ABSTRACT

Title of dissertation: DESCRIBING AND MODELING REPETITIVE SEQUENCES IN DNA

Suzanne S. Sindi, Doctor of Philosophy, 2006

Dissertation directed by: Professor James A. Yorke
Department of Mathematics
Department of Physics
and
Professor Brian R. Hunt
Department of Mathematics

A significant fraction of the **genome**, i.e. the complete DNA sequence, of most organisms is comprised of sequences for which there are similar copies somewhere within the genome. While most repetitive DNA was originally thought to have no function, there is a growing body of literature to suggest that repetitive sequences are vital to the genome.

The goal of this dissertation is to analyze statistical properties of repetitive sequences in the genomes of a variety of organisms. We find a variety of striking features of repetitive sequence in the human genome and the genomes of *C. elegans* (worm), *A. thaliana* (mustard seed) and *D. melanogaster* (fruit fly) with some comparison to *S. cerevisiae* (yeast) and *E. coli* (a bacteria). We find that the number of times each 40-mer (sequence of 40 bases) occurs in a genome is approximated by a power law distribution. We analyze in detail the separation between copies of 40-mers that occur exactly twice in a chromosome and observe that a significant

portion of these pairs, that we call "proximal", have extremely small separations, while the remaining "distant" pairs have a distribution more consistent with being uniformly distributed throughout the chromosome. We introduce a type of exactly repetitive region, which we call a "repeat string," and find the distribution of lengths of repeat strings is roughly a power law.

Since these properties have been verified for the genomes of a variety of organisms there may be a common explanation of their origin. When possible, we suggest evolutionary mechanisms that could cause the emergence of such statistical properties. In particular, we developed a model of the evolution of repeat strings in a genome. We find that, under quite general conditions, the stationary distribution of our evolutionary model is the Pareto distribution, a close relative of the power law distribution.

DESCRIBING AND MODELING REPEATED SEQUENCES IN DNA

by

Suzanne S. Sindi

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2006

Advisory Committee:
Dr. James A. Yorke, Chair/Advisor
Dr. Brian R. Hunt, Co-Advisor
Dr. Stephen Altschul
Dr. David Mount
Dr. Steven Mount

ACKNOWLEDGMENTS

There are many at this University and elsewhere that I owe a tremendous debt of gratitude. I know that I will certainly forget to list some, and for that I apologize in advance. While it "takes a village to raise a child", it takes at least that many to help someone earn a PhD.

I would first like to acknowledge my advisors James A. Yorke and Brian R. Hunt. Over the years they, more than anyone, have helped shape me into the researcher that I am today. Through them I learned that, surprisingly, the most difficult part of research can be simply formulating an interesting question that you can answer.

The support I have received from my fellow graduate students has been so important to my progress through graduate school and general sanity. I would especially like to thank Cathy Jones and D.J. Patil; they really looked after me and their love and attention have been so important to me. I was also lucky enough to have many wonderful students enter graduate school around the same time as me. This was such an amazing group of people and I am so pleased to call them friends. I would especially like to thank Brandy Rapatski who was my first Maryland friend; Greggo Johnson and Blake Pelzer for many nights of television viewing; Richard Hallquist for being such a good friend; JT Halbert and Aaron Lott for the many outdoor excursions and frequent conversations about "life, the universe and

everything".

In addition, I was lucky to get to know a wonderful person, Somantika Datta. Because of her stories and view on life my thoughts on the world have been broadened. I will never forget one of the first things we did together; she drew a map of India and told me about food, culture and history of some of the different regions. I feel very special to gotten to know some of her family as well.

I thank my family: Maria, Abdullah and Fred, for providing me with encouragement and reminding me just why I wanted to do this PhD thing anyway.

A few years ago I was fortunate enough to meet Ben Jewell. In addition to being an amazing cook, Ben has been a wonderful companion and partner. His daily support and positive attitude have been vital to my progress, especially in this last year. I feel so fortunate to get to share the rest of my life with him.

I have also benefited tremendously from discussions about my research with Nathan Edwards, Stephen Altschul and Steven Salzberg. Their comments and criticism has helped shape the direction of my research. Throughout graduate school, my research was funded in part by NSF Grants DMS 0104087 and DMS 0312360 and under NIH Grant 1R01HG0294501.

I must also thank the employees and owners of The Bagel Place for providing a location conducive to pondering coutour integrals. It is in part because of their reasonably priced breakfasts and endless coffee refills that I passed the Analysis qualifying examination.

(Thanks Dennis, Ruth, Patrick, Abbey, Alex, Helena, Jen, Joe, Magda, Maria and all the Leaders - I did it!)

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Introduction

1.1 Repetitive DNA

The availability of complete genomes for a variety of organisms has resulted in an explosion of work in analyzing the statistics of DNA sequences. One area of scientific and mathematical interest has been studying sequences of DNA with highly similar copies in the DNA sequence. Such repetitive subsequences constitute a significant fraction of the genomes of many organisms (for example, [3, 7, 19, 24, 47, 56, 77, 79, 80]). In fact more than half of the human genome is repetitive [46, 82]. Numerous studies (for example, [5, 34, 38, 42, 49, 50, 69]) have found evidence that repetitive DNA is involved with important processes. In the case of humans, some repetitive sequences have been linked to genetic disorders [23]. As James Shapiro wrote, "...the distribution of repetitive DNA sequence elements is a key determinant of how a particular genome functions (i.e., replicates, transmits to future generations, and encodes phenotypic traits.)" [74]

The goal of this dissertation is to analyze the structure of repetitive sequences in the DNA sequences of a variety of organisms. We study the human genome and the genomes of *C. elegans* (worm), *A. thaliana* (mustard seed), *D. melanogaster* (fruit fly) with some comparison to the genomes of *S. cerevisiae* (yeast) and *E. coli* (a bacteria). The genomes we study have been sequenced and analyzed beyond the

draft phase we will describe in the next section. These organisms were selected, in part, because their published sequences are among the most complete and accurate. Through studying the structure of these repetitive sequences in these genomes we seek to develop evolutionary models that provide plausible explanations for the structures we observe.

1.2 Whole Genome Shotgun Assembly

One of our motivations for studying repetitive sequences in DNA is the complications repetitive DNA poses to **genome assembly**, the process of determining the DNA sequence of an organism. Throughout this dissertation we will make reference to the relationship of our results to genome assembly. There are a number of different genome assembly programs in use (for example [8,29,35,54,63]). Although there are differences in the details of their algorithms, the underlying procedure is similar to what we next describe.

In the dominant method of genome assembly, termed **shotgun assembly** many copies of a genome are broken into millions of overlapping **reads** of about 700 bases (on average). No information about the location of the reads in the genome is obtained. The sequence of the reads has an error rate of roughly 1%; that is, typically several bases from each read are incorrect. Given this information, one must computationally assemble these reads into an estimate of the sequence of the genome. The first step in this process is typically to determine which reads come from overlapping parts of the genome; two such reads are said to **overlap**. Further

steps combine the reads into larger segments and the collection of all such segments is the **draft assembly**. (For further details on genome assembly refer to [55,65,66].)

Repetitive DNA is the major obstacle to shotgun assembly because it causes reads from different parts of the genome to appear to overlap. This makes it difficult to determine the correct location of each read relative to other reads. Generally, repetitive regions shorter than the read length are not so problematic because a single read can span a repeat, connecting the nonrepetitive sequence on each side. Repeat copies that are less than 98% similar are unlikely to pose a problem for a high quality assembler. Indeed, a small number of differences ($\approx 1/2\%$) are often sufficient to distinguish between repeat copies. High fidelity repeats are the major cause of difficulty in genome assembly as well as the major source of errors in draft assemblies (see [25,72]).

1.3 Outline of Thesis

In Chapter 2, we observe and discuss an approximate power law in the distribution of counts (number of occurrences) of length 40 words, **40-mers** in the human genome and the genomes of *C. elegans*, *A. thaliana*, and *D. melanogaster*. Our results are not sensitive to the word length $k = 40$; we obtain similar distributions for $20 \leq k \leq 100$. Previous studies have examined count distributions of much shorter words - those of length k for $3 \leq k \leq 10$ - in DNA sequences and sometimes utilized ranked distributions. We discuss how using longer word lengths provides a distribution more consistent with the counts of repetitive sequences in the genome.

In Chapter 3, we discuss a preference for small separations between copies of 40-mers that occur exactly twice in a single chromosome. We classify these pairs as either "proximal" or "distant". **Proximal pairs** are those that are separated by less than 0.3% of the chromosome length. In all five of the cases we examined, over 20% of the pairs are proximal, and in three cases over half are proximal. We explore extending proximal repeat pairs to an inexact match to study if both 40-mers are actually part of the same repeat region. We find that for most organisms a significant fraction of the cases we studied proximal repeat pairs belong to the same inexact repeat.

In Chapter 4, we study the distribution of lengths of repetitive sequences in a variety of genomes. We say a k-mer is **repetitive**, in a specified genome, if it occurs at least twice in the genome. We say a sequence in the genome of length $\geq k$ is a **repeat string** if each k-mer in it is repetitive and the string is maximal.[1] In all cases we observed that the distribution of repeat string lengths can be approximated by a power law. We develop a model of the evolution of repeat strings by random segmental duplications and point mutations. We show that under general conditions the distribution produced by this model is a **Pareto Distribution**[2], a distribution closely related to the power law [59]. The convergence of our model suggests distributions we observe in the genomes could be related to

[1] By **maximal** we mean that the subsequence is not contained in a longer sequence that is repetitive.

[2] A power law distribution is a function of the form $f(x) = ax^b$ while a Pareto distribution is of the form $g(x) = a(x + c)^b$ where $a, b, c, x \in \mathbb{R}$ and $x > 0$.

evolution through segmental duplications.

Our results are computed from GenBank [11] sequences of the human genome and the genomes of *C. elegans*, *A. thaliana*, *D. melanogaster*, *S. cerevisiae* and *E. coli*. The genome sequences and gene annotations we use in this research were current as of March 2006.

Chapter 2

Distribution of Long Word Counts in DNA Sequences

The term **genome** refers to the complete DNA sequence of an organism. DNA can be represented as one or more sequences of bases denoted A,C,G and T. Some recent papers have studied the number of occurrences (or counts) of short words (10 or fewer letters) in genomes [4, 22, 28, 32, 48, 51, 52, 60, 84]. Since a genome contains information complex enough to determine an entire organism, there should be no expectation that such a sequence would have statistics consistent with a randomly generated sequence of A, C, G and Ts. One of the features of DNA that distinguishes it from a randomly generated sequence is the presence of subsequences that occur repeatedly throughout the genome. Repetitive DNA comprises a significant fraction of the genomes of many organisms [3, 7, 19, 46, 47, 77, 79, 80, 82].

In this investigation we determine how many times each length–k word, or k**-mer**, occurs in a genome. If a k-mer occurs c times in the genome, we say it has **count** c. DNA consists of a double strand,

Because DNA is a double stranded helix where one strand is the complement of the other, the count of a k-mer includes all occurrences of its reverse complement as well.[1] We analyze the distribution of the number of k-mers with count c. We

[1]DNA consists of two complementary strands (or sequences); As are paired with Ts and Gs with Cs. Furthermore, the two strands are read in reverse direction and are called **reverse complements**.Hence, AAC is the reverse complement of the string GTT. When searching for a particular

generally study words of length 40, **40-mers**, because in a random sequence (each base equally likely) of length similar to the genomes we analyze, 40 is sufficiently large so that we expect no repeated words of length ≥ 40 (see Appendix). However, the results we present are not sensitive to this value and similar results hold for a broad range of word lengths between 20 and 100.

We study these distributions for the human genome and the genomes of *C. elegans*, *A. thaliana* and *D. melanogaster*. For all the genomes we study, we observe a power-law-like distribution in the counts of repetitive length 40 words over a significant range of counts, from 3 to roughly 70. When the data are approximated by a power law distribution, we find that the power law exponent, ≈ -2.5, is similar for all of the genomes analyzed. In addition, this exponent is preserved over different word lengths.

Previous studies, such as [48], have observed a power law distribution in the counts of significantly shorter words, k-mers where $k \leq 10$. Several other studies analyzed the distribution of ranked word counts; that is, the counts of k-mers plotted in decreasing order (such as [51]). Both of these types of analysis reflect the distribution of the most frequently occurring words, those with counts in the hundreds or thousands. Some properties reported in [51] have been found to hold for randomly generated sequences [14]. In this paper we characterize the distribution of counts of longer words, 40-mers, and show they provide results over a range of counts more consistent with the majority of repetitive DNA for the organisms we

string, we also look for its reverse complement. The count of a k-mer is actually the number of occurrences of it or its reverse complement.

study.

2.1 Low Count Repetitive Sequence

We now demonstrate that repetitive sequences in the genomes we study typically have a count less than 70 (for two different ways of counting repetitions). This is in the range of counts that is emphasized by our 40-mer distributions but not by previous studies [48,51]. For three of the genomes we study, over half of the repetitive 40-mers have a count of exactly two.

There is also a position-based way to count repetitions. Each position (or base) in the genome is the beginning of a 40-mer (except near the end of a chromosome), so we can refer to the **count of the position** (meaning the count of its 40-mer). We say a position is **repetitive** if its 40-mer is repetitive. For all our genomes, we find a typical repetitive position has a relatively low count. Over 2/3 of the positions, for each genome we study, had count at most 70.

To illustrate the difference between the distribution of 40-mer counts and position counts, imagine that a genome has only 101 repetitive 40-mers, 100 occurring twice and one 100 times. The counts by position are dominated by the most frequently occurring 40-mers. The average count of a repetitive 40-mer is $\frac{(1)100+(100)2}{101} \approx 3$ while the average count of the 300 repetitive positions is $\frac{(1)100^2+(100)2^2}{300} \approx 35$. Notice that in this case the median repetitive position count is two. The significant fraction of repetitive positions having high counts in the genomes we study can be illustrated as follows.

- For the human genome, the average repetitive 40-mer count is 5.08, while the average repetitive position count is 726. The median repetitive position count is 9. The maximum count of a 40-mer (and thus of a position) is 53,022.

- For *C. elegans*, the count of the average repetitive 40-mer is 3.34, while the average repetitive position count is 15.6. The median repetitive position count is 3. The maximum count of a 40-mer (and position) is 1,890.

The graphs in this paper show the distributions of repetitive 40-mer counts, not repetitive position counts. The distributions of 40-mer counts and position counts are related as follows. If $N(c)$ is the number of 40-mers with count c, then since each of these 40-mers occurs in c positions, the number of positions with count c is $cN(c)$. In this paper we graph $N(c)$, but the power law behavior we observe applies equally well to $cN(c)$ with an exponent one greater, i.e. ≈ -1.5.

2.2 Distribution of 40-mer Counts.

Recall, by the **count** of a 40-mer we mean the number of times a 40-mer or its reverse complement appears in a genome. In Figure 2.1 we graph the number $N(c)$ of 40-mers with count c for the genomes we considered. A power law distribution, $N(c) \approx ac^b$, for numbers a and b, is reflected by a linear relationship on the log-log scale, $\log(N(c)) \approx b\log(c) + \log(a)$. We indicate with a straight line segment the range best approximated by a power law for three of the genomes we study. The line segment approximates the counts from $3 \leq c \leq 70$ and has a slope ≈ -2.5. This indicates that for all our genomes, $N(c)$ is approximately proportional to $c^{-2.5}$

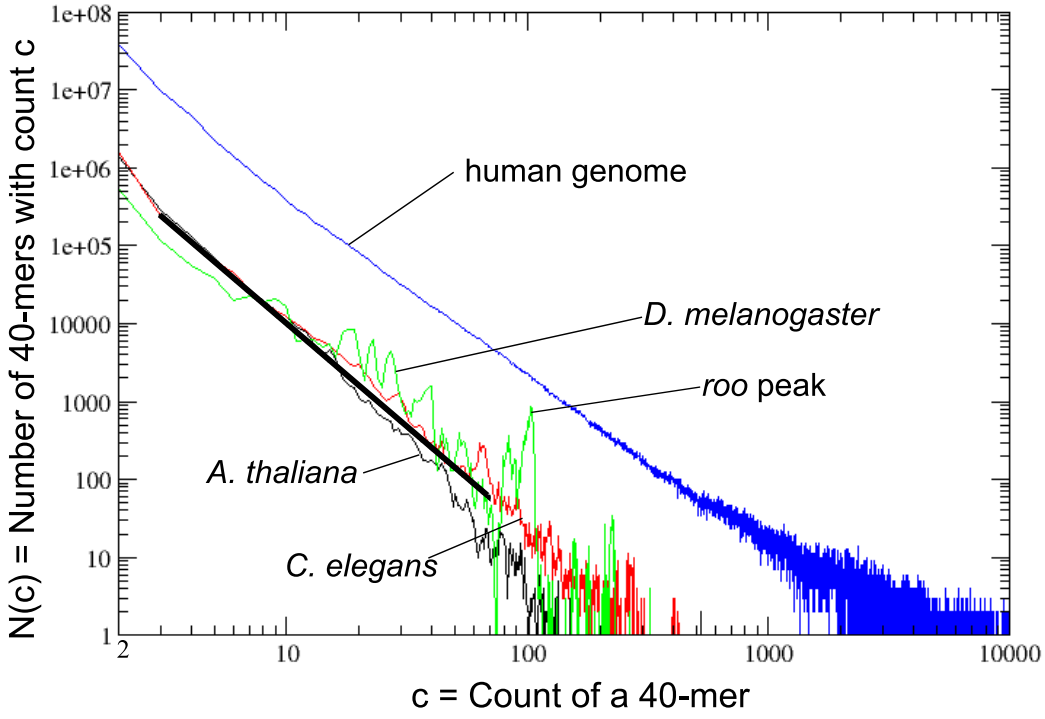

Figure 2.1: **Distribution of 40-mer Counts.** The number, $N(c)$, of distinct 40-mers with count c for *C. elegans* and *A. thaliana* is roughly a power law over the range $3 \leq c \leq 70$; that is, the distributions can be approximated by the straight line segment shown above. The plot for the human genome is nearly a straight line on the range $10 \leq c \leq 700$. The plot for *D. melanogaster* fluctuates substantially but follows the trend of the line segment approximating *C. elegans* and *A. thaliana*. The peak in the distribution of *D. melanogaster* denoted "*roo* peak" is caused by the *roo* transposon (see text). The power law exponent for each genome studied is roughly, -2.5.

in this range. For *C. elegans*, $N(c)$ remains close to this line over the range, while for *A. thaliana* the distribution drops more quickly for $c > 30$. The distribution for *D. melanogaster* fluctuates considerably more than the others but follows the general trend of the line. The data for the human genome follows a power law closely over the range $10 \leq c \leq 700$.

We further investigated instances where 40-mer counts for *D. melanogaster* deviate from the power law (see Figure 2.1). We find that these peaks are primarily due to the presence of high fidelity copies of transposons (i.e. transposable elements[2]) whose sequence can be found at the Berkeley *Drosophila* Genome Project [12]. For example, the peak at count $c = 104$ is due to the *roo* elements, which is the transposable element in *D. melanogaster* having the highest copy number. It also has high sequence conservation as described in [37]. We can similarly relate some of the other peaks in the distribution to high fidelity transposons in *D. melanogaster*. Since the distribution of 40-mer counts is significantly smoother for the other genomes, we believe this may indicate that they have comparatively fewer high fidelity copies of transposable elements.

In all cases we study, the data deviates significantly from the power law for the 40-mers with the highest counts. These 40-mers consist largely of low complexity 40-mers that comprise microsatellites[3] and other types of tandem sequence.

[2]Transposable elements are a class of repetitive sequences in DNA. These sequences can create a copy of themselves and insert the copy at another location in the genome.

[3]Microsatellites are subsequences of DNA composed of the same short word, typically less than length 6, repeated many times in succession. For example, CATCATCATCATCATCATCAT.

2.3 Sensitivity to Word Length

We compared the distribution of word counts for a variety word lengths, k, for the genomes we study. There is a significant qualitative change in the distribution of short word lengths, i.e., words that are sufficiently short that they will occur repeatedly by chance, compared to longer words where the number of words in the genome is significantly less than the number of words of that length. Figure 2.2 shows the distribution of k-mers in *C. elegans* for a variety of word lengths. For $k = 10$ only the tail of the distribution, $c \geq 300$, is consistent with a power law. Notice that the peak is near $c = 40$, which is somewhat lower than the value of $c \approx 190$ we would see for a purely random genome. For $k \geq 20$ the full distribution is more consistent with a power law. The behavior of the distribution is similar for these higher values of k and the range of counts from $3 \leq c \leq 70$ is well approximated by a power law with exponent ≈ -2.5. As seen in Figure 2.2, as k increases beyond 40 the data become more sparse. However, the distributions are still characterized by a power law with similar slope as for the lower values of k, but over a shorter range.

2.4 40-mer Counts in Coding DNA

We analyzed the distributions of counts of 40-mers in the genes and exons of the genomes of *C. elegans*, *A. thaliana* and *D. melanogaster*. We find that the distribution of word counts in the genes and exons is characterized by a power law of similar exponent to the power law for the entire genome. The behavior for *C. elegans*

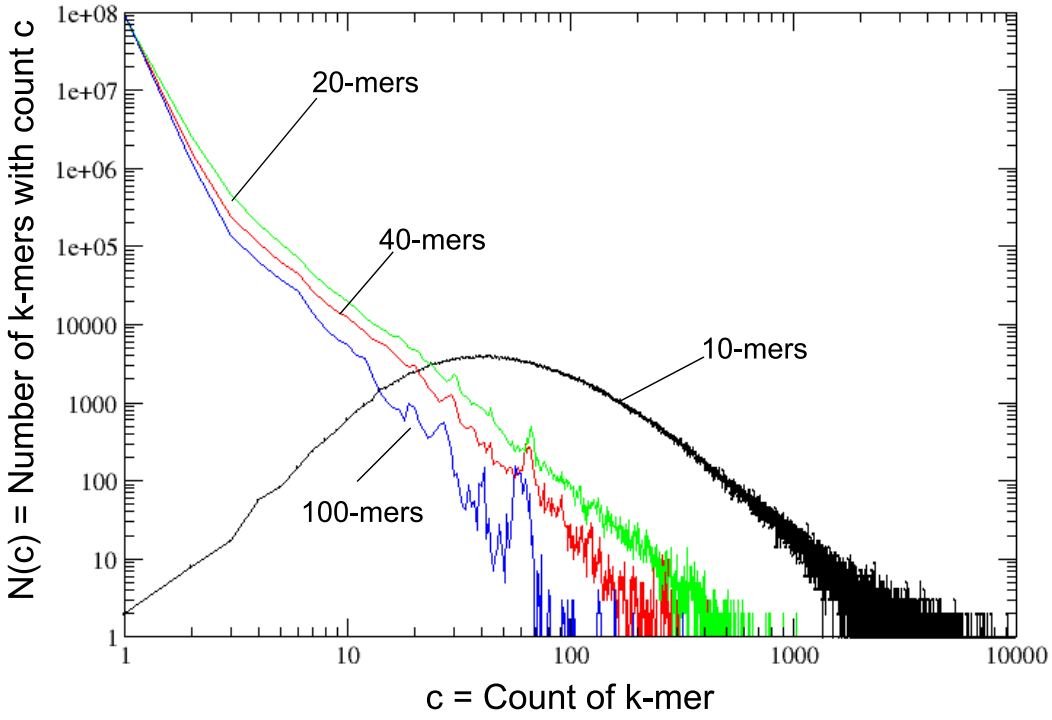

Figure 2.2: **Distribution of Counts in *C.elegans* for Different Word Lengths.** The distribution of k-mer counts for *C. elegans* follows a similar power law for all the values of k shown. In a random sequence of {A,C,G,T} of the same length as *C. elegans*, we expect every 10-mer to occur roughly 190 times. Because of this effect, the power law behavior for $k = 10$ does not emerge until well beyond a count of 200.

is typical for the other two genomes studied and therefore we present results for only *C. elegans*.

The distribution of counts of 40-mers in the genes, exons and the entire genome for *C. elegans* is shown in Figure 2.3. The distribution of the 40-mers in the genes is similar to the power law behavior seen for the entire genome. There is significantly less sequence in the exons, but the distribution of counts is reasonably consistent with a power-law-like distribution over a much shorter range (Figure 2.3).

Genes often occur in families that code for similar proteins. A power law has been shown to hold for the distribution of the number of members of gene families [43] as well as for other quantities related to coding DNA such as protein folds and pseudogene families [48]. Since genes in the same family may have highly similar sequence and the size of families of genes is distributed according to a power law, one may expect the distribution of 40-mers in coding regions of DNA to follow a power law as well. As is evident from Figure 2.3, the 40-mer counts for genes represent only a fraction of the overall counts. Thus, the distribution of 40-mer counts is preserved throughout the genome and cannot be explained by the distribution in only the coding regions of DNA.

2.5 Relation to Previous Work

Several papers have described power laws in the counts of short ($k \leq 10$) words (see for example, [48] and [51]). Other work has suggested that the short word data is better fit by a distribution with more parameters such as the Yule distribution

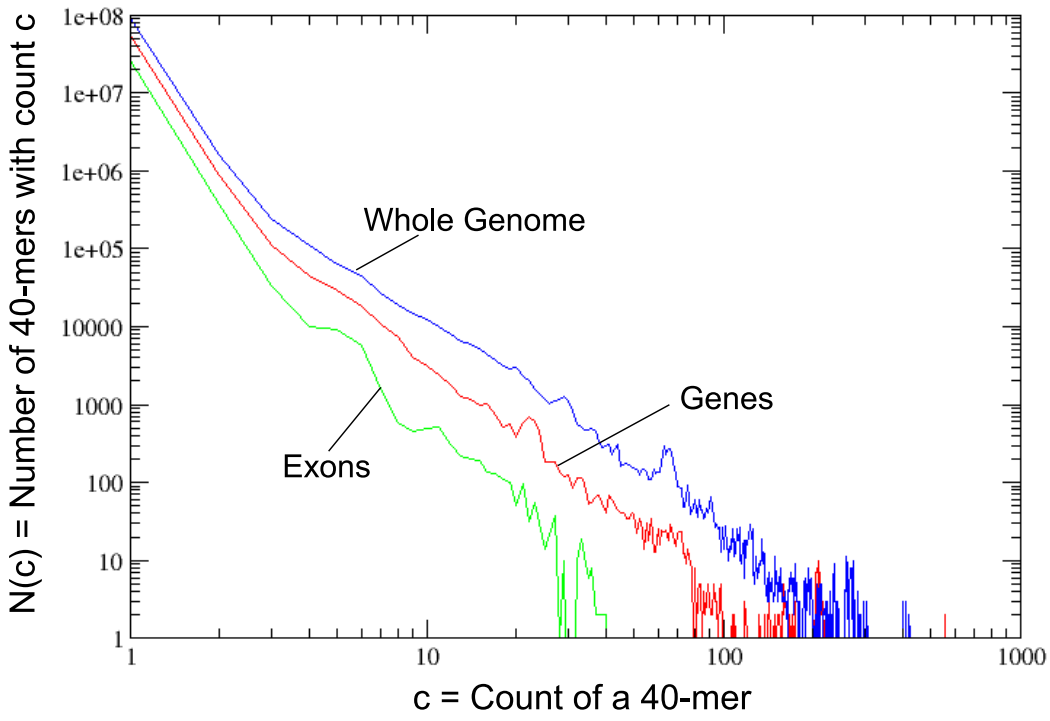

Figure 2.3: **Distribution of 40-mers in the Coding DNA of *C. elegans*.** We present the distribution of 40-mer counts in the genes of *C.elegans* as well in as the distribution of 40-mer counts restricted to the exons. For comparison, we plot the distribution of 40-mers for the entire genome of *C.elegans*. The gene counts follow an approximate power law similar to that for the whole genome, while the exon counts decrease with a similar slope over a shorter range.

[52]. We do not claim that a pure power law distribution is the best fit to the distribution of 40-mers counts. Rather, we claim that the power law distribution provides a reasonable characterization of the distribution of 40-mers in the range of relatively low copy repeats, such as $c \leq 100$.

As described earlier, for a random sequence having the same length as the genome of *C. elegans*, the expected count of each 10-mer is roughly 190. Thus, repeat regions with a count that is significantly less than 200 will not have a noticeable effect on the distribution of 10-mer counts. Indeed, the power laws observed in [48] all occur for counts well over 200.

Recall that in the genomes we studied, most of the bases in repetitive 40-mers (and repetitive positions in the genome) have count less than 100. Thus, the power law we observe reflects the majority of the repeat content in these genomes with high enough fidelity (similarity between copies) that many of the 40-mers they contain are identical. On the other hand, lower fidelity repeat regions (e.g., those with 90% similarity) will contain relatively few repetitive 40-mers. But then, based on our discussion above, repeat regions with sufficiently low fidelity and low count will not contribute significantly to the distribution of k-mer counts for any k.

Recent papers describe models of the evolution of gene families [39,68]. These papers may provide an explanation of the power law in genes (see Figure 2.3) [39,68]. However, because most repetitive 40-mers are not in genes these models as developed do not provide a complete explanation of the behavior in the counts of long words.

Chapter 3

Separations Between Repeats that Occur Exactly Twice

Repetitive sequence is typically classified according to sequence composition, copy number and mode of propagation in the genome. Some types of repetitive DNA, such as ALUs and LINEs, have millions of copies throughout the genome [67]. Other types of repetitive DNA have copies that occur adjacently and their method of propagation is through slippage or unequal crossing over (see [20, 44, 76]). The evolutionary dynamics of some types of repetitive DNA have been studied through direct observation and modeling (see references [10, 15, 16, 17, 18, 21, 30, 73]). However, there are other types of repetitive DNA whose mechanisms for propagation remain less certain. In this chapter we explore the distribution of a type of repetitive sequence that is not well characterized by previous classification methods.

In this chapter we analyze repetitive sequence having exactly two copies in the genome where both copies are within the same chromosome. Our goal is to see how these are distributed. There is a technical problem in identifying these repeats when the copies are of low fidelity (similarity). If the repeats are of low fidelity they can be hard to find, and if we find two such copies, we do not know if there are other copies that are of such low fidelity that they can not be detected. To avoid this problem, we select a certain high fidelity subset of the duplicates. We identify these high fidelity regions by first finding 40-mers that occur exactly twice in the genome.

We generally use length 40 because 40 is sufficiently large that the probability of any 40-mer occurring more than once by chance in a random sequence the length of the genomes (or chromosomes) is small (see Appendix).

Recall from Chapter 2 that we refer to the number of times a 40-mer, or its reverse complement, appear in a genome as the **count** of the 40-mer. A 40-mer with a count of at least two is called a **repeat** 40-mer because it has at least one other copy. The most common copy number for a repeat 40-mer in the genomes we study is exactly two (see Chapter 2). Strikingly, 40-mers with count two typically have both copies on the same chromosome (see Table 3.1). We call 40-mers with this property a **repeat pair**. We refer to the length of the sequence separating the 40-mers in a repeat pair as the **separation**. We analyze in greater detail **repeat pairs** in the human genome and the genomes of *C. elegans*, *A. thaliana* and *D. melanogaster* with some comparison to *E. coli*. In addition to presenting results for the complete genomes, we focus specifically on a typical chromosome from each genome; chromosomes 1 of *C. elegans* and *A. thaliana*, chromosome arm 2L of *D. melanogaster* and human chromosome 21.

In the next section, we classify these pairs as either "proximal" or "distant"; proximal pairs are those that are separated by less than 0.3% of the chromosome length. For all five of the genomes we examined, over 40% of the pairs are proximal, and in three cases over half are proximal. We also find a preference for proximal repeat pairs to occur in the same orientation[1].

[1]The 40-mers in a repeat pair are said to occur in the same orientation if they are not reverse complements of one another.

We extend the match between proximal repeat pairs in the same orientation and find that for some organisms roughly half of these belong to two "nearly adjacent" repeats. In other words, some lie in an approximate tandem repeat as we will discuss later.

A recent paper discusses separations between intra-chromosomal duplications [78]. The goal of this work was to analyze the separations between short exact duplications that were not part of a larger inexact duplication. In contrast, we seek to study segmental duplications in a more general setting. In particular, we wish to extend exact matches to larger inexact repeats, in the case of proximal repeats, to study if these could have originated from tandem duplication.

Although we can not identify the mechanisms responsible for the creation of such repeats in DNA we are able to observe their effects. By examining these intra-chromosomal repeat pairs we can determine characteristics of the mechanisms that created them.

3.1 Repeat Pairs and Proximal Separations

In Table 3.1 we describe the lengths of the genomes we study and how many 40-mers with count two have both copies on the same chromosome. In all cases we find that a significant fraction of 40-mers with count two occur on the same chromosome.

Figure 3.1 shows the fraction of the repeat pairs in human chromosome 21 that appear within a particular separation. The graph seems to hit the vertical axis at

about 0.73 because 73% of the repeat pairs have separations less than 0.003 times the length of the chromosome. These are the pairs we call proximal for this chromosome. There are over 120 times as many proximal pairs as would be expected if repeat pair locations were uniformly distributed. (For human chromosome 21, most of the first 13 million bases were undetermined, so we consider only the final 34.2 million bases. [81])

In all the organisms studied we find that repeat pairs with small separations are over represented; we define a **proximal cutoff** of 0.003 times the chromosome length to designate repeat pairs that are proximal. Repeat pairs that occur at a separation larger than the proximal cutoff are called **distant**. While our choice of the cutoff between proximal and distant pairs is somewhat arbitrary we generally see a natural division in the distribution of separations somewhere close to 0.003 times the chromosome length. In Table 3.2, we describe the percent of repeat pairs that are proximal for the selected chromosomes from the organisms we study.

The Proximal Cutoff. We now illustrate the distribution of proximal separations more explicitly. Figure 3.2 shows the fraction, $F(s)$, of repeat pairs with separation less than s in the chosen chromosomes of *C. elegans*, *A. thaliana* and *D. menalogaster*. The horizontal axis has a logarithmic scale, unlike Figure 3.1. Table 3.3 lists the proximal cutoffs for these chromosomes and the fraction of repeat pairs that are proximal. For each chromosome in Figure 3.2, there are about 100 times more proximal pairs than if the repeat pair locations were uniformly distributed.

Organism	Genome Length (in millions)	# of Count 2 40-mers (in thousands)	# of Repeat Pairs (in thousands)	# of Proximal Repeat Pairs (in thousands)
C. elegans	100	1,618	1,286	909
A. thaliana	119	1,404	924	611
D. melanogaster	120	551	472	407
human	3,000	38,767	18,104	8,321

Table 3.1: **Repeat Pairs.** We list for each organism the length of its genome and the number of 40-mers with count two. Strikingly, most 40-mers with count two have both their copies on the same chromosome. We call such 40-mers a **repeat pair**.

Organism	Genome Length (in millions)	≈% Proximal Repeat Pairs
C. elegans	100	70.7
A. thaliana	119	66.2
D. melanogaster	120	86.2
human	3,000	46.0

Table 3.2: **Proximal Repeat Pairs.** We say a repeat pair is **proximal** if the separation between the 40-mers is less than 0.003 times the length of the chromosome it belongs to. We list the percent of repeat pairs that are proximal in each chromosome averaged over all chromosomes in a genome.

Chromosome	Length (in millions)	Proximal Cutoff (in thousands)	≈% Proximal
C. elegans chr 1	15	45	80.7
A. thaliana chr 1	30	91	64.1
D. melanogaster arm 2L	28	67	97.9
human chr 21	34.2	141	73.1

Table 3.3: **Proximal Cutoffs for Selected Chromosomes.** We analyze human chromosome 21 because it is similar in length to the chromosomes for the other organisms. In all cases we find that proximal separations make up the majority of repeat pairs, the most extreme case being *D. melanogaster*, which had less than 3,000 distant repeat pairs. Notice that in contrast to the 46% proximal fraction for the human genome in Table 3.1, here over 70% of the repeat pairs is proximal for human chromosome 21.

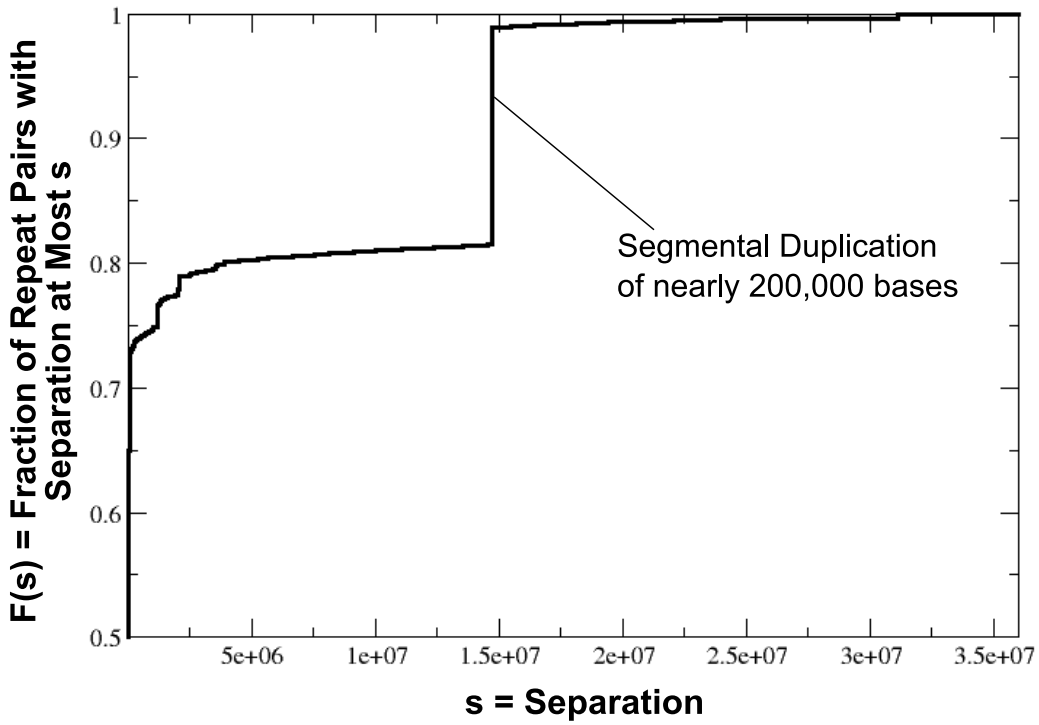

Figure 3.1: **Repeat Pair Separations - Human Chromosome 21.** The cumulative fraction $F(s)$ of repeat pairs that have separation less than s for human chromosome 21 appears to hit the vertical axis at roughly 0.73. The vertical jump that we label "Segmental Duplication" is due to a large segmental duplication in the genome; a region of approximately 200,000 bases is duplicated (with some local rearrangements) at a separation of ≈15 million bases [81].

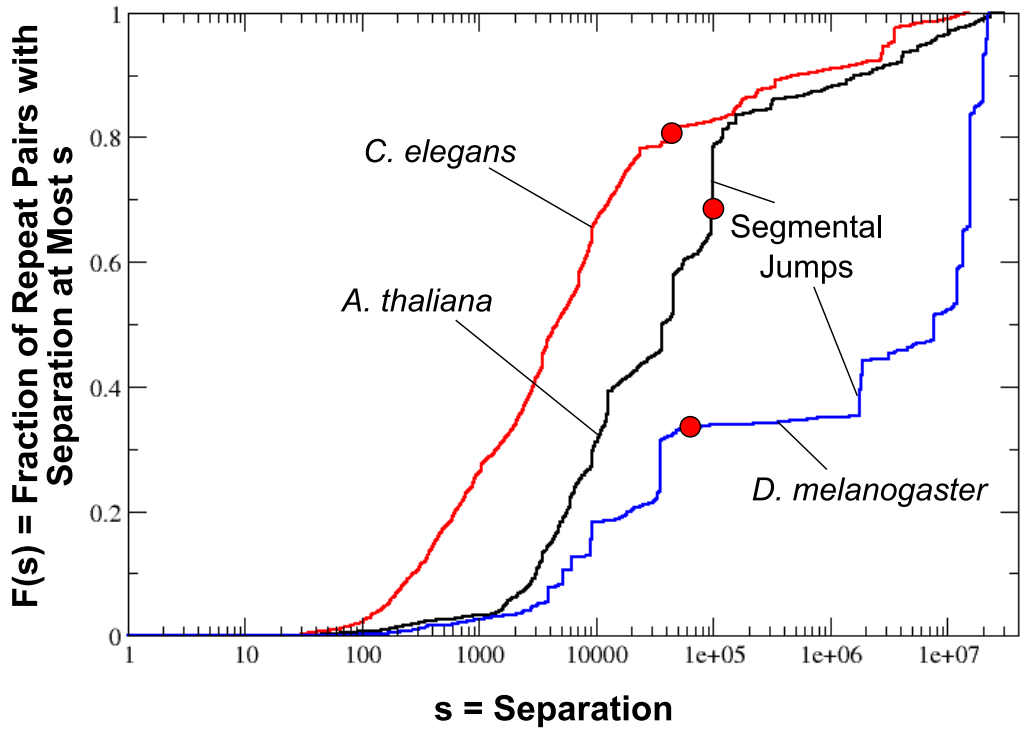

Figure 3.2: **Repeat Pair Separations - *C. elegans*, *A. thaliana* and *D. melanogaster*.** We plot the cumulative fraction $F(s)$ of separations less than s for repeat pairs in the first chromosomes of *C. elegans* and *A. thaliana* and chromosome arm 2L for *D. melanogaster*. The dots on the distributions indicate the separation chosen as proximal cutoff for each chromosome. Notice that if this were on a linear scale, the dots would appear to lie on the vertical axis.

Distant 40**-mer Pairs.** The cumulative distribution in Figure 3.1, for human chromosome 21, looks like a quadratic distribution with a vertical discontinuity. To investigate the separations of distant repeat pairs, we plot the distribution for the selected chromosomes after removing the proximal repeat pairs.

In Figure 3.4, we show the cumulative distributions of separations for distant repeat pairs in the chromosomes we studied. These distributions are shown along with the quadratic that is the cumulative distribution for the uniform random model in which the locations of each copy of a 40-mer in a repeat pair is determined by a uniform random variable.

The distant separations for chromosomes 1 of *C. elegans* and *A. thaliana* are significantly more likely to be small than when compared to the uniform random model. In contrast, the distant separations for *D. melanogaster* arm 2L are larger than would be expected under a uniform random model. For human chromosome 21 the cumulative distribution is more consistent with being determined by a uniform random model, with the exception of the single large segmental duplication. If we remove repeat pairs from the large segmental distribution for this chromosome, we find a distant distribution that can be reasonably approximated by the uniform random model.

Segmental Jumps. As we have said, the large jump in Figure 3.2 is due to a large **segmental duplication** corresponding to many repeat pairs at approximately the same separation. Segmental jumps indicate that there is a relatively long highly conserved repeat (not necessarily exact). The size of the jump will be related both to the length and fidelity of the repeat. Such a jump can occur within either proximal

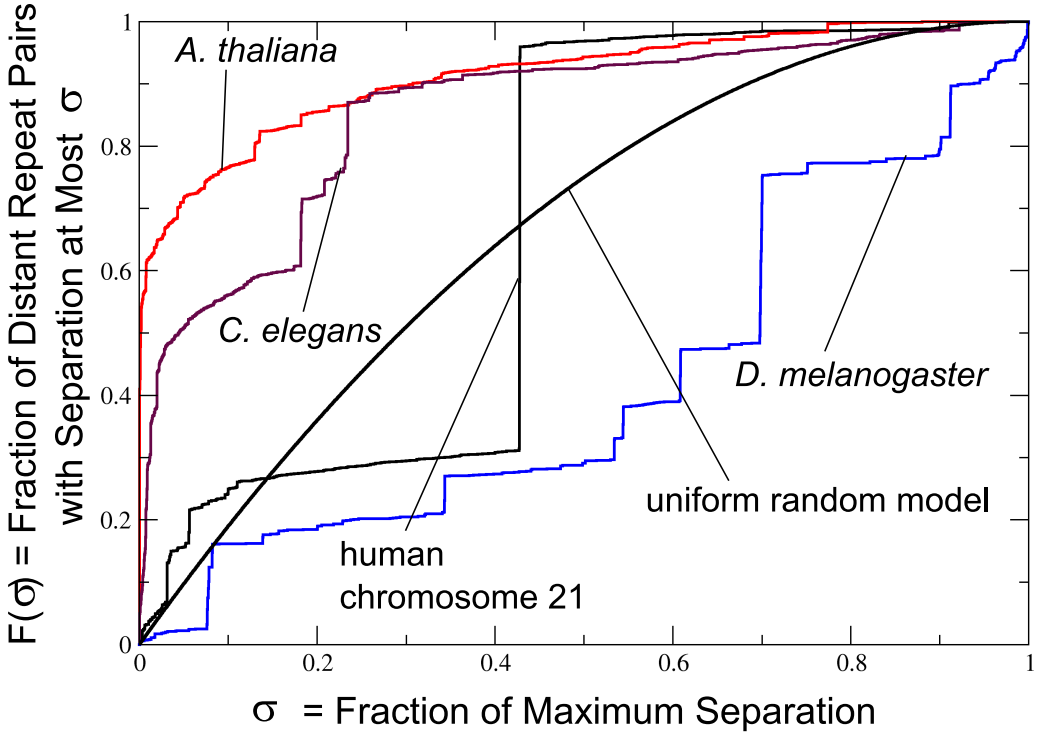

Figure 3.3: **Distant Separations.** For the same chromosomes as in Figures 3.2 and 3.3, we plot the cumulative fraction of separations between 0.003 and σ. That is, proximal repeat pairs are removed and we have normalized the chromosomes to have unit length. In addition, we show the separation distribution under a model when we pick locations of each 40-mer in a repeat pair from a uniform random distribution, namely $F(\sigma) = 2\sigma - \sigma^2$. While none of the distributions are well matched by the uniform random model we see that the shape of the distribution is more characteristic of the uniform random model than the distant separations, especially for human chromosome 21. For chromosome 1 of *C. elegans* and *A. thaliana* small separations are still over-represented and for arm 2L of *D. melanogaster* small separations are under-represented.

or distant separations. The large segmental jump in Figure 3.1 is due to a sequence of approximately 200,000 bases duplicated with a separation of about 15 million bases. The two copies are 96% identical, with some local rearrangements [81]. If the 4% differences were distributed at random, then about 0.96^{40}, or 20%, of the 40-mers in each segment would have a duplicate in the corresponding position of the other segment copy.

Orientation of Repeat Pairs. A repeat pair is in the **reverse** orientation if the two copies of the repeated 40-mer are reverse complements; otherwise they are in the same orientation. In general, we find that proximal repeat pairs are significantly more likely to occur in the same orientation than distant repeat pairs. The only exception is *C. elegans* for which we see essentially no correlation between proximality and orientation.

Similarly, over all repeat pairs, repeat pairs with the same orientation have smaller separations than repeat pairs with reverse orientation. For *D. melanogaster* arm 2L the median separation for pairs in the same orientation is $\approx 3,000$ bases and $\approx 5.8 \times 10^6$ bases for reverse oriented pairs. For *A. thaliana* chromosome 1 the median separation for pairs with the same orientation is approximately 12,500 bases and 320,000 bases for reverse oriented pairs. For human chromosome 21 the median separation for pairs with the same orientation is $\approx 260,000$ bases and $\approx 7 \times 10^6$ bases for pairs with reverse orientation. However, for *C. elegans* the median separation both types of repeat pairs was similar, approximately 8,000 bases.

As an additional anomaly, *C. elegans* chromosome 1 had approximately the same number of pairs in each orientation while the other chromosomes studied had

a preference for repeat pairs with the same orientation.

3.2 Repeat Pair Separations for *E. coli*

For a comparison to the repeat structure of the human genome and then genomes of *C. elegans*, *A. thaliana* and *D. melanogaster* we investigated repeat pair separations in the genome of *E. coli*. Since *E. coli* has a single (circular) chromosome [13], all 40-mers with count two will be, by default, repeat pairs. As with the other genomes, we find that a significantly higher fraction of 40-mer repeat pairs have small separation.

The separation between repeat pairs for *E. coli* is shown in Figure 3.4. The proximal cutoff is 11,500 bases and, roughly 20% of the pairs are proximal. Nearly all proximal repeat pairs are in the same orientation. The median separation between repeat pairs with the same orientation is roughly 140,000 bases and just over 700,000 bases for pairs with reverse orientation. This confirms that repeat pairs in the same orientation occur closer together than would be expected if both types had the same distribution.

3.3 Extension of Proximal Repeat Pairs

In this section we analyze in more detail the proximal repeat pairs. In particular, we want to determine if each repeat pair is from two adjacent repeat regions, i.e., a **tandem repeat**. We find in many cases the length of the sequence separating proximal repeat pairs can be decreased substantially by extending the match

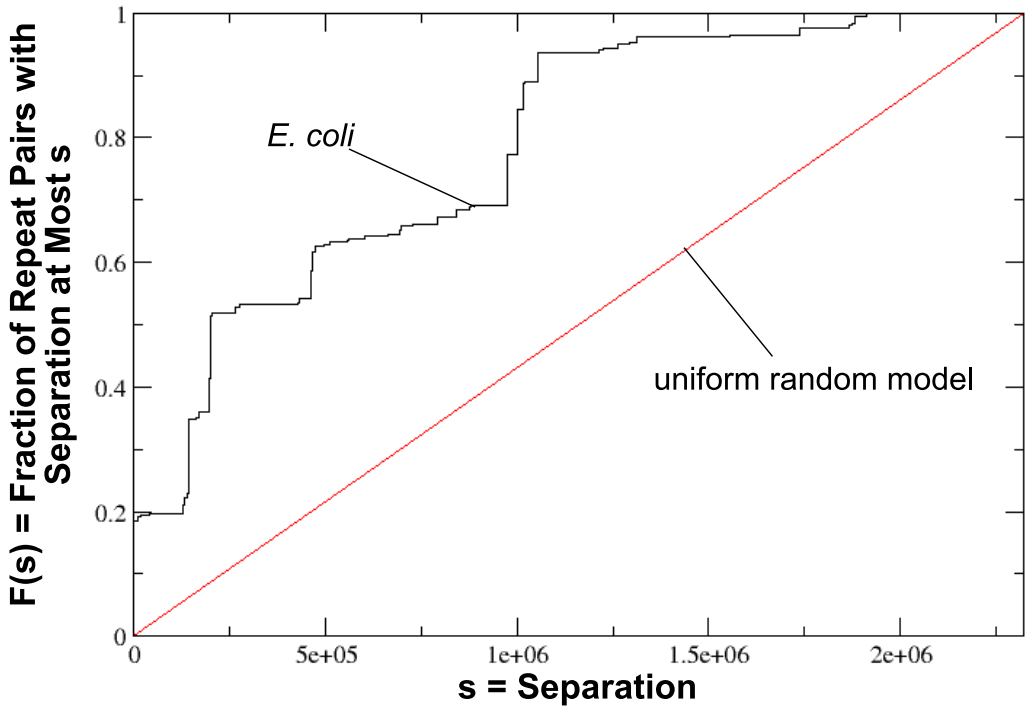

Figure 3.4: **Repeat Separations - *E. coli*.** The cumulative fraction of separations between repeat pairs in *E. coli* appears to hit the vertical at 0.20 and 20% of the repeat pairs are proximal. Since the genome of *E. coli* is circular, the largest separation possible between repeat pairs is half the genome length, or approximately 2.65 million letters. The straight line shown with the distribution is the cumulative distribution of separations expected from a uniform random model.

between them as an inexact repeat.

In this study we use the two 40-mers in a proximal repeat pair as seeds and attempt to extend the match (see Figure 3.5). We present results for extending the match for repeat pairs with the same orientation only. We sample the collection of repeat pairs to avoid "discovering" the same repeat more than once as follows. Of the 40-mers in a repeat pair, we select the one closest to the beginning of the chromosome and attempt to extend the alignment. We proceed to the next proximal repeat pair along the chromosome, ignoring any repeat pairs having a 40-mer in between the copies of previously considered repeat pairs.

We attempt to extend the sequence (in both directions) around both 40-mers in the repeat pair by running the NCBI-BLAST algorithm[2][57] between the sequences labeled L and S and then R and S in Figure 3.5. We determine what fraction of the separation sequence, indicated S in Figure 3.5, matches (part of) the sequence in L and R.

For *C. elegans* we sample 2,816 proximal repeat pairs with the same orientation. We find that for 1,410 cases over 90% of the sequence between the repeat pairs can be matched by sequence in L or R. In this case we say the repeat pairs could be **joined**. Repeat pairs with this property are likely to belong to the same inexact repeat region that will exhibit some tandem-like structure. A similar fraction of repeat pairs could be similar joined with this method for *D. melanogaster*. Using our sampling method we select 814 repeat pairs. Of these, 411 were joined.

[2]The program blastall was used to run blastn between two sequences. The minimum word length was $W = 20$ and every match identified by blastn was utilized.

For *A. thaliana* the situation is significantly different. Our sampling method selected 2,022 same oriented proximal repeat pairs; only 540 cases could be joined.

For the case of the human genome, we applied a different sampling process. Because the proximal cutoff for some human chromosomes was significantly larger than for the other genomes, we reduced the execution time by eliminating all proximal repeat pairs having separation larger than 10^6 bases. We then continued sampling repeat pairs in the same manner as described above. Of the 20,678 repeat pairs in our sample 10,395 of them were joined.

This suggests that about 50% (25% for *A. thaliana*) of the proximal repeat pairs belong to the same repeat region that seem to have originated from a tandem duplication event.

3.4 Discussion

In this section we discuss some possible consequences and significance of our findings. One of our main motivations for studying repetitive DNA is its importance in genome assembly, and in this section we discuss the implication of our results in this context.

3.4.1 Mechanisms Creating Proximal and Distant Repeat Pairs

As stated in the introduction, there are many mechanisms that create new repeat regions in the genome [17,41]. Some, such as slippage [44] and unequal crossing over [20,76] cause the formation of repetitive regions that are in tandem.

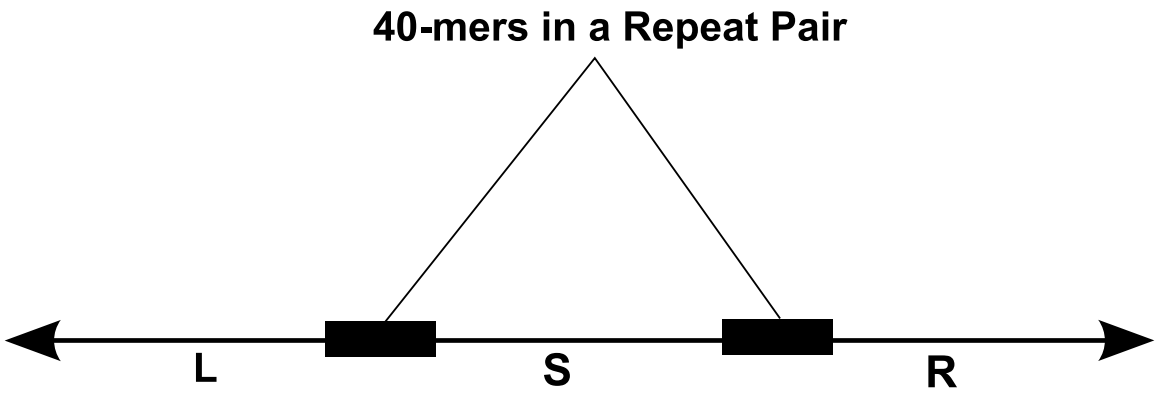

Figure 3.5: **Extending the Match for a Proximal Repeat Pair.** We attempt to extend the match between proximal 40-mers in both directions to determine if the repeat pair belongs to the same inexact repeat region. We align the sequences labeled **R** and **L** with the sequence between the repeat pair, labeled **S**. For the cases we sampled from the various genomes we find roughly half the time over 90% of the separation **S** has a high quality match in either **L** or **R**.

As time passes, identical tandem repeats become less similar due to the accumulation of mutations. Other mechanisms, such as the mechanisms of transposable elements and chromosome/genome duplication, create copies that can be separated by large distances. We can broadly classify these mechanisms to be either **proximal**, creating nearby copies, or **distant** creating copies with little bias regarding location on the chromosome.

Our studies suggest that for different organisms, distant and proximal mechanisms for creating intra-chromosomal repeats occur with different relative frequency. In all cases, we find the local mechanisms occur more than the global mechanisms. Between different species, the same ratio between these mechanisms is not preserved. The **proximal fraction**, that is fraction of repeat pairs that are proximal, is much greater for *D. melanogaster* than our other genomes (see Table 3.1), while human has by far the smallest proximal fraction.

Within individual genomes, the proximal fraction between chromosomes varies. In the case of human and *C. elegans* the X chromosome had the greatest proximal fraction over all other chromosomes in these genomes. The proximal fractions are 88.85% for chromosome X of human and 85.50% for *C. elegans*. However, for *D. melanogaster* the situation is reversed. With the exception of the smallest chromosome, which is chromosome arm 4, the X chromosome had the smallest proximal fraction 71.98%.

3.4.2 Selecting Regions with Count Exactly Two

Notice that a repeat pair might in fact belong to a repeat with a count higher than two due to minor differences in other copies. For example, the two 40-mers in the repeat pair may be part of a low-fidelity repeat of count three where one of the copies has a mutation causing some 40-mers to have count two. We remedy this situation by looking at the counts of 20-mers within the 40-mers in the repeat pair. We require that each 20-mer in the 40-mer occur only twice in the genome.

We find that requiring all 20-mers within the repeat pair to have count exactly two eliminates roughly half of all repeat pairs. This suggests that the eliminated repeat pairs are part of an inexact repeat in the genome with count greater than two. Repeat pairs that are proximal are preferentially eliminated by this process.

3.4.3 Implication for Genome Assembly: BAC Assembly

One method that assembly groups have undertaken to combat the presence of repetitive DNA in genome assembly is to first create a tiling of the genome by Bacterial Artificial Chromosomes (BACs) which are regions of about $150,000$ bases [9]. Then each of these BACs is sequenced according to the shotgun procedure described in Chapter 1. While this reduces the problems caused by repetitive DNA occurring on different chromosomes or with separations greater than the length of a BAC, the problem of "proximal" repetitive DNA still exists. Further study of problems caused by repetitive DNA reveal that repeat regions with a nearby copy are especially problematic for genome assembly [66,72]. As we have found in our

analysis of repeat pairs in many cases the majority of repetitive DNA with two copies in the genome occurs nearby and in the same chromosome. Thus, simply switching to a BAC by BAC shotgun procedure will not provide a solution for the problem of repetitive DNA.

Chapter 4

Evolution by Random Segmental Duplication Accurately Predicts Repeat String Distribution

The DNA sequence, or **genome**, of an organism is one or more sequences of four nucleic acids (or bases) represented by letters from the alphabet {A,C,G,T}. The length of the genome is several million letters for typical bacteria and about three billion for a mammalian genome. Because DNA sequences come from a small alphabet we expect to see subsequences that occur more than once. If a sequence of DNA is longer than $4^k + k$ for some positive integer k, then the sequence must have at least one repeated k-letter subsequence or k-**mer**. However, in practice there are many sequences in a genome that have highly similar copies more often than would be predicted by a random sequence of {A,C,G,T}. Such **repetitive regions** compose a major part of the DNA sequence of many organisms [3, 7, 19, 79, 80].

Although there have been numerous papers describing algorithms for finding repeats, such as [6, 36, 45, 64, 83], the concept of a repetitive region is typically left imprecise. This is in part because the term includes both exact and inexact, but similar, copies. In the latter case distinct copies of repeat regions have differences between them. After accumulation of many differences, two copies of a repeat region may not appear to be related at all. Identifying all copies of a repeated region may involve finding regions of varying similarity and thus depends on the methods used

to define similarity.

Another difficulty is the complicated structure of repeat regions. Repetitive sequences can overlap by arbitrary amounts and rearrangements in copies of a repeat region further complicate their identification [64]. In these situations each part of a subsequence of the genome will have a matching subsequence elsewhere in the genome; however, the entire subsequence will not have a complete match anywhere. For these reasons and others, the notion of length distribution of repeat regions is not defined in general.

To identify and study the length distribution of repetitive DNA in an unambiguous fashion, we develop a formulation we call a repeat string. We say a k-mer is a **repetitive k-mer**, in a specified genome, if it (or its reverse complement) appears at least twice in the genome. We say a subsequence is **repetitive** if, for a fixed k, each k-mer within it is repetitive. For simplicity in this dissertation we fix $k = 40$ because in a random sequence of {A,C,G,T} the length of a mammalian genome we do not expect any repetitive 40-mers (see Appendix). However, the results we present are valid over a range of sufficiently large values of k.

We define a sequence, S, to be a **repeat string** if it is repetitive and not contained in a longer repetitive sequence. Notice that for $k \geq 40$, each k-mer with an exact duplicate in the genome must lie in a repeat string.

In this chapter we find an approximate power law describes the distribution of lengths of repeat strings in the human genome and the genomes of *C. elegans*, *A. thaliana*, *D. melanogaster* and *S. cerevisiae*. Because the same distribution occurs for a variety of organisms the processes responsible the emergence of these

distributions should be quite general. We aim at determining general mechanisms that can create such power laws.

In this chapter, we develop a model of the genome evolution employing point mutation and segmental duplications and deletions. We show that under quite general conditions the stationary distribution of lengths of repeat strings for our model is a Pareto Distribution, a close relative of the power law [59]. The consistency of our model results with the repeat string length distributions we find in several genomes suggest the distributions we observe could have emerged through evolution by random segmental duplications and deletions.

4.1 Distribution of Lengths of Repeat Strings

We find an approximate power law relationship between the length, L, of a repeat string and the number, $N(L)$, of repeat strings of that length. Figure 4.1 shows the cumulative number, $C(L) = N(L) + N(L+1) + L(N+2)\ldots$, of repeat strings with length at least L for the human genome and the genomes of *C. elegans*, *A. thaliana*, *D. melanogaster* and *S. cerevisiae*. For all genomes, the distribution in the log-log plot of Figure 4.1 can be well approximated by a line segment over a significant range of the repeat string lengths. This linear fit corresponds to a distribution of lengths having the form $C(L) \approx aL^b$, where $a > 0$ and $b < 0$. Values of the exponent, b, of $C(L)$ were determined by a least squares fit and given in Table 4.1. While the different distributions are qualitative similar for the various organisms, namely a roughly power law distribution, there are significant

quantitative differences.

The exponents for the genomes we study seem to fall into two classes. The exponents for human and *C. elegans* are similar to one another, approximately -1.7. Similarly, the exponents for *D. melanogaster* and *S. cerevisiae* are roughly -1. The distribution for *A. thaliana* is similar to that of *C. elegans*, but the log-log graph for *A. thaliana* is significantly less straight than *C. elegans*. The exponent -1.36 for *A. thaliana* is most representative of the range $1,000 \leq L \leq 10,000$.

Another difference in the distributions is the range approximated by a power law. For human, *C. elegans*, *A. thaliana* nearly the entire range of the distribution in Figure 4.1 is well approximated by a line segment. This illustrates that the longest repeat string in a genome does not seem to grow linearly with the genome length.

For *S. cerevisiae* the distribution decays quite rapidly after $L = 5,000$, but the genome is significantly shorter than the others. The length of the genome of *D. melanogaster* is similar to *C. elegans* and *A. thaliana*; however, the distribution decays more rapidly than a power law beyond $L = 5,000$. With the exception of *D. melanogaster*, the range approximated by a power law increases with the length of the genome.

4.2 Models of Genome Evolution Through Random Segmental Duplications and Deletions

Because the distribution of repeat strings is qualitatively similar for the genomes we study, we expect that there are common mechanisms influencing the distribu-

Table 4.1: **Repeat String Distribution Slope.** The distribution of repeat string length L for the organisms we study roughly follows a power law. The slope of the logarithm of the cumulative distribution versus $\log(L)$ was calculated, using a least squares fit, over a range determined appropriate for each organism. We find a diversity of slopes, but the distributions appear to be split into two categories. Distributions for the human genome, *C. elegans* have slopes fairly close together. The slope of the cumulative distribution for *D. melanogaster* and *S. cerevisiae* are similar and significantly more shallow that the slope for the other three genomes. With the exception of *D. melanogaster*, the range of the power law is correlated with genome length.

Genome	Length (in millions)	Slope	Range of Power Law Approximation
human	3,000	-1.70	$40 \leq L \leq 30,000$
A. thaliana	119	-1.36	$40 \leq L \leq 10,000$
D. melanogaster	120	-0.99	$40 \leq L \leq 5,000$
C. elegans	100	-1.69	$40 \leq L \leq 10,000$
S. cerevisiae	12	-0.99	$40 \leq L \leq 6,000$

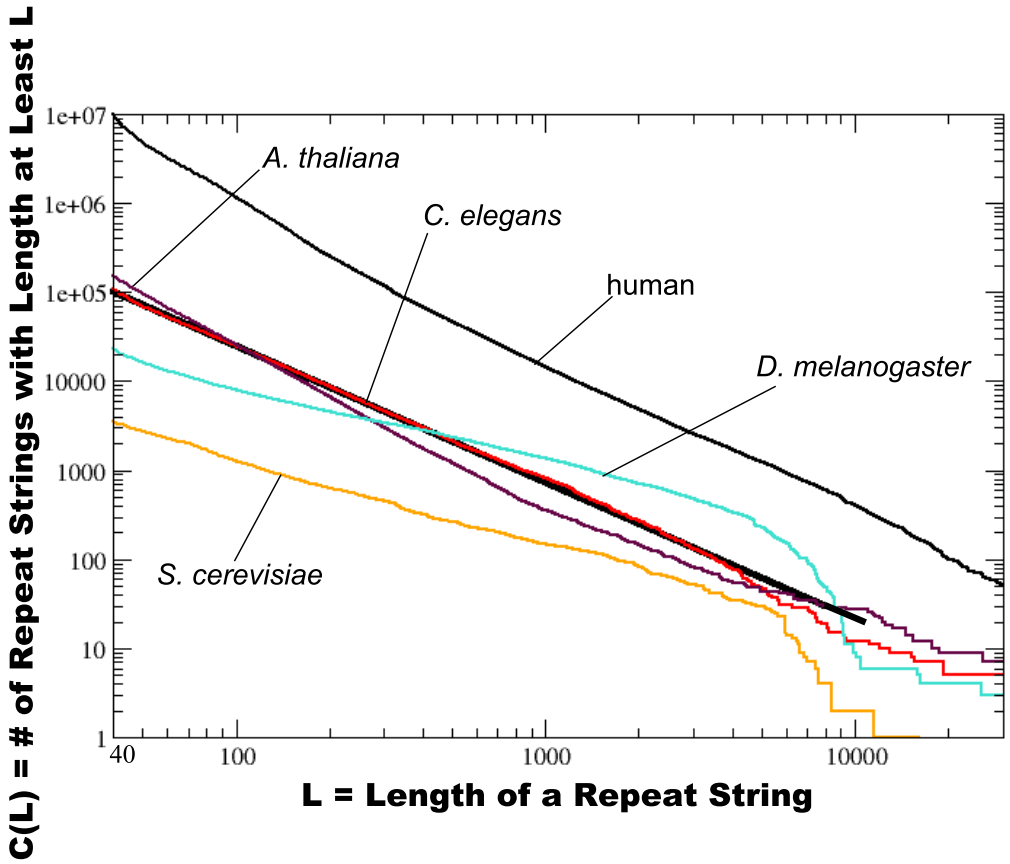

Figure 4.1: **Distribution of Lengths of Repeat Strings.** We plot the cumulative distribution of lengths of repeat strings for a variety of organisms. In each case we find this distribution to be approximated by a power law for a part of the range of L. Along with the data, we plot a line segment illustrating the approximate range and slope of the power law for *C. elegans*. For *D. melanogaster* and *S. cerevisiae* the range of the power law is shorter than for human, *C. elegans* and *A. thaliana*. Additionally, the exponent of the power law, corresponding to the slope on this log-log plot, for *D. melanogaster* and *S. cerevisiae* is shallower than for the other organisms.

tion. Since a genome represents only a snap-shot in evolution it is not possible to observe the processes responsible for creating repetitive DNA from the sequence data alone. We view the genome as the current state of a dynamical system describing evolutionary mechanisms. Our goal is, by observing the current distribution, to determine evolutionary mechanisms that could have lead to the emergence of behavior we observe.

Distinct types of repetitive DNA have different modes of duplication [1, 2, 21, 24, 41]. As described in [10, 34, 44, 76] there are many mechanisms for the growth of existing repeat regions and the creation of new repeat regions and repeat strings. Our goal is not to model specific types of repetitive DNA such as microsatellites and transposable elements; the dynamics of these sequences have been modeled extensively (see [30, 44, 61, 70, 71, 73]). Our aim is to understand more broadly the evolution of the genome. We develop a simplified model of segmental duplications and deletions that occur in two copies.

4.2.1 Fixed Length Model

In this initial model, we maintain a fixed length for segmental duplications/deletions in the genome. Additionally, we fix equal rates of segmental duplication and deletion to ensure that the length of the genome stays roughly the same. Our model has three parameters: W – the genome length; S – the segmental duplication/deletion length; and M – a constant that relates the point mutation rate to the segmental mutations rate.

Our model begins with a random genome of length W where each letter in the sequence is generated independently and is equally likely to be an A, C, G or T. Then, for W on the scale of the lengths of the genomes we study, we do not expect any repetitive 40-mers, and consequently repeat strings (see Appendix).

At each step in the model we induce a mutation, either a segmental mutation (which is a segmental duplication or deletion) or a point mutation. With probability $1/(M+2)$ we create a segmental duplication: a random subsequence of length S is chosen from the genome and inserted at a random location. A segmental deletion occurs with the same probability, $1/(M+2)$. We choose a subsequence of length S from the genome at random and delete it. With the remaining probability, $M/(M+2)$, we introduce a point mutation by changing the letter at a randomly chosen location in the genome.

We implemented our model as a computer simulation and ran experiments over many parameter values. After a specified amount of time, typically several million iterations, the simulation was halted and the repeat strings in the resulting genome were determined. We provide the results of some simulations in Figure 4.2 along with the stationary distribution for the model, which we will derive in the next section. Figure 4.2 indicates that the distribution from the simulation is qualitatively similar to the distributions observed for the genomes we study. That is, the distributions produced by our simulation can be roughly approximated by a line on a log-log plot.

Notice that nearly all repeat strings have length less than the copy length S. The exception is a repeat string introduced by copying a segment of the genome

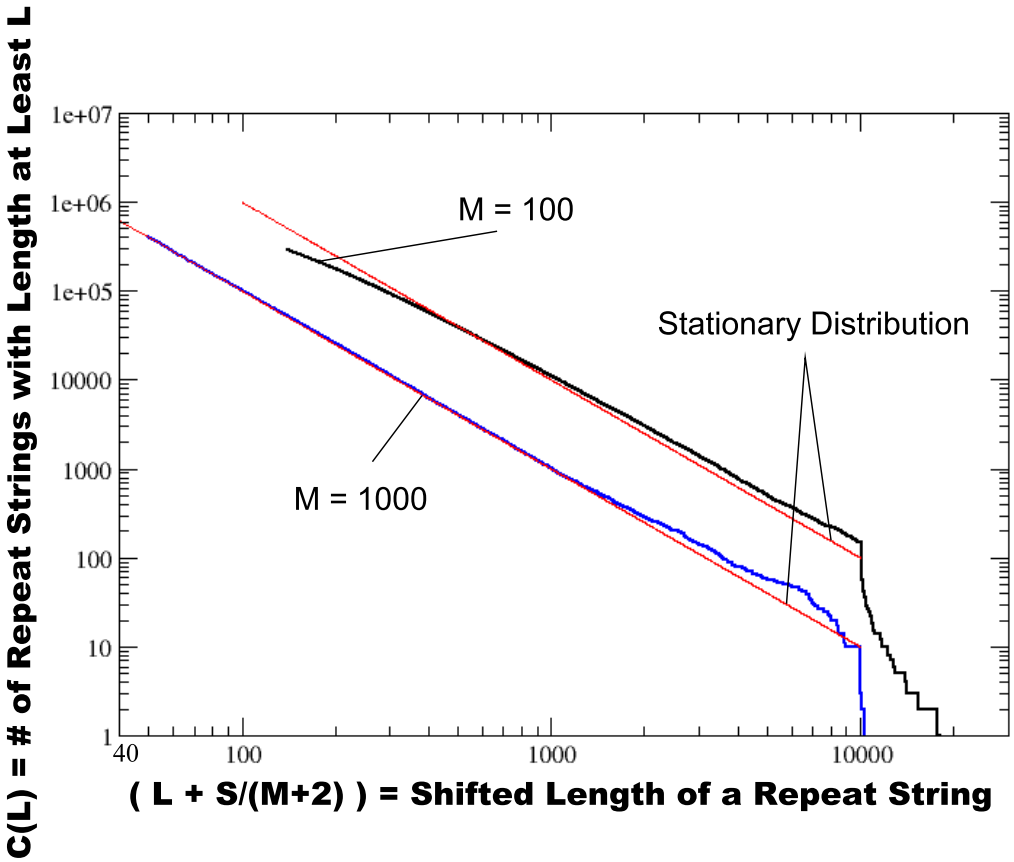

Figure 4.2: **Simulation of Fixed Length Genome Evolution Model.** We fix parameters $W = 10^8$, $S = 10^4$ and show results for $M = 100$ and $M = 1,000$. The simulation for $M = 100$ was carried out for 1 million steps and for $M = 1,000$ for 10 million steps, so that we made approximately 10,000 segmental duplications in each simulation. We overlay the shifted length distribution with the stationary distribution, a Pareto distribution, predicted by our analysis. Notice that over a significant range the computer simulations agree well with our theoretical findings, especially for $M = 1000$.

with both ends in existing repeat strings. However, repeat strings of this form occur relatively infrequently and make up no more than 1% of the total number of repeat strings produced by the model as long as $S/(M+2)$ is not too large.

4.2.2 Stationary Distribution of Fixed Length Model

We can explicitly analyze the stationary distribution of an approximation to the Fixed Length Model by making additional assumptions. Using a continuous approximation we let $x \in [0, S]$ represent the possible lengths of repeat strings and denote by $f(x)$ the density of repeat strings of length x at a given time. As discussed above we may create a repeat string longer than S but since these will be very uncommon we assume, for simplicity, that repeat strings have a maximum length of S.

We further assume that all repeat strings contain only repetitive 40-mers with count two. In doing so we are able to model a point mutation in an existing repeat string as splitting it into two repeat strings, the sum of whose length is equal to the length of the original repeat strings (here we ignore the 1 base that is lost due to the point mutation). If a repeat string had a 40-mer with count greater than two, this assumption would be violated.

We analyze the distribution of repeat strings produced by this process. Since our model begins with no repetitive sequences, all repeats are introduced by segmental duplications. Changes in the lengths of previously duplicated segments are caused by: (1) deleting (or partially deleting) an existing repeat string; (2) intro-

ducing a point mutation to an existing repeat string; (3) inserting a segmental duplication into an existing repeat string. Thus to analyze the distribution of lengths of repeat strings we consider the probability a repeat string is deleted or is mutated during a step of our simulation.

Let Lf be the expected density after one step of the model. We want to identify the stationary distribution, that is, the limit as n goes to infinity of $L^n f$. We begin by identifying the explicit form of Lf. We claim that Lf is an affine transformation, whose linear part we call Hf:

$$(Lf)(x) = (Hf)(x) + \frac{1}{M+2}\delta(x - S) \tag{4.1}$$

where

$$(Hf)(x) = \left(1 - \frac{1}{W}\left(x + \frac{S}{(M+2)}\right)\right) f(x) + \int_x^S \frac{2}{W} f(y)dy. \tag{4.2}$$

We will derive equations (4.1) and (4.2) by considering the expected action of each of our mutations. We first consider the expected action of introducing a segmental duplication to the distribution of repeat strings. Each time we introduce a duplication we create a repeat string of length exactly S. Because the length of repeat strings are continuous in the range $[0, S]$ this corresponds to changing f by adding a delta function centered at S.[1]. The other operation that is part of a segmental duplication is inserting the repeat string to a randomly chosen position in the genome. If the location of the insertion were in an existing repeat string of length x, this would split the repeat string into two separate repeat strings whose

[1]The Dirac delta function is denoted $\delta(x)$. This "function" is defined as having unit integral, $\int_{-\infty}^{+\infty} \delta(x)dx = 1$ but $\delta(x) = 0$ when $x \neq 0$.

total lengths sum to x. Since the insertion point is chosen random from the W positions in the genome, a position in a repeat string of length x will be chosen with probability $\frac{x}{W}$. After an insertion the fraction of repeat strings with length x that will still have length x is $(1 - \frac{x}{W})$.

If a repeat string has length $y > x$ there are exactly two insertion points that would create a repeat string of length x. Similarly the probability of choosing the insertion point in a string of length y is simply y/W for each repeat string of length y, which is $f(y)$. Combining these ideas, the expected action on of one segmental duplication is given by the operator C:

$$(Cf)(x) = \left(1 - \frac{x}{W}\right) f(x) + \int_x^S \frac{2}{W} f(y)dy + \delta(x - S). \tag{4.3}$$

A segmental deletion selects at random a position in the genome and deletes the S bases following the position. An existing repeat string may be entirely or partially deleted by this process. For each repeat string of length x the probability that the length of the repeat string is changed by the segmental deletion is, $\frac{S+x}{W}$. The fraction of repeat strings with length x that are not partially deleted by the deletion operation is $(1 - \frac{S+x}{W})$.

Similarly to the duplication process a repeat string of length x can be created if the deletion point is in a repeat string of length $y > x$. For each repeat string there are two choices for deletion positions that would create a repeat string of length x. Thus, the expected action on f of a segmental deletion of length S is given by the operator D:

$$(Df)(x) = \left(1 - \frac{x + S}{W}\right) f(x) + \int_x^S \frac{2}{W} f(y)dy. \tag{4.4}$$

Finally, we consider the action of a single point mutation. A point mutation will change the repeat structure only if it lands in a repeat string. As described above, the repeat string will split into two parts. Similar to the insertion mutation, the fraction of repeat strings of length x that will survive a point mutation is $(1-\frac{x}{W})$. Additionally, as in our two previous operations, a mutation occurring in a repeat string of length $y > x$ can create a repeat string of length x if the mutation chosen to be in one of two possible positions. The expected action of a single point mutation is given by E:

$$(Ef)(x) = \left(1 - \frac{x}{W}\right) f(x) + \int_x^S \frac{2}{W} f(y)dy. \tag{4.5}$$

Notice that that the operators C and E differ by only the term from creating a repeat string of length S. This is because the effect of an insertion on existing repeat strings is essentially equivalent to that of a point mutation.

The expected action on f overall is given by a linear combination of these operators:

$$(Lf)(x) = \frac{1}{M+2}\left((Cf)(x) + (Df)(x)\right) + \frac{M}{M+2}(Ef)(x).$$

This can be simplified to the form given in equations (4.1) and (4.2):

$$(Lf)(x) = \left(1 - \frac{1}{W}\left(x + \frac{S}{(M+2)}\right)\right) f(x) + \int_x^S \frac{2}{W} f(y)dy + \frac{1}{M+2}\delta(x-S).$$

Let Hf be the first two terms in the above expression for Lf. Observe that Hf is linear in f:

$$(Hf)(x) = \left(1 - \frac{1}{W}\left(x + \frac{S}{(M+2)}\right)\right) f(x) + \int_x^S \frac{2}{W} f(y)dy.$$

We claim that Hf is a contraction in the norm

$$\|f\|_* := \int_0^S |xf(x)|dx. \tag{4.6}$$

Indeed, a simple calculation (see Appendix A.2) shows that H is a contraction:

$$\|(H(f))\|_* = \left(1 - \frac{S}{W(M+2)}\right)\|f\|_*$$

Because the linear part of the operator L is a contraction, the entire operator is a contraction and must have a unique attracting fixed point g. By inspection we determine this fixed point, the stationary distribution, to be

$$g(x) = \alpha\left(x + \frac{S}{M+2}\right)^{-3} + \beta\delta(x - S). \tag{4.7}$$

α and β are constants that depend on W, M and S.

(See Appendix A.3 for a derivation of constants α and β.)

Observe that g is a global attractor, and thus the stationary distribution does not depend on the initial distribution of repeat intervals. That is, our assumption that there were no initial repeat intervals can be removed and not impact the stationary distribution.

The stationary distribution, given in equation (4.7), is a Pareto distribution (a shifted power-law) plus a multiple of a delta function. We refer to $\frac{S}{M+2}$ as the **shift**. The delta function component of $g(x)$ is an artifact of all segmental duplications created with the same length S in this model. The power -3 in the density function corresponds to a power -2 in the cumulative distribution function. Notice that some of the slopes in Figure 4.1 are roughly -1.7, which is near our theoretical

value of -2. In a later section we discuss a refinement of the notion of repeat string that more closely matches the types of repeat strings our model represents.

Notice in our analysis we assumed that S is the maximum length of a repeat string. As discussed previously, there may be repeat strings longer than S created. We find that when approximating our theoretical distribution of repeat strings from our model with the actual distribution of repeat strings produced by a computer simulation that the two distributions are quite similar (see Figure 4.2).

Recall we also assumed the count of a repeat string was two. We note that the results from our simulation agree better with our theoretical distribution when $M = 1,000$ that for $M = 100$ (see Figure 4.2). We find that 40.3% of repetitive 40-mers had count greater than two when $M = 100$ as opposed to only 13.9% when $M = 1,000$. Larger values of M correspond to lower 40-mer counts; this suggests that our stationary distribution is more accurate for larger values of M.

Notice that the length of two different repeat strings will change if the location of an insertion, deletion or point mutation occurs within a repeat string. However, we only explicitly describe the impact of these mutations on one repeat string. Changing Lf to describe changes in the length of both repeat strings multiplies some of the terms by a factor of two, but does not change the stationary distribution.

4.2.3 Variable Length Model

Our Fixed Length Model has several assumptions that we loosen to analyze how general the convergence to a Pareto distribution is. We first change the length

distribution of segmental duplications and deletions. We now model the length of segmental mutations as coming from a probability distribution with mean S. We find that the distribution of repeat strings remains approximated by the same Pareto distribution as found for the constant length model.

We present results in Figure 4.3 for $W = 10^8$, $M = 100$ and the segmental mutations as an exponential, normal and uniform distribution with mean $S = 10^4$. Note we plot the length distribution with the same shift as predicted by our analysis of the Fixed Length Model. We find that the stationary distribution derived in the previous section is consistent with the repeat string distribution from the Variable Length Model.

4.2.4 Growing Genome Model

Another restriction of our Fixed Length Model was that, on average, the length of the genome remained constant. Certainly the length of the genome is not a fixed value. There are frequent insertions and even whole genome duplications especially in plants [27,40]. In this model we allow the length of the genome to vary, but so the length of the genome does not go to zero we have a preference for duplications over deletions.

We developed a simulation, based on our Fixed and Variable Length Models, which allows the length of a genome to increase. The model has the same parameters; W – the initial genome length; S – the average length of a segmental duplication or deletion and M – a constant that relates the point mutation rate to

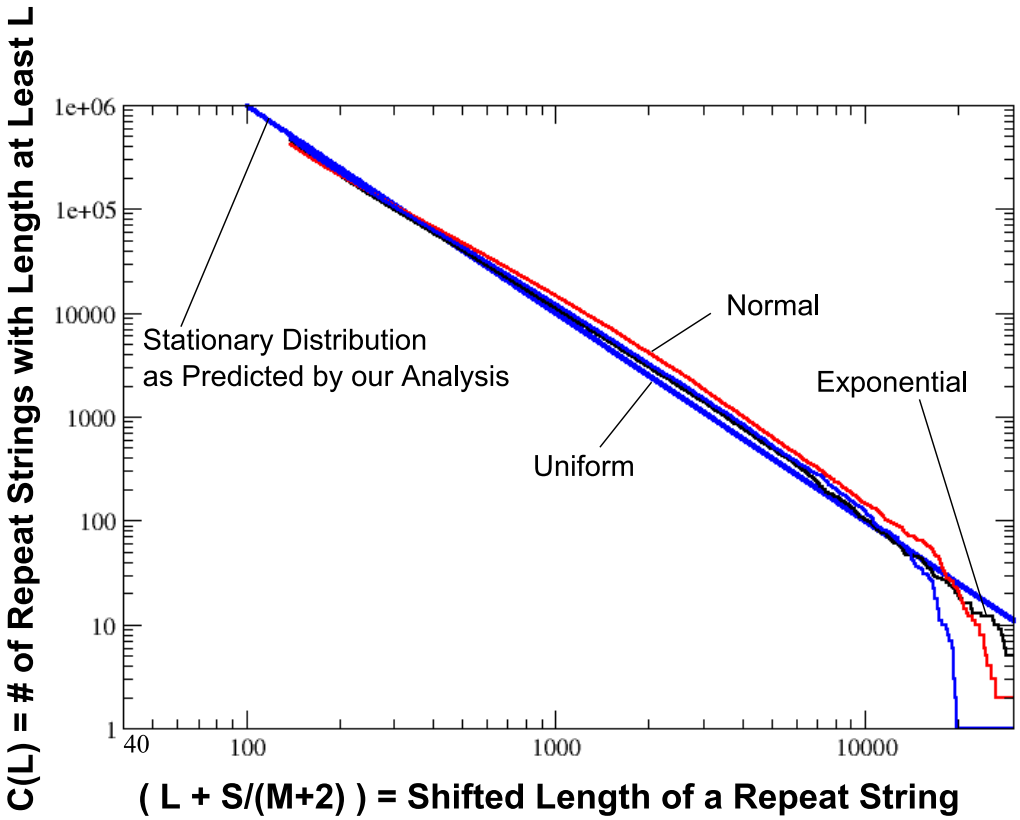

Figure 4.3: **Simulation of Variable Length Genome Evolution Model.** We plot the length distribution from runs of our Variable Length Model where we have fixed $W = 10^8$ and $M = 100$. We show the result from three runs where different probability distributions, all having mean $S = 10^4$, were used as the length distribution for segmental duplications and deletions. We use the exponential distribution with mean S, uniform distribution on $[0, 2 \times 10^4]$ and normal distribution with mean and standard deviation S. In the case that the length determined by the normal distribution is negative the value is ignored and a new length selected from the normal distribution. In this manner, the actual mean of the distribution is somewhat larger than S. In all three sample cases shown, the distribution of repeat strings converges to a distribution consistent with the stationary distribution derived for modeling the length of segmental mutations as constant.

the segmental mutations rate. We perform a point mutation with the same probability $M/(M+2)$, but the probabilities of segmental duplications and deletions are no longer symmetric. With probability $1.1/(M+2)$ we perform a segmental duplication and probability $0.9/(M+2)$ we perform a segmental deletion. The lengths of the segmental duplications and deletions are selected from a probability distribution with mean S. Instead of halting the simulation after a specified number of steps we halt when the length of the genome has reached a specified value.

We show the output of two simulations in Figure 4.4. Note that we plot the length distributions with the same shift as for the Fixed and Variable Models. For these simulations the lengths of segmental duplication and deletions were selected from a uniform distribution on $[0, 2 \times 10^4]$.

We note that in Figure 4.4 the distributions of repeat string lengths is characterized by a relatively linear region (on a log-log scale) and then a decaying tail. The decay of the tail is more consistent with the decay in the distribution of *D. melanogaster* (see Figure 4.1).

The distinguishing feature of the distribution produced by the Growing Genome Model from the Fixed and Variable Length Models is the slope in the linear region of the distribution. The slope produced by the Growing Genome model is significantly shallower. For the two simulations shown in Figure 4.4 the slope of the linear regions are roughly -1.5 and -1.8 when determined using a least-squares fit over the intervals $[1040, 10^6]$ and $[300, 7000]$ respectively. Notice that these slopes are closer to the values for the slope determined for *D. melanogaster* and *S. cereviaise* than produced by the other two models. Because the genome of *S. cerevisiae* has

undergone genome duplication [40], the dynamics of the Growing Genome Model are perhaps more consistent with its evolutionary history than the dynamics of the other two models we developed.

4.3 Unitary Repeat Strings

We have shown the distributions produced by our Fixed and Variable Models converge to a Pareto distribution with a constant exponent of -3, corresponding to a cumulative slope of -2. However, as shown from Figure 4.1 and Table 4.1, none of our organisms have a slope of -2. In this section, we consider a refinement of the definition of a "repeat string" and find that our Fixed and Variable Length Models model produces results that are more consistent with this refinement for three of the genomes we study.

Repetitive sequence can be created through **tandem duplication**. This is when a duplicate copy is created adjacent to the original sequence. In this case, the length of a repeat string would increase and a new repeat string may not be created. The models we developed in the previous section do not include the dynamics of tandem duplication.

To attempt to remove repeat strings that may have been created through a tandem duplication we define a (n, k) unitary repeat string. Recall that a repeat string is a sequence where, for a fixed k, every k-mer in the sequence occurs more than once in the genome. We now define a (n, k) **unitary repeat string** as a repeat string where, for a fixed $n \leq k$, every n-mer within the repeat string occurs only once

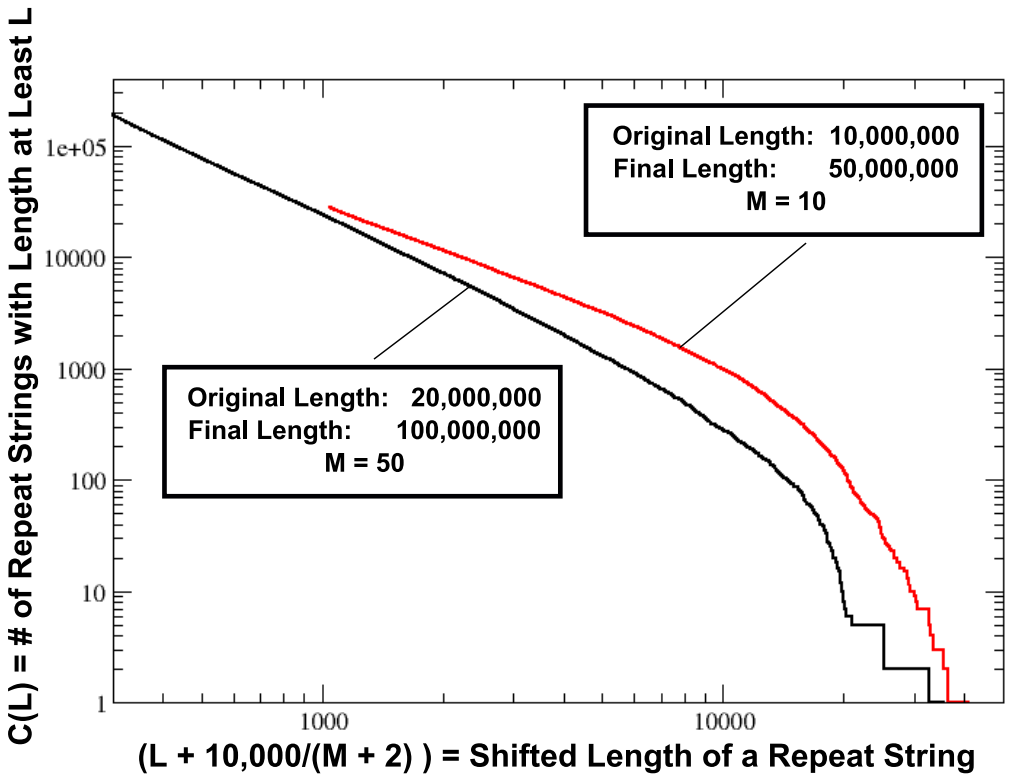

Figure 4.4: **Simulation of Growing Genome Evolutionary Model.** We plot the output for two runs of our Growing Genome Simulation. In each case the distribution of segmental mutations was determined from a uniform random variable on $[0, 2 \times 10^4]$. The model was halted when the genome reached the specified length. In both cases the distribution can be approximated a power law over a part of the range. In the case that the genome grew from length 10^6 to 5×10^6, the power law exponent determined by a least square fit was -1.5 over the range $[1040, 10^6]$. For the other simulation, where the genome grew from 2×10^6 to 10^7, the power law exponent -1.8 over $[300, 7000]$.

Table 4.2: **Unitary Repeat Strings.** We select a subset of repeat strings we call unitary repeat strings. For all genomes we considered, unitary repeat strings were the majority of repeat strings.

Genome	Length (in millions)	% of Non-Unitary Repeat Strings
human genome	3,000	2.0%
A. thaliana	119	2.2%
D. melanogaster	120	12.5%
C. elegans	100	8.0%
S. cerevisiae	12	5.8%

in the repeat string. Since the unitary repeat strings do not contain any duplicate n-mers, we are able to say with some confidence that they do not represent a tandem repeat. (The examples in this dissertation are for the distribution of $(20, 40)$ unitary repeat strings.)

Notice that for all genomes the unitary repeat strings are the majority of repeat strings (see Table 4.2). In all cases at least 87% (and as much as 98%) of the repeat strings were preserved.

After determining the unitary repeat strings for the various organisms we plot the resulting distribution of repeat strings (see Figure 4.5). As in Figure 4.1, we plot the length L of a repeat string against $C(L)$ the number of repeat strings in the genome with length at least L. We find that the distributions are again consistent with an approximate power law.

Notice from Figure 4.5, most of the longest repeat strings in the genomes we study are not $(20, 40)$ unitary repeat strings. This is consistent with our observations that the longest repeat strings for the various organisms were tandem repeats. In particular, the longest repeat string for *C. elegans* was a near-perfect tandem repeat of over 200 copies of a 72 letter sequence.

Comparing the distributions in Figure 4.5 to those in Figure 4.1, we note a number of differences. In some cases, such as *A. thaliana*, the distribution is straighter. The slope of the distribution of unitary repeat strings for each genome differs somewhat from the slope for repeat strings (compare Table 4.1 with Table 4.3). For three organisms, human, *C. elegans* and *A. thaliana*, the slope determined for the distribution of unitary repeat strings is close to the slope of the stationary distribution of the Fixed and Variable Length Models, -2. In addition, the tail of the distribution produced by the Growing Genome Model is more consistent with the decay of the distribution of *D. melanogaster*.

The stationary distribution from our Fixed and Variable Length Models provide a reasonable match to the distribution of unitary repeat strings in *C. elegans*, *A. thaliana* and the human genome. While we have not determined parameters for our Growing Genome Model that would produce the same slope as the unitary repeat string distribution for *D. melanogaster* and *S. cerevisiae* we believe that their distributions are consistent with the results provided by our model.

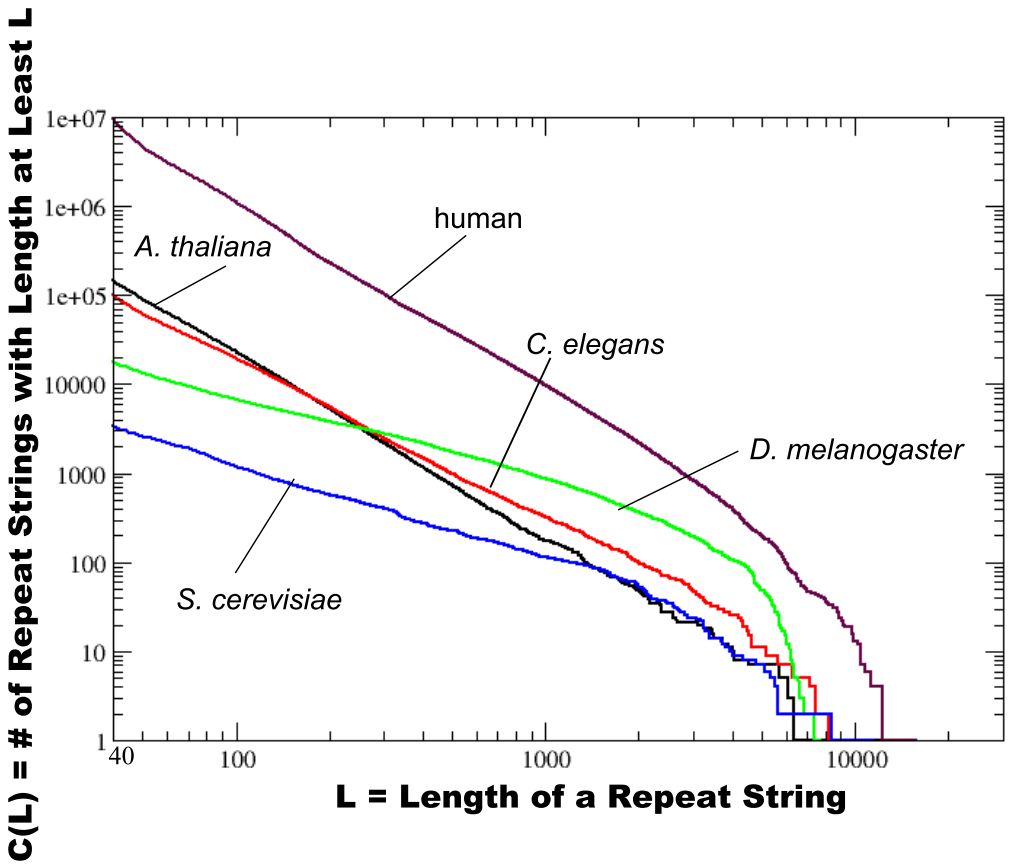

Figure 4.5: **Distribution of Unitary Repeat Strings.** The distributions of unitary repeat strings for the organisms we consider are consistent with a power law distribution and have steeper slopes that for all repeat strings. The unitary repeat strings are more consistent with the evolutionary dynamics described by our model. The distributions for the human genome, *C. elegans* and *A. thaliana* are consistent with the slope predicted by our Fixed and Variable Length Models. Since the slopes of the distributions of *D. melanogaster* and *S. cerevisiae* are shallower than -2, the Growing Genome Model provides the best characterization of their distributions.

Table 4.3: **Unitary Repeat String Distribution Slope.** The power law slope, as determined by a least square fit, for the unitary repeat strings more closely matches the slope predicted by our Fixed and Variable Length Models for three of our genomes. The distributions of the two, *D. melanogaster* and *S. cerevisiae*, have a slope that is still more shallow than predicted by our Fixed and Variable Length Models, but that could be consistent with the Growing Genome Model. Notice that the range approximated by a power law has decreased from the distributions of repeat strings.

Genome	Length (in millions)	Slope	Range of Power Law Approximation
human genome	3,000	-2.20	$40 \leq L \leq 8,000$
A. thaliana	119	-2.03	$40 \leq L \leq 5,000$
D. melanogaster	120	-1.16	$40 \leq L \leq 4,000$
C. elegans	100	-1.83	$40 \leq L \leq 5,000$
S. cerevisiae	12	-1.03	$40 \leq L \leq 1,000$

4.4 Discussion

In this section we provide more general interpretations of our results. We discuss some implications of our models for genome evolution. We close by discussing the relation of our findings to the process of genome assembly.

4.4.1 Chromosome Repeat Strings

The genomes of organisms we study are composed of several chromosomes. Additionally, there are several mechanisms related to the propagation of repeats that have a preference for creating copies in the same chromosome [1,20,75,76]. To study the repeat structure at the chromosome level, we define a **chromosome repeat string** to be a subsequence of a chromosome where, for a fixed k, every k-mer is repetitive within that chromosome. The collection of genome repeat strings implicitly contains every chromosome repeat string.

We plot in Figure 4.6 the distribution of chromosome repeat strings for human chromosomes 3, 9 along with the repeat string distribution for the entire human genome. Although all three distributions have a power law approximation, the exponent of the power law for these various distributions is different. This result suggests that there may be different evolutionary models operating on different chromosomes.

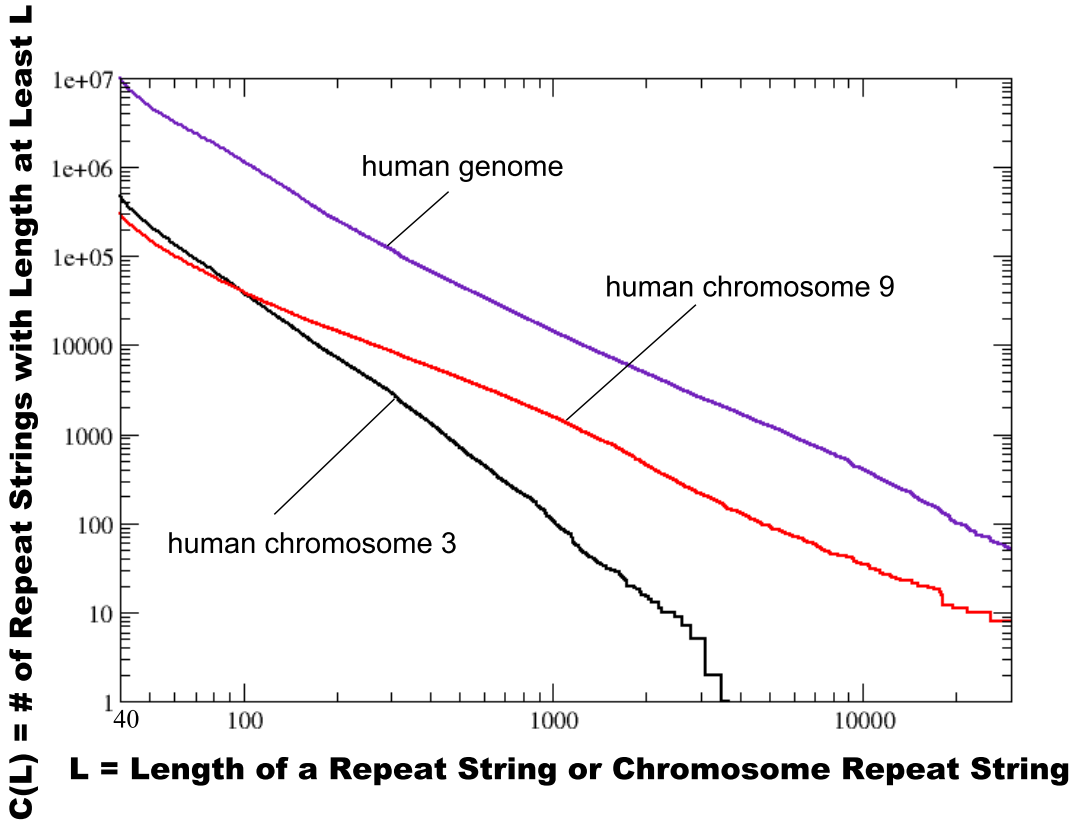

Figure 4.6: **Chromosome Repeat Strings.** We plot the distribution of repeat strings as determined for only the chromosome for human chromosomes 3 and 9 along with the entire human genome. We notice that while the distribution of repeat strings in each case can be approximated by a power law ofer a range of the distribution the slopes of the distribution are different.

4.4.2 Implications of Model

The convergence of our computer simulations and the convergence of our abstracted model strongly support the idea that simple mechanisms of repeat evolution – point mutation and sequence duplication – can explain the behavior seen in the distributions of lengths of repeat strings seen in the genomes we study. In particular, our findings suggest that the genomes of *A. thaliana*, *C. elegans* and human have had an evolutionary history consistent with our Fixed and Variable Length Models. While *D. melanogaster* and *S. cerevisiae* have a structure of repeat strings that is more consistent with a evolutionary past characterized greater duplications than deletions, such as our Growing Genome Model.

However, there were several other statistical properties of repetitive DNA discussed in this dissertation. A natural question to ask of our evolutionary models is if the other features of repetitive DNA were demonstrated.

Proximal Repeat Separations. Since there is no preference for inserting duplications near the original copy in our models, the distribution of repeat pairs is similar to that of the uniform random model discussed in Chapter 3. A more sophisticated model would include several methods of repeat creation each of which could have preferences related to the separation from the original copy.

Counts of 40-mers. The distribution of 40-mer counts in human genome and the genomes of *C. elegans*, *A. thaliana* and *D. melanogaster* was shown to roughly follow power law distribution with exponent -2.5 (see Chapter 2).

The models we have developed fail to duplicate the power law relationship

in counts of 40-mers. In the output of the Fixed and Variable Length models, the distribution 40-mer counts follows a roughly exponential distribution. One of the reasons is a lack of preference for which sequence is duplicated. When the model begins all 40-mers in the sequence have count one, thus they are all equally likely to be duplicated by the first step in the duplication process.

Our Growing Genome model produces a distribution more similar to a power law. However, the reasons for this are somewhat artificial because the final genome was produced by a smaller pool of 40-mers than the final genome from the other models. The Growing Genome model creates greater heterogeneity of counts of 40-mers as opposed to the other two models.

Similar models to our Growing Genome Model were developed in [33] and [85] others here were developed. For example, in [33] Hsieh and Lee developed a model for a bacterial genome where the initial genome length of 1000 bases grew to 1,000,000 bases (the length increased to 1,000 times the original length). Their model was shown to produce a similar distribution to that found in the word counts of short words, $k \leq 10$.

4.4.3 Implication for Genome Assembly

As stated in Chapter 1, one of our reasons for investigating repetitive DNA was the difficulties repetitive DNA poses for genome assembly. The power law relationship we find between the length L of a repeat string and the number, $N(L)$ of such strings that exist in a genome suggests a relationship between the read length

R and the number of repeat strings that are problematic for assembly. Namely if $C(L) \sim L^b$, the number of repeat strings that are longer than a read scales like R^b. If we assume that the slope of the repeat string distribution is $b = -2$, then for every factor of $\sqrt{2}$ that the read length is increased, the number of repeat strings greater than the read length will be halved. This provides a method of comparing read length to the computational complexity of the assembly problem.

Appendix A

Mathematical Formulations

A.1 Choosing a Word Length

If this dissertation we study repetitive DNA by identifying repetitive 40-mers. As mentioned, the value of 40 is not an essential part of our results and similar behavior holds for sufficiently long k-mers. That is, k sufficiently large that we do not expect to find any repeated k-mers by chance in a sequence the length of the genomes we study. To make more precise what we mean by "k sufficiently large" we compute the expected number of pairs of locations in a random genome that are the starting position of identical k-mers.

Consider a sequence W of length N from the alphabet {A,C,G,T} with the letter at each position determined independently with equal probability of being any of the four letters. Select two locations in the genome at random, we consider the k-mers that begin at those locations. The probability that the letter at both locations is the same is simply:

$$\frac{1}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16} = \frac{1}{4}$$

Since the letter at each position is chosen independently, the probability that the k-mers at both locations are identical is $\frac{1}{4}^k = 4^{-k}$. When we also count reverse complements as matches the probability is doubled, $2(4^{-k})$.

The number of pairs of locations in W is approximately $\frac{N^2}{2}$ when $N \gg k$. Because expectation is linear, the expected number, E, of pairs of locations with identical k-mers is

$$E = 4^{-k}N^2.$$

The expected number E is less than 1 when

$$k > \frac{2\log(N)}{\log(4)} = \frac{\log(N)}{\log(2)}$$

When $N = 3 \times 10^9$, as in the case of the human genome, we require $k > 31.5$ to ensure that we do not expect any repeated k-mers in the sequence W. (Of course k must be integer valued.)

We now investigate how this value changes when the occurrence of each letter is not equally likely. Assume that A occurs with probability P_A and similarly for C, G and T with the condition that $P_A = P_T$, $P_C = P_G$ and $P_A + P_C + P_G + P_T = 1$. Then the probability of the letter at two randomly chosen positions in W matching is given by, P where

$$P = {P_A}^2 + {P_C}^2 + {P_G}^2 + {P_T}^2.$$

Whatever P is, the preceeding calculation of the expected number of pairs of locations with identical k-mers becomes,

$$E = P^k N^2.$$

We have $E < 1$ when $k > 2\log(N)/\log(1/P)$. The more skewed the base composition, the greater the minimum value of k such that $E < 1$ is. For example, with $N = 3 \times 10^9$, $P_A = P_T = \frac{3}{8}$ and $P_C = P_G = \frac{1}{8}$ ($P = 5/16$) we require $k > 38.2$.

Sequence composition varies throughout a genome. Various papers have classified the varying CG content throughout the genome, such as [58] and [62]. These findings indicate that the most skewed regions are when the GC content is roughly 30%. Our above calculation with $P = 5/16$ corresponds to a GC content of 25%. Selecting $k = 40$ ensures that, even for regions of low GC content, we expect no k-mers to have an identical copy in the genome.

Genome assembly (see Chapter 1) often uses seeds of length 20, a choice that is related to our choice of 40-mers. In genome assembly the length of the seed is chosen, in part, so that the number of occurrences of a typical k-mer is 1.

Consider a random genome of length N, where $P_A = P_T$, $P_C = P_G$ and $P_A + P_C + P_G + P_T = 1$. Pick an arbitrary k-mer and let $E'(k)$ denote the expected number of occurrences of that k-mer in a random genome of length N where $N \gg k$.

From our analysis above,

$$E'(k) = P^k N$$

and

$$E(k) = P^k N^2$$

.

Now, we select a value for k' so that $E'(k') = 1$ and k so that $E(k) = 1$, (allowing the fiction that k and k' need not be integers), yields

$$2k' = k$$

Thus, when $k' = 20$, $k = 40$. This is an additional reason for selecting $k = 40$ to illustrate our results.

A.2 Showing $H(f)$ is a Contraction

In Section 4.2.2 we derived the stationary distribution for our Fixed Length Model. Here we provide more details of the calculation of the stationary distribution. In Section 4.2.2 we show that, under appropriate assumptions and approximations, the expected action of one mutation on the distribution of repeat strings of length $x \in [0, S]$, which we denote $f(x)$ is given by

$$(Lf)(x) = \left(1 - \frac{1}{W}\left(x + \frac{S}{(M+2)}\right)\right) f(x) + \int_x^S \frac{2}{W} f(y)dy + \frac{1}{M+2}\delta(x-S).$$

We denote by H the terms in L that are linear in f:

$$(Hf)(x) = \left(1 - \frac{1}{W}\left(x + \frac{S}{(M+2)}\right)\right) f(x) + \int_x^S \frac{2}{W} f(y)dy.$$

We now show by calculation that H is a contraction in the norm

$$\|f\|_* := \int_0^S |xf(x)|dx$$

.

By definition we have,

$$\|(H(f))\|_* = \int_0^S |x(Hf)(x)|dx$$

$$= \int_0^S \left| x\left(1 - \frac{1}{W}\left(x + \frac{S}{(M+2)}\right)\right) f(x) + \int_x^S \frac{2}{W} f(y)dy \right| dx$$

Because $f(x) \geq 0$, $x \geq 0$ and $W > \left(S + \frac{S}{M+2}\right)$ for reasonable values of M, S, and W, we can remove the absolute value signs.

$$= \int_0^S x\left(1 - \frac{1}{W}\left(x + \frac{S}{(M+2)}\right)\right) f(x)dx + \int_0^S x \int_x^S \frac{2}{W} f(y)dydx$$

We interchange the bounds of integration using Fubini's Theorem:

$$= \int_0^S xf(x)dx - \int_0^S x^2 \frac{1}{W} f(x)dx - \int_0^S \frac{xS}{W(M+2)} f(x)dx + \int_0^S \int_0^y x \frac{2}{W} f(y)dxdy$$

$$= \|f\|_* - \int_0^S \frac{x^2}{W} f(x)dx - \frac{S}{W(M+2)} \|f\|_* + \int_0^S \frac{y^2}{W} f(y)dy$$

$$= \left(1 - \frac{S}{W(M+2)}\right) \|f\|_*$$

Thus, $\|(H(f))\|_* = \left(1 - \frac{S}{W(M+2)}\right) \|f\|_*$. Because $0 < \frac{S}{W(M+2)} < 1$ the operator H is a contraction.

A.3 Deriving the Stationary Distribution for the General Model

In Section 4.2.2 we state that the stationary distribution from the approximation to the Fixed Length Model is a Pareto Distribution $g(x)$:

$$g(x) = \alpha \left(x + \frac{S}{M+2}\right)^{-3} + \beta\delta(x - S).$$

We now explicitly determine the values of constants A and B. We begin by studying $L(g(x)) = g(x)$:

$$L(g(x)) = H(g(x)) + \frac{1}{M+2}\delta(x - S) \tag{A.1}$$

Because H is a linear operator we can write $H(g(x))$ in terms of each component of $g(x)$:

$$H(g(x)) = \alpha H\left(\left(x + \frac{S}{M+2}\right)^{-3}\right) + \beta H(\delta(x - S)) \tag{A.2}$$

We now consider each component separately:

$$H\left(\left(x+\frac{S}{M+2}\right)^{-3}\right)=\left(1-\frac{1}{W}\left(x+\frac{S}{(M+2)}\right)\right)\left(x+\frac{S}{M+2}\right)^{-3}$$
$$+\int_x^S \frac{2}{W}\left(y+\frac{S}{M+2}\right)^{-3}dy$$

$$=\left(x+\frac{S}{M+2}\right)^{-3}-\frac{1}{W}\left(x+\frac{S}{M+2}\right)^{-2}-\frac{1}{W}\left(y+\frac{S}{M+2}\right)\Bigg|_x^S$$

$$=\left(x+\frac{S}{M+2}\right)^{-3}-\frac{1}{W}\left(x+\frac{S}{M+2}\right)^{-2}+\frac{1}{W}\left(x+\frac{S}{M+2}\right)^{-2}-\frac{1}{W}\left(S+\frac{S}{M+2}\right)^{-2}$$

Thus we have,

$$\alpha H\left(\left(x+\frac{S}{M+2}\right)^{-3}\right)=\alpha\left(x+\frac{S}{M+2}\right)^{-3}-\frac{\alpha}{W}\left(S+\frac{S}{M+2}\right)^{-2} \qquad \text{(A.3)}$$

Similarly for the term involving δ in equation (A.2),

$$H(\delta(x-S))=\left(1-\frac{1}{W}\left(x+\frac{S}{(M+2)}\right)\right)\delta(x-S)+\int_x^S\frac{2}{W}\delta(x-S)dy$$

$$=\left(1-\frac{1}{W}\left(x+\frac{S}{(M+2)}\right)\right)\delta(x-S)+\frac{2}{W}$$

So we can express,

$$\beta H(\delta(x-S))=\beta\left(1-\frac{1}{W}\left(x+\frac{S}{(M+2)}\right)\right)\delta(x-S)+\frac{2\beta}{W} \qquad \text{(A.4)}$$

We combine equations (A.1)-(A.4) to solve for α and β so that

$$\alpha\left(x+\frac{S}{M+2}\right)^{-3}+\beta\delta(x-S)=\alpha\left(x+\frac{S}{M+2}\right)^{-3}-\frac{\alpha}{W}\left(S+\frac{S}{M+2}\right)^{-2}$$

$$+\beta\left(1-\frac{1}{W}\left(x+\frac{S}{(M+2)}\right)\right)\delta(x-S)+\frac{2\beta}{W}+\frac{1}{M+2}\delta(x-S)$$

When $x < S$, $\delta(x-S)=0$ so the above equation becomes:

$$\alpha\left(x+\frac{S}{M+2}\right)^{-3}=\alpha\left(x+\frac{S}{M+2}\right)^{-3}-\frac{\alpha}{W}\left(S+\frac{S}{M+2}\right)^{-2}+\frac{2\beta}{W}$$

The terms involving x cancel and we can solve for α in terms of β.

$$\alpha=2\beta\left(S+\frac{S}{M+2}\right)^{2} \tag{A.5}$$

When $x = S$ we can simplify our expression to involve only β:

$$\beta\delta(x-S)=\beta\left(1-\frac{1}{W}\left(x+\frac{S}{(M+2)}\right)\right)\delta(x-S)+\frac{1}{M+2}\delta(x-S)$$

Equating the coefficients on both sides yields:

$$\beta=\frac{W}{(M+2)\left(S+\frac{S}{M+2}\right)} \tag{A.6}$$

.

We can rewrite equation (A.5) for α as:

$$\alpha=\frac{2W}{M+2}\left(S+\frac{S}{M+2}\right) \tag{A.7}$$

[1] Achaz G, Netter P and Coissac E. 2001. Study of intrachromosomal duplications among the eukaryote genomes. *Mol. Biol. Evol.* **18(12)**:2280-2288.

[2] Achaz G, Rocha E, Netter P and Cossiac E. 2002. Origina and fate of repeats in bacteria. *Nucleic Acids Research.* **30(13):**2987-2994.

[3] Adams M et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science.* **287**:1185-1195.

[4] Anh V, Lau K, Yu Z. 2001. Multifractal characterization of complete genomes. *J. Phys A: Math. Gen.* **34**:7127-7139.

[5] Aminetzach Y, Macpherson M and Petrov D. 2005. Pesticide resistance via transposition-mediated adaptive gene truncation in *Drosophila*. *Science.* **309**:764-767.

[6] Bao, Z., and Eddy, S. Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes. Genome Research. 12:1269-1276. (2002).

[7] Bartolome C, Maside X and Charlesworth B. 2002. On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. *Mol. Biol. Evol.* **19(6)**:926-937.

[8] Batzoglou S, et al. 2002. ARACHNE: a whole-genome shotgun assembler. *Genome Research.* **12(1)**:177-189.

[9] Batzoglou S, et al. 1999. Sequencing a genome by walking with clone-end sequences: A mathematical analysis. *Genome Research.* **9**:1163-1174.

[10] Bell G and Jurka J. 1997. The length distribution of perfect dimmer repetitive DNA is consistent with its evolution by an unbiased single step mutation process. *J. Mol. Evol.* **44**:414-421.

[11] Benson et al. 2006. GenBank. *Nucleic Acids Res.* **33**:D34-D38.

[12] Berkeley Drosophila Genome Project. http://www.flybase.org.

[13] Blattner F, et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science.* **277**:1453-1462.

[14] Bonhoeffer S, Herz A, Boerlijst M, Nee S, Nowak M, and May R. 1996. No Signs of Hidden Language in Noncoding DNA. *Phys. Rev. Lett.* **76**:1977.

[15] Borstnik B and Pumpernik D. 2004. Mutational dynamics of short tandem repeats in human genome. *Eurpohysics Letters*. **65(2)**:290-296.

[16] Borstnih B and Pumpernik D. 2005. Evidence on DNA slippage step-length distribution. *Phy Rev E*. **71**:031913.

[17] Brookfield J. 2005. The ecology of the genome mobile DNA elements and their hosts. *Nature Reviews Genetics*. **6**:128-136.

[18] Brookfield J and Badge R. 1997. Populations genetics models of transposable elements. *Genetica*. **100**:281-294.

[19] Celniker S et al. 2002. Finishing a whole-genome shotgun: Release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biology*. **3**:research0079.1-0079.14.

[20] Charlesworth B, Langley C and Wolfgang S. 1986. The evolution of restricted recombination and the accumulation of repeated DNA sequences. *Genetics*. **112**:947-962.

[21] Charlesworth, B., Sniegowski, P. and Wolfgang S., The Evolutionary Dynamics of Repetitive DNA in Eukaryotes. Nature 371:215-220. (1994).

[22] Chaudhuri P and Das S. 2001. Statistical analysis of large DNA sequences using distribution of DNA words. *Current Science*. **80(9):**1161-1166.

[23] Djian P. 1998. Evolution of simple repeats in DNA and their relation to human disease. *Cell*. **94**:155-160.

[24] Edler J and Turner B. 1995. Concerted evolution of repetitive DNA sequences in eukaryotes. *Quarterly Review of Biology*. **70(3)**:297-320.

[25] Eichler E. 2001. Segmental duplications: whats missing, misassigned and misassembled and should we care? *Genome Research*. **11**:653-656.

[26] Eickbush T and Furano A. 2002. Fruit flies and humans respond differently to retrotransposons. *Curr Opin in Genetics and Development*. **12**:669-674.

[27] Gregory T. 2005. *The Evolution of the Genome*. Pp3 151. Elsevier Academic Press.

[28] Hao B, Lee H, Zhang S. 2000. Fractals related to long DNA sequences and complete genomes. *Chaos, Solitons and Fractals*. **11**:825-836.

[29] Havlak P, et al. 2001. The Atlas whole-genome assembler. *Currents Trends in Computational Molecular Biology, Montreal, Canada, 2001.*

[30] Hedges D, Cordaux R, Xing J, Witherspoon D, Rogers A, Jorde L and Batzer M. 2005. Modeling the amplification dynamics of the human *Alu* retrotransposons. *PLoS Computational Biology.* **1(4)**:333-340.

[31] Holste D, Grosse I, Beirer S, Schieg P and Herzel H. 2003. Repeats and correlations in human DNA sequences. *Phy Rev E.* **67**:061913.

[32] Hsieh L et al. 2003. Minimal model for genome evolution and growth. *PRL.* **90**(1):018101(4).

[33] Hsieh L and Lee H. 2002. Model for the growth of bacterial genomes. *Modern Physics Letters B.* **16(22)**:821-827.

[34] Hurst G and Werren J. 2001. The role of selfish genetic elements in eukaryotic evolution. *Nature Reviews Genetics.* **2**:597-606.

[35] Jaffe D, et al. 2003. Whole-genome sequence assembly for mammalian genomes: arachne 2. *Genome Research.* **13(1)**:91-96.

[36] Jurka J. 194. Approaches to identification and analysis of interspersed repetitive DNA sequences. In *Automated DNA Sequencing and Analysis* (Adams et al, eds.) p. 294-298, Academic Press.

[37] Kaminker J et al. 2002. The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biology.* **3(12)**:research0084.1-0084.20.

[38] Kazazian H. 2004. Mobile elements: drivers of genome evolution. *Science.* **303**:1626-1632.

[39] Karev G, Wolf Y, Rzhetzky A, Berezovskaya, F and Koonon E. 2002. Birth and death of protein domains: a simple model of evolution explains power law behavior. *BMC Evolutionary Biology.* **2**:18.

[40] Kellis M, Birren B and Lander E. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae. Nature.* **428**:617-624.

[41] Kent W, Baertsch R, Hinrichs A, Miller W and Haussler D. 2003. Evolutions cauldron: duplication, deletion and rearrangement in the mouse and human genomes. *PNAS*. **100(20)**:11484-11489.

[42] Kidwell M and Lisch D. 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution*. **55(1)**:1-24.

[43] Koonin E, Wolf Y and Karev G. 2002. The structure of the protein universe and genome evolution. *Nature*. **420**:218-223.

[44] Kruglyak S, Durrett R, Schut M and Aquadro C. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *PNAS*. **95**:10774-10778.

[45] Kurtz S, Choudhuri J, Ohlebusch E, Schleiermacher C, Stoye J and Gierich R. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nuc. Acids Res.* **29(22)**:4633-4642.

[46] Lander E et al. 2001. Initial sequencing and analysis of the human genome. *Nature*. **409**:860-921.

[47] Liu G, NISC Comparative Sequencing Program, Zhao S, Bailey J, Sahinalp C, Alkan C, Tuzun E, Green E and Eichler E. 2003. Analysis of primate genomic variation reveals a repeat-drive expansion of the human genome. *Genome Research*. **13**:358-368.

[48] Luscombe N et al. 2002. The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties.*Genome Biology* **3(8)**:research0040.1-0040.7.

[49] Majewski J and Ott J. 2000. GT repeats are associated with recombination on human chromosome 22. *Genome Research*. **10**:1108-1114.

[50] Makalowski W. 2000. Genomic scrap yard: how genomes utilize all that junk. *Gene*. **259**:61-67.

[51] Mantegna R et al. 1994. Linguistic features of noncoding DNA sequences. *Physical Review Letters*. **73(23)**:3169-3172.

[52] Martindale C and Konopka A. 1996. Oligonucleotide frequencies in DNA follow a Yule distribution. *Computers and Chemistry*. **20**(1):35-38.

[53] McNeil J, Smith K, Hall L and Lawrence J. 2006. Word frequency analysis reveals enrichment of dinucleotide repeats on the human X chromosome and $[GATA]_n$ in the X escape region. *Genome Research*. **16**:477-484.

[54] Mullikin J and Ning Z. 2003. The phusion assembler. *Genome Research*. **13(1)**:81-90.

[55] Myers G. 1999. Whole-genome DNA sequencing. **IEEE: Computational Biology**. 33-42.

[56] Myers E et al. A whole-genome assembly of Drosophila. 2000. *Science*. **287**:2196-2204.

[57] NCBI-BLAST ftp://ftp.ncbi.nlm.nih.gov/blast/executables/

[58] Nekrutenko A and Li W. 2000. Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome Research*. **10**:1986-1995.

[59] Newman M. 2005. Power laws, Pareto distributions and Zipfs law. *Contemporary Physics*. **46(5)**:323-351.

[60] Nikolaou C and Almirantia Y. 2005. "Word" preference in the genomic text and genome evolution: different modes of n-tuplet usage in coding and noncoding sequences. *J. Mole. Evol.* **61**:23-35.

[61] Ohta T and Kimura M. 1981. Some calculations on the amount of selfish DNA. *PNAS*. **78(2)**:1129-1132.

[62] Oliver J, Bernaola-Galvan P, Carpena P and Roman-Roldan R. 2001. Isochore chromosome maps of the eukaryotic genomes. *Gene*. **276**:47-56.

[63] Pevzner P, Thang Haixu and Waterman M. 2001. An Eulerian path approach to DNA fragment assembly. *PNAS*. **98(17):**9748=9753.

[64] Pevzner P, Tang H and Tesler G. 2004. De novo repeat classification and fragment assembly. *Genome Research*. **14**:1786-1796.

[65] Pop M, Salzberg S and Shumway M. 2002. Genome sequence assembly: algorithms and issues. *IEEE Computer*. **35(7)**:47-54.

[66] Pop M. 2004. Shotgun Sequence Assembly. *Advances in Computers*. **60**:193-245.

[67] Price A, Eskin E, and Pevzner P. 2004. Whole-genome analysis of ALU repeat elements reveals complex evolutionary history. *Genome Research*. **14**:2245-2253.

[68] Qian J, Luscombe N and Gerstein M. 2001. Protein family and fold occurrence in genomes: power-law behavior and evolutionary model. *J. Mol. Biol.* **313**:673-681.

[69] Rigoutsos I, Huynh T, Miranda K, Tsirigos A, McHardy A and Platt D. 2006. Short blocks from the noncoding parts of the human genome have instances within nearly all known genes and relate to biological processes. *PNAS*. **103(17):**6605-6610.

[70] Le Rouzic A and Deceliere G. 2005. Models of the population genetics of transposable elements. *Genet. Res. Camb.* **85**:171-181.

[71] Rudd M, Wray G and Willard H. 2006. The evolutionary dynamics of α-satellite. *Genome Research*. **16**:88-96.

[72] Salzberg SL and Yorke JA. 2005. Beware of mis-assembled genomes. *Bioinformatics*. **21**(24):4320-4321.

[73] Schlotterer C. 2000. Evolutionary dynamics of microsatellite DNA. *Chromosoma*. **109**:365-371.

[74] Shapiro JA. (2002) Genome organization and reorganization in evolution: formatting for computation and function. *Ann NY Acad Sci*. **981**:111-134.

[75] Slamovits, C. and Rossi M. Satellite DNA: Agent of Chromosomal Evolution in Mammals. J. Neotrop. Mammal 9(2):297-308. (2002).

[76] Smith, G. Evolution of Repeated DNA Sequences by Unequal Crossingover. Science. 191:528-535. (1976).

[77] Stein L et al. 2003. The genome sequence of *C. briggsae*: a platform for comparative genomics. *PLoS Biol.* **1(2)**:166-192.

[78] Thomas E, Srebro N, Sebat J, Navlin N, Healey J, Mishra B and Wigler M. 2004. Distribution of short paired duplications in mammalian genomes. *PNAS*. **101(28)**:10349-10354.

[79] The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*. **408**:796-815.

[80] The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*. **282**:2012-2018.

[81] The chromosome 21 mapping and sequencing consortium. 2000. The DNA sequence of human chromosome 21. *Nature*. **405**:311-319.

[82] Venter J et al. 2001. The sequence of the human genome. *Nature*. **291**(5507):1304-1351.

[83] Volfovsky, N., Hass, B., and Salzberg, S. A Clustering Method for Repeat Analysis in DNA Sequences. Genome Biology 2(8):1-11(2001).

[84] Yu Z, Anh V, Lau K. 2003. Iterated function system and multifractal analysis of biological sequences. *Int. J. of Modern Physics B*. **17**:4367-4375.

[85] Zhou Y and Mishra B. 2004. Models of genome evolution. *Modeling in Molecular Biology*, G. Ciobanu, G. Rozenberg (Eds.), Natural Computing Series, pages 287-304, Lecture Notes in Computer Science.