STATISTICAL INFERENCE OF MICROSATELLITE MODELS: AN APPLICATION TO HUMANS AND CHIMPANZEES

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Master of Science

by

Raazesh Sainudiin

August 2003

ABSTRACT

We compare and test several models of microsatellite evolution found in the literature in a likelihood framework using genomic data from homologous microsatellite loci in humans and chimpanzees. Hypotheses regarding the relative significance of many qualitative features are tested using classical likelihood ratio tests and the information-theoretic Akaike information criterion. These may include: (i) *proportionality* in the mutation rate; whereby long microsatellites mutate more frequently than shorter ones, (ii) *bias* in the mutational process; whereby the probability of a contraction upon mutation may depend on repeat length and (iii) *phase* of the mutations, whereby mutations result in the instantaneous loss or gain of 1 (one-phase) or more (two-phase) repeat units.

A proportional-rate, linear-biased, one-phase model emerges as the best model. A focal length towards which the mutational process is linearly biased in the presence of rate proportionality, is a crucial feature of microsatellite evolution. Such a focal bias may be due to the counteracting forces of replication slippage and repair and/or natural selection against longer microsatellites. We find little support for a two-phase model, in which more than one unit of repeat length may be gained or lost by a single mutation. We also assess the performance of these models based on the fit of their stationary distributions to the empirical distribution of microsatellite lengths in the human genome and find the results to be consistent with those based on the human and chimp comparison.

The mutational mechanisms of AC-repeats are found to differ significantly from those of AT or AG-repeats. Microsatellites interrupted by even a single point mutation exhibit a two-fold decrease in their mutation rate and a four-fold decrease in the per-repeat unit slippage rate when compared to pure AC repeats. Some competing theories explaining the phenomenon of longer repeat length in humans relative to those in the chimpanzees are explored. In general, models that allow chimps to have a larger per-repeat unit slippage rate and/or a shorter focal allele compared to humans give a better fit to the human-chimp data as well as the human genomic data. Species-specific differences in mutational mechanisms as well as relaxation of selection against longer microsatellites in humans are compatible with the data.

BIOGRAPHICAL SKETCH

Raazesh Sainudiin was born in Chennai (Madras), Tamil Nadu, India on November 15, 1973 to S. Nagalakshmi and A. K. Sainuddin. He grew up in Tamil Nadu and Minnesota. In Tamil Nadu he developed a deep appreciation for Euclidean geometry and calculus under the guidance of Ms. R. Balamani and Mr. M. Sivanu Pandian. He is most thankful to Mr. Pandian for illuminating why $\frac{d}{dx}x^3$ is $3x^2$ with a view from the origin toward the three faces of an expanding cube and guiding him through a dynamic geometric understanding of Calculus. Dr. Rodney O. Davis of Gustavus Adolphus College taught him to articulate in English and comprehend others' thoughts with a keen awareness of the implicit biases. Dr. Keith K. Klien of Minnesota State University, Mankato showed him the beauty of evolutionary genetics. In the Spring of 1999, he graduated B.S. summa cum laude with a double major in Biology and Mathematics from Minnesota State University, Mankato. He is most thankful to the tax-payers of Minnesotta for heavily subsidizing the cost of his undergraduate education. In the Summer of 2003, he received his M.S. in Biometry from Cornell University.

To four wonderful women; grandmother, mother, wife and daughter.

ACKNOWLEDGMENTS

Raazesh would like to thank Rasmus Nielsen for being a great advisor and guiding him with the utmost enthusiasm through stochastic molecular evolution, Rick Durrett for always being there to crack his fabrications upon, and Chip Aquadro for keeping him biologically focussed. Krishna Athreya, Tessa Bauer, Dave Capella, Michael Nussbaum, Sydney Resnick, Laurent Saloff-Coste, Gennady Samorodnitsky, James Signorovitch, Robert Strawderman, Floyd Reed, Guy Reeves, and Krithika Yogeeswaran have been instrumental in strengthening his intuitions. He is thankful to every American tax-payer for supporting him through an IGERT grant from NSF DGE-9870631, a computational biology grant from NSF to Carlos Castillo-Chavez, and grants to Rasmus Nielsen.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

Microsatellites are tandem repeats of short DNA motifs between two and five base pairs, usually characterized by their repeat length. Their high length variability, genome-wide distribution, and abundance make them useful for evolutionary and population genetic inference in areas as diverse as molecular forensics, parentage testing, molecular anthropology, conservation genetics, and in studies of human evolutionary history [19]. Population genetic inferences may be sensitive to the assumed model of microsatellite evolution. Therefore, much focus has centered on the development of biologically realistic models. However, there has been relatively little focus on testing and comparing these models using real data.

The simplest popular model of microsatellite evolution is the classical stepwise mutation model (*SMM*) of Ohta and Kimura [24] in which, upon a mutation, one repeat unit is either gained resulting in an expansion or lost resulting in a contraction. However, a small proportion of mutations have been observed to change the repeat length by more than one unit [17] [16]. The two-phase model (*TPM*) of Di Rienzo *et al* [9] addresses this by allowing mutations of 1 repeat unit (one-phase) with probability p, and mutations of $\geq$1 unit(s) (two-phase) with probability $1-p$, while the distribution of the lengths of multi-unit mutations is geometric. In a simpler two-phase model of Fu and Chakraborty [13] mutations of length ≥ 1 are geometrically distributed. Under the *SMM* and the *TPM*, a microsatellite is assumed to mutate at a constant rate, irrespective of its repeat length, which may be any integer. Moreover, under these models there is no bias toward an expansion or a contraction, and thus the microsatellites are expected to grow or contract unconstrained over time. While constraining the range of repeat lengths through a model with reflecting boundaries [12] can circumvent this problem of unbounded growth, the biological reality of such a defined boundary is unclear.

Evidence for length dependent effects on mutation rate [10], whereby longer microsatellites mutate more often than shorter ones, and the presence of point mutations in some repeats, make the proportional slippage (*PS*) model of Kruglyak *et al* [20] and its extensions by Calabrese *et al* [6] attractive. In the symmetric *PS* model, an equilibrium distribution of repeat lengths exists through a balance between slippage events and point mutations [20]. Various mutational biases have been observed including an upward bias favoring expansions in humans [1] and barn swallows [29], an excess of contractions in long microsatellites of yeast [35] and fruit fly [15], and the rate of contractions increasing exponentially with repeat length in humans [37]. Thus, models that incorporate mutational bias are biologically appealing. In the presence of a linear bias toward a target or focal length, as proposed by Garza, Slatkin, and Freimer [14], microsatellites below the focal length tend to expand, and those above it tend to contract. Other models emphasize mutational bias by allowing the probability of an expansion upon mutation to be independent of repeat length [13] or be dependent on it exponentially, linearly, quadratically, or piece-wise linearly [5].

Thus, broadly speaking, there are at least three qualitatively contrasting features in the existing models of microsatellite evolution. The first is one-phase versus two-phase mutations. The second is mutation rate proportionality, the proportional dependence of mutation rate on repeat length, versus rate equality. The final contrasting feature is the presence or absence of mutational bias, whereby the probability of expansion upon mutation may depend on the repeat length of the mutating microsatellite in one form or another. We only address constant bias, where the probability that a mutation results in an expansion is constant for all alleles, and linear bias, where this probability varies linearly with repeat length.

We test the relative significance of these contrasting features, as embodied by variants of some popular models and their hybrids, using data from dinucleotide loci homologous between humans (*Homo sapiens*) and chimps (*Pan troglodytes*), through likelihood ratio tests (LRTs) and the Akaike information criterion (AIC). Complications to the mutational process from variation in repeat motif as well as interruptions by point mutations are also explored. We address the question of longer repeat length in humans compared to chimps through a lineage-specific analysis.

Comparison of models in the past has often been limited to establishing the supremacy of some particular class of models over another simpler class. Using homologous loci from humans and chimps, Webster *et al* [34] have shed light on the heterogeneity in the mutation process through a descriptive analysis. We provide a rigorous statistical framework to compare several popular microsatellite models used by biologists as well as test motif-specific and lineage-specific hypotheses about mechanisms underlying microsatellite evolution using species-pair data.

Chapter 2

THEORY

For mathematical convenience, most models of microsatellite evolution assume that a microsatellite can attain any integer in repeat length. We analyze only those models whose behavior can be approximated by Markov Chains on a truncated state space $\mathbb{S} = \{\kappa, \kappa + 1, \cdots, \Omega\}$, the set containing all possible repeat lengths a microsatellite is allowed to attain. We denote a microsatellite allele by its repeat length i. Truncation of the state space from above is biologically reasonable, as microsatellites are rarely longer than Ω (a few tens of repeat units), and that from below ensures that κ is greater than the threshold repeat length above which mutations in length occur that are characteristic of microsatellites [30].

The data D is a $2 \times N$ matrix of microsatellite allele lengths from N loci homologous in humans and chimps. We model the distribution of D by superimposing three Markov chains, $\mathbf{X}^{(a)}$, $\mathbf{X}^{(c)}$, and $\mathbf{X}^{(h)}$, on the ancestral, chimp, and human branches, respectively, of the two taxa tree $\boldsymbol{\tau}$, as shown in Figure 2.1. In $\boldsymbol{\tau}$, each of the two terminal branch lengths, λ_c and λ_h, represents the product of mutation rate at allele κ and number of generations along the chimp and human lineages, respectively. We assume that the time to coalescence for a pair of homologous alleles, within the ancestral population, is negligible relative to the time since the human-chimp speciation.

Let $\Theta^{(a)}$, $\Theta^{(c)}$, and $\Theta^{(h)}$ be parameters of the Markov chains $\mathbf{X}^{(a)}$, $\mathbf{X}^{(c)}$, and $\mathbf{X}^{(h)}$, with transition probability matrices, $\mathbf{P}^{(a)}$, $\mathbf{P}^{(c)}$, and $\mathbf{P}^{(h)}$, respectively. For an ergodic continuous time Markov chain, its transition probability matrix $\mathbf{P}(\lambda)$ $:= (P_{i,j})_{i,j=\kappa}^{\Omega} = \exp\{\mathbf{Q}\lambda\}$, where $\mathbf{Q} := (q_{i,j})_{i,j=\kappa}^{\Omega}$ is its infinitesimal generator or rate matrix. The stationary distribution of such a Markov Chain, denoted by $\boldsymbol{\pi} = (\pi_\kappa, \pi_{\kappa+1}, \cdots, \pi_\Omega)$, is the unique probability distribution on $\mathbb{S}$ satisfying the matrix equation $\boldsymbol{\pi}\mathbf{Q} = \mathbf{0} = (0, 0, \cdots, 0)$ (see for e.g. [3]). Interest in $\mathbf{P}(\lambda)$ and $\boldsymbol{\pi}$ arises because they determine the likelihood function L_i in Equation (2.1).

Let $\boldsymbol{\pi}^{(\boldsymbol{a})}$ be the stationary distribution of the ancestral chain. Defining $\boldsymbol{\Theta} := (\Theta^{(a)}, \Theta^{(c)}, \Theta^{(h)})$, and $\boldsymbol{\lambda} := (\lambda_c, \lambda_h)$, the likelihood, given homologous allele length data $D_i = (C_i, H_i)$ at locus i is:

$$L_i(\boldsymbol{\Theta}, \boldsymbol{\lambda}|D_i) := \sum_{j \,\in\, \mathbb{S}} \pi_j^{(a)} \; P_{j,C_i}^{(c)}(\lambda_c) \; P_{j,H_i}^{(h)}(\lambda_h). \tag{2.1}$$

Since we do not know the ancestral state, the likelihood may be thought of as a weighted sum over all possible ancestral states, where the weights come from the stationary distribution of the ancestral chain. Assuming independence (free recombination) among the N loci, the likelihood, given the entire data D, is obtained by multiplication.

$$L(\boldsymbol{\Theta}, \boldsymbol{\lambda}|D) := \prod_{i=1}^{N} L_i(\boldsymbol{\Theta}, \boldsymbol{\lambda}|D_i). \tag{2.2}$$

2.1 Model M

We start by defining a general model $\mathbf{M}$, in which all other models of interest are nested. A continuous time Markov Chain $\mathbf{X}^M$ on $\mathbb{S}$ is defined with an

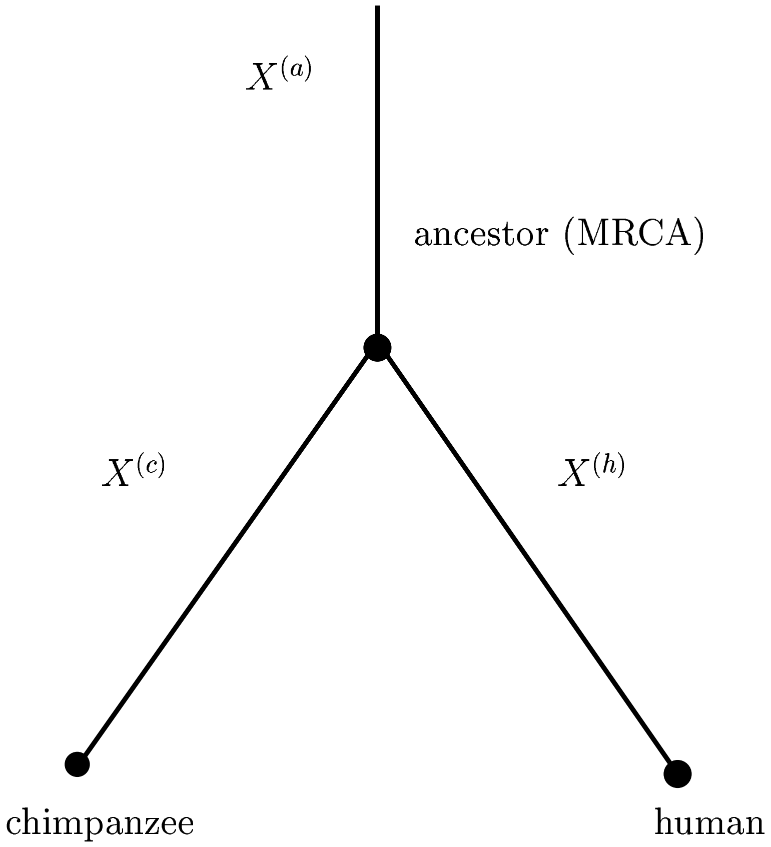

Figure 2.1: Markov Chains on the branch leading to the ancestor ($X^{(a)}$), chimpanzee ($X^{(c)}$), and human ($X^{(h)}$)

infinitesimal Generator $\mathbf{Q}^M$ given by,

$$q^M_{i,j} = \begin{cases} \beta(i,s)\ \alpha(u,v,i)\ (p\ + (1-p)\ \gamma(m,i,j))\ , & i = j-1 \\ \beta(i,s)\ \alpha(u,v,i)\ (1-p)\ \gamma(m,i,j)\ , & i < j-1 \\ \beta(i,s)\ (1-\alpha(u,v,i))\ (p\ + (1-p)\ \gamma(m,i,j))\ , & i = j+1 \\ \beta(i,s)\ (1-\alpha(u,v,i))\ (1-p)\ \gamma(m,i,j)\ , & i > j+1 \\ -\sum_{i \neq j} q^M_{i,j}\ , & i = j. \end{cases} \tag{2.3}$$

where, the functions α, β and γ are defined as follows,

$$\begin{aligned} \beta(i,s) &= \mu(1 + (i-\kappa)s), \\ \alpha(u,v,i) &= \max\{0, \min\{1, u - v(i-\kappa)\}\}, \end{aligned}$$

$$\gamma(m,i,j) = \begin{cases} \frac{m(1-m)^{|i-j|-1}}{1-(1-m)^{\Omega-i}}, & \kappa \leq i < j \leq \Omega \\ \\ \frac{m(1-m)^{|i-j|-1}}{1-(1-m)^{i-\kappa}}, & \kappa \leq j < i \leq \Omega. \end{cases}$$

Allele i mutates at rate $\beta(i,s)$. The proportional dependence of mutation rate on repeat length is captured by the proportional rate parameter $s \in (-\frac{1}{\Omega-\kappa+1}, \infty)$. When s=0, alleles of all lengths have the same mutation rate $\mu \in (0,\infty)$ of allele κ. Observe that $1/\beta(i,s)$ is the average amount of time spent by a microsatellite locus in an allele of repeat length i (mean holding time in allele i).

Upon a mutation at allele i, the probability that it is an expansion is $\alpha(u,v,i)$, and that it is a contraction is $1-\alpha(u,v,i)$. In the function $\alpha(u,v,i)$, the constant bias parameter is $u \in [0,1]$ and the linear bias parameter is $v \in (-\infty, +\infty)$. If u=0.5 and v=0, we have a symmetric unbiased mutational process in which the probability that a mutation is an expansion or a contraction is equal. If v=0, then $\alpha(u,v,i)$=$u \in [0,1]$ for any allele i, and we have a model with constant mutational bias. Furthermore, we have a linear mutational bias when $v{\neq}0$. If $0.5 < u < 1$ and $\frac{u-0.5}{\Omega-\kappa} < v < \infty$, we have a *focal* allele near $((u-0.5)/v)+\kappa$, at which the probability of contraction equals that of expansion ($\alpha(u,v,f)$=0.5), and towards which the mutational process is linearly biased. So, when $i < f$, the mutational bias is upwards, towards f, since $\alpha(u,v,i) > 0.5$, and when $i > f$, the bias is downwards, towards f, as $\alpha(u,v,i) < 0.5$.

When p=1, any microsatellite allele mutates (*i.e.* expands or contracts) by only one unit of repeat length, but when p is less than 1, it mutates by one or more unit(s) of length with probability $1-p$ and by one unit of length with probability p. Given that an allele i, undegoes a multi-step mutation, the probability of expanding or contracting by k units is given by $\gamma(m,i,j)$, a conditional geometric distribution with success probability m.

Observe that for every allele i, $\sum_{j=i+1}^{\Omega}\gamma(m,i,j) = \sum_{j=\kappa}^{i-1}\gamma(m,i,j) = 1$. The probability of a transition from allele i to j in t generations, under model **M**, with a mutation rate μ at κ, is given by $P_{i,j}^{M}(\lambda)$, where $\lambda = \mu\ t$.

Below we describe how some of the common models of microsatellite evolution arise as special cases of this more general model.

2.2 Submodels of M

The equal-rate unbiased one-phase model (*EU1*) is a truncated version of one of the simplest models of microsatellite evolution, namely, the *SMM* of Ohta and Kimura. The equal-rate, constant-biased, one-phase model (*EC1*) embodies constant bias towards expansion in the mutation process by constraining $\alpha(u,0,i)$=u for any allele i. Observe that u does not vary with allele length in the *EC1* model, as v, a linear bias parameter, is set at 0. Freeing v allows a linear mutational bias as embodied by the equal-rate, linear-biased, one-phase model (*EL1*), with a mutational bias towards a focal allele f, akin in spirit to the mutation scheme introduced by Garza, Slatkin, and Freimer [14]. Note that *EL1* is related to the simplest version of the *PLBias* model of Calabrese and Durrett [5].

The equal-rate, one-phase models, *EU1*, *EC1*, and *EL1*, have s set to 0, making the mutation rate, $\beta(i, s)=\mu$, equal for all alleles, unlike their proportional-rate, one-phase cousins, *PU1*, *PC1*, and *PL1*, respectively, which allow s to take values in $(-\frac{1}{\Omega-\kappa+1}, \infty)$. The proportional-rate, unbiased, one-phase model (*PU1*) is akin to the proportional slippage without point mutations model (*PS\0M*) of Calabrese *et al* [6]. Our proportional rate models differ from those in the literature because we use an affine function $(1+(i-\kappa)s))$, instead of a linear function $((i-\kappa)s)$, to relate a microsatellite's mutation rate to its length i, in order to embed the equal-rate model within the proportional-rate model. Note that the proportional-rate, constant-biased, one-phase model (*PC1*) and the proportional-rate, linear-biased, one-phase model (*PL1*) address the effects of mutational bias and rate proportionality simultaneously in a nested setting.

In all six models discussed so far, alleles mutate by only one unit of repeat length, since p and m are set at 1. When $p < 1$ and $m < 1$, we have their two-phase cousins in the spirit of DiRienzo *et al*, namely, *EU2**, *EC2**, *EL2**, *PU2**, *PC2**, and *PL2**, respectively, which allow both single-step and multi-step mutations instantaneously. However, in these two-phase models, the parameters p and m are nonidentifiabile at the boundaries ($p = 1$ or $m = 1$). We rectify this by a single-valued transformation $T(p, m)=(p^*, m^*)$ as described in the appendix. Henceforth, p and m in these models will denote the identifiable p^* and m^*, respectively, for notational simplicity.

It is also possible to obtain the six one-phase models from model $\mathbf{M}$ by setting p at 0 to allow mutations of length ≥ 1 and setting m at 1 to force the geometric distribution to put all its mass on one-step mutations. When, $m < 1$, we have their two phase cousins in the spirit of Fu and Chakraborty, namely *EU2*, *EC2*, *EL2*, *PU2*, *PC2*, and *PL2*. Since these models capture the qualitative features of one-phase and two-phase in a simpler and identifiable manner, we will preferentially employ these models for inference.

The equal-rate, unbiased, two-phase model (*EU2**) is a truncated version of the *TPM* of DiRienzo *et al.* Observe that for every state i, as m approaches 0, $\gamma(i, j)$ approaches two uniform distributions on $\{i+1, \cdots, \Omega\}$ and on $\{\kappa, \cdots, i-1\}$, above and below i, respectively. Therefore, our *EU2** model approaches a $p{:}(1-p)$ mixture of the truncated version of Ohta and Kimura's *SMM* and Crow and Kimura's K-allele model (*KAM*) [8], which allows mutations uniformly between finitely many states, on each side of i. The equal-rate, constant-biased, two-phase models (*EC2* and *EC2**), and the equal-rate, linear-biased, two-phase models (*EL2* and *EL2**) add constant and linear bias, respectively to their unbiased cousins (*EU2* and *EU2**).

The proportional-rate, two-phase models, *PU2* or *PU2**, *PC2* or *PC2**, and *PL2* or *PL2**, add rate proportionality to their equal-rate, two-phase cousins, *EU2* or *EU2**, *EC2* or *EC2**, and *EL2* or *EL2**, respectively. The proportional-rate, unbiased, two-phase model (*PU2**) is a hybrid of the truncated versions of Ohta and Kimura's *SMM*, *TPM*, and a variant of *PS\0M*. The most general of this nested family of models, is the proportional-rate, linear-biased, two-phase model (*PL2**), which is exactly our model $\mathbf{M}$. See Table 2.1 for a description of the various homogeneous models described above.

Table 2.1: Model Description

Models	$i \to i+1$	$i \to i-1$
EU1	0.5	0.5
EU2	$0.5\ \gamma(m,i,i+1)$	$0.5\ \gamma(m,i,i-1)$
EU2*	$0.5\ (p+(1-p)\ \gamma(m,i,i+1))$	$0.5\ (p+(1-p)\ \gamma(m,i,i-1))$
EC1	u	$(1-u)$
EC2	$u\ \gamma(m,i,i+1)$	$(1-u)\ \gamma(m,i,i-1)$
PU1	$\beta(i,s)\ 0.5$	$\beta(i,s)\ 0.5$
PU2	$\beta(i,s)\ 0.5\ \gamma(m,i,i+1)$	$\beta(i,s)\ 0.5\ \gamma(m,i,i-1)$
PC1	$\beta(i,s)\ u$	$\beta(i,s)\ (1-u)$
PC2	$\beta(i,s)\ u\ \gamma(m,i,i+1)$	$\beta(i,s)\ (1-u)\ \gamma(m,i,i+1)$
EL1	$\alpha(u,v,i)$	$(1-\alpha(u,v,i))$
EL2	$\alpha(u,v,i)\ \gamma(m,i,i+1)$	$(1-\alpha(u,v,i))\ \gamma(m,i,i+1)$
PL1	$\beta(i,s)\ \alpha(u,v,i)\ \gamma(m,i,i+1)$	$\beta(i,s)\ (1-\alpha(u,v,i))\ \gamma(m,i,i+1)$
PL2	$\beta(i,s)\ \alpha(u,v,i)\ \gamma(m,i,i+1)$	$\beta(i,s)\ (1-\alpha(u,v,i))\ \gamma(m,i,i+1)$

Models	$i \to i+j$ †	$i \to i-k$ ‡
EU1	0	0
EU2	$0.5\ \gamma(m,i,i+j)$	$0.5\ \gamma(m,i,i-k)$
EU2*	$0.5\ (1-p)\ \gamma(m,i,i+j)$	$0.5\ (1-p)\ \gamma(m,i,i-k)$
EC1	0	0
EC2	$u\ \gamma(m,i,i+j)$	$(1-u)\ \gamma(m,i,i-k)$
PU1	0	0
PU2	$\beta(i,s)\ 0.5\ \gamma(m,i,i+j)$	$\beta(i,s)\ 0.5\ \gamma(m,i,i-k)$
PC1	0	0
PC2	$\beta(i,s)\ u\ \gamma(m,i,i+j)$	$\beta(i,s)\ (1-u)\ \gamma(m,i,i-k)$
EL1	0	0
EL2	$\alpha(u,v,i)\ \gamma(m,i,i+j)$	$(1-\alpha(u,v,i))\ \gamma(m,i,i-k)$
PL1	0	0
PL2	$\beta(i,s)\ \alpha(u,v,i)\ \gamma(m,i,i+j)$	$\beta(i,s)\ (1-\alpha(u,v,i))\ \gamma(m,i,i-k)$

† where, $\kappa \leq i \leq \Omega$ and $1 \leq j \leq \Omega - i$

‡ where, $\kappa \leq i \leq \Omega$ and $1 \leq k \leq i - \kappa$

2.3 Stationary Distribution of one-phase models

Observe that all the one-phase models including *PL1* are special cases of the general birth-death chain with birth and death rates b_i and d_i representing the rate of expansion and contraction, respectively, of allele i by one repeat unit. Using the convention $\prod_{j=\kappa}^{\kappa-1}(\cdot) = 1$, the stationary distribution π_i, up to a normalizing factor, is given by,

$$\pi_i \propto \prod_{j=\kappa}^{i-1} \frac{b_j}{d_{j+1}}$$

Thus, for the *PL1* model with birth rate $\alpha(u,v,i) \times \beta(i,s)$ and death rate $(1-\alpha(u,v,i)) \times \beta(i,s)$,

$$\begin{aligned}
\pi_i &\propto \prod_{j=\kappa}^{i-1} \frac{\alpha(u,v,j)\ \beta(j,s)}{(1-\alpha(u,v,j+1))\ \beta(j+1,s)} \qquad (2.4)\\
&\propto \prod_{j=\kappa}^{i-1} \frac{\alpha(u,v,j)}{(1-\alpha(u,v,j+1))} \prod_{j=\kappa}^{i-1} \frac{\beta(j,s)}{\beta(j+1,s)}\\
&\propto \frac{1}{1+(i-\kappa)s} \prod_{j=\kappa}^{i-1} \frac{\alpha(u,v,j)}{(1-\alpha(u,v,j+1))}.
\end{aligned}$$

2.4 Repeat-Specific Models

The presence or absence of any significant difference between the mutational mechanisms of two distinct types of dinucleotide repeats can be investigated. The distribution of D^{I}, the data of type I, is modeled by superimposing a markov chain model $\mathbf{X}^{\mathrm{I}}$ with parameters Θ^{I} on the three branches of τ with terminal branches of equal length λ^{I}. D^{II}, the data of type II, is modeled in a similar manner, by $\mathbf{X}^{\mathrm{II}}$ with its respective parameters and branch length. Thus, akin to equation (2.2), our likelihood function for the data $(D^{\mathrm{I}}, D^{\mathrm{II}})$, where D^{I} is a

$2 \times N^{\text{I}}$ matrix, and D^{II} is a $2 \times N^{\text{II}}$ matrix, is as follows:

$$L(\Theta^{\text{I}}, \Theta^{\text{II}}, \lambda^{\text{I}}, \lambda^{\text{II}} | (D^{\text{I}}, D^{\text{II}})) := \qquad (2.5)$$
$$\textstyle\prod_{i=1}^{N^{\text{I}}} \; L_i(\Theta^{\text{I}}, \lambda^{\text{I}} | D_i^{\text{I}}) \; \prod_{i=1}^{N^{\text{II}}} \; L_i(\Theta^{\text{II}}, \lambda^{\text{II}} | D_i^{\text{II}})$$

2.5 Likelihood Ratio Test ($\boldsymbol{LRT}$)

Suppose we want to test the null hypothesis H_0:$(\boldsymbol{\Theta}, \boldsymbol{\lambda}) = (\boldsymbol{\Theta_0}, \boldsymbol{\lambda_0})$, against the alternative hypothesis, H_1:$(\boldsymbol{\Theta}, \boldsymbol{\lambda}) \in (\boldsymbol{\Theta_1}, \boldsymbol{\lambda_1})$. The likelihood ratio test statistic ($LRTS$) given by,

$$-2 \ \log \frac{L(\boldsymbol{\Theta_0}, \boldsymbol{\lambda_0} | D)}{\sup\limits_{(\boldsymbol{\Theta}, \boldsymbol{\lambda})} \ L(\boldsymbol{\Theta}, \boldsymbol{\lambda} | D)} \qquad (2.6)$$

is asymptotically χ^2_{z} distributed under the null hypothesis, where z is the difference in the number of parameters between the two hypotheses, under standard conditions [7]. We reject the null hypothesis if the observed $LRTS$ is extreme enough to give a P value ≤ 0.01.

2.6 Model Selection

Given an *a priori* set of candidate models, they can be ranked from the best to the worst, in an information-theoretic paradigm through AIC_c, a second-order Akaike Information Criterion. This ranking can help distinguish models that are nearly equally good fits versus those that are poor explanations for the given data D of sample size N. The best candidate model with a total of K parameters in $(\boldsymbol{\Theta}, \boldsymbol{\lambda})$, is the one which minimizes the quantity,

$$AIC_c := -2 \log L(\boldsymbol{\Theta}, \boldsymbol{\lambda} | D) + 2 \left(K + \frac{K(K+1)}{N-K-1} \right). \qquad (2.7)$$

We use AIC_c (see [32] and [18]), the second-order estimator of the Kullback-Liebeler information, instead of the first order estimator AIC or the asymptotically unbiased estimator TIC, because $N/\max\{K\}$ is small in our study (for discussion see [4]). It is worth highlighting that AIC_c is an estimator subject to stochastic noise in the data.

Chapter 3

DATA

In order to find the most number of homologous loci in the pair of primates, while minimizing ascertainment bias and sequencing error, we first obtained 21.4 million base pairs of the *Pan troglodytes* (chimp) sequences in HTGS (high throughput genomic sequence) [25] phase 3, available by March 4, 2003, through the *Entrez* retrieval system of NCBI (http://www.ncbi.nlm.nih.gov/Entrez/). The sequences in HTGS phases 0, 1, and 2, were excluded to minimize sequencing error and circumvent the problem of aligning the unordered pieces. For all analyses in this study we set the lower bound κ=10. Chimp microsatellites of dinucleotide motifs with repeat length $\geq$ 10 were mined. To assure some level of independence, all microsatellites within 200 base pairs of another were discarded.

Each selected chimp microsatellite, with 200 base pairs of flanking sequence upstream and downstream, was used to perform an extremely stringent (E-value $\leq 1 \times 10^{-100}$) unfiltered BLAST search against the human contigs downloaded from the August 23, 2002 NCBI release at ftp://ncbi.nlm.nih.gov/genomes /H_sapiens/, using formatdb and blastall (2.2.3 release) of the NCBI Toolkit in ftp://ftp.ncbi.nih.gov/toolbox/ncbi_tools/. Thus we obtained 644 candidate microsatellite loci homologous between the two primates.

Each such microsatellite locus was retained if it had a flanking sequence of length $\geq$ 200 base pairs on at least one side of the di-nucleotide repeat in both species, and a flanking sequence of length $\geq$ 50 base pairs on the other side in both species. A compound repeat is defined to have more than one motif, each of repeat length $\geq$ 10, within a 50-basepair radius. 30% of loci contained compound repeats in at least one of the homologs and were excluded from further analysis. Finally, those loci whose simple repeats in at least one species, were interrupted by two or more point mutations were omitted. Thus 383 candidate loci were obtained. About 70% of these loci occurred in human chromosome 7 (see Figure 3.1). 15% of these 383 loci were ommitted as their human homologs were $\leq$ 9 units in repeat length. Among the remaining 321 loci 78% were AC-repeats (namely, AC,CA,TG, and GT repeats), 13% were AT-repeats (namely, AT and TA repeats), 9% were AG-repeats (namely, AG, GA, TC, and CT repeats), and 0% were CG-repeats (namely, CG and GC repeats).

Among these 321 loci, 18% contained homologous pairs of once-interrupted dinucleotide repeats, which have exactly one point mutation interrupting an otherwise pure stretch of the repeat in either or both species. We count the repeat length of a once-interrupted AC-repeat (iAC-repeat) ignoring the interruption. For instance, the iAC-repeat ‘ACACATACAC’ is taken to be of length 5. The common practice in the literature of directly extrapolating the repeat length of a microsatellite from its PCR fragment length is the motivation behind such a characterization of repeat length for an interrupted microsatellite.

Thus we found 321 homologous pairs of simple dinucleotide repeats with at most one interruption, of which 264 were uninterrupted or pure dinucleotide repeats and 235 were pure AC-repeats. This constitutes our basic human-chimp data set. The empirical joint and marginal distributions of homologous pure AC-

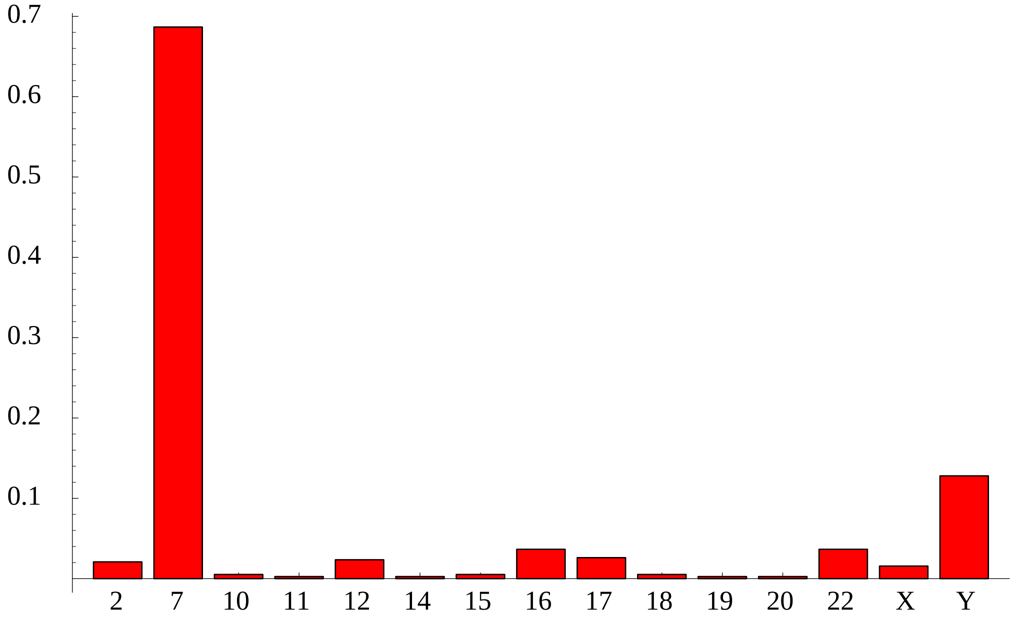

Figure 3.1: Distribution of candidate dinucleotide microsatellites across human chromosomes.

repeat data is shown in Figures 3.2 and 3.3 respectively. We also obtained the human genomic data of perfect (devoid of interruptions) and isolated (at least 50 basepairs from the nearest dinucleotide microsatellite of length more than 4 repeat units) AC-repeats as described in Calabrese *et al* [5] for comparative purposes.

To maximize the likelihood L, we transform the constrained parameter space to an unconstrained one, and perform an unconstrained optimization using the function *Findminimum* of *Mathematica* [36]. We explore most of the support of the parameter space by partitioning it into small hypercubes which are used as initial conditions to find local optima. Finally, the Broydon-Fletcher-Goldfarb-Shanno algorithm [28] is started near the best local optimum to obtain the global optimum.

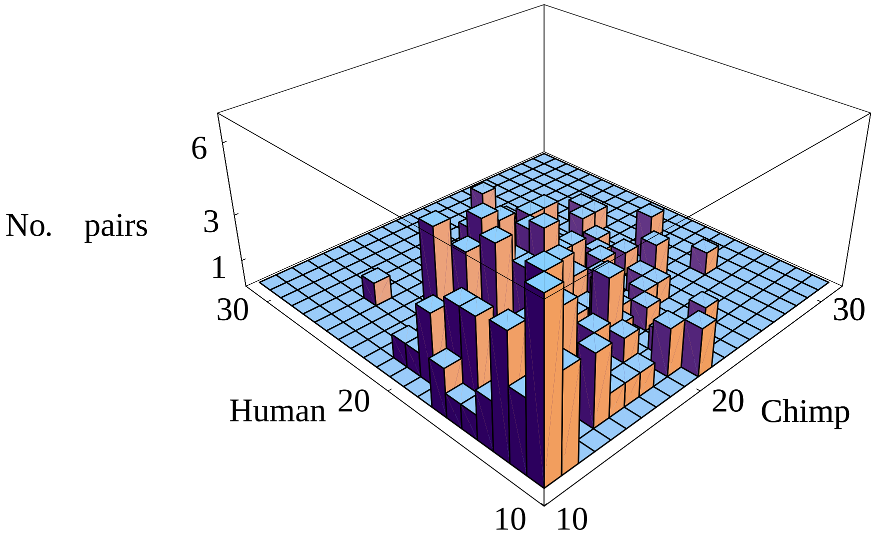

Figure 3.2: Raw counts of homologous microsatellites of pure AC-repeats between human and chimp (D_{AC}).

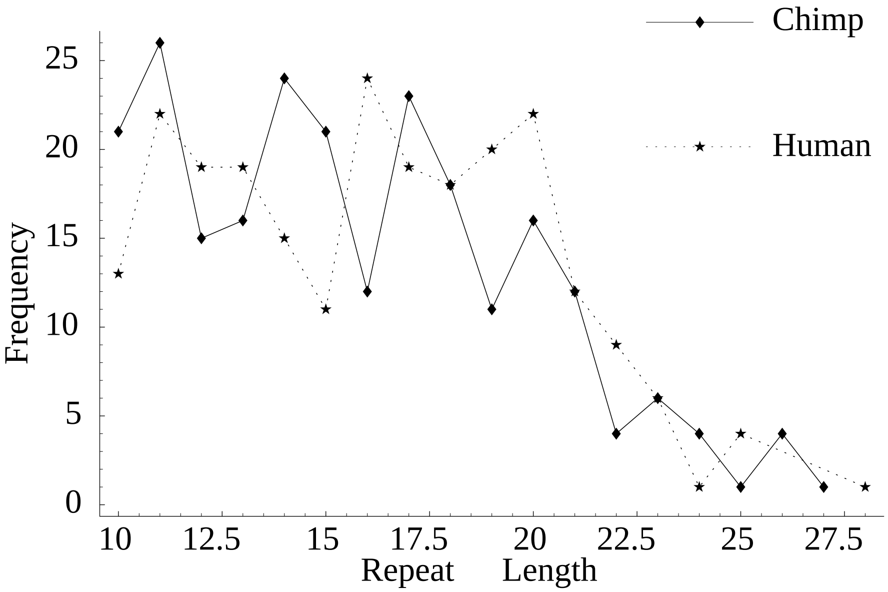

Figure 3.3: Marginal frequencies of the chimp and human microsatellites of pure AC-repeats (D_{AC}).

Chapter 4

RESULTS

We initially assume a lineage-homogeneous mutational process to model the distribution of the 235 homologous pairs of pure AC-repeats. Thus the same Markov chain model (*i.e.*, $\Theta_a{=}\Theta_c{=}\Theta_h{=}\Theta$) is superimposed on the three branches of $\boldsymbol{\tau}$ whose terminal branches are of equal length (*i.e.*, $\lambda_c{=}\lambda_h{=}\lambda$). Observe that for time reversible Markov chains, such as *PL1*, we can only estimate the sum of the terminal branch lengths (2λ) along with Θ. This is because the per-locus likelihood given by Equation 2.1 becomes $\sum_{j\in\mathbb{S}} \pi_j \; P_{j,C_i}(\lambda) \; P_{j,H_i}(\lambda)$ due to lineage homogeneity and further simplifies to $\pi_{C_i} \; P_{C_i,H_i}(2\lambda)$ due to time reversibility. We relax, and even test, these homogeneity assumptions later when we study repeat-specific and lineage-specific processes.

4.1 Ranking Submodels of M

The submodels of **M** define the set of candidate models to be ranked from best to worst according to their AIC_c values using equation 2.7, based on data D_{AC} (see Table 4.1). Five groupings of models are found. The best group contains the proportional-rate linear-bias models, *PL1* and *PL2*, where longer microsatellites mutate more often than shorter ones towards an attracting focal allele. The second best group comprises of *EL1* and *EL2*. In these models, all microsatellites, irrespective of their repeat length, mutate at the same rate towards a focal allele. The third best group comprises of the constant-bias models, namely, *PC1*, *PC2*, *EC1*, and *EC2*. In the presence of a constant downward bias in the mutational process none of the other features seem to matter very much. The proportional-rate, unbiased models, *PU1* and *PU2*, constituting the fourth best group, outperform their equal-rate, unbiased cousins, *EU1* and *EU2*. Observe that the model ranking is unaffected by variation in the upper bound Ω except for the worst group.

Another ranking of the submodels of $\mathbf{M}$ is performed as shown in Table 4.2 based on the fit of their stationary distributions to the empirical distribution of pure and isolated AC-repeat lengths in the human genome as described by Calabrese *et al.* [5]. These results are consistent with those based on the human and chimp comparison. However, when fitting a model's stationary distribution, due to the large sample size, any increase in the degrees of freedom toward a multinomial model greatly increases its likelihood.

4.2 One Phase Vs. Two Phase

The null hypothesis of the simplest, one-phase model *EU1* is tested against its two-phase cousin *EU2**, through a *LRT*. The *LRTS* under this null hypothesis has a nontrivial mixture of χ^2_0, χ^2_1, and χ^2_2 for its asymptotic distribution, since both p and m lie on the boundary of the parameter space under the null hypothesis [31]. Instead of analytically pursuing this asymptotic distribution under such nonstandard boundary conditions, we resort to parametric bootstrap to obtain an approximation to the finite sample distribution of the *LRTS*

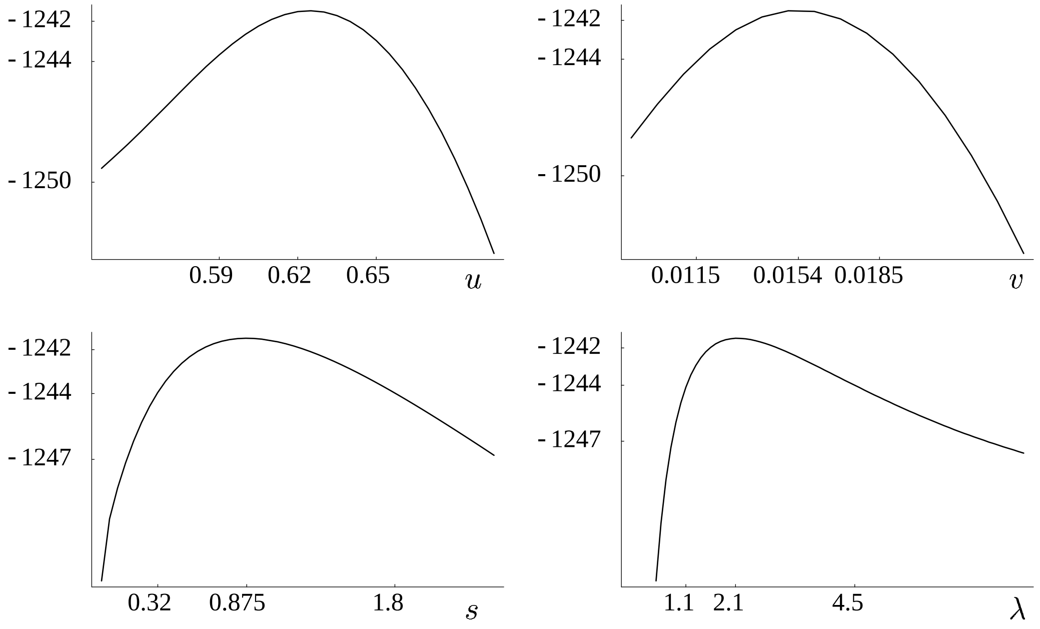

Figure 4.1: Profile Log Likelihood of the parameters, u, v, s and λ of the best model *PL1*

(see part i. of Figure 4.3). Based on the simulations the one-phase hypothesis prevails ($P = 0.16$). One is unable to reject *EU1* in favor of the simpler equal-rate two-phase unbiased model *EU2* as well, since the *LRTS* which is asymptotically $0.5{+}0.5\chi_1^2$ distributed is observed to be 0.084 ($P = 0.39$). Similarly, we are unable to reject the null hypothesis of every other one-phase model, in favor of its two-phase cousin, except in the equal-rate linear-biased case where one-phase is marginally rejected ($P = 0.013$). The *EC2* model with p=0, akin to a truncated version of Fu and Chakraborty's *SMM* [13], as well as, *PC2* and *PU2* assign almost all of the probability mass to single-step jumps. Hence, in these cases, we fail to reject the one-phase hypothesis that m=1 in favor of a two-phase hypothesis that $0 < m < 1$. Among the best group of models, *PL1* and *PL2*, the hypothesis of one-phase prevails as $P = 0.06$ (see part ii. of

Table 4.1: Parameter estimation, maximum likelihood, and model ranking using species-pair data from 235 loci of AC-repeats.

Models	K [a]	**Fixed** parameters [b] and MLEs of free parameters [c]					$\log L$	AIC$_c$		
		u	v	m	s	λ	$\Omega = 40$	$\Omega = 40$	$\Omega = 60$	$\Omega = 100$
PL2	5	0.8158	0.03947	0.5475	0.7638	0.5646	-1240.2079	0.00	0.00	0.00
PL1	4	0.6246	0.01542	**1.0000**	0.8752	2.1441	-1241.4802	0.46	0.41	0.44
EL2	4	0.6774	0.03701	0.4317	**0.0000**	1.7153	-1247.9369	13.37	13.36	13.40
EL1	3	0.5416	0.009464	**1.0000**	**0.0000**	12.2643	-1250.3987	16.22	16.17	16.21
EC1	2	0.4650	**0.0000**	**1.0000**	**0.0000**	11.0898	-1294.5354	102.44	107.54	107.83
PC1	3	0.4654	**0.0000**	**1.0000**	-0.0048	10.9308	-1294.5243	104.47	109.56	109.85
EC2	3	0.4650	**0.0000**	0.9999	**0.0000**	11.0898	-1294.5354	104.50	109.59	109.88
PC2	4	0.4654	**0.0000**	0.9999	-0.0048	10.9308	-1294.5243	106.54	111.63	111.91
PU1	2	**0.5000**	**0.0000**	**1.0000**	0.2802	3.6773	-1342.4756	198.325	275.91	347.17
PU2	3	**0.5000**	**0.0000**	0.9999	0.2802	3.6773	-1342.4756	200.38	277.96	349.22
EU1	1	**0.5000**	**0.0000**	**1.0000**	**0.0000**	10.3296	-1432.3527	376.04	609.98	882.16
EU2	2	**0.5000**	**0.0000**	0.9398	**0.0000**	8.6285	-1432.3107	377.99	609.33	877.11
		$[0,1]^d$	$(-\infty,+\infty)$	$[0,1]$	$(-\frac{1}{31},\infty)$	$(0,\infty)$		+2490.68	+2490.73	+2490.70

[a] K denotes the number of free parameters.

[b] The parameters of model **M** that are fixed for a given submodel are shown in bold.

[c] Free parameters take their maximum likelihood estimates (MLEs) when $\Omega = 40$

[d] The range of each parameter under model **M** is given in the last row.

Table 4.2: Parameter estimation, maximum likelihood, and model ranking using human genomic data with 33298 loci of AC-repeats.

Models	K [a]	**Fixed** values [b] and MLEs of parameters [c]				$\log L$	AIC		
		u	v	m	s	$\Omega = 40$	$\Omega = 40$	$\Omega = 60$	$\Omega = 100$
PL2	4	0.9310	0.0455	0.4224	1.0325	-95087.0592	0.00	0.00	0.00
PL1	3	0.6148	0.0121	**1.0000**	4.9252	-95095.9213	15.72	285.79	288.15
EL2	3	0.6726	0.0295	0.4310	**0.0000**	-95251.0395	325.96	321.510	321.24
EL1	2	0.5437	0.0078	**1.0000**	**0.0000**	-95371.0969	564.07	681.77	684.13
EC1	1	0.4702	**0.0000**	**1.0000**	**0.0000**	-100116.6430	10053.17	11162.38	11247.39
PC1	2	0.4702	**0.0000**	**1.0000**	0.0000	-100116.6430	10055.17	11164.38	11249.39
EC2	2	0.4702	**0.0000**	0.9999	**0.0000**	-100116.6430	10055.18	11164.40	11249.42
PC2	3	0.4702	**0.0000**	0.9999	0.0000	-100116.6430	10057.17	11166.38	11251.39
PU1	1	**0.5000**	**1.0000**	**1.0000**	0.3166	-104475.0674	18770.02	29951.28	40172.42
PU2	2	**0.5000**	**0.0000**	0.9999	0.3166	-104475.0672	18772.02	29953.28	40174.42
EU2	1	**0.5000**	**0.0000**	0.9315	**0.0000**	-114340.2766	38500.44	68484.04	90335.52
EU1	0	**0.5000**	**0.0000**	**1.0000**	**0.0000**	-114344.9059	38507.70	71446.12	110019.08
		$[0,1]$ [d]	$(-\infty, +\infty)$	$[0,1]$	$(-\frac{1}{31}, \infty)$		+190182.12	+190460.65	+190458.29

[a] K denotes the number of free parameters.

[b] The parameters of model **M** that are fixed for a given submodel are shown in bold.

[c] Free parameters take their maximum likelihood estimates (MLEs) with $\Omega = 40$

[d] The range of each parameter under model **M** is given in the last row.

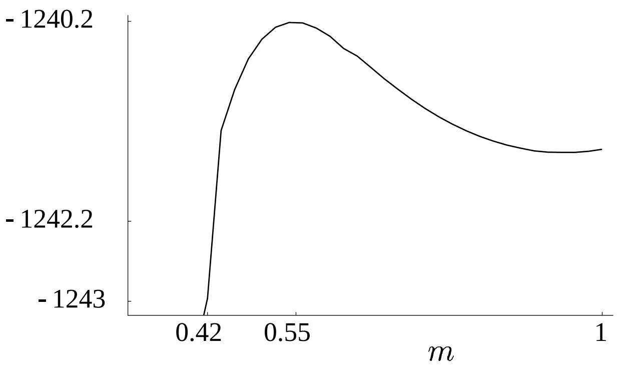

Figure 4.2: Profile Log Likelihood of the two-phase parameter m of the model *PL2*

Figure 4.3). Furthermore, there is even less evidence in the data to reject *PL1* in favor of *PL2** ($P = 0.23$). The fairly flat profile log likelihood of m under the *PL2** model is shown in Figure 4.2. The confidence interval containing 2 loglikelihood units from the median is [0.42,1].

4.3 Mutational Bias

The absence of any mutational bias as embodied by *EU1*, is first rejected in favour of the constant bias model *EC1*. The maximum likelihood estimate (*MLE*) of the constant upward bias parameter $\hat{u}$=0.4650. *EU1* is also rejected in favor of the linear bias model *EL1*.

The hypothesis of constant mutational bias for all alleles, *i.e.* *EC1*, is rejected in favor of the linear bias model *EL1* in the absence of rate proportionality. This *LRTS* is asymptotically distributed as χ_1^2 under the null hypothesis (see part iii. of Figure 4.3). The *MLE* of the attracting focal allele for the linear bias model *EL1* was $\lfloor((\hat{u} - 0.5)/\hat{v}) + \kappa\rfloor$ =$\lfloor((0.5416 - 0.5)/0.009464) + 10\rfloor$=14.

In order to investigate the nature of mutational bias in the presence of rate proportionality we conducted similar *LRT*s. Once again absence of bias (*PU1*)

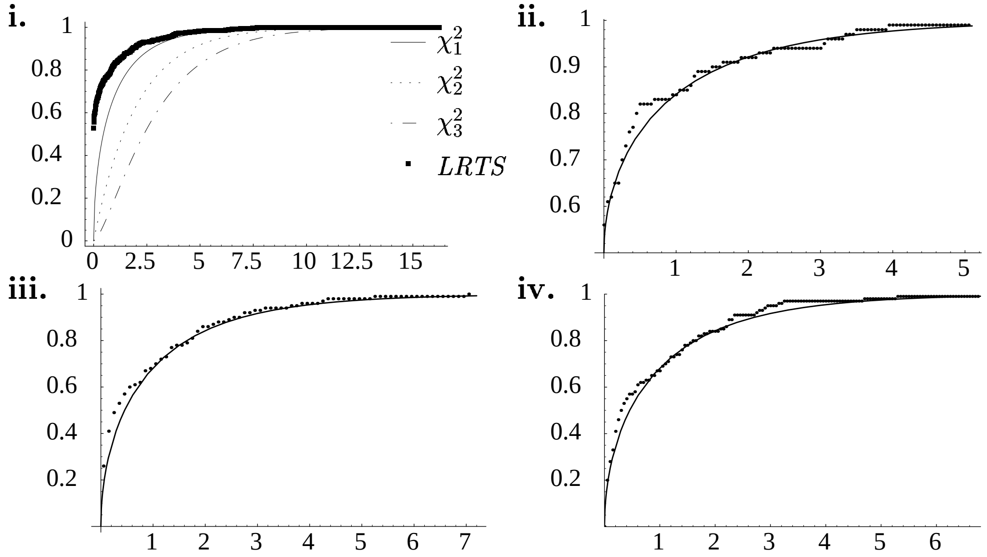

Figure 4.3: **i.** 500 simulations of the finite sample *LRTS* under the null hypothesis for *EU1* Vs. *EU2**, and 100 simulations each for, **ii.** *PL1* Vs. *PL2* $\sim 0.5{+}0.5\chi_1^2$, **iii.** *EC1* Vs. *EL1* $\sim \chi_1^2$ and **iv.** *EL1* Vs. *PL1* $\sim \chi_1^2$. The asymptotically expected distribution in each case is the solid line.

was rejected in favor of its presence (*PC1* and *PL1*) and the hypothesis of constant bias (*PC1*) was rejected in favor of linear bias (*PL1*). The *MLE* of the attracting focal allele for the proportional-rate linear-bias model *PL1* was $\lfloor((.6246-0.5)/0.01542)+10\rfloor=18$. When more general functional forms, such as, piece-wise linear, quadratic, or cubic, were employed to model the dependence of mutational bias on repeat length, the likelihood did not improve significantly enough to reject the linear bias model (results not shown).

4.4 Rate Equality Vs. Proportionality

We test the hypothesis of equal mutation rates for all alleles (*EU1*) against a hypothesis of proportional rates (*PU1*). This *LRTS* is asymptotically χ_1^2 distributed under the null hypothesis. Thus, the null hypothesis of rate constancy among alleles is rejected, in favour of a directly proportional relationship between mutation rate and repeat length ($\hat{s}$=0.2556) in the presence of an unbiased mutation process.

In order to determine the relevance of rate proportionality in the presence of mutational bias two more *LRT*s are performed. In the presence of a constant bias, we failed to reject the null hypothesis of rate equality among alleles in favor of rate proportionality ($P = 0.022$). In the presence of linear bias, the *LRTS* is asymptotically distributed as χ^2_1 under the null hypothesis (see part iv. of Figure 4.3). We were able to reject rate equality (*EL1*) in favor of rate proportionality (*PL1*). Thus, for pure AC-repeats, the proportional-rate linear-bias model (*PL1*) explains the data best.

When performing multiple *LRT*s in a nested setting, the order in which such tests are done could affect the final conclusions drawn. We are assured, however, that this order has not influenced our conclusions, since the results of model selection are consistent with those of the hypothesis tests. All conclusions drawn above using the *LRT*s are robust to changes in the upper bound Ω (results not shown).

So far we have only used D_{AC} for inference and assumed homogeneity in the mutational mechanisms of all these loci. In doing so, we have ignored inter-locus variation, and could not address possible motif-specific and interruption-induced complications. Such issues are examined below using *PL1*, which emerged earlier as the best model.

4.5 Inter-Locus Rate Variation

The possible presence of variation in mutation rate among loci of pure AC-repeats is investigated next. Since λ is estimable as the product of μ and

Table 4.3: Some hypothesis tests of time homogeneous models through likelihood ratios.

LRT	$\mathbf{H}_0$ Vs. $\mathbf{H}_1$	*Asym. Dist.* [a]	*LRTS* [b]	*P* [c]
1	*EU1* Vs. *EU2*	$0.5\chi_0^2 + 0.5\chi_1^2$	0.084	0.39
2	*EU1* Vs. *EU2**	simul. [d]	1.060	0.16
3	*EL1* Vs. *EL2*	$0.5\chi_0^2 + 0.5\chi_1^2$	4.92	0.013
4	*PL1* Vs. *PL2*	$0.5\chi_0^2 + 0.5\chi_1^2$	2.54	0.055
5	*EU1* Vs. *EC1*	χ_1^2	275.62	$\ll 0.01$
6	*EU1* Vs. *EL1*	χ_2^2	363.91	$\ll 0.01$
7	*EC1* Vs. *EL1*	χ_1^2	88.27	$\ll 0.01$
8	*EU1* Vs. *PU1*	χ_1^2	179.75	$\ll 0.01$
9	*EC1* Vs. *PC1*	χ_1^2	0.022	0.88
10	*EL1* Vs. *PL1*	χ_1^2	17.84	$\ll 0.01$

[a] The expected asymptotic behavior of likelihood ratio test statistic (*LRTS*) under the null hypothesis $\mathbf{H}_0$.

[b] Ω is set at 40.

[c] P values < 0.01 are considered significant

[d] Simulated finite sample distribution (part **i.** of Figure 4.3)

t, variation in mutation rate (μ) translates to variation in λ, as the number of generations (t) remains identical for all loci. We model three equi-proportionate classes of loci, 1, 2, and 3, with distinct mutation rates reflected by, λ_1, λ_2, and λ_3 respectively. We are unable to reject the null hypothesis of equal rates across loci, H_0: $\lambda = \lambda_1 = \lambda_2 = \lambda_3$, in favor of inter-locus rate variation, H_1: $0 < \lambda_1 \leq \lambda_2$, $0 < \lambda_2 < \infty$, and $\lambda_2 \leq \lambda_3 < \infty$, as the asymptotically χ^2_2 distributed *LRTS* is observed to be 0.67 ($P = 0.73$). We also rejected a simpler variable rates model with only 2 classes ($P = 0.46$).

4.6 Motif-Specific Variation

As there are no pure GC-repeats, and only 29 pure AT-repeats or AG-repeats in our data, only differences in the mutational mechanism between pure AC-repeats and pure AT-repeats or AG-repeats (AT\G-repeats) are investigated. The evolution of pure AC-repeats is modeled by a proportional-rate linear-biased one-phase model with parameters u^{AC}, v^{AC}, s and λ, and that of pure AT\G-repeats is modeled similarly with parameters $u^{\mathrm{AT\backslash G}}$, $v^{\mathrm{AT\backslash G}}$, s, and λ. By calculating the likelihood according to equation 2.5, we test the null hypothesis of identical parameters, H_0: $u = u^{\mathrm{AC}} = u^{\mathrm{AT\backslash G}}$ and $v = v^{\mathrm{AC}} = v^{\mathrm{AT\backslash G}}$, against the alternative of possibly distinct bias parameters, H_1: $u^{\mathrm{AC}} \neq u^{\mathrm{AT\backslash G}}$ and $v^{\mathrm{AC}} \neq v^{\mathrm{AT\backslash G}}$. The *LRTS*, which is asymptotically χ^2_2 distributed, is observed to be 13.72 ($P \leq 0.01$). We thus reject the null hypothesis of identical mutational mechanisms for AC-repeats and AT\G-repeats. Figure 4.4 plots the probability of an expansion upon mutation ($\alpha(\hat{u}, \hat{v}, i)$) as a function of repeat length i based on the *MLE*s of the bias parameters for AC and AT\G motif types. The *MLE* of the focal allele for AC-repeats is 18 while that for the AT\G-repeats is 20.

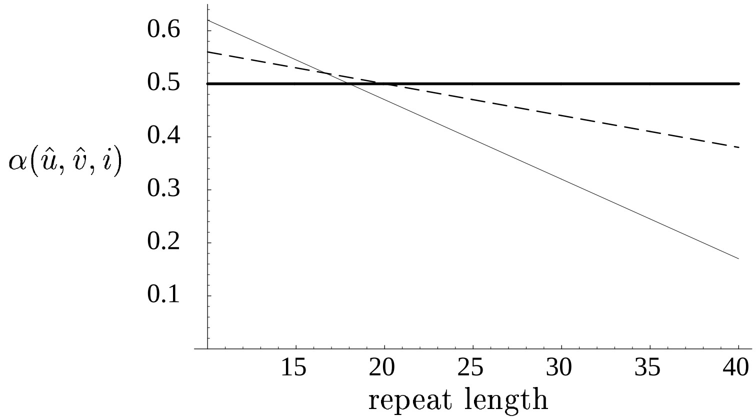

Figure 4.4: Probability of expansion upon mutation, $\alpha(\hat{u}, \hat{v}, i)$, as a function of repeat length based on the *MLE*s for pure AC-repeats (thin line) pure AT\G-repeats (dashed line)

However, we were unable to reject the null hypothesis H_0 of identical mutational mechanisms in favor of an alternative which allowed distinct proportional-rate parameters (s^{AC}, $s^{\mathrm{AT\setminus G}}$) but identical bias parameters ($P = 0.74$). Furthermore, the distinctness of bias parameters alone seems to matter as we are unable to reject this hypothesis (H_1) in favor of a more general hypothesis which allowed distinct proportional-rate parameters in addition to distinct bias parameters ($P = 0.62$).

4.7 Interruption-Induced Variation

We study possible effects of an interruption by a point mutation on the evolution of otherwise pure AC-repeats. Recall that the repeat length of a once-interrupted AC-repeat (iAC) is counted ignoring the interruption. As in the previous section, the stochastic dynamics of pure AC-repeats is described by a proportional-rate linear-biased one-phase model with parameters u^{AC}, v^{AC}, s^{AC} and λ, and that of the iAC-repeats is described by another such model with parameters u^{iAC}, v^{iAC} and s^{iAC} and λ. By calculating the likelihood according to

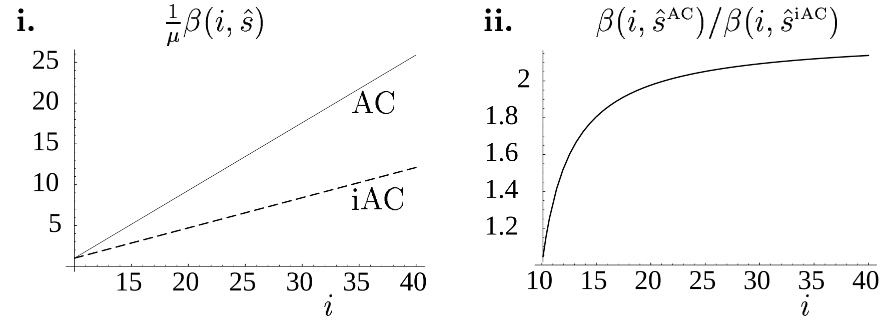

Figure 4.5: **i.** Scaled mutation rates; $\frac{1}{\mu}\beta(i,\hat{s}^{\mathrm{AC}})$ for pure AC-repeats (solid) and $\frac{1}{\mu}\beta(i,\hat{s}^{\mathrm{iAC}})$ for once-interrupted iAC-repeats (dashed), as a function of repeat length. **ii.** The rate ratio $\beta(i,\hat{s}^{\mathrm{AC}})/\beta(i,\hat{s}^{\mathrm{iAC}})$ as a function of repeat length i.

equation 2.5, we test hypotheses through *LRT*s.

The null hypothesis of an identical mutational mechanism between pure AC-repeats and iAC-repeats, H_0: $u = u^{\mathrm{AC}} = u^{\mathrm{iAC}}$, $v = v^{\mathrm{AC}} = v^{\mathrm{iAC}}$ and $s = s^{\mathrm{AC}} = s^{\mathrm{iAC}}$, is successfully rejected in favor of the alternative which allows distinct mutational mechanisms; H_1: $u^{\mathrm{AC}} \neq u^{\mathrm{iAC}}$, $v^{\mathrm{AC}} \neq v^{\mathrm{iAC}}$ and $s^{\mathrm{AC}} \neq s^{\mathrm{iAC}}$, since the asymptotically χ^2_3 distributed *LRTS* is observed to be 26.27. The *MLE* of the focal allele for AC-repeats is still 18 but that of the iAC-repeats is longer at 21.

The scaled mutation rate $\frac{1}{\mu}\beta(i,s)$ is plotted as a function of repeat length using the *MLE*s of the proportional-rate parameters for pure AC-repeats ($\hat{s}^{\mathrm{AC}} = 0.83$) and iAC-repeats ($\hat{s}^{\mathrm{iAC}} = 0.37$) in part **i.** of Figure 4.5. The ratio of the *MLE* of mutation rate of AC-repeats over that of iAC-repeats which asymptotes to $0.83/0.37 = 2.24$ is plotted in part **ii.** of Figure 4.5. Unlike the case of AC vs. AT\G-repeats, the null hypothesis H_0 is strongly rejected in favor of a simpler alternative which assumes identical bias parameters u and v but distinct proportional-rate parameters s^{AC} and s^{iAC}. For this test the *LRTS* which is asymptotically distributed as χ^2_1 is observed to be 14.56.

4.8 Lineage-Specific Variation

Here, we relax the assumption of lineage homogeneity that Θ_a=Θ_c=Θ_h=Θ, and allow distinct Markov chain models to be superimposed on distinct branches of τ. We study lineage-specific differences in the mutational mechanism only for the *PL1* model. By superimposing a proportional-rate linear-biased one-phase model with parameters u_a, v_a and s_a upon the ancestral branch, another such model with parameters u_c, v_c and s_c upon the chimp branch, and finally another with parameters u_h, v_h and s_h upon the human branch we address lineage-specific differences in the mutational mechanism of pure AC-repeats. Naturally, the lineage-homogeneous models studied thus far, in which all three branches have superimposed upon them three Markov chain models with identical parameters ($u = u_a = u_c = u_h$, $v = v_a = v_c = v_h$ and $s = s_a = s_c = s_h$), embody the essence of identical mutational mechanisms along the three lineages and constitute our null hypothesis of lineage homogeneity in the mutational process. However, there are numerous ways to model lineage-specific differences in the mutational process. In fact, the scenario of biased microsatellite expansion along the human lineage is indiscernable from that of a biased contraction along the chimp lineage without repeat length data at homologous loci in an additional outgroup species. Owing to the nature and sample size of our species-pair data and in light of a human-chimpanzee-baboon study by Webster *et al* [34], we introduce non-homogeneity by constraining the ancestral mutational mechanism to be identical to that of chimp. Moreover, the non-homogeneous models which impose an identical mutational mechanisms between the human and the ancestral species do not have better AIC_c scores (results not shown).

We marginally reject ($P = 0.018$) the null hypothesis of identical mutational mechanism for the ancestor, chimp, and human microsatellites of the pure AC-repeats (*PL1* model) in favor of an alternative hypothesis of an almost identical mechanism for the three lineages with the exception of a distinct proportional-rate parameter s_h for the human lineage ($PL1^1$). Since the various alternatives are not nested we resort to AIC_c to rank the models. The better performing non-homogeneous models decrease the mutation rate (by decreasing s_h) for longer human microsatellites relative to that of the chimps and/or increase the focal allele of humans by one or two repeat units as evident from Table 4.4. Similar two-phase non-homogeneous models did not perform better than $PL1^1$ (results not shown).

We were also able to fit non-homogeneous models much better to the empirical distribution of isolated pure AC-repeats from human genomic data. A nonhomogeneous *PL1* model with 7 parameters had a log likelihood value of -95050.02 and outperformed the time homogeneous *PL2* model from Table 4.2 by 96 *AIC* units. The *MLE*s suggest a scenario of ongoing repeat expansion in humans. Figure 4.6 shows the fits of the homogeneous and non-homogeneous *PL1* model to the empirical distribution of the AC-repeats found in the human genome.

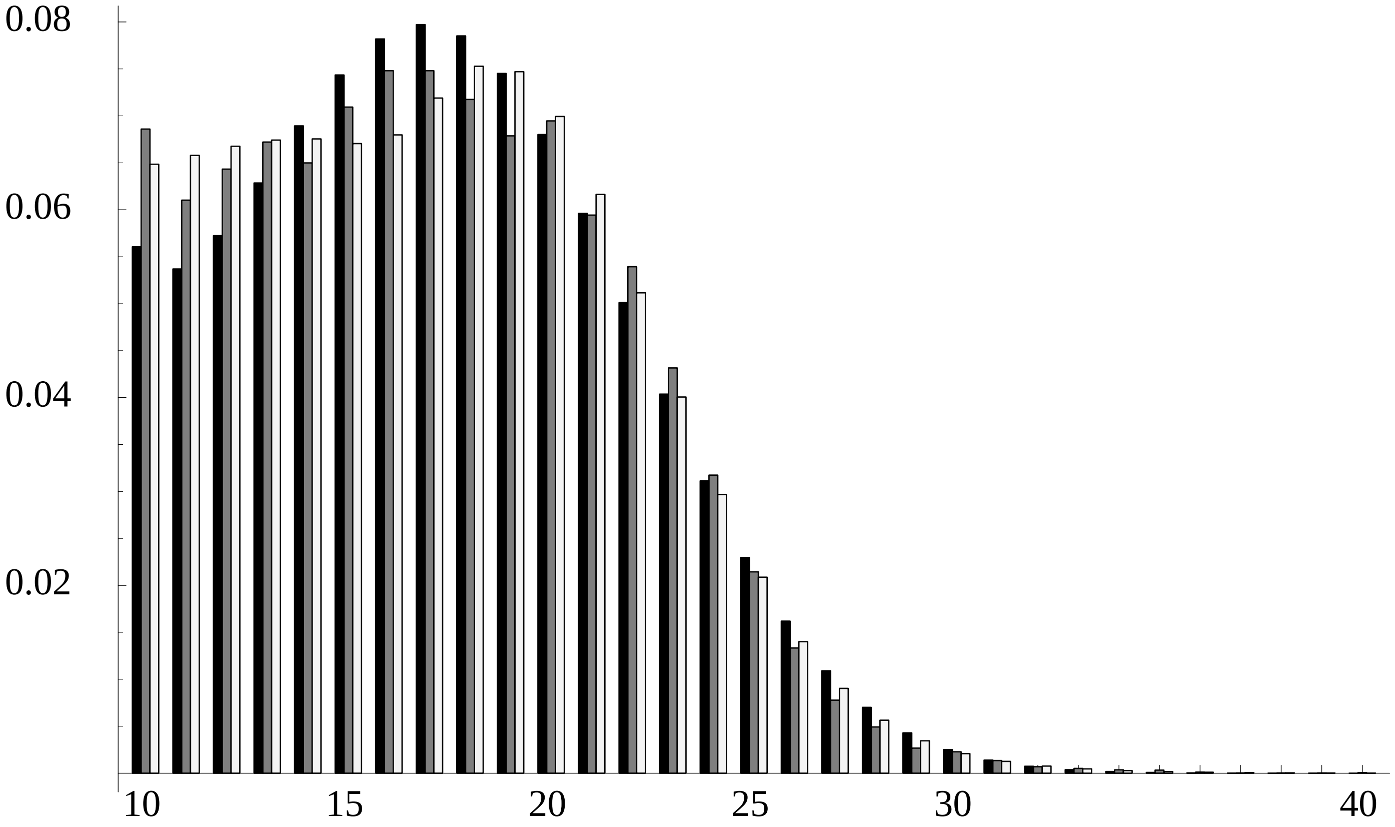

Figure 4.6: Stationary distribution of homogeneous (black bars) *PL1* model, non-stationary distribution of non-homogeneous (white bars) *PL1* model and empirical distribution of the isolated AC-repeats (grey bars) in the human genome.

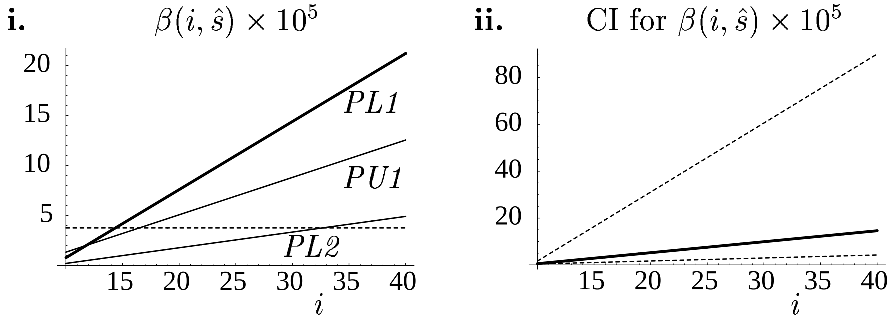

Figure 4.7: **i.** *ML* estimates of the allele-specific mutation rate $\beta(i,\hat{s}) \times 10^5$ for models *PL1*, *PL2*, *PU1* and *EU1* (dotted line). **ii.** Confidence interval (CI) for the allele-specific mutation rate $\beta(i,\hat{s}) \times 10^5$ for *PL1* model.

4.9 Mutation Rate Estimation

Assuming 5.5×10^6 years since human-chimp speciation and an average lifetime of 20 years for the two species leads to an estimate of 2.75×10^5 generations since speciation. Since $\mu = \lambda/t$ in our formulation, its *MLE* is $\hat{\mu} = \hat{\lambda}/(2.75 \times 10^5)$. Thus the *MLE* of the allele-specific mutation rate $\beta(i,\hat{s})=\hat{\mu}(1+(i-10)\hat{s})$ is obtained.

In order to compare it with the estimates of mutation rates in the literature we obtain an average rate $\beta^* = \sum_i \hat{\pi}_i \beta(i,\hat{s})$, where $\hat{\pi}_i$ is the stationary probability of allele i under the *MLE*s of the model. For the best model (*PL1*) β^* is 4.87×10^{-5} per locus per generation and for the worst model (*EU1*) it is 23% less at 3.76×10^{-5}. The mutation rate estimates are fairly similar for different models as shown in Figure 4.7.

Moreover, by walking 2 log likelihood units on either side of the median along the profile log likelihoods we obtain a confidence interval of $[1.1, 4.5]$ for λ and $[0.32, 1.8]$ for s. This translates to a confidence interval for the average per-locus per-generation mutation rate of $[1.3 \times 10^{-5}, 1.8 \times 10^{-4}]$ for pure AC-repeats under the *PL1* model.

Table 4.4: Lineage-specific model ranking.

Models	K [a]	Lineage-Specific Parameters								$\log L$ [b]	AIC$_c$
		u	u_h	v	v_h	s	s_h	m	λ		- 2487.70
PL1^1	5	**0.63**	**0.63**	**0.016**	**0.016**	**1.40**	0.0184	**1**	2.62	-1238.72	0.0
PL1^2	7	0.63	0.68	0.016	0.020	1.26	0.42	**1**	2.23	-1237.83	2.5
PL1^3	5	0.63	0.65	**0.016**	**0.016**	**0.88**	**0.88**	**1**	2.10	-1240.12	2.8
PL2	5	**0.82**	**0.82**	**0.04**	**0.04**	**0.76**	**0.76**	0.55	0.56	-1240.21	3.0
PL1	4	**0.62**	**0.62**	**0.015**	**0.015**	**0.88**	**0.88**	**1**	2.14	-1241.48	3.4

[a] K denotes the number of parameters in a model.

[b] The MLEs and $\log L$ when Ω was set at 40.

Chapter 5

DISCUSSION

Species-pair data from humans and chimps provides opportunities for analyzing microsatellite evolution not found in population genetic data or genomic data from a single species. A population's demography determines the distribution of its genealogy, which in turn accounts for the correlation among homologous alleles in a population sample . Thus strong demographic assumptions have to be made [23] to reject one model of microsatellite evolution in favor of another . Since our inferences are based on a sample of size one from each population, they do not rely on assumptions regarding the demographic history of the analyzed populations. Different models can give rise to similar equilibrium distributions despite distinct finite time transition probabilities. Thus any inference based on genomic data from a species is limited to parametric families of models whose members have distinct equilibrium distributions [22]. However, this approach currently has the advantage of larger data sets over our species-pair approach, as the chimp genome is not yet fully sequenced. We provide a framework for hypothesis tests directed at a mechanistic understanding of the mutational process of microsatellites using information about their divergence.

Our analysis indicates that bias in the mutational process and proportionality in mutation rate are vital for realistic stochastic models of evolution of pure dinucleotide repeats. All models imbued with an unbiased mutational process perform poorly compared to their biased counterparts, both in terms of having high AIC_c values, and thereby being ranked the worst, as well as, being rejected with extreme significance in favor of their biased cousins through LRTs.

The best group of models with the lowest AIC_c values are the proportional-rate linear-biased one and two-phase models, *PL1* and *PL2* respectively. To decide between *PL1* and *PL2* we resort to a LRT. The hypothesis of one-phase embodied by *PL1* remains unrejected.

The models with a linear bias toward a focal allele constitute the top four models. This suggests a primal role for linear bias in microsatellite evolution and further affirms the findings of Calabrese *et al.* that proportional slippage along with point mutations is not sufficient in the absence of mutational bias to explain the human genomic microsatellite length distribution [5].

The linear-biased models partly achieve a better fit to the data by producing better approximations to the variances in the marginal distributions of human and chimp microsatellites. Such a linear bias may be a signature of underlying counteracting forces in the mutational mechanism. The empirical findings that upward mutation bias of primary slippage mutations could be countered by the downward mutation bias at longer alleles due to the mismatch repair system [16] further strengthen the primacy of linear bias toward a focal allele. Natural selection could also be contributing to the downward bias by acting directly against longer microsatellites if they confer some selective disadvantage or by acting indirectly on the mismatch-repair machinery itself.

The biased models are robust to variation in the upper bound Ω, as is evident from their asymptotic AIC_c values, due to the presence of a downward or focal bias. The unbiased models, on the other hand, do considerably worse for larger values of Ω, because as microsatellites mutate without preferring contractions over expansions, they distribute themselves uniformly over the entire state space as time progresses. Thus, when Ω is large, microsatellites can attain unrealistically large repeat lengths under the unbiased models.

Among the one-phase models, rate proportionality gives a better fit to the data than rate equality among alleles in the presence of an unbiased or a linear-biased mutational process. However it does not do so in the presence of a strong constant downward bias ($\hat{u}$=0.46). Under a constant downward bias, most of the probability mass under stationarity is already piled over shorter alleles, and thus any increase in rate proportionality will only exacerbate this trend by reducing the mean holding time of longer alleles and thereby further reducing their stationary probability. In fact, the small negative value taken by the proportional-rate parameter ($\hat{s} = -0.0048$) reflects some level of restoration of probability mass to longer alleles countering the effects of geometric decay caused by constant bias. In the absence of any mutational bias, on the other hand, the ratio term $\alpha(u, v, j)/(1-\alpha(u, v, j+1))$ in the finite product of equation 2.4 simplifies to 1 for all alleles and thus makes the effects of proportionality pronounced. Any increase from 0 in the proportional-rate parameter s shifts the probability mass away from being uniformly distributed among all alleles toward shorter alleles reflecting their increased mean holding times relative to longer alleles. Similarly, under linear bias, the effects of proportionality are pronounced as this finite product has terms both ≥ 1 and < 1 for longer alleles.

The truncated *TPM* of DiRienzo *et al.* fits the pure AC-repeat data by essentially mimicking the truncated version of Ohta and Kimura's *SMM*. The two-phase models generally mimick their one-phase cousins in order to minimize variances in marginal distributions of chimp and human repeats. Even the slighest deviation from one-phase increases these variances. Our inability to reject one-phase in favor of two-phase using human-chimp data is in contrast with experimental observations of multi-step mutations. There are several explanations for this. First, noise in repeat length estimates due to indel activity in the flanking region may be at least partly responsible for elevating the experimentally observed proportion of multi-step mutations. Empirical studies usually keep track of the length of a microsatellite repeat along with its flanking sequence (PCR fragment length), rather than the actual repeat length. Studies have found both inter-specific and intra-specific fragment length polymorphism to be caused by indels in the flanking regions [2] [21]. Thus, on a cautionary note, indels in the flanking sequence could be construed as multi-unit microsatellite mutations if repeat lengths are directly extrapolated from the PCR fragment length.

Second, in the interest of not introducing any new models, except hybrids of existing ones, we forged our two-phase models in the image of *TPM* and a well-identified simplification of it in the spirit of Fu and Chakraborty. However, other formulations of a two-phase mutational mechanism, particularly, those which allow the probability p of single-step mutations and/or the success probability m of the conditional geometric distribution specifying the lengths of multi-step jumps to decrease with repeat length, may be more realistic, especially in light of empirical evidence for large contractions being more common among long alleles in yeast [35] and fruit fly [16]. For instance, the proportional-rate, linear-biased, two-phase model (*PL2*) corresponding to a linear-biased, truncated version of Fu and Chakraborty's *SMM*, can be modified further to incarnate a varying two-phase model whose geometric parameter m is further allowed to decrease with repeat length. As more of the chimp genome gets sequenced such varying two-phase models should be tested to further evaluate the importance of multi-step mutations. In this light, our rejection of two-phase should really be seen as the rejection of a homogeneous two-phase mechanism that is insensitive to repeat length in favor of a homogeneous one-phase mechanism.

Calabrese *et al.* [5] found significant motif-specific differences in equilibrium distributions obtained from human genomic data of pure dinucleotides repeats. We find such differences using human-chimp data. Motif-specific differences in efficiency of the mismatch repair system are manifested through the differences in *MLE*s of various model parameters for AC versus AT or AG-repeats (AT\G). The larger focal allele, along with the weaker downward bias for longer alleles, of pure AT\G-repeats compared to those of pure AC-repeats suggests that the mismatch-repair machinery is more efficient at repairing primary slippage mutations at longer AC-repeats. Interestingly, AC-repeats are also the most abundant of all dinucoletide repeats in humans and chimps. The absence of any significant differences in the slippage rates (proportional-rate parameters) between AC and AT\G-repeats suggests that the slippage machinery is not sensitive to the motif type.

On the other hand the slippage machinery is sensitive to point mutations as evidenced by a two-fold decrease in the slippage rate of an AC-repeat interrupted by just one point mutation relative to a pure repeat. This is not surprising as a point mutation is expected to create fewer opportunities for polymerase slippage and thereby decreases mutation rate as demonstrated in yeast [26]. There are differences in the mutational mechanisms of pure and interrupted repeats. First, the focal allele estimate of interrupted repeats is 3 units longer than that of pure repeats. Second, there is about a two-fold decrease in the mutation rate for longer interrupted repeats relative to the pure ones. This suggest that longer repeats, which are more prone to getting hit by point mutations, upon interruption, are less likely to mutate and thereby contract, due to linear bias toward the focal allele, as much as the pure repeats.

There is evidence in the human-chimp data as well as in human genomic data to reject lineage homogeneity in favor of lineage-specific variation in the mutational mechanisms of AC-repeats. There is an increase in the focal allele length along the human lineage relative to that along the chimp lineage. One possible explanation is that the human mismatch-repair system is not as efficient as that of the chimp. As has been pointed out by Harr *et al.* [16], subtle differences in the mismatch-repair system between two species could easily give rise to distinct mutational biases. The human AC-repeats also show a relative decrease in the mutation rate for longer alleles. This is compatible with a reduction in $N_e \times s$, the product of the effective population size and the selection coefficient against longer alleles in humans. Additional data are required to distinguish lineage specific differences in mismatch-repair efficiency and selection.

Our mutation rate estimates are not significantly different from often accepted rate of 6×10^{-4} for autosomal dinucleotide repeats in humans [11]. Empirical overestimates of the true mutation rate may result from sampling bias toward highly polymorphic loci which are typically also the fastest mutating. If the loci chosen to estimate mutation rate empirically have longer alleles on average, then an overestimation of the true average may result. The sample in our study is small for reliable mutation rate estimates as reflected in the large confidence interval of $[1.3 \times 10^{-5}, 1.8 \times 10^{-4}]$.

Chapter 6

FUTURE DIRECTIONS

These methods can be extended to more species as more primate sequences become available. One can test hypotheses and estimate parameters in a locus-specific as well as lineage-specific manner simultaneously. In particular, as data for primates accrue, it would be biologically relevant to test more general functional forms to model mutational bias as well as the nature of two-phase mutations. One may further use such species-specific and motif-specific parameter estimates in various population genetic inferences. The impact of model misspecification on signals of selective sweeps needs to be investigated.

Appendix A

Rectifying Nonidentifiability

The truncated *TPM* is nonidentifiable. When p=1, m is nonidentifiable and when m=1, p is nonidentifiable. In other words, when m=1, any value of p will produce the same probability distribution of the data, for all data, and vice versa [38]. Nonidentifiability guarantees inconsistency of MLE or any other estimator [33]. We propose the following single-valued transformation $T(p,m)$=(p^*,m^*), taking ordered pairs (p,m) in the square $[0,1]\times[0,1]$ to ordered pairs (p^*,m^*) in the kite $\mathcal{K}^\epsilon$, in order to rectify this. [27].

$$T(p,m) := \begin{cases} (p,m), & p,m \leq 1-\epsilon, \text{ or } p=m \\ (p(\frac{m-m_\epsilon}{m})+m_\epsilon, m), & m > \max\{1-\epsilon, p\} \\ (p, m(\frac{p-p_\epsilon}{p})+p_\epsilon), & p > \max\{1-\epsilon, m\} \end{cases} \tag{A.1}$$

where, $m_\epsilon := (m+\epsilon-1)/\epsilon$, and $p_\epsilon := (p+\epsilon-1)/\epsilon$. In our computations, we fix ϵ at 0.001. Thus, for all biological purposes, one may interpret p^* as p and m^* as m. We set $T^{-1}(1,1) = (1,1)$, so that the inverse image, $T^{-1}(p^*,m^*)$: $\mathcal{K}^{0.001} \rightarrow [0,1]\times[0,1]$ becomes well defined.

REFERENCES

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

[2] B. Angers and L. Bernatchez. Complex evolution of a salmonid microsatellite locus and its consequences in inferring alleleic divergence from size information. *Mol. Biol. Evol.* , 14:230–238, 1997.

[3] P. Brémaud. *Markov Chains, Gibbs Fields, Montecarlo Simulations and Queues.* Springer-Verlag, New York, 1999.

[4] K. P. Burnham and D. R. Anderson. *Model Selection and Inference A Practical Information-Theoretic Approach.* Springer-Verlag, New York, 1998.

[5] P. P. Calabrese and R. T. Durrett. Dinucloetide repeats in the drosophila and human genomes have complex, length-dependent mutation processes. *Mol. Biol. Evol.* , 20:715–725, 2003.

[6] P. P. Calabrese, Durrett R. T., and C. F. Aquadro. Dynamics of microsatellite divergence. *Genetics*, 159:839–852, 2001.

[7] H. Chernoff. On the distribution of likelihood ratio. *Ann. Math. Stats.* , 25:573–578, 1954.

[8] J. F. Crow and M. Kimura. *An Introduction to population genetics theory.* Harper and Row, New York, 1970.

[9] A. DiRienzo, A. C. Peterson, J. C. Garza, A. M. Valdes, M. Slatkin, and N. B. Freimer. Mutational processes of simple-sequence repeat loci in human populations. *Proc. Natl. Acad. Sci. USA*, 91:3166–70, 1994.

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

[11] H. Ellegren. Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in Genetics*, 16:551–558, 2000.

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

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

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

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

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

[17] Q. Huang, F. Xu, H. Shen, Q. Deng, Y. Liu, J. Li, R. R. Recker, and H. Deng. Mutation patterns at dinucleotide microsatellite loci in humans. *Am. Jnl. Hum. Gen.* , 70:625–634, 2002.

[18] C. Hurvich and C-L. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76:297–307, 1989.

[19] P. Jarne and P.J.L. Lagoda. Microsatellites, from molecules to populations and back. *Trends in Ecol. and Evol.* , 11:424–429, 1996.

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

[21] Y. Matsuoka, S. E. Mitchell, S. Kresovich, M. Goodman, and J. Doebley. Microsatellites in *Zea* - variability, patterns of mutations, and use for evolutionary studies. *Theor. Appl. Genet.* , 104:436–450, 2002.

[22] M. L. Menendez, D. Morales, L. Pardo, and I. Vajda. Inference about stationary distributions of Markov chains based on divergences with observed frequencies. *Kybernetika*, 35:265–268, 1999.

[23] R. Nielsen. A likelihood approach to population samples of microsatellite alleles. *Genetics*, 146:711–716, 1997.

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

[25] B. F. F. Ouellette and M. S. Boguski. Database divisions and homology search files: a guide for the perplexed. *Genome Res.* , 7:952–955, 1997.

[26] T. D. Petes, P. W. Greenwell, and M. Dominska. Stabilization of microsatellite sequences by variant repeats in the yeast *Saccharomyces cerevisiae. Genetics*, 146:491–498, 1997.

[27] B. L. S. Prakasa Rao. *Identifiability in stochastic models.* Academic Press, London, 1992.

[28] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press, Cambridge, UK, 1992.

[29] C. G. Primmer, H. Ellengren, N. Saino, and A. P. Moller. Directional evolution in germline microsatellite mutations. *Nat. Gen.* , 13:391–393, 1996.

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

[31] S. G. Self and K. Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Jnl. Am. Stat. Assoc.* , 82:605–610, 1987.

[32] N. Sugiura. Further analysis of the data by akaike's information criterion and the finite corrections. *Comm. stats. theory and methods*, A7:13–26, 1978.

[33] A. W. Van der Vart. *Asymptotic statistics.* Cambridge University Press, New York, 1998.

[34] M. T. Webster, N. G. C. Smith, and H. Ellegren. Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. *Proc. Natl. Acad. Sci. USA*, 99:8748–8753, 2002.

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

[36] S. Wolfram. *The Mathematica Book, 4th ed.* . Cambridge University Press, Cambridge, UK, 1999.

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

[38] S. J. Yakowitz and J. D. Spragins. On the identifiability of finite mixtures. *Ann. Math. Stats.* , 39:209–214, 1968.