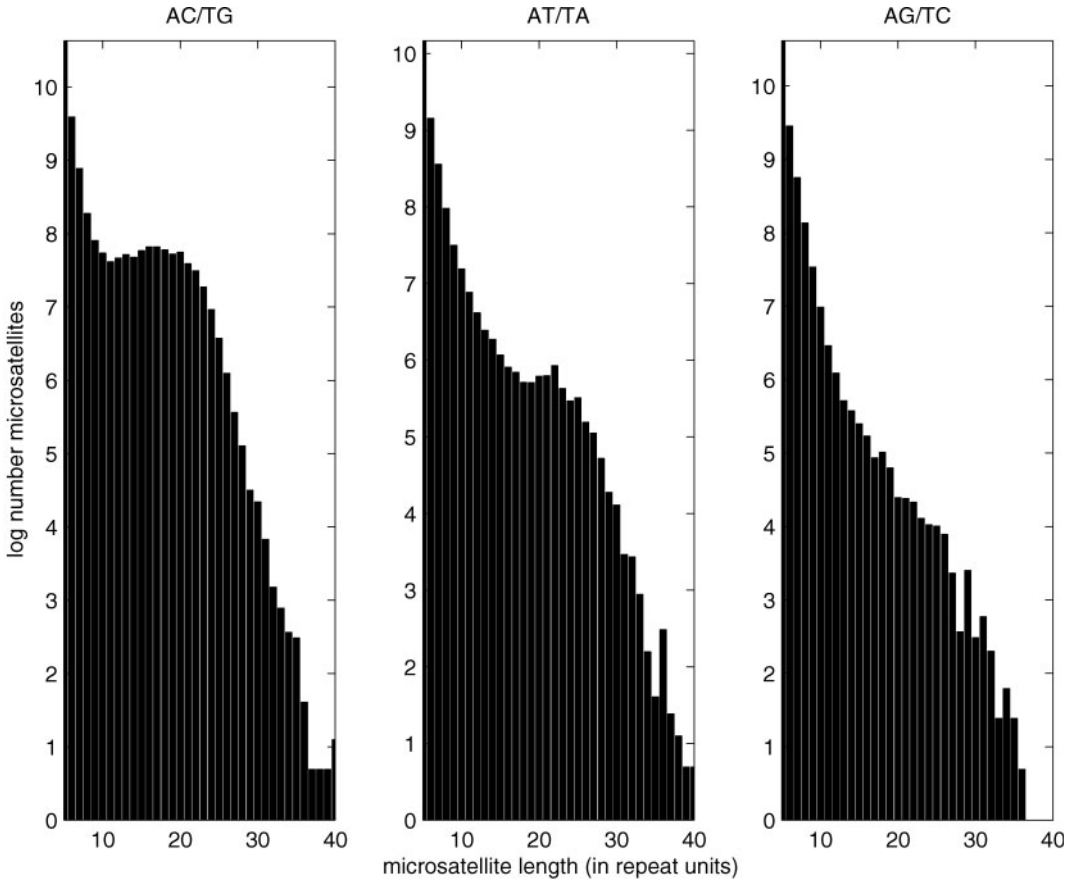

FIG. 3.—Human microsatellite distribution: Natural logarithm of one plus the number of isolated, perfect microsatellites of different lengths.

For the human genome, we solved for the stationary distribution of microsatellite lengths $lo = 10, \ldots, hi = 45$ to 10 for two reasons: (1) we are most interested in the mutational bias of long microsatellites because there is evidence that the slippage threshold may be as high as 8 for the AC/TG motif in humans (Weber 1990), and (2) the large number of dinucleotide repeats in the human genome allows us to restrict our attention to those with more than 10 repeats and still have a large amount of data.

For all models and motifs, the AIC score differences are listed in Table 5. For each column, the actual AIC score is obtained by adding the term in the last row. For the PSwK model the maximum-likelihood κ is only non-zero for the AG/TC motif; therefore only for this motif is the model different from the PS model. Only for the AT/TA motif do the PS and PSwK models have lower AIC scores than the SMM. None of these symmetric models have lower scores than the multinomial model.

For all three motifs, the asymmetric models have lower AIC scores than the symmetric models, and long microsatellites have a bias toward contractions. For the AsyLin model, for the AC/TG motif, microsatellites with length greater than 24 have a bias down; for the AT/TA motif, all microsatellites have a bias down; and for the AG/TC motif, microsatellites longer than 14 have a bias down. For all three motifs, the two best models are the AsyQuad and PLBias models. For the AsyQuad (PLBias) model, for the AC/TG motif, microsatellites longer than 23 (18) have a bias down; for the AT/TA motif, microsatellites shorter than 14 (17) and longer than 26 (22) have a bias down; and for the AG/TC motif, all microsatellites have a bias down. For both of these models, figure 4 shows the probability of contraction as a function of microsatellite length. With one exception (AsyQuad model, AC/TG motif), for all three motifs, these two models score better than the multinomial model.

Table 5
Human AIC Model Scores

Model	# parameters	AIC		
		AC/TG	AT/TA	AG/TC
Multinomial	35	0	0	0
SMM	1	8,522	701	320
PS	1	13,104	939	163
PSwK	2	13,106	941	110
ConExp	3	210	664	211
AsyLin	4	200	673	60
AsyQuad	6	8	−21	−19
PLBias	7	−12	−11	−22
		+190,083	+48,014	+20,551

NOTE.—To make the scores easier to compare, for each column we have subtracted the number listed in the last row.

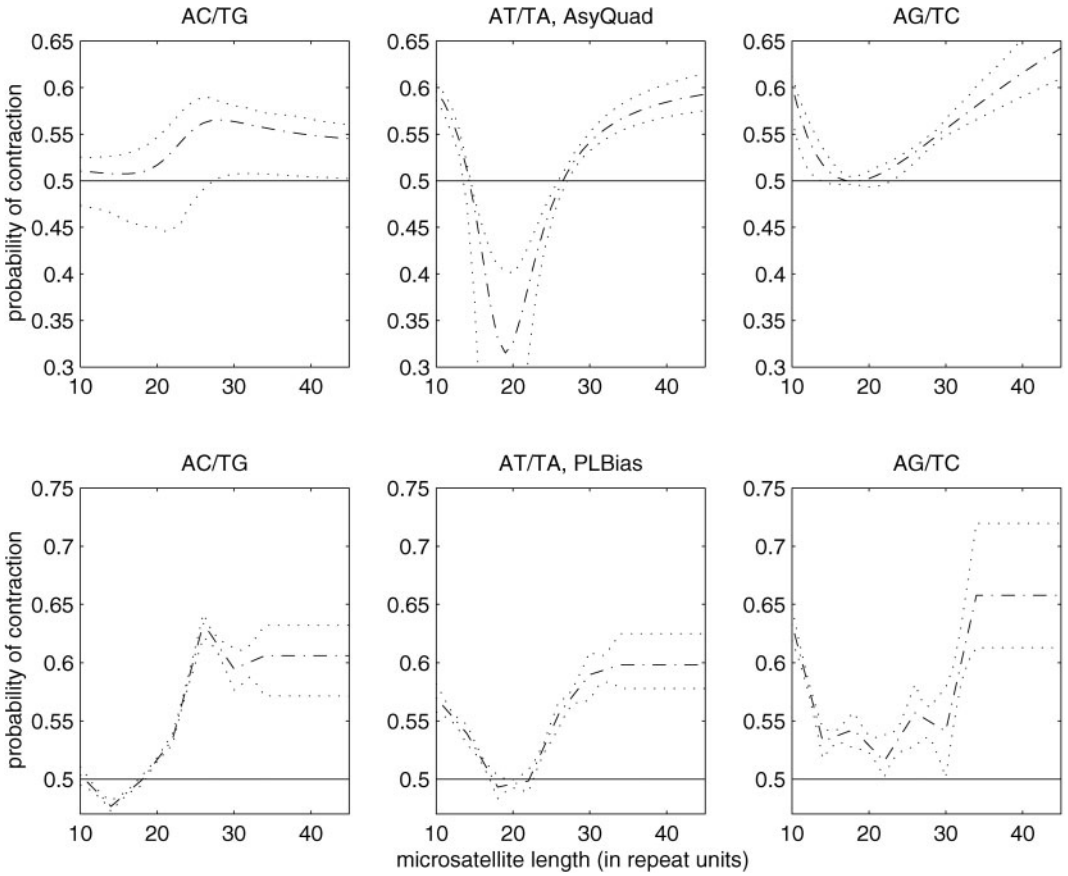

FIG. 4.—Human: Probability of contraction as a function of microsatellite length in AsyQuad and PLBias model fits. The first row is the AsyQuad model, and the second row is the PLBias model. The dotted and dashed line is the model fit; the dotted lines form the 95% confidence envelope.

Possible Contributing Factors

In an attempt to explain the observation that dinucleotide repeats in humans have very different distributions, we investigated several possible contributing factors: local recombination rates, proximity to genes, local GC content, location on the chromosome, and proximity to *Alu* repeats identified by Smit and Green's RepeatMasker algorithm (http://ftp.genome.washington.edu/RM/RepeatMasker.html). As in the previous section, we analyze each motif separately, and we only consider dinucleotide microsatellites in the human genome that are isolated, perfect, and have length longer than 10.

We compare pairs of microsatellite distributions from regions of the genome with contrasting properties, for example, recombination hot and cold spots. Let $\{l_i(j)\}$, $i = 1, 2$, be the two samples where in sample i there are $l_i(j)$ microsatellites of length j. Define $Z_i = \sum_j l_i(j)$, the number of microsatellites in sample i, and $p_i(j) = l_i(j)/Z_i$, the percentage of microsatellites in sample i with length j. Motivated by the χ^2 statistic, we introduce

$$X^2 = \sum_j \frac{(.5(Z_1 + Z_2) \times (p_1(j) - p_2(j)))^2}{.5(Z_1 + Z_2) \times .5(p_1(j) + p_2(j))}.$$

After normalizing for the different size of the two samples, each term in the sum represents, for a given length, the squared difference of the number of microsatellites in the two samples, divided by the average number of microsatellites. Thus if the two samples are drawn from a similar distribution, X^2 will be small; otherwise; it will be large. To determine confidence intervals we use the bootstrap (see e.g., Efron and Tibshirani 1993): for $i = 1, 2$ choose Z_i values with replacement from the combined sample $\{l_1(j) + l_2(j)\}$, recompute X^2, and repeat. This test determines whether distributions are significantly different, but it does not specify what properties differentiate the distributions.

For recombination, we compare the distributional shape of two extremes: recombination hot and cold spots. Yu et al. (2001) identified 11 recombination hot spots totaling 37 million bases on chromosomes 3, 4, 5, 10, 14, 15, 16, 17, 18, and X, and 19 recombination cold spots totaling 57 million bases on chromosomes 3, 4, 5, 6, 8, 10, 11, 12, 13, 18, and 20. For all comparisons in this section, the breakdown by motif is listed in table 6. For none of the motifs were the distributions significantly different.

For the next four comparisons we only consider Chromosome 1. Chromosome 1 contains roughly 280 million bases, or 9% of the human genome. The first comparison is proximity to genes. We downloaded an annotated version of Chromosome 1 from the National Center for Biotechnology Information: ftp://ncbi.nlm.nih.gov/genomes/H_sapiens/CHR_01/. This version contains 1,239 genes. If we include the entire region labeled

''gene,'' which is the coding sequence including introns and a 500-base pair flanking window, this ''genic region'' numbers 48 million bases. For the AC/TG motif, the two distributions are significantly different ($P = 0.003$), however, as figure 5 shows, the differences are fairly subtle. There is a tendency for microsatellites in genic regions to be shorter. For the other two motifs, the distributions are not significantly different.

Next we examine the GC content of flanking sequences. For each microsatellite, we measure the GC content in 100-base and 1,000-base windows on both sides of the microsatellite. We further restrict the set of microsatellites so that flanking windows do not overlap. For each window size, we separate the microsatellites into three sets based on the GC content of their flanking windows, and then compare all three combinations of these three sets. The three sets are ''low GC content,'' one standard deviation less than the mean; ''medium GC content,'' within one standard deviation of the mean; and ''high GC content,'' one standard deviation greater than the mean. For the 100-base window and the AC/TG motif, two comparisons are significant: low GC content versus medium GC content ($P < 0.001$), and low GC content versus high GC content ($P < 0.001$). The distributions are shown in figure 6. Comparing the histograms suggests that microsatellites in regions with low GC content tend to be shorter. This is somewhat surprising in view of the previous comparison, because genes tend to be in GC-rich regions. None of the other comparisons are significant for the 100-base window, and none are significant for the 1,000-base window.

For the question of chromosome location, we break Chromosome 1 into two halves, each having an equal number of base pairs. For none of the motifs do the different halves have significantly different distributions. Then we break one of the halves into thirds, each piece having an equal number of base pairs. For each motif, we compare all three pairs, and again there are no significant differences.

Finally we consider proximity to *Alu* repeats identified by the RepeatMasker algorithm. For each microsatellite, we consider a 500-base window on each side of the microsatellite. We run the RepeatMasker algorithm on each window, and if there is an *Alu* within 50 bases of the microsatellite we consider this microsatellite close to a repeat; otherwise, we do not. For none of the motifs are the two distributions significantly different.

Table 6
Human Comparisons

	Number Microsatellites Isolated, Perfect, Length $\geq$ 10		
Comparison	AC/TG	AT/TA	AG/TC
Recombination			
hot spots, 37×10^6 bases	426	95	61
cold spots, 57×10^6 bases	562	154	65
On chromosome 1			
Genes			
Genic region, 48×10^6 bases	530	164	77
Nongenic region, 232×10^6 bases	2,416	672	394
GC content 100-base windows			
$< 32\%$	429	161	41
$> 32\%, < 48\%$	1,798	504	307
$> 48\%$	519	82	98
1,000-base windows:			
$< 35\%$	320	112	48
$> 35\%, < 47\%$	1,521	445	233
$> 47\%$	363	54	71
Chromosome location			
One half	1,338	332	168
Other half	1,252	321	179
Subdivide one half			
Extreme end	408	97	61
Middle third	448	108	45
Centromere end	482	127	62
RepeatMasker			
Alu	460	270	104
no Alu	2,030	413	293

Discussion

For both Drosophila and humans, and for all motifs, the simple, symmetric models had AIC scores that were higher than the multinomial model. This contrasts with the study of Sibly, Whitaker, and Talbot (2001), which used a much smaller data set including only 186 dinuceotide microsatellites (all motifs combined) and found that the PSwK model is a significant improvement over the multinomial model. The asymmetric models we have considered fit much better than the symmetric ones, and in some cases they have lower AIC scores than the multinomial model. These asymmetric models allow for biases up or down, and for the bias to change as a function of microsatellite length. However, for all asymmetric models, motifs, and both organisms, the optimal parameters always imply a bias down for long microsatellites. This bias could explain two mysteries from our previous work (Calabrese, Durrett, and Aquadro 2001): the dearth of long microsatellites and the underestimation of divergence times. Throughout, we have only considered models with single-step mutations; however, if a model allows for larger mutations it will still have a stationary distribution and our approach will apply. Moreover, we have assumed that the same model and parameters apply throughout the genome.

In humans, there is a striking difference between the distributions of various dinucleotide and trinucleotide motifs (see figs. 3 and 7). To our knowledge, this is the first report of this observation. An appealing explanation that is apparently not valid is that the different distributions result from different strength bonds. The A-T opposing strand pair contains two hydrogen bonds, and the G-C pair contains three hydrogen bonds. However, the AC/TG and AG/TC motifs each have one strong bond and one weak bond, and their distributions are the most different from each other, whereas the AT/TA distribution (two weak bonds) appears to be in between these two. A second possible explanation for the dependence on the repeat type is that the DNA mismatch repair system is sensitive to motif content. There is experimental evidence for this explanation both in Drosophila (Harr, Todorova, and Schlötterer 2002) and in yeast (Sia et al. 2000). A third

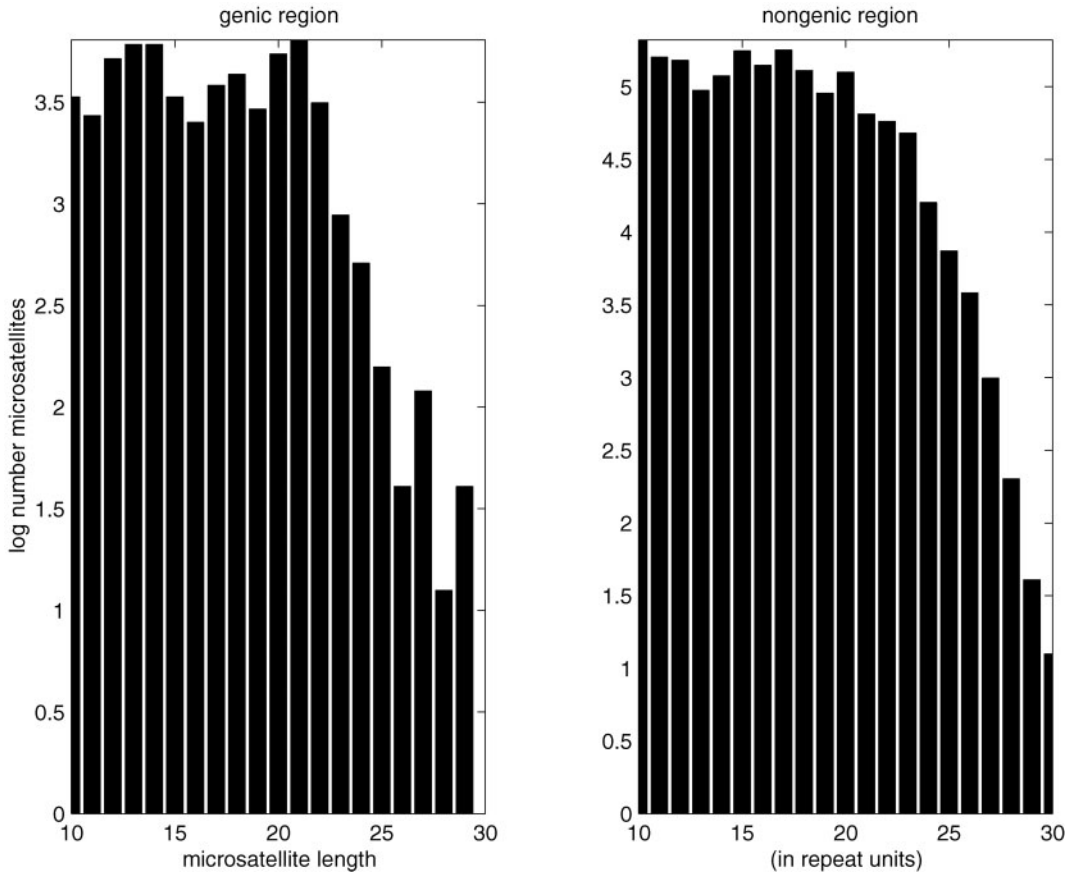

FIG. 5.—Genic and nongenic regions: Natural logarithm of one plus the number of human, AC/TG, isolated, and perfect microsatellites on Chromosome 1 of different lengths.

possible explanation is selection. For the AC/TG motif, we have reported a significant difference in the distribution of microsatellites in genic and nongenic regions. A fourth explanation could involve the GC content of the flanking regions, because for the AC/TG motif, we have reported a significant difference in distributional shape for sets with different flanking region GC content.

It is interesting that only the AC/TG motif had a significantly different distribution for any of the comparisons we performed. However, it should be noted that the differences in the distribution of the lengths of AC/TG repeats, while statistically significant is visually rather subtle. The striking difference in the distributions of the three repeat types must have some definite cause, but we have not been able to determine the underlying mechanism.

Appendix

In the usual queuing theory terminology, microsatellites correspond to customers, microsatellite lengths correspond to stations, and at each length or station there are an infinite number of servers. The stations of the network are the microsatellite lengths $lo, lo+1, \ldots, hi \ldots, top$. The "$M/M/\infty$" translates (1) that microsatellites are born at length lo at the times of a Poisson process, (2) that microsatellites stay at a given length for an independent exponential amount of time before either mutating to another length or being interrupted by a point mutation and exiting the network of queues entirely, and (3) that each microsatellite moves through the various lengths independently. Because it affects only the number of microsatellites and not their distribution, the rate λ of the Poisson birth process does not matter. For each length j, the rate $\mu(j)$ of the exponential holding time, and the probability $p(j, i)$ a microsatellite will mutate next to length i or exit the network $q(j) = 1 - \sum_{i=lo}^{top} p(j, i)$, depend on the mutation model and its parameters.

Because all microsatellites have a positive probability of leaving the network, there exists a stationary distribution. Define arrival rates,

$$r(lo) = \lambda + \sum_{i=lo}^{top} r(i)p(i, lo) \quad (1)$$

$$r(j) = \sum_{i=lo}^{top} r(i)p(i,j), \; j > lo. \quad (2)$$

Then the stationary distribution is for all j the number of microsatellites with length j are independent and Poisson distributed with mean $r(j)/\mu(j)$ (see e.g., Durrett 1999, p. 192). Let $\{l(j)\}_{j=lo}^{hi}$ be the data, $l(j)$ microsatellites of length j, and define the normalizing constant

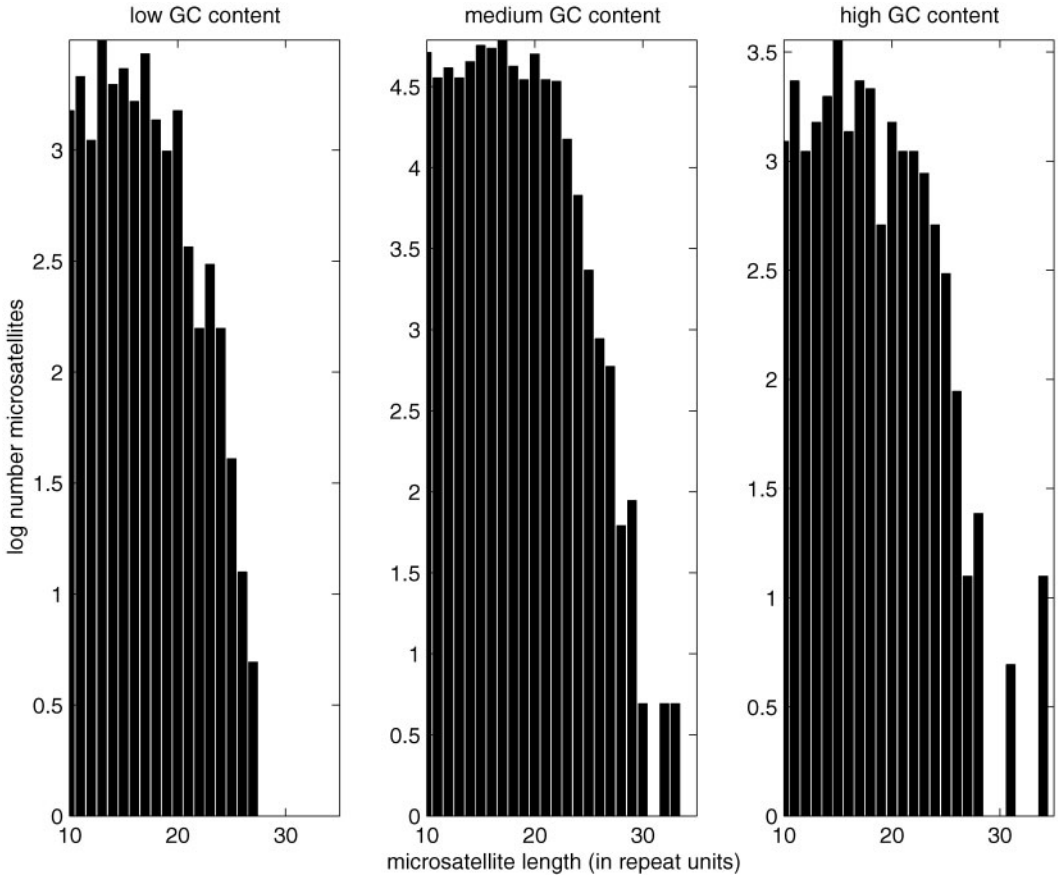

FIG. 6.—Different flanking region GC content: Natural logarithm of one plus the number of human, AC/TG, isolated, and perfect microsatellites on Chromosome 1 of different lengths.

$$Z = \sum_{j=lo}^{hi} \frac{r(j)}{\mu(j)}.$$

Because we have assumed that the microsatellites evolve independently, conditioning on the number of microsatellites the likelihood of the data is,

$$\prod_{j=lo}^{hi} \left(\frac{r(j)}{Z\mu(j)} \right)^{l(j)}. \quad (3)$$

For each model, we numerically solve the linear system of equations (1), (2) to determine $r(j)$, and then numerically maximize the likelihood (3) over the space of parameters.

One last detail concerns the boundary points *lo*, *hi*, and *top*. Conditioning on microsatellites having lengths between *lo* and *hi*, we solve for the stationary distribution of this range. First we discuss the lower boundary. For the models we consider, we only detail the slippage law of microsatellites with lengths greater than or equal to *lo*. While shorter microsatellites may mutate, we only begin to watch them when their length is greater than or equal to *lo*. We assume that from the pool of microsatellites of length less that *lo*, new microsatellites of length *lo* are produced, either by slippage or by point mutation, at the times of a Poisson process. Next we discuss the upper boundary. To compute the stationary distribution as outlined above, we need to solve the *finite* system of coupled linear equations (1), (2). So we only detail the slippage law of microsatellites with length less than or equal to *top*, where $top > hi$. But the models we want to consider have no such upper boundary. Therefore we actually calculate the stationary distribution of two related networks which bound the stationary distribution for the model of interest. The laws for these two networks are identical to the true model except in the way they handle microsatellites of length greater than *top*. If a microsatellite of length less than or equal to *top* mutates to a length greater than *top* in the true model, in the lower-bound network it exits the network entirely. Because in the true model these long microsatellites conceivably could mutate back downward, this bounding network is a lower bound. Likewise if a microsatellite of length less than or equal to *top* mutates to a length greater than *top* in the true model, in the upper-bound network it goes to length *top*. Because all models considered only allow single steps, they have the monotonic property $p(top, j) \geq p(x, j)$ for all $top \geq j$, $x > top$; and this bounding network is an upper bound. If the stationary distribution for microsatellite lengths $lo, lo+1, \ldots, hi$ of the two bounding networks is sufficiently close, we use their average; otherwise, we increase *top* until it is sufficiently close. To be ''suffi-

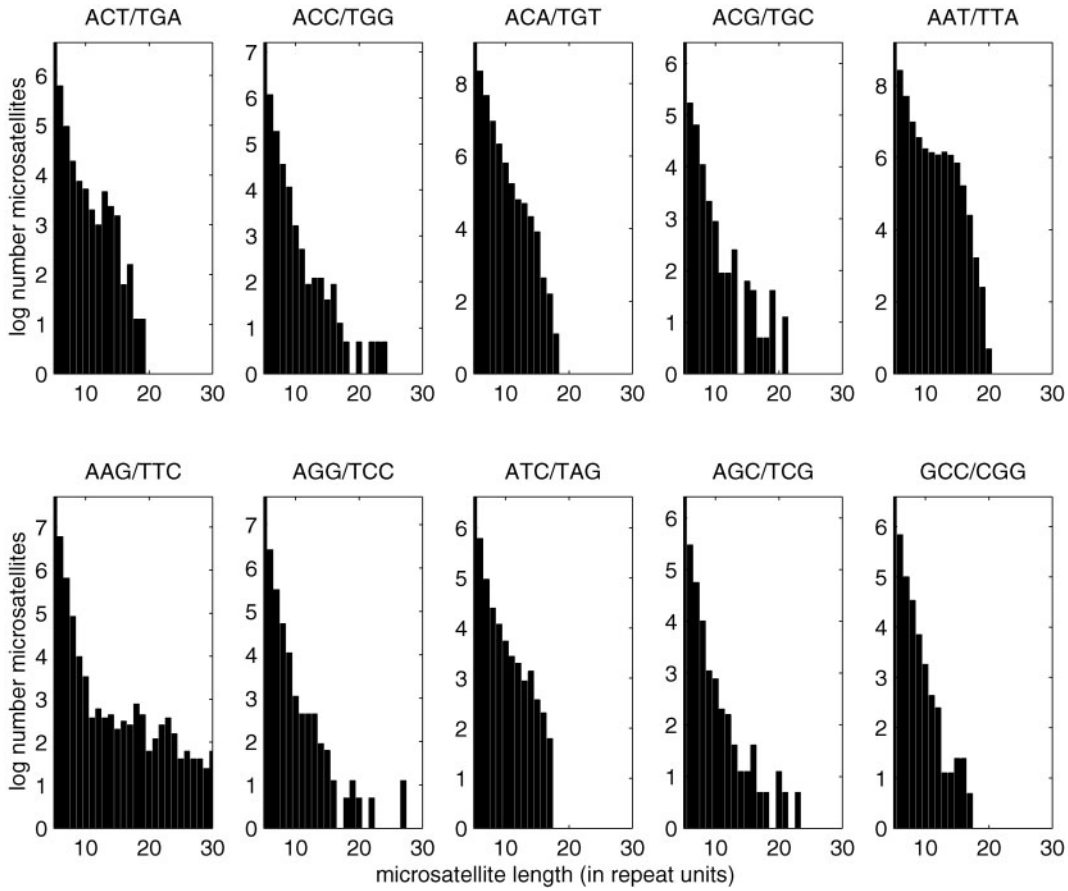

FIG. 7.—Trinucleotides, human microsatellite distribution: Natural logarithm of one plus the number of isolated, perfect microsatellites of different lengths.

ciently close,'' for each of the lengths lo, $lo+1, \ldots, hi$, we require that the relative difference in probabilities between the two networks be less than 10^{-5}.

Acknowledgments

P.C. is supported by a National Science Foundation mathematics postdoctoral fellowship. R.D. is supported by a grant from the probability program at the National Science Foundation. The authors thank Rasmus Nielsen for numerous discussions, and the referees for their suggestions.

Literature Cited

Akaike, H. 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control **19**:716–723.

Amos, W., S. J. Sawcer, R. W. Feakes, and D. C. Rubinsztein. 1996. Microsatellites show mutational bias and heterozygote instability. Nat. Genet. **13**:390–391.

Calabrese, P. P., R. T. Durrett, and C. F. Aquadro. 2001. Dynamics of microsatellite divergence and proportional slippage/point mutation models. Genetics **159**:839–852.

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

Durrett, R. 1999. Essentials of stochastic processes. Springer, New York.

Efron, B., and R. J. Tibshirani. 1993. An introduction to the bootstrap. Chapman & Hall, New York.

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.

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., J. Todorova, and C. Schlötterer. 2002. Mismatch repair-driven mutational bias in *D. melanogaster*. Mol. Cell **10**:199–205.

Huang, Q-Y., F-H. Xu, H. Shen, H-Y. Deng, Y-J. Liu, Y-Z. Liu, J-L. Li, R. R. Recker, and H-W. Deng. 2002. Mutational patterns at dinucleotide microsatellite loci in humans. Am. J. Hum. Genet. **70**:625–634.

Huelsenbeck, J. P., and B. Rannala. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science **276**:227–232.

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 an be explained by a balance between slippage events and point mutations. Mol. Biol. Evol. **17**:1210–1219.

Primmer, C. G., H. Ellegren, N. Saino, and A. P. Moller. 1996. Directional evolution in germline microsatellite mutations. Nat. Genet. **13**:391–393.

Sia, E. A., C. A. Butler, M. Dominska, P. Greenwell, T. D. Fox, and T. D. Petes. 2000. Analysis of microsatellite mutations in the mitochondrial DNA of *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA **97**:250–255.

Sibly, R. M., J. C. Whittaker, and M. Talbot. 2001. A maximum-likelihood approach to fitting equilibrium models of microsatellite evolution. Mol. Biol. Evol. **18**:413–417.

Weber, J. L. 1990. Informativeness of human $(dC - dA)_n \cdot (dG - dT)_n$ polymorphisms. Genomics **7**:524–530.

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

Yu, A., C. Zhao, Y. Fan et al. (11 co-authors). 2001. Comparison of human genetic and sequence-based physical maps. Nature **409**:951–953.

Wolfgang Stephen, Associate Editor

Accepted December 20, 2002