TEL-AVIV UNIVERSITY
RAYMOND AND BEVERLY SACKLER
FACULTY OF EXACT SCIENCES

Common, Rare, and Surprising Words: Some Unexpected Results on the Distribution of Genomic DNA k-mers

Thesis submitted in partial fulfillment of the requirements
for the degree of M.Sc.
in
Tel -Aviv University
The Department of Computer Science

by

Yaron Levy

The research work for this thesis
has been carried out at Tel-Aviv University
under the supervision of Prof. Benny Chor and Prof. David Horn

January 2008

Abstract

The probability distribution of DNA k-mers in whole genome sequences provides an interesting perspective of the complexity of these systems. Whereas previous research concentrated on *missing* k-mers, we study the overall k-mer distribution of more than 100 species from archaea, bacteria, and eukarya (including mammals). We focus on low order Markov models, which capture short range correlations between nucleotides in the DNA sequence. In particular, they enable us to decide if rare, missing, and common k-mers are *surprising* or not. We show that various local and global properties of DNA k-mers can be modeled fairly well by low order chains.

While exploring these empirical k-mer distributions, we discovered that a few species, including all mammals, have multi-modal histograms, while most species exhibit unimodal distributions. From an evolutionary perspective, these multi-modal distributions are exactly the tetrapods. These distributions are characterized by specific values of `C+G` contents and `CpG` dinucleotide suppression, but not by any one of these factors alone. Again, we provide an explanation for this phenomenon, using low order Markov models.

Finally, we have investigated the k-mer distributions of specific functional elements of the human genome, like exons, introns, and promoters. We found, for example, that the k-mer distribution for human exons is unimodal, while for introns and long promoter regions it is multi-modal.

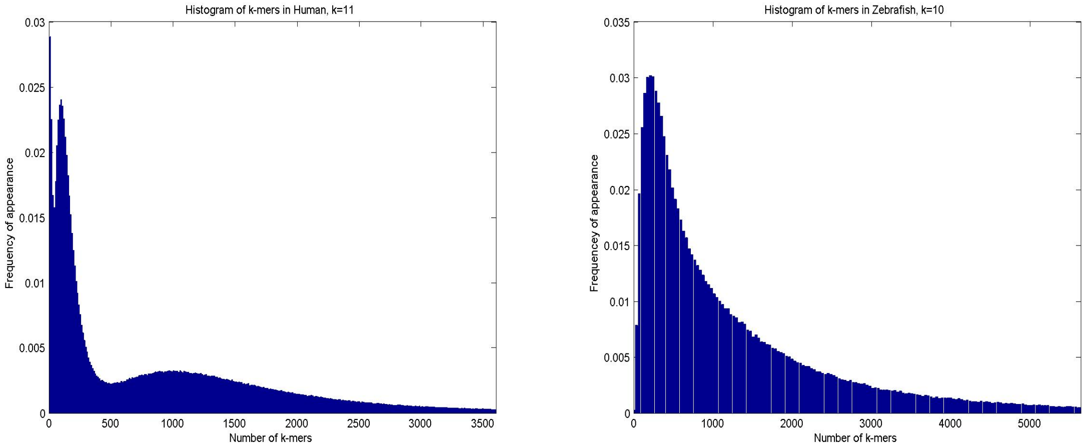

Figure 0.1: 11-mer distribution. Multi-modal for Human (left), and unimodal for Zebrafish (right).

Acknowledgments

I would like to thank my supervisors, Prof. Benny Chor and Prof. David Horn, for their guidance and support throughout my thesis work. I have learned tremendously from their brilliance, high standards and passion for knowledge and science. I feel privileged to have had the opportunity to work closely and learn so much from them.

I would like to thank my fellow researchers Dr. Nick Goldman and Tim Massingham of the European Bioinformatics Institute in Hinxton, Cambridge, UK, who contributed in the research regarding the mammalian portion of this thesis.

I would like to thank the members of the lab: Vered, Zach, Roy, Assaf, Liat, Yasmine, Uri and Yair for their help, interesting conversations and insights, and friendship. Lots of thanks and appreciation to Michal Finkelman-Reuven for her constant care and help.

Special thanks to my parents Uri and Ziva, and my brothers Nir and Dan for their constant support and involvement.

I am eternally grateful to my wife Gali for her patience with me through one of the most challenging, somewhat frustrating, and pivotal times in our lives.

And of course to my daughter Yarden, who unknowingly provided the last straw to finally get it done.

List of Figures

List of Tables

Contents

1 Introduction

This thesis deals with the distribution of short DNA words (k-mers) in about one hundred species that are representatives of all genomes that were sequenced to date. We checked the histogram which describes the number of words as a function of the number of appearances (namely for any $m \geq 0$, how many k-mers appear *exactly* m times in the given genome). We found an interesting phenomenon which correlates the shape of the histogram with the phylogenetic placement of the species. We also investigated the connection between the shape of a histogram and both the CpG-suppression phenomenon and the GC-content in the relevant genome.

1.1 Short DNA words (k-mers)

Short DNA words (k-mers) are a fundamental entity in studying DNA, but for values of k exceeding 2 or 3 they were not studied intensively. A number of statistical/theoretical models for the distribution of such k-mers have been proposed. For example, Robin and Schbath compare several approximate k-mer distributions, and also analyze the empirical k-mer distributions of the phage Lambda (49K bp long genome) [19]. Reinert *et al.* discussed various plausible k-mer distributions [18]. They show that the distribution for the number of occurrences of a particular k-mer has two distinct large sample regimes: a normal distribution for abundant k-mers, and a Poisson or compound Poisson distribution for extremely rare k-mers. With the sequencing of more and more complete genomes, it becomes possible to move from theoretical to empirical studies, examine the properties of these DNA words, and how their distributions vary with different species or genome elements.

1.2 Related Work

Maybe the simplest question to answer is that of *missing* DNA k-mers. Earlier works have studied non-existent short amino acid (AA) k-mers [16, 23], and have attributed them mainly to chemical constrains (like hydrophobic and hydrophilic AAs). DNA does not have the complicated three dimensional structure and chemical constraints of proteins. So intuitively, if k is not too large, compared to the genome or chromosome length, we expect that all k-mers will be present. This expectation turns out to be incorrect. Fofanov *et al.* studied correlations between present and absent short DNA k-mers in over $1,500$ species, and observed short k-mers that are missing [7]. A systematic study of missing k-mers, termed nullomers, was recently carried out by Hampikian and Andersen [8]. They reported the complete lists of missing k-mers ($8 \leq k \leq 13$) in 12 species, including human. For example, the human genome has 80 missing 11-mers, $39,852$ missing 12-mers, and $2,232,448$ missing 13-mers. Furthermore, they described several possible uses of these nullomers, concluding by "These absent sequences define the maximum set of potentially lethal oligomers,..., and identify potential targets for therapeutic intervention and suicide markers".

1.3 Our Approach

In this work, we initiate a systematic investigation and modeling not only of the *missing* k-mers, but of the entire empirical distributions of the number of k-mer ($4 \leq k \leq 13$) occurrences in the whole genomes of more than 100 species, including archaea, bacteria, and eukarya. Our analysis reveals that a number of higher eukaryotes, including all mammals, exhibit a rather unusual distribution, with a number of local maxima, and unusually high numbers of rare k-mers (those not appearing at all, or appearing very few times) and abundant k-mers (those appearing very many times). We show that while these observed distributions are poorly matched by an independently, identical distribution on the nucleotides, they are matched reasonably well by low order Markov models. Thus it is quite conceivable that unlike short missing peptides, missing DNA k-mers, as well as rare and abundant ones, are not an outcome of structural or functional constraints, but instead are explained by these simple probabilistic models. This has consequences for our expectations of the significance and even potential use of nullomers and short, infrequent words.

Our starting points were the k-mer distributions of the DNA sequences in the whole human genome, as well as all individual chromosomes. For each, we examined both the single strands and the double strands (namely considering all k-mers on one strand and their reverse complements on the other strand). Quite unexpectedly, we discovered that the k-mers distributions in all these sequences are *multi modal*. We went on to explore the k-mer distributions of all mammals with sequenced genomes. These describe a wide evolutionary range within mammalia, including chimp, mouse, dog, cow, opossum (non-placental mammal), and platypus (a monotreme, egg laying mammal). The same phenomena we observed for human still occurred here. This has motivated us to explore about 90 additional species, including archea, bacteria, and many (but not all) other sequenced eukarya. Finally, we examined different functional regions of the human genome – introns, exons, 3'UTRs, 5'UTRs and promoter regions of different lengths (600, 1000, and 5000 bases upstream).

The k-mer distributions for different organisms no longer followed the mammalian pattern. All archea and bacteria exhibited *unimodal* k-mer histograms. For eukarya, the findings are more involved. Non-mammalian species exhibiting multi-modal, “mammalian like” distributions include chicken, lizard, and frog. All these are representatives of the tetrapod clade. The next branch up the tree of life is the bony fish. Its 5 sequenced representatives are the zebrafish, fugu, tetraodon, stickleback, and Japanese medaka, exhibiting unimodal k-mer distributions.

We continued to investigate the extent to which the low order Markov models capture properties of k-mer DNA distributions, such as the whole empirical distribution, the “heavy tail” phenomenon, and differences between the number of occurrences of k-mers and their reverse complements on a single strand. In all cases, we found that the low order Markov models provide a fairly good fit. This enabled us to quantitatively define the notion of *surprising* k-mers.

Due to the non-homogenoeous nature of genomes, and the existence of long stretches with distinct compositional biases, the fit is far from being perfect. In light of this, it is somewhat surprising that these low order models are capable of predicting whether the empirical k-mer distribution is unimodal or not, and furthermore to predict the actual shape of the empirical histogram and other properties.

1.4 Outline

The outline of this thesis is as follows. In chapter 2 we provide information about the datasets we used in our work, the analysis applied to the data in order to create the histograms, and a method of simulating the results using a low order Markov chain.

Chapter 3 describes in detail the results regarding the different species analyzed, including mammalian and non-mammalian genomes, and various human genomic regions. Following are the Markov model simulation results, and the results defining “surprising” k-mers, with respect to these models. Finally, we present some interesting observations that we encountered regarding specific results.

Finally, chapter 4 contains concluding remarks as well as a discussion regarding some aspects of this work, with a wink to the future.

2 Methods

The goal of our research was to analyze the different k-mer distributions in a large variety of species, and to try to deduce two major conclusions - a simple statistical model which can correlate to the empirical data, and an attempt to gain biological insight regarding the results of our k-mer analysis.

2.1 Whole Genome Analysis

We computed the empirical k-mer distributions of genomic DNA for more than 100 genomes, taken from various online sources (Ensembl, NCBI, TIGR). Most sequences are complete, but some are only assembled on scaffolds, with no chromosomal assignment. See the last three tables in the appendix for the list of species we considered, their genome length (measured as number of k-mers on a single strand), and additional properties of interest. For each genome, we separately computed the k-mer distribution for the range $4 \leq k \leq 13$, taking k-mers from both strands into account. In addition to the "raw" sequence, we have also applied a repeat mask filter [20], in order to avoid low complexity repetitive elements and enable the identification of "interesting", over-represented k-mer sequences. The value of k in the histograms that we show was taken as the nearest integer to $c \cdot \log_4 \ell$, where ℓ is the length of the genome, and $c = 0.7$ is a constant, which was chosen based on $k = 11$ as a model-graph for human. For any species, the histograms for values of k that are close to the one that is shown, exhibit very similar characteristics.

For increasing values of k, the number of k-mers that do not appear at all increases, and soon it becomes dominant, namely the largest among all k-mer counts. For moderately large values of k, 4^k is much larger than the length of the genome, but the "nullomer dominance" phenomenon occurs well before $4^k > \ell$. For species with relatively small genomes, like bacteria and archaea (0.5-10 million nucleotides), this phenomenon already occurs at values of k that are typically 8 or smaller, and therefore their corresponding graphs may look skewed "to the left" for these values of k.

Each genome in our dataset was analyzed using a simple program that tabulates the k-mers in a large array, indexed by k-mers, and outputs useful summary information like the total number scanned, those missing, and transition frequencies for different orders of Markov model. On standard machines using 1GB RAM, the entire human genome can be analyzed for k-mers with k up to 14 in no more than 15 minutes, so the analysis is not computationally bound and is limited only by the availability of genomic data. The time required to scan and produce statistics about a genome is proportional to its length, although finding which k-mers are missing requires time proportional to the total number of possible k-mers (i.e. 4^k); storage increases as the number of possible k-mers, although alternative methods such as storing only those k-mers that are present can considerably reduce the storage requirements for large k.

2.2 Low Order Markov Model Simulations

Perhaps the simplest probabilistic models to describe strings like genomes and chromosomes are low order Markov models [15], such as those commonly used for a genomic "background" comparison when detecting regulatory elements (e.g. [12, 14]). A zero order Markov model simply describes the frequencies of each nucleotide (when we consider both strands, the frequencies of A, T are equal, and so are those of C, G). A first order Markov model describes the frequencies of individual nucleotides, given the nucleotide immediately preceding it, a second order Markov model describes the frequencies of individual nucleotides given the *pair* of nucleotides immediately preceding it, etc. These transition probabilities can easily be estimated from counts of transitions observed [5], and actually provide maximum likelihood estimates. The number of model parameters for a Markov model of order m is $4^{m+1} - 1$ (there are 4^m m-tuples, each determining one stationary probability and four transition probabilities: the reduced number of parameters follows from the stationary probabilities, and the transition probabilities dependent on each m-tuple, each summing up to 1).

For Markov models of order m lower than the length of the k-mer of interest, the sequence of k-mers emitted is itself a *first-order* Markov chain – the relevant history of the chain is entirely encoded in the current k-mer. From each k-mer there are 4 possible transitions, depending on which nucleotide is emitted next. The transition probabilities are given by the embedded Markov chain. If this k-mer chain has a stationary distribution, which it does, provided that the embedded Markov chain satisfies some weak properties, the frequency with which a particular k-mer occurs converges to its stationary frequency (by the Ergodic theorem, e.g. [15]). This gives us a direct method of calculating how many times a particular k-mer is expected to occur, based on a lower order Markov model, M. For example, a first order model M is specified by the stationary frequencies $\pi_M(a)$ and the transition probabilities $Pr(a \,|\, b)$ for all $a, b \in \{A, C, G, T\}$. By repeatedly applying these formulae, we can find, for any k-mer $a_1 a_2 \ldots a_k \in \{A, C, G, T\}^k$, the value $\pi_K(a_1 a_2 \ldots a_k)$, where π_{K} describes the stationary frequencies of the k-mer embedded model chains.

Given $\pi_K(a_1 a_2 \ldots a_k)$, the expected number of occurrences of this k-mer in a genome of length ℓ is simply $\ell \cdot \pi_K(a_1 a_2 \ldots a_k)$. We can thus compute all these probabilities (dynamic programming will speed up the computation), and consequently all expected frequencies directly (without resorting, *e.g.* to simulations). Comparing the empirical distribution to the one predicted from the model, we can determine how well the model describes the data.

Being able to fit a model leaves the question of how well it describes the data observed or, to reverse the question, what order of model is adequate? Reinert et al. [18] consider a chi-square test of model adequacy, testing whether the model explains transitions that are an order higher, although they advise caution about using this test for high-order chains. Here we calculate the log-likelihood of each model, and compare them using Akaike's Information Criteria (AIC) [1]:

$$\mathrm{AIC} = 2K - 2 \ln L(\mathcal{S} \,|\, \hat{\theta})$$

where K is the number of model parameters, $\hat{\theta}$ is the maximum likelihood estimate of the model parameters (stationary and transition probabilities), $\mathcal{S}$ is the sequence data (genome, chromosome, etc.), and $L(\mathcal{S} \,|\, \hat{\theta})$ is the likelihood of the model.

It should be mentioned that even fairly high-order Markov models will not perfectly model the genome, because they do not take into account compositional heterogeneity. This may be especially important when looking at missing k-mers, since a small patch of biased composition could significantly increase the probability that a few instances of the k-mer are observed. This is a fundamental limitation of any Markov model in this context.

3 Results

In this section we describe the empirical k-mer DNA distributions, examine whether they are unimodal or multi-modal, and how multi-modality is related to properties like C+G content and CpG suppression. We describe the fit of low order Markov models to these k-mer distributions, and analyse *surprising* k-mers in the human genome. In addition to whole genomes and chromosomes, we have also considered various functional regions in the human genome.

3.1 Mammalian Genomes: Empirical Distributions

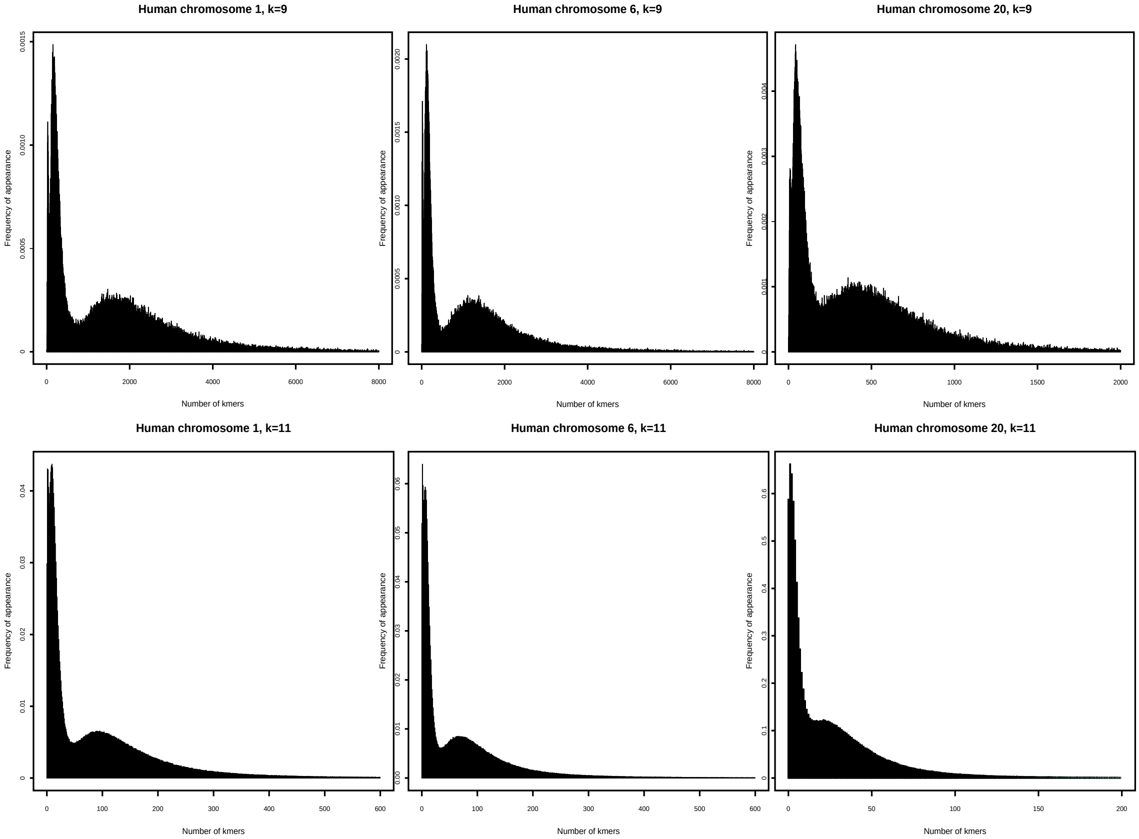

Figure 3.1: Multi-modal distributions for human chromosomes.
The figures describe the frequency of appearance (y-axis) of the k-mers which appear the same number of times (x-axis), for different DNA sequences and values of k. Human chromosomes (left to right) 1, 6, 20, both 9-mers (top) and 11-mers (bottom). All exhibiting multi-modal 9-mer and 11-mer distributions.

The empirical distribution of k-mers in all mammalian single chromosomes and whole genomes that we examined are all multi-modal. Figure 3.1 depicts the multi modal histograms for human chromosomes 1 (a long chromosome), 6 (medium) and 20 (short), for both 9-mers and 11-mers. Figure 3.2 depicts it for the complete human and opossum genomes, again for both 9-mers and 11-mers.

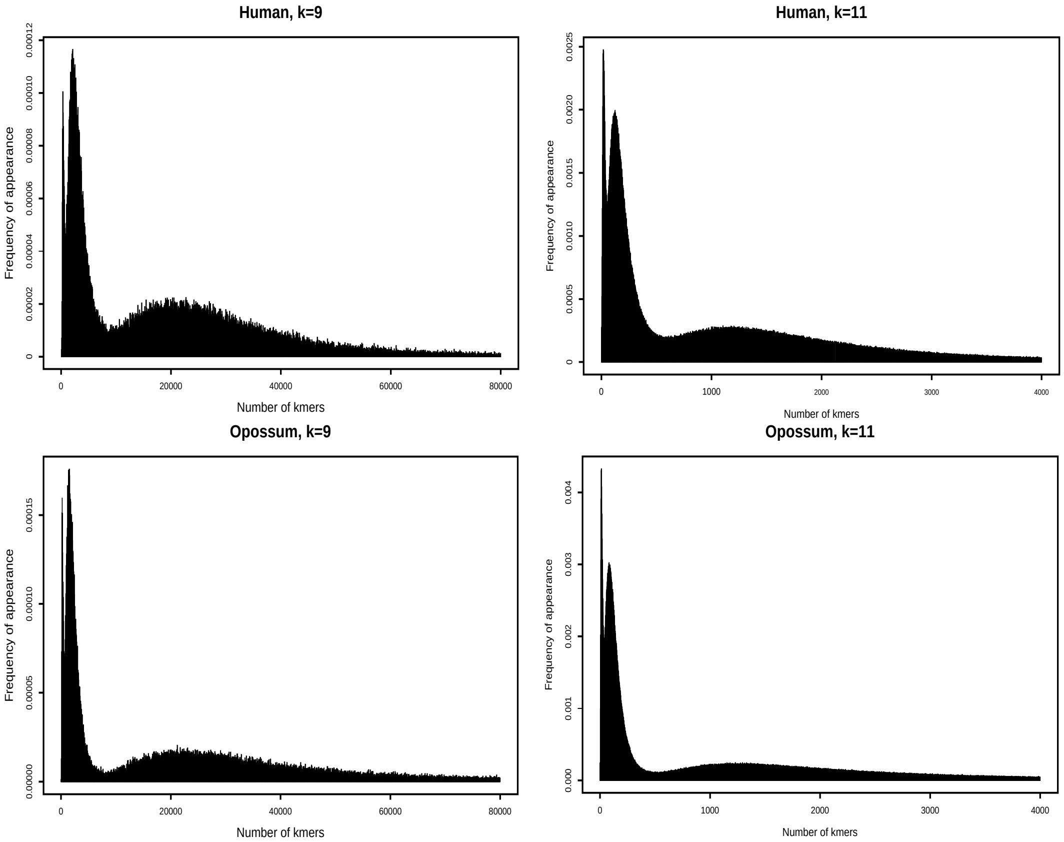

Figure 3.2: Multi-modal distributions for human and opossum.
Human (top) and opossum (bottom) whole genome, 9-mers (left) and 11-mers (right). All exhibiting multi-modal 9-mer and 11-mer distributions.

We chose to demonstrate the results for human (for obvious reasons), and for opossum, which being a non placental mammal represents an outgroup to the placental mammalian species. The three specific chromosomes per species were chosen as representatives of a long, medium, and short chromosomes. We chose 11-mers since $k = 11$ is the smallest for which there are missing k-mers in the human genome.

Denote by ℓ the length of the mammalian genomes or individual chromosomes. There is typically a high peak (or more) close to 0, corresponding to a large number of k-mers that are either missing or rare (a low number of appearances). Then there is a second, shallower local peak around the average number of occurrences ($\ell/4^k$), from where the numbers decrease monotonically. The high peak close to 0 flattens when ℓ grows larger, compared to 4^k. It gains more mass as 4^k grows with respect to ℓ. The number of over-abundant k-mers is also substantially higher than expected by a zero order model, and the decay is slower than exponential.

3.2 Non Mammalian Genomes: Empirical Distributions

We analyzed the k-mer distributions for 89 non-mammalian genomes: 33 archaea, 36 bacteria, and 20 non-mammalian eukarya, including 8 vertebrates. These distributions can be divided into two main categories:

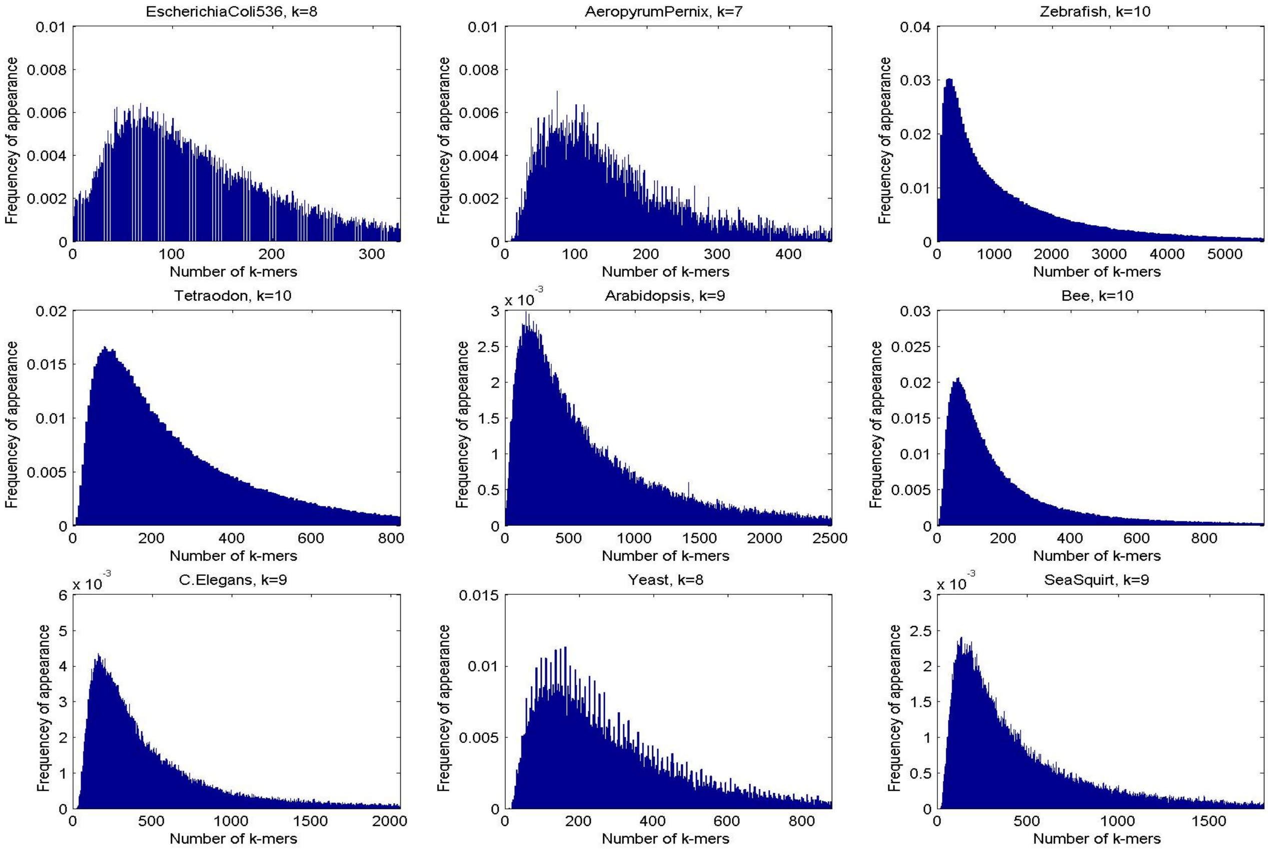

Figure 3.3: Unimodal k-mer distributions in various species.
From top-left: E.Coli, Aeropyrum Pernix, Zebrafish, Tetraodon, Arabidopsis, Bee, C. Elegans, Yeast (Saccharomyces cerevisiae), Sea Squirt.

1. ***Unimodal distributions***, where the corresponding k-mer histograms have a single maximum, usually at a realtively low number of k-mers. Figure 3.3 depicts typical unimodal k-mer histograms for 9 species. (In all cases k is the nearest integer to $c \cdot \log_4 \ell$.)

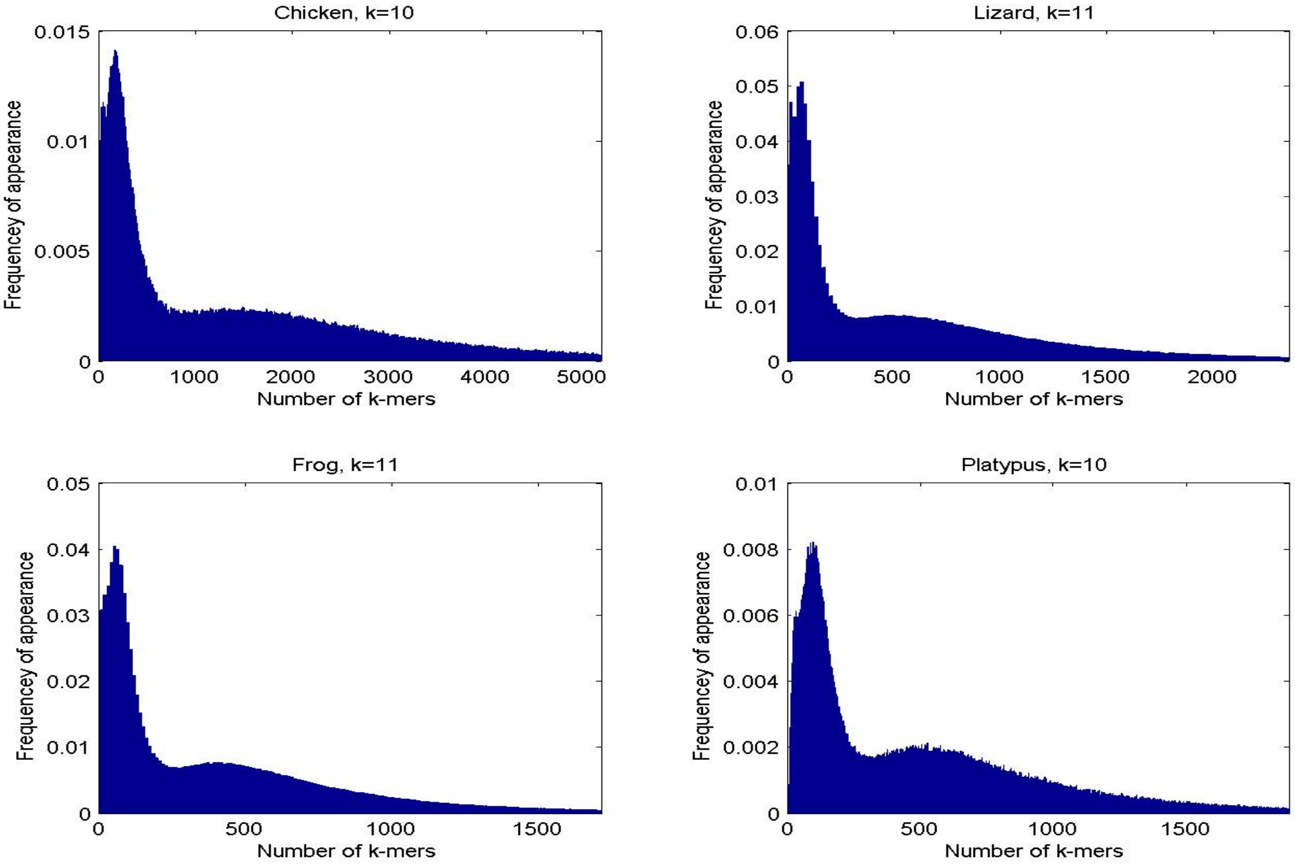

Figure 3.4: Multi-modal k-mers distributions in four non-mammals. These species are part of the tetrapod clade.
From top-left: chicken ($k = 10$), lizard ($k = 11$), frog ($k = 11$), platypus ($k = 10$).

2. ***Multi-modal distributions***, where the corresponding k-mer histograms have two or more maxima. Usually one (or more) high maximum is at a very low number of k-mers, and another, shallower one at a larger number. We found that only a small and well characterized group of species exhibits this distribution (figure 3.4). This group includes *Gallus gallus* (Chicken), *Anolis carolinesis* (Green Anole Lizard), and *Xenopus tropicalis* (Frog), all in the tetradon clade. Notice that the five bonny fish (vertebrates) are *not* a part of this group.

3.3 Human Genomic Regions

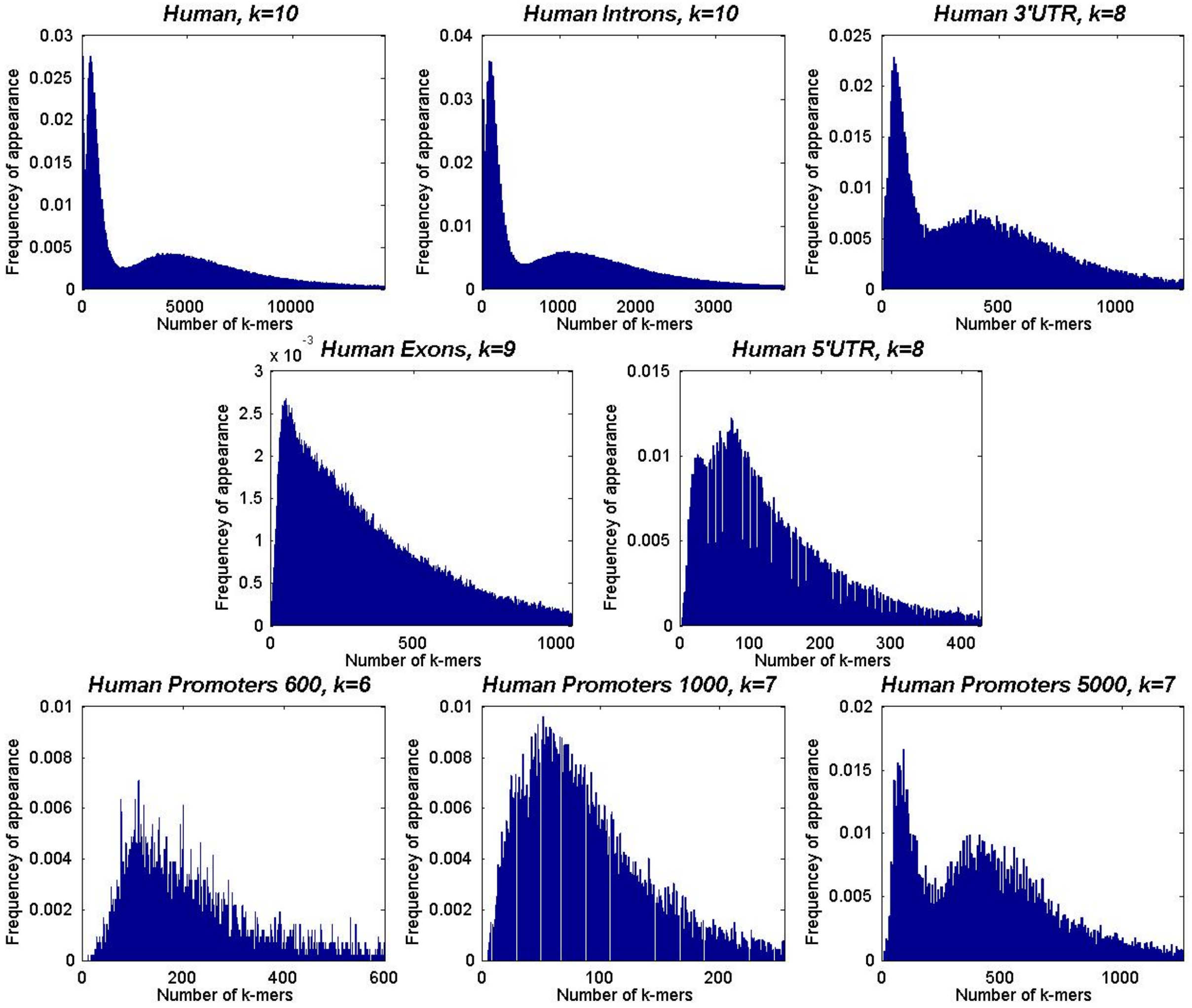

Figure 3.5: k-mer distributions among the different human genomic regions. From top-left: The whole human genome, both strands, k=10; Introns, single strand, k=10; 3'UTRs, single strand, k=8; Coding regions (exons), single strand, k=9; 5'UTRs, single strand, k=8; Gene promoter region 600 bases, single strand, k=6; Gene promoter region 1000 bases, single strand, k=6; Gene promoter region 5000 bases, single strand, k=7.

Figure 3.5 depicts the differences of k-mer distributions *within* human genomic regions. The regions analyzed were *coding regions* (exons), *introns*, *3'UTRs*, *5'UTRs* and *gene promoter regions*. The *gene promoter regions* were separately analyzed three times, corresponding to varying lengths of the promoter region (600, 1000, and 5000 nucleotide bases upstream of the 5'UTR of the gene). The most striking empirical observation is that the *coding regions*, the *5'UTRs*, and the shorter lengths of *gene promoter regions* exhibit unimodal k-mer distribution, while the *introns*, the *3'UTRs* and the *gene promoter regions* of length 5000 bases exhibit multi-mode k-mer distributions.

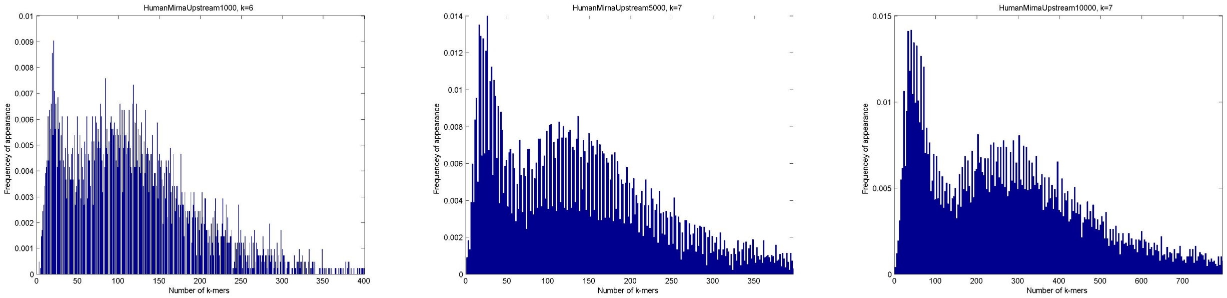

Figure 3.6: k-mer distributions in miRNA gene promoter regions.
Length of promoter regions: 1000 bases (left, $k = 6$), 5000 bases (middle, $k = 7$), and 10000 bases (right, $k = 7$).

Figure 3.6 shows the k-mer histograms of the promoter regions upstream of 529 known human micro RNA (miRNA) genes. Although all three graphs exhibit some sort of multi-modality, it is clear that this multi-modality increases its effect as the length of the sequences increases. This leads to the claim that there may be some change in the distribution in the regions closer to the miRNA genes. This claim has been supported by Lee et al. [11] who found that, for example, the 6-mer `CGCGCG` is over-represented in these regions and acts as a transcription factor binding site (TFBS) for several miRNA genes. We note that the promoter regions of miRNA genes are quite different than those of regular genes, and are closer to introns, with respect to the properties we analyzed (see 5.3).

3.4 Low Order Markov Models

Low order Markov models were fitted to the genome of *Homo sapiens* by counting the transitions required to emit the sequence. Since these are simple (non-hidden) Markov models, this is equivalent to maximum likelihood estimation [5]. Because of the lengths of sequences being analyzed, extremely complex, namely high order, models still show a significant improvement in their fit (as measured by the AIC).

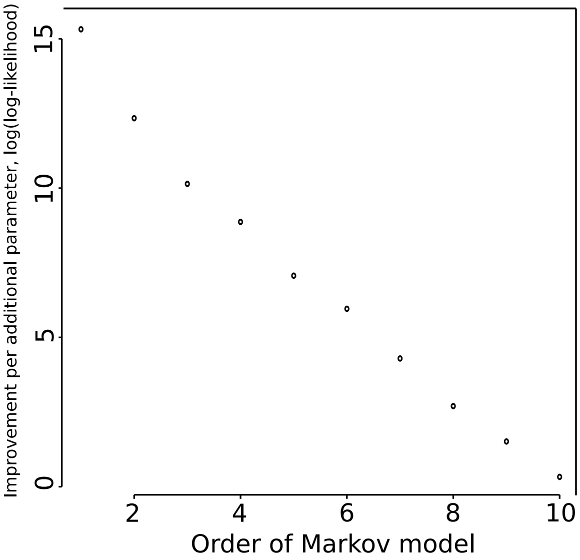

Figure 3.7: Improvement in log-likelihood score of human genome per additional parameter. Improvement in log-likelihood score of human genome sequence per additional parameter, as the order of the model increases (log scale). The improvement decreases exponentially per parameter as the order of the model increases.

Figure 3.7 shows the log of the improvement per additional parameter as the order of the model increases, showing that there is an exponential decrease in the improvement per parameter as the order of the model increases. For the difference between the models to be insignificant, the improvement per parameter would have to drop below 1, and we can see this is not the case.

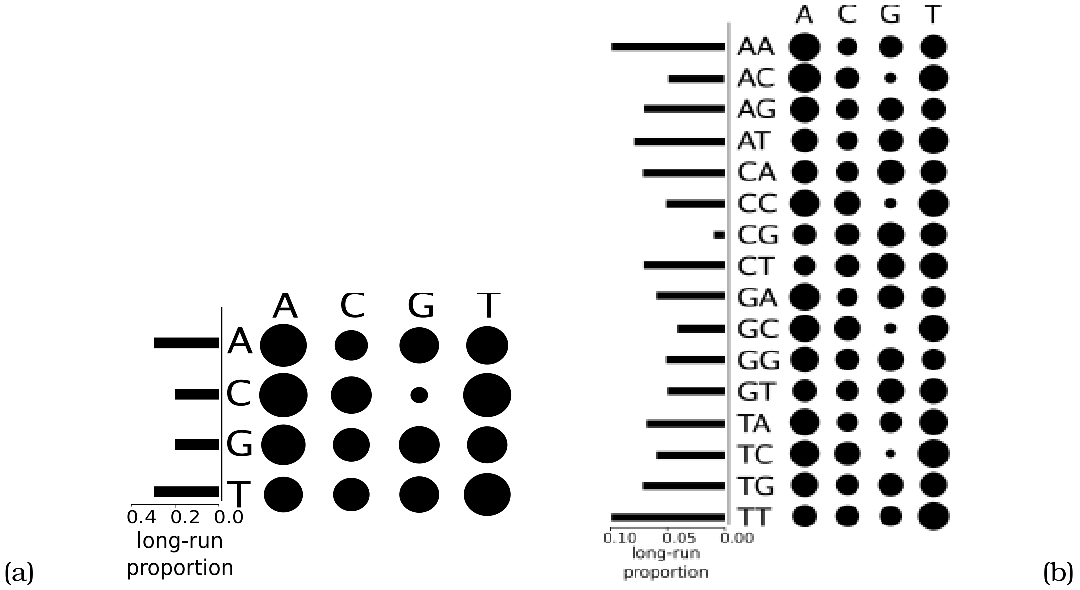

Figure 3.8: Graphical representation of the first order (a) and second order (b) Markov models of the human genome.
Bars are proportional to stationary probabilities. Areas of disks are proportional to transition probabilities. Note the CpG suppression.

Despite more complex models being significantly better, even the first order model captures effects like CpG suppression. Figure 3.8(a) depicts the first order model graphically. The relative small probability for the transition from C to G is readily seen. Comparing this to a similar graph of the transition probabilities for a second-order model (figure 3.8(b)), it is almost like the first order transition probabilities have been repeated four-fold. This suggests that improvements by increasing the order further come from fitting observed poly-nucleotide frequencies better, rather than transitions.

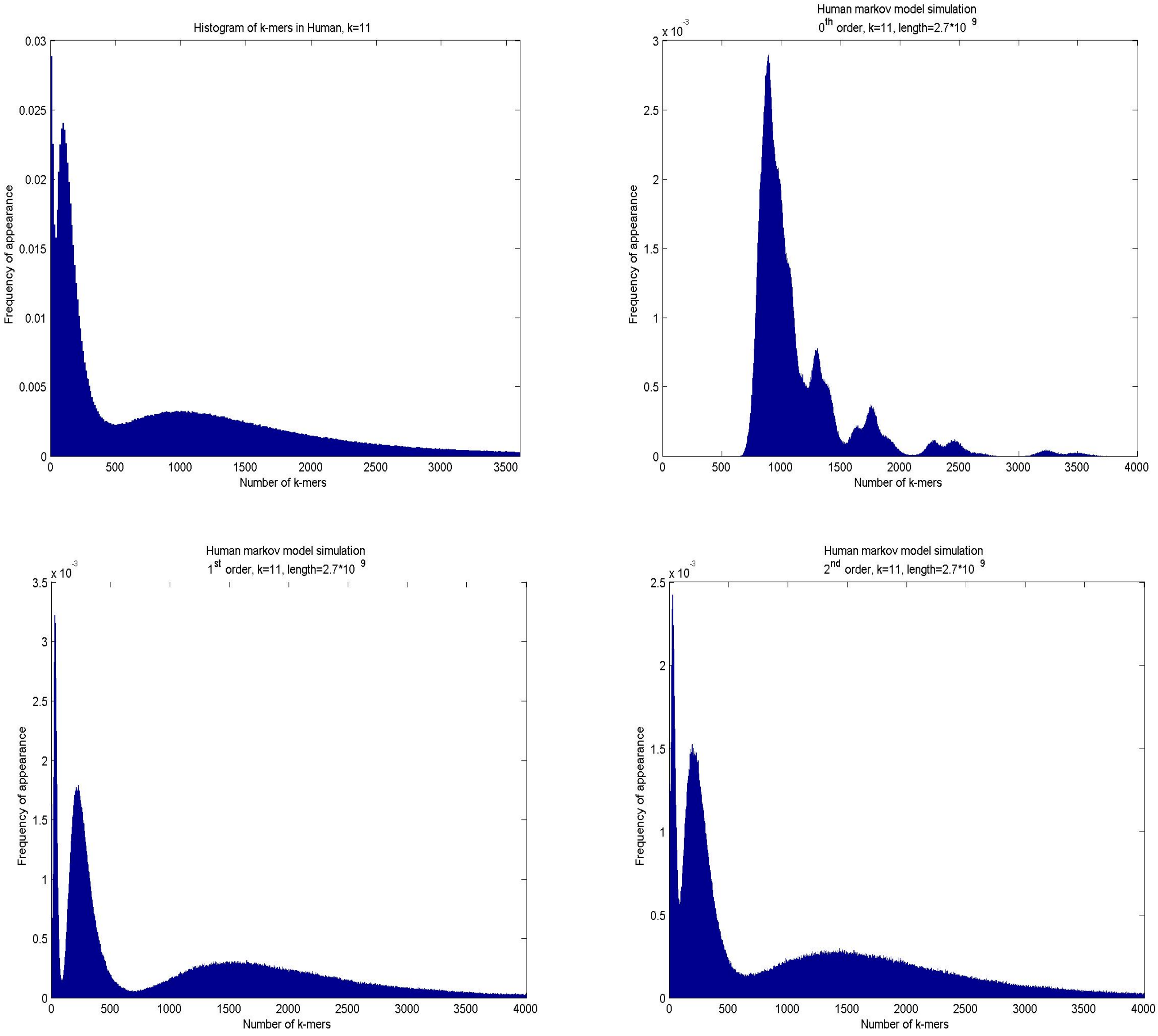

Figure 3.9: Multi-modal histogram Markov model simulations results (human, $k = 11$). The empirical histogram (upper-left), 0^{th} order Markov model (upper-right), 1^{st} order Markov model (lower-left), and 2^{nd} order Markov model (lower-right). Simulation lengths were chosen as the length of the original genome.

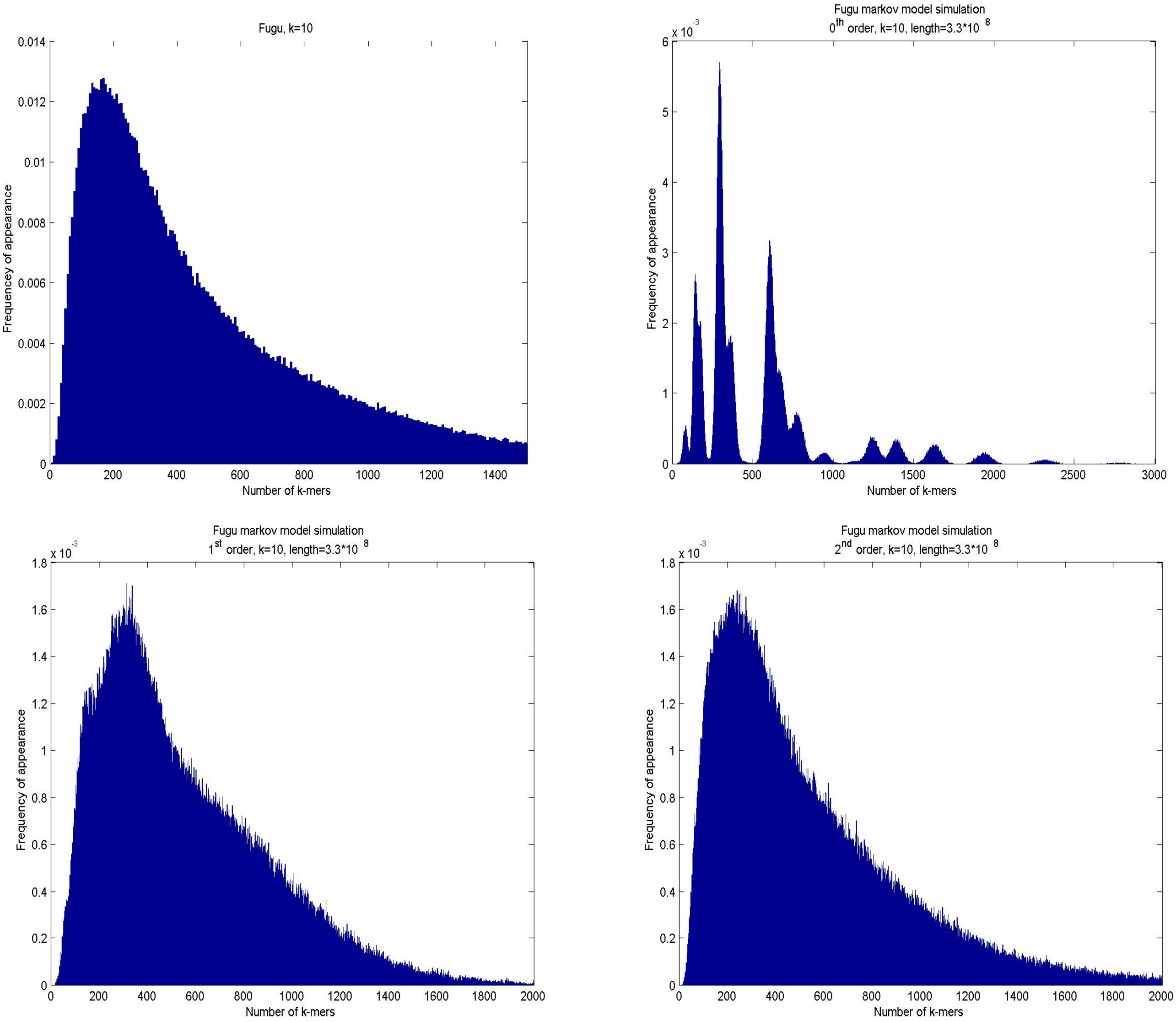

Figure 3.10: Unimodal histogram Markov model simulations results (fugu, $k = 10$). The empirical histogram (upper-left), 0^{th} order Markov model (upper-right), 1^{st} order Markov model (lower-left), and 2^{nd} order Markov model (lower-right). Simulation lengths were chosen as the length of the original genome.

In figures 3.9 and 3.10 we show the empirical k-mer histograms, and those induced from the zero order Markov model, first order Markov model, and second order Markov model, for 2 eukaryotes: human ($k = 11$) and fugu ($k = 10$). It can be seen that while the zero order model poorly captures the empirical histograms, first and second orders provide a much better fit. In particular, the second order Markov model captures the modality of the distribution quite well.

3.5 Tail Weight Distribution

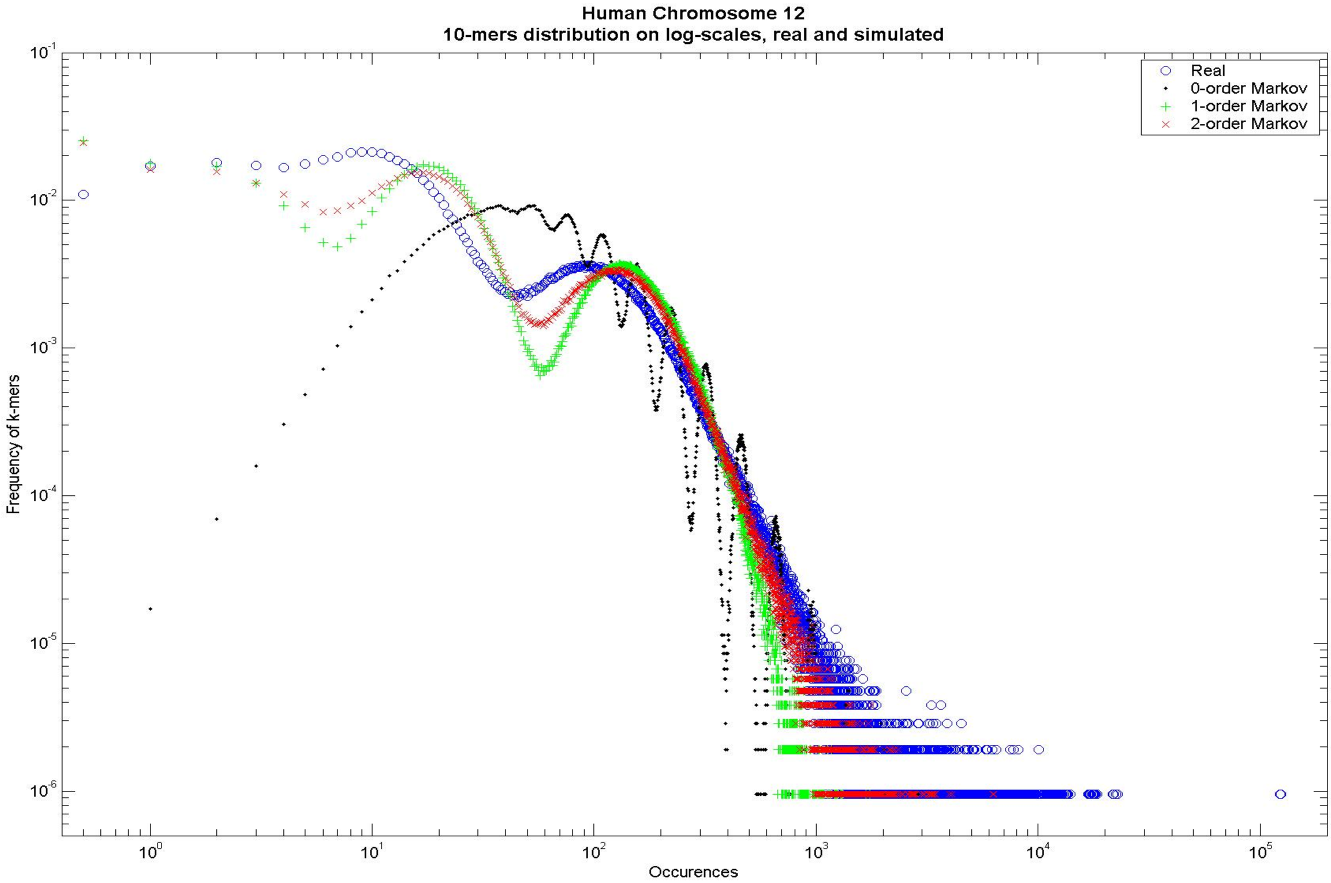

Figure 3.11: Human chromosome 12: Log-scaled distribution of 10-mers in Markov models and real data.
Markov model simulations versus the real data. Real data is marked by blue circlets; Markov models: 0-order (black dots), 1-order (green +) and 2-order (red x).

During the completion of this thesis, we were informed of a related analysis by Csuros et al. [4]. In their article, they claim that low order Markov chains do not give a fit to the empirical genomic data, especially in the "heavy tail" part of the histograms. Instead, they propose a model where CpG's are first removed, and the remaining is modeled by a double Pareto distribution [17]. We argue that this is true for the 0^{th}-order Markov chain (i.e. transition matrix, preserving only single nucleotide statistics). However, if we increase the order of the model even just to 2, we achieve a fairly good fit in the "heavy tail" as well. We exhibit this using log axes histograms for human chromosome 12 and repeat masked human chromosome 5 (figures 3.11, 3.12).

We note that the differences between the real data and our modeled data regarding the "heavy tail" include mainly the long repeat sequences, such as poly-A and poly-T. It can be seen that the fit for the repeat-masked chromosome is better, where the repeat-masker masks these types of k-mers.

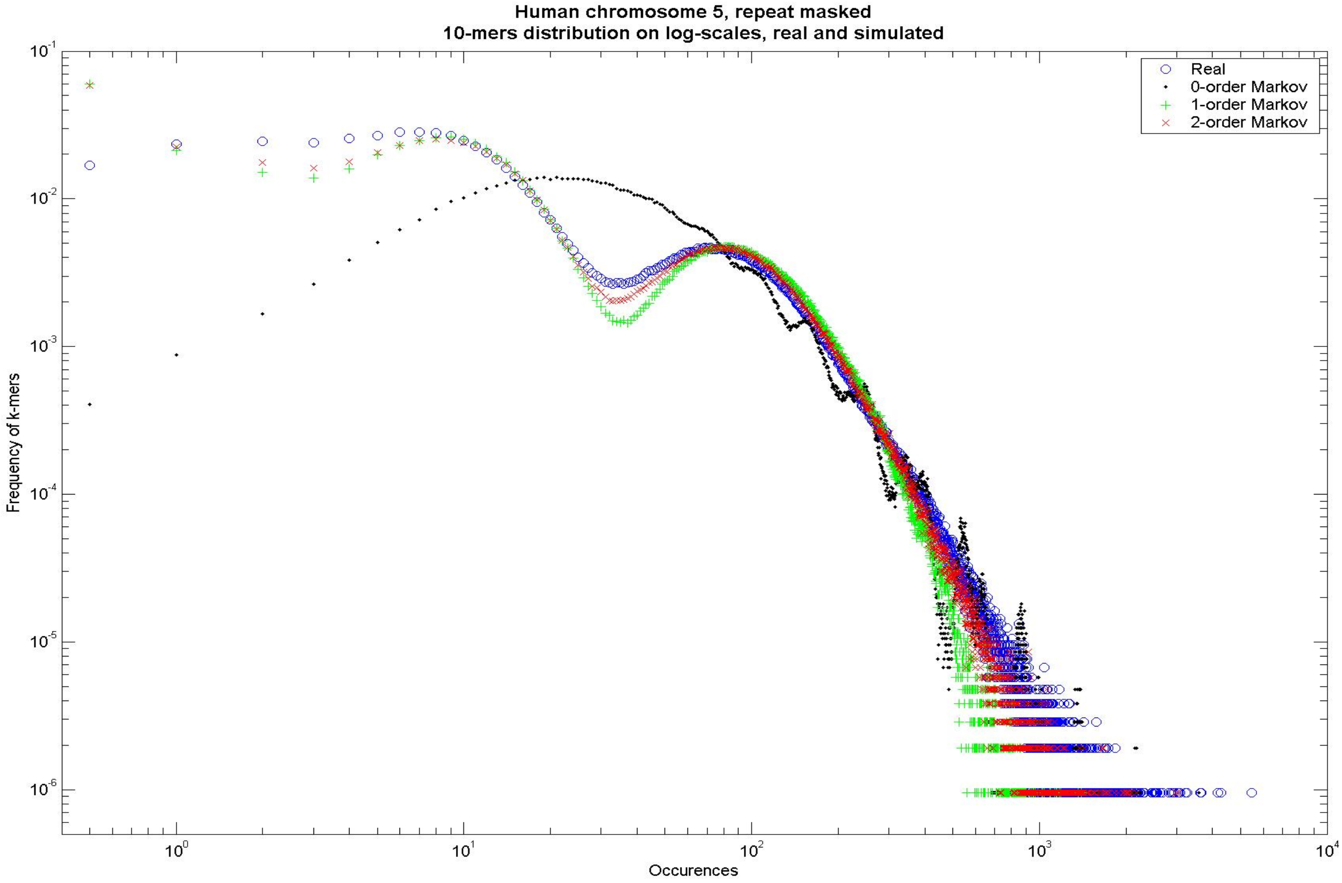

Figure 3.12: Repeat-masked human chromosome 5: Log-scaled distribution of 10-mers in Markov models and real data.
Markov model simulations versus the real data. Real data is marked by blue circlets; Markov models: 0-order (black dots), 1-order (green +) and 2-order (red x).

Figures 3.11 and 3.12 depict the Markov models of orders 0 (black dots), 1 (green +), and 2 (red x) versus the real data (blue circlets) for human chromosome 12, and for repeat-masked human chromosome 5, for 10-mers. We display the results using a log scale for both axes (as used in [4]), though we note that the data is the same as shown in our previous histogram-bars style display. The Markov models of order 1 and 2 provide a fairly good fit to the real data. Moreover, these low order Markov models manage to capture the majority of the so-called “heavy tail” of the distributions, without the need to disregard parts of the data (such as CpG enriched k-mers), as suggested in [4].

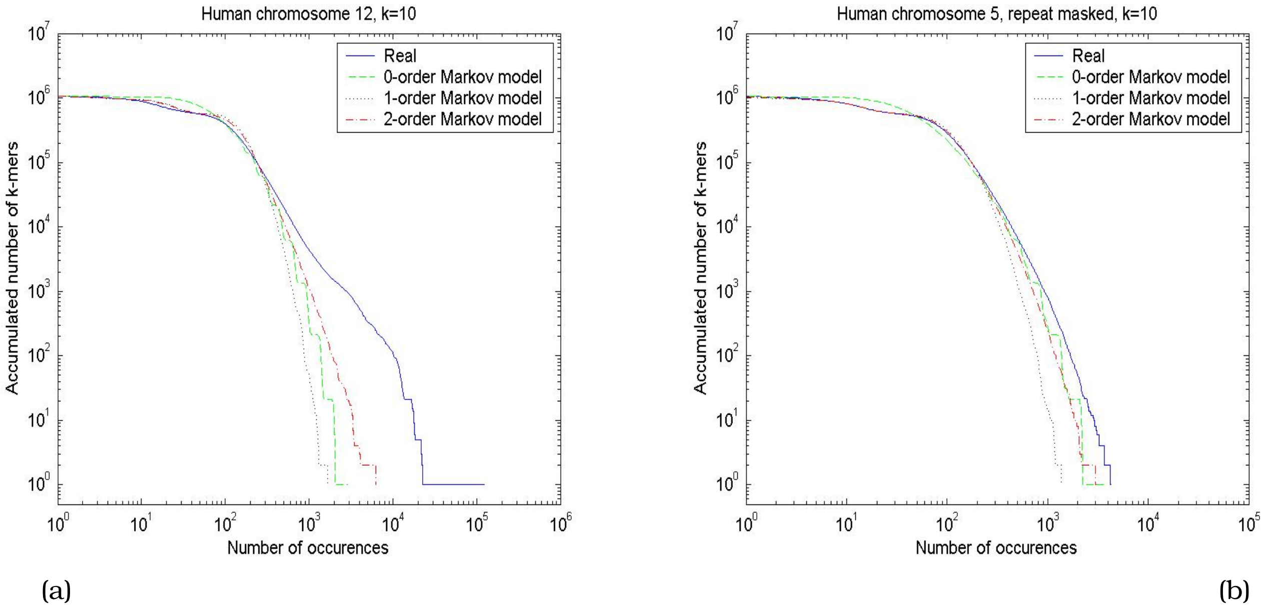

Figure 3.13: Tail weight distribution.
Markov model simulations versus the real data. Real data is marked by a full blue line; Markov models: 0-order (green dashed line), 1-order (black dotted line) and 2-order (red dash-dotted line). Shown for (a) human chromosome 12, and (b) repeat-masked human chromosome 5, for $k = 10$.

Chromosome 12								
# of	% of all	Real	0-order		1-order		2-order	
occur.	k-mers	# of k-mers	# of k-mers	(d)/(c)	# of k-mers	(f)/(c)	# of k-mers	(h)/(c)
(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
≥ 7	90	946136	1046733	1.10	955065	1.01	954300	1.01
≥ 14	75	796815	1030725	1.29	899363	1.13	878222	1.10
≥ 64	50	524330	622401	1.18	562153	1.07	551253	1.05
≥ 123	30	314938	283911	0.90	422315	1.34	386321	1.23
≥ 232	10	105172	84343	0.80	112916	1.07	110924	1.05
≥ 322	5	52556	39863	0.75	34258	0.65	43496	0.83
≥ 661	1	10514	3766	0.35	777	0.07	4716	0.45
≥ 2711	0.1	1049	1	0.001	0	0.00	30	0.03
≥ 17870	0.001	11	0	0.00	0	0.00	0	0.00

Table 1: Real data fit using low order Markov models, human chromosome 12, $k = 10$

Chromosome 5, Repeat-masked								
# of	% of all	Real	0-order		1-order		2-order	
occur.	k-mers	# of k-mers	# of k-mers	(d)/(c)	# of k-mers	(f)/(c)	# of k-mers	(h)/(c)
(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
≥ 5	90	955528	1042716	1.09	933181	0.98	928382	0.97
≥ 10	75	811978	1010372	1.24	818559	1.01	811780	1.00
≥ 46	50	526332	548490	1.04	548674	1.04	537336	1.02
≥ 93	30	316552	247553	0.78	352583	1.11	332627	1.05
≥ 171	10	105378	83695	0.79	106338	1.01	104400	0.99
≥ 230	5	52860	47285	0.89	43730	0.83	48556	0.92
≥ 430	1	10540	7204	0.68	3219	0.31	6965	0.66
≥ 914	0.1	1049	436	0.42	23	0.02	378	0.36
≥ 2881	0.001	11	1	0.09	0	0.00	2	0.18

Table 2: Real data fit using low order Markov models, repeat masked human chromosome 5, $k = 10$

The columns represent: (a) the number of occurrences of k-mers; (b) the percentage of k-mers which appear at least the number of times as in column (a); (c) the total number of k-mers which appear at least the number of times as in column (a), in the real data; (d) the total number of k-mers which appear at least the number of times as in column (a), in the Markov model simulation of order 0; (e) the percentage of fitness between the number of k-mers in the Markov model simulation and the real data $\frac{(d)}{(c)}$; (f),(g) same as (d) and (e), for Markov model order 1; (h),(i) same as (d) and (e), for Markov model order 2.

Figure 3.13 along with some actual values (tables 1, 2) exhibit how close our low order Markov models are to capturing the essence of the real data. The figure and tables show how many k-mers appear *at least* x number of times. As we can see, the tail is indeed quite heavy (notice that the axes in the figure are log scaled), and yet the fit is quite good. It is clear that the fit for the repeat masked chromosome is better, and that the 2^{nd} order Markov model provides a better fit than the lower order models. For the non-repeat masked chromosome 12, there appears to be a skew in the fit for all models for the last about 1,000 k-mers (0.1% of all 10-mers), which is magnified in the log-scaled display. These are the most abundant 10-mers, most of which are removed by the repeat masker, thus allowing for a better fit for the repeat masked chromosome. The relative fit of the models to these most abundant 10-mers decreases drastically on the non-repeat masked chromosome.

3.6 Surprising k-mers

Rather than looking at the absolute abundance of a k-mer, it is better to ask how surprising the observed abundance is. The level of surprise can be measured using the odds-ratio (OR) [21]: how over- or under- represented a k-mer is in a genome given our expectations.

$$\text{OR} = \frac{(1-e)/e}{(1-p)/p}$$

where e is the empirical frequency at which a given k-mer is observed in a genome and p is the proportion of time that it is expected to occur. In practice, an empirical log of the odds-ratio is used because of problems with unobserved k-mers (see 'empirical logits', [21]).

Over-abundant k-mers							
Homo sapiens genome				Repeat masked *Homo sapiens* genome			
k-mer	Count	Expected	OR	k-mer	Count	Expected	OR
aaaaaaaaaaa	2488487	1366	182	tttttaaaaaa	32998	4578	7.21
ttttttttttt	2472795	1338	185	ttttttaaaaa	32862	4569	7.19
tgtgtgtgtgt	536538	687	781	atttttaaaaa	31604	3685	8.58
acacacacaca	534130	697	767	tttttaaaaat	31519	3680	8.57
gtgtgtgtgtg	500834	476	1053	ttttaaaaaat	28465	3687	7.72
cacacacacac	498270	482	1034	attttttaaaa	28341	3678	7.71
ctgtaatccca	404242	1087	372	tatttttaaaa	23708	2553	9.29
tgggattacag	402211	1085	371	ttttaaaaata	23681	2558	9.26
tatatatatat	394726	560	705	ttttaaaaatt	22181	3680	6.03
atatatatata	394600	561	703	aattttaaaaa	22134	3685	6.01

Table 3: Top ten most abundant k-mers in the human genome.
Top ten most abundant k-mers found in the *Homo Sapiens* genome (Ensembl release 46) for both unprocessed and repeat-masked sequences, as measured by the total number of occurrences (Count). 'Expected' is the number of times it would be expected to occur under a first order model and 'OR' is the Odds Ratio describing how surprising the observed occurrence is compared to expectation.

Table 3 lists the 10 most abundant 11-mers in the human genome, and also the most abundant ones in the genome under the repeat mask filter [20]. The unprocessed genome tends to contain many mono-nucleotide and di-nucleotide repeats, whereas the repeat masked genome is rich in interspersed poly-t and poly-a repeats. As may be predicted from the removal of simple repeats, the repeat masked *Homo sapiens* genome has an under-abundance of k-mers containing long mono-nucleotide repeats and some of these are also under-abundant in the unmasked genome.

Surprisingly Over-abundant k-mers — Surprisingly Under-abundant k-mers

Homo sapiens genome

k-mer	Count	Expected	OR	k-mer	Count	Expected	OR
cgcgcgcgcgc	2448	0.061	40098	tcgaaattcgc	0	46.0	0.011
gcgcgcgcgcg	2437	0.060	39949	cccccccctat	9	817.2	0.012
cgccgccgccg	4642	0.481	9650	attgcgaacga	0	41.9	0.012
cggcggcggcg	4601	0.480	9566	tcgcgagttaa	0	34.2	0.015
ccgcgcccggc	19701	2.541	7753	atcttcgcgag	0	33.9	0.015
gccgggcgcgg	19556	2.539	7703	tcagggggggg	15	1003.5	0.015
caccgcgcccg	19907	2.895	6876	atcgcaacgga	0	32.3	0.015
cgggcgcggtg	19848	2.888	6873	tatgtttcgcg	0	31.9	0.016
accgcgcccgg	19655	3.002	6547	tgcaacgatcg	0	31.1	0.016
ccgggcgcggt	19539	2.996	6522	agtccgcgcaa	0	30.3	0.016

Repeat masked *Homo sapiens* genome

k-mer	Count	Expected	OR	k-mer	Count	Expected	OR, $\times 10^{-3}$
cgcgcgcgcgc	601	0.022	26760	taagggggggg	0	481.8	1.038
gcgcgcgcgcg	576	0.022	25673	cccccccctaa	0	481.7	1.038
cggcggcggcg	1208	0.183	6609	ttagggggggg	0	480.9	1.040
cgccgccgccg	1073	0.183	5874	tcagggggggg	0	450.2	1.111
cgcgcgcgcgt	152	0.027	5660	cccccccctca	0	450.1	1.111
cgcgccgcgcg	150	0.028	5466	cccccccctga	0	449.4	1.113
acgcgcgcgcg	144	0.027	5352	tagggggggga	0	432.1	1.157
cgcgcggcgcg	138	0.028	5029	tccccccccta	0	431.5	1.159
cgcggcgcgcg	138	0.028	5029	cccccccccat	0	430.3	1.162
cggcgcgcgcg	133	0.028	4847	cccccccctat	0	387.2	1.291

Table 4: Top ten most surprising over- and under- abundant k-mers in the human genome. Top ten most surprising over- and under- abundant k-mers found in the *Homo Sapiens* genome (Ensembl release 46) for both unprocessed and repeated masked sequences, as measured by the Odds Ratio (OR). 'Count' is the number of times the k-mer occurs, whereas 'Expected' is the number of times it would be expected to occur under a first order model.

Table 4 describes the 10 *most surprising* 11-mers in the human genome – both rare and abundant ones (those with smallest and largest odds ratios, correspondingly). Those surprisingly over-abundant k-mers tend to be CpG rich, in contrast to those with high absolute abundance, which do not contain CpG di-nucleotides; the over-abundance may be due to the hypothesized effect on the major groove in DsDNA. While there is a tendency for under-abundant k-mers to be missing, 8 of the top 10 shown, the relationship is not strong. Out of the first 1000 least absolutely abundant k-mers, 982 are missing whereas only 192 are missing from the first 1000 most under-abundant k-mers; for the repeat masked genome, all of the first 1000 least absolutely abundant are missing whereas only 331 of first 1000 most under-abundant k-mers as missing. Notice that rare CpG-rich k-mers are *not* surprising. We remark that other low-order models give fairly similar results. Our results show that the distribution of k-mer frequencies is consistent with CpG suppression and many k-mers are just too rare to be observed in the data.

3.7 CpG Suppression

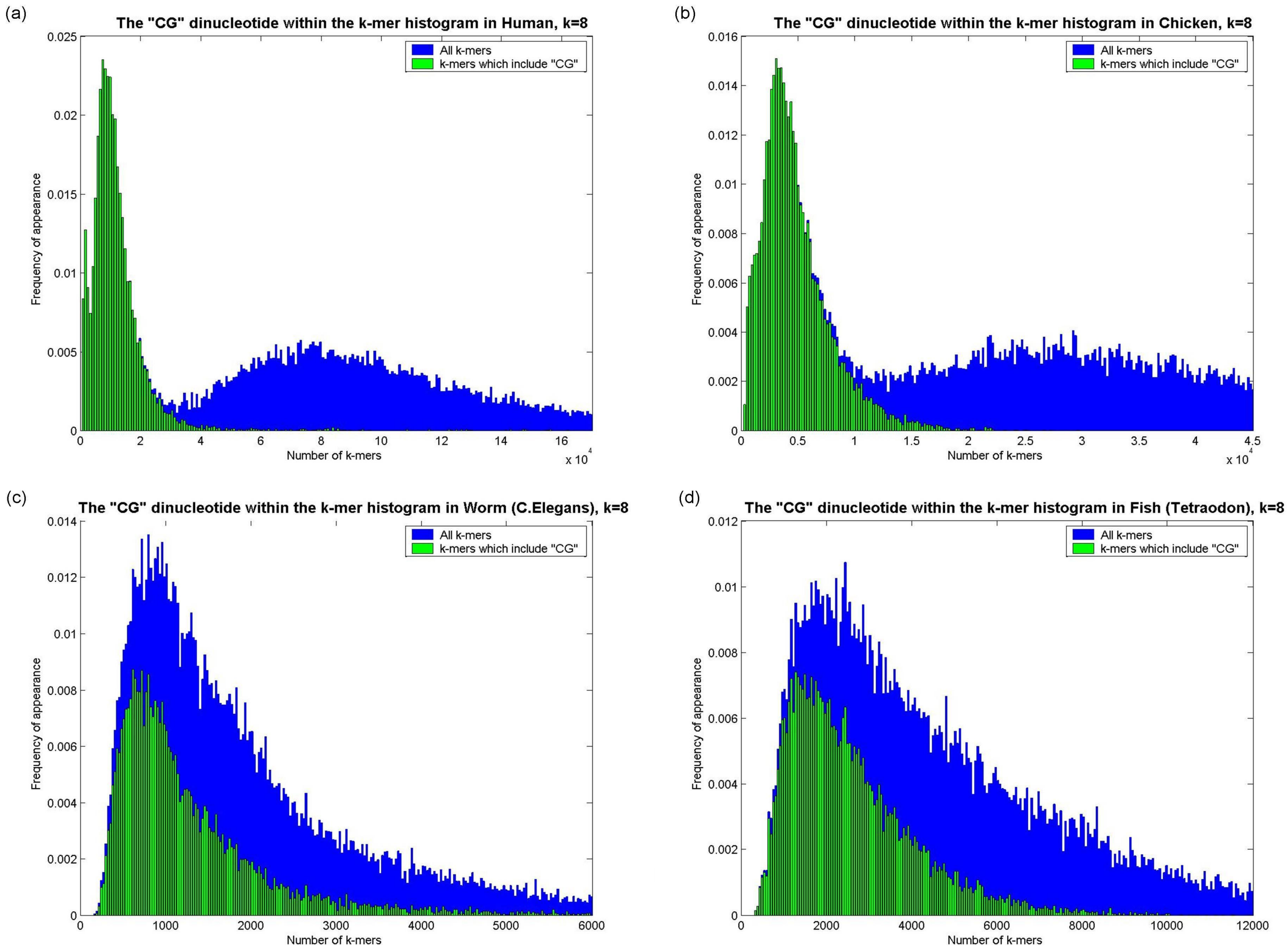

Figure 3.14: The CpG-dimer within the distributions.
k-mers which contain the CpG-dimer comprise the left-most part in the multi-modal histograms: (a) Human, (b) Chicken. In other species it has no such effect: (c) C. Elegans (Worm), (d) Tetraodon (Fish).

Figure 3.14 shows histograms of the k-mer distribution for (a) Human (*Homo sapiens*) and (b) Chicken (*Gallus gallus*). k-mers which include the dimer CpG are colored green, while all others are blue. For clarity, the right hand side of the histograms are truncated, but those truncated regions are essentially “CG-free”. It is evident that these highlighted k-mers wholly comprise the left-most areas of the multi-mode histograms. The same histograms and highlights are shown for species with unimodal k-mer distributions – (c) the Worm (*Caenorhabditis elegans*) and (d) Fish (*Tetraodon nigroviridis*). In these examples it is clear that the CpG dimer does not have the same dramatic effect.

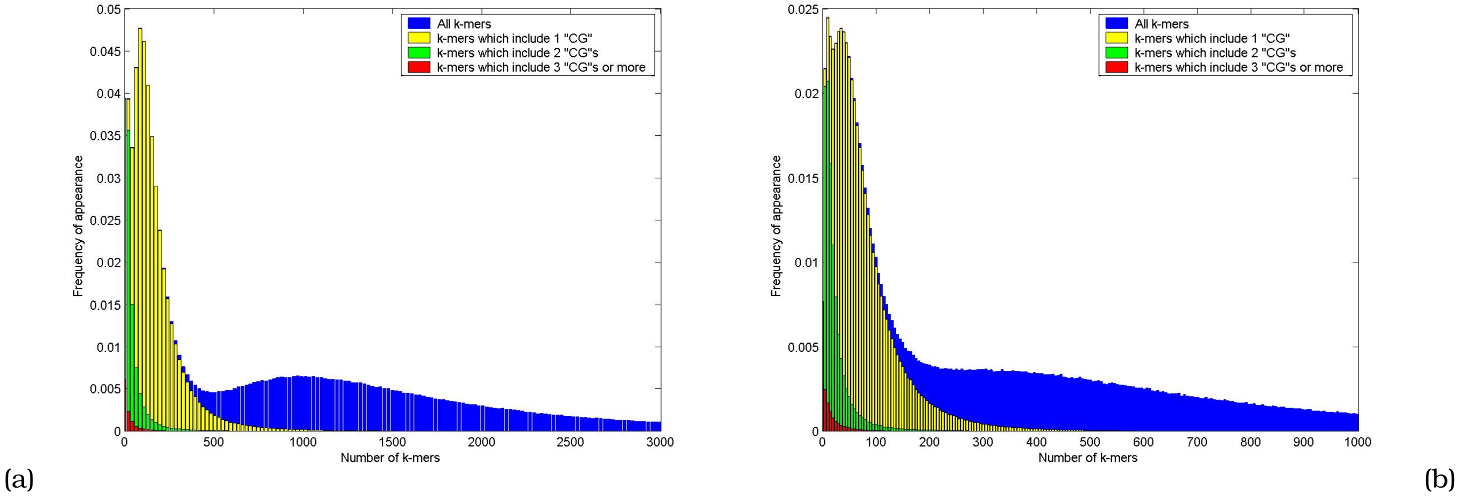

Figure 3.15: Enrichment of the `CpG`-dimer within rare 11-mers: (a) Human, (b) Chicken.

Those k-mers containing even more `CpG` dimers appear more and more to the left of the histogram, as seen in figure 3.15. In these histograms those k-mers which include 3 or more instances of `CpG` are colored red, those which include 2 instances are colored green, 1 instance are yellow, and the rest (0 instances) are blue.

Karlin et al. [10] defined ρ as the ratio between the empirically found probability of a dimer, and the expected combination of its monomers, according to their respective probabilities. Thus, for the `CpG` dimer, the relevant ratio is

$$\rho_{\texttt{CG}} = \frac{Pr(\texttt{CpG})}{Pr(\texttt{C})Pr(\texttt{G})}$$

If the occurrences of `C` and `G` were independent, $\rho_{\texttt{CG}}$ would be 1. For human, $\rho_{\texttt{CG}} = 0.24$, for opossum it is 0.13, for lizard it is 0.3 and for frog it is 0.34. Low values are also attained for some bacteria (*e.g.* Entamoeba Histolytica, 0.3) and archea (*e.g.* Methanococcus Jannaschii, 0.32, and for Methanosphaera stadtmanae, 0.27). Values of $\rho_{\texttt{CG}}$ exceeding 1 are less frequent but nonetheless exist, *e.g.* for beetle it is 1.15 and for bee it is 1.64.

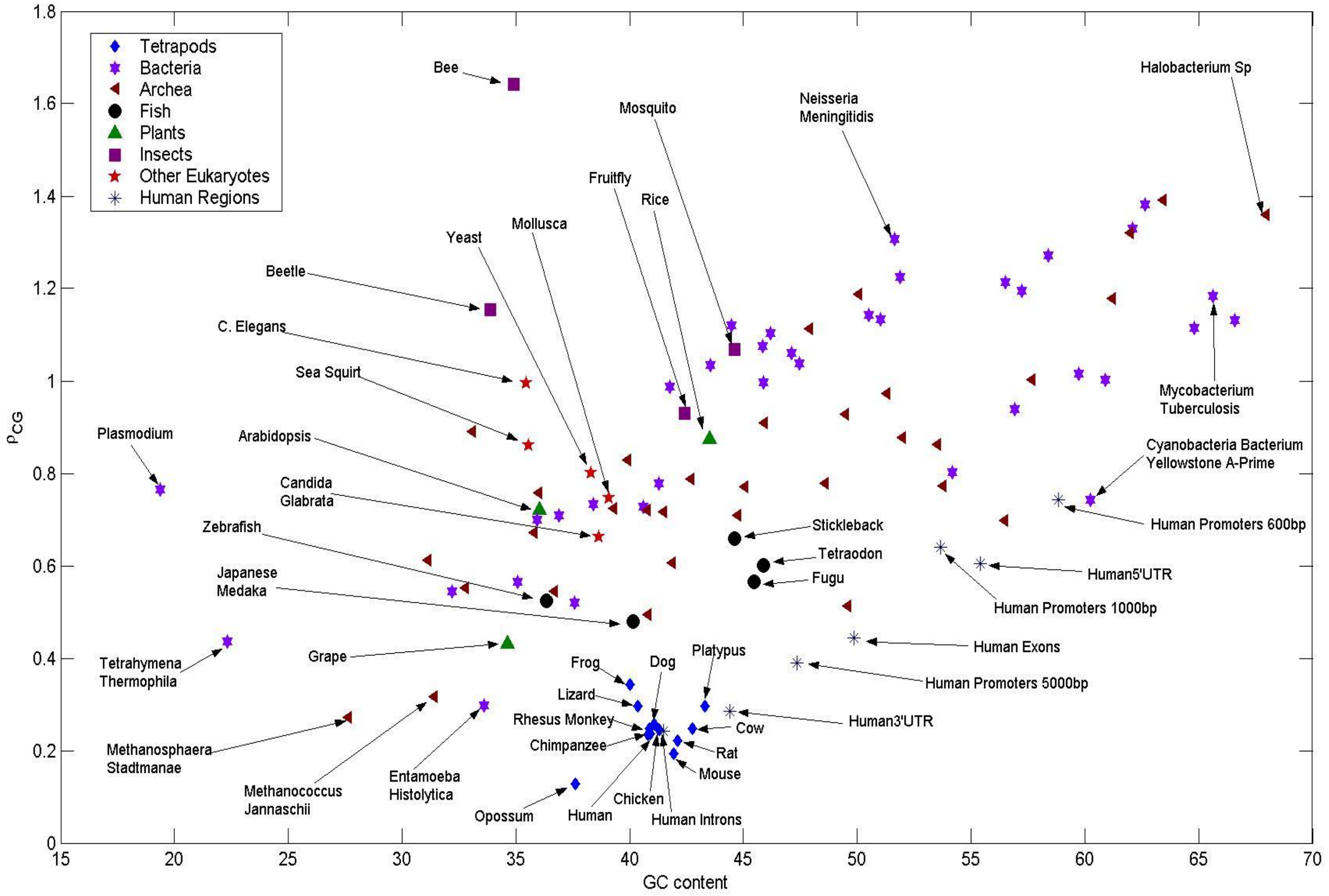

Figure 3.16: GC content vs. ρ_{CG}.
Notice the tetrapods cluster.

Figure 3.16 shows a 2-D graph, where the x-axis represents the GC-content (in %), and the y-axis represents ρ_{CG}. The tetrapods, which exhibit a multi-modal k-mer histograms, cluster quite closely. We can also see that some species, such as the archaea *Methanosphaera stadtmanae* and *Methanococcus Jannaschi*, the protozoan bacteria *Entamoeba hystolytica* and *Tetrahymena thermophila* all have $\rho_{\text{CG}} < 0.33$, a value similar to the lizard and smaller than the frog.

3.8 Simulated Multi-modal Borders

We tried to establish an empirical border to the multiple modality phenomenon, with respect to the GC-content and ρ_{CG}. In order to do this we use a convex combination between the dimer distributions of human and all other species.

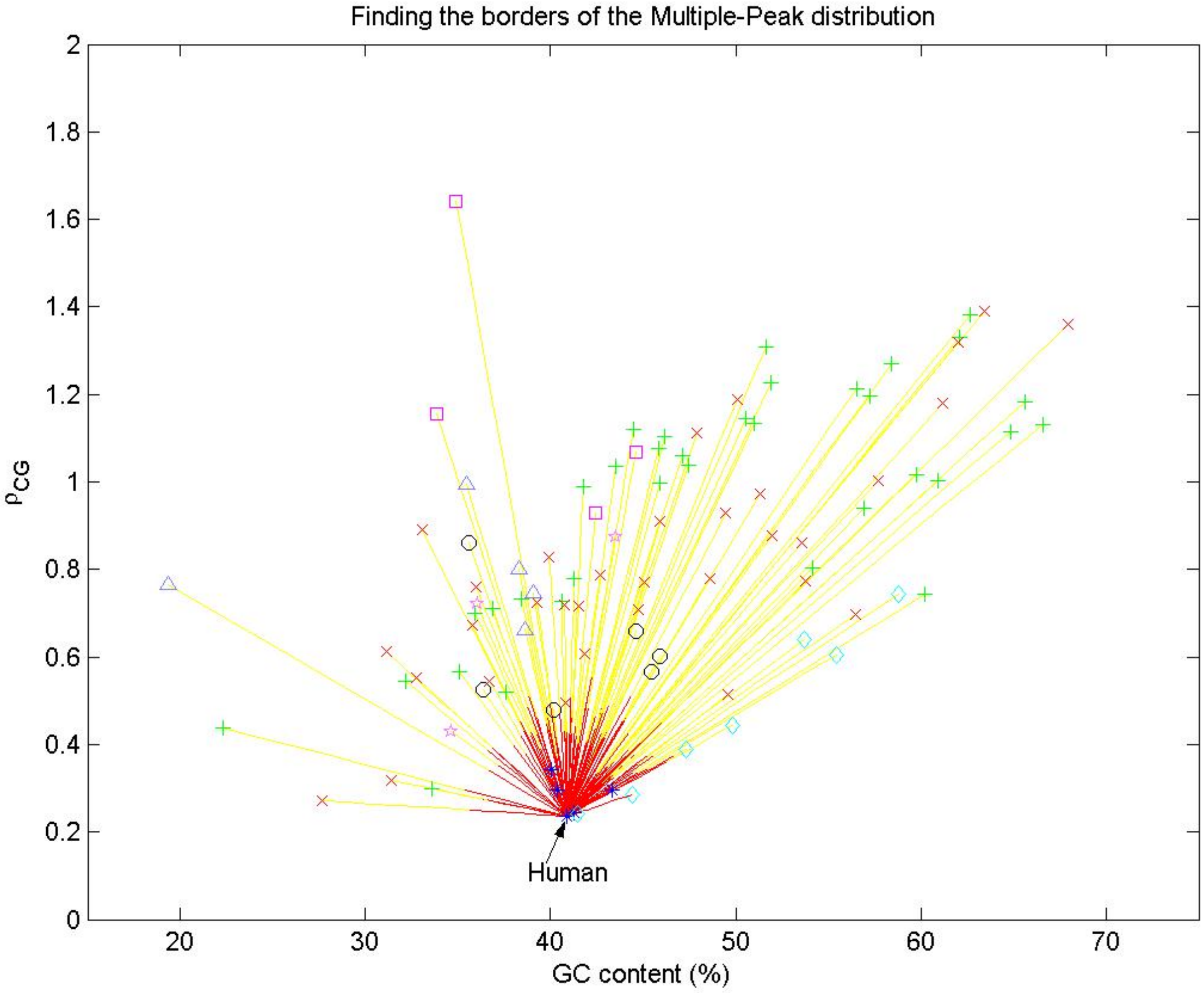

Figure 3.17: The simulated border of the multi-modal distribution.

Figure 3.17 shows the simulated borders of the multi-modal k-mer distribution phenomenon, with respect to the GC-content, ρ_{CG} and α_s, using the human distribution as the basic multi-modal distribution.

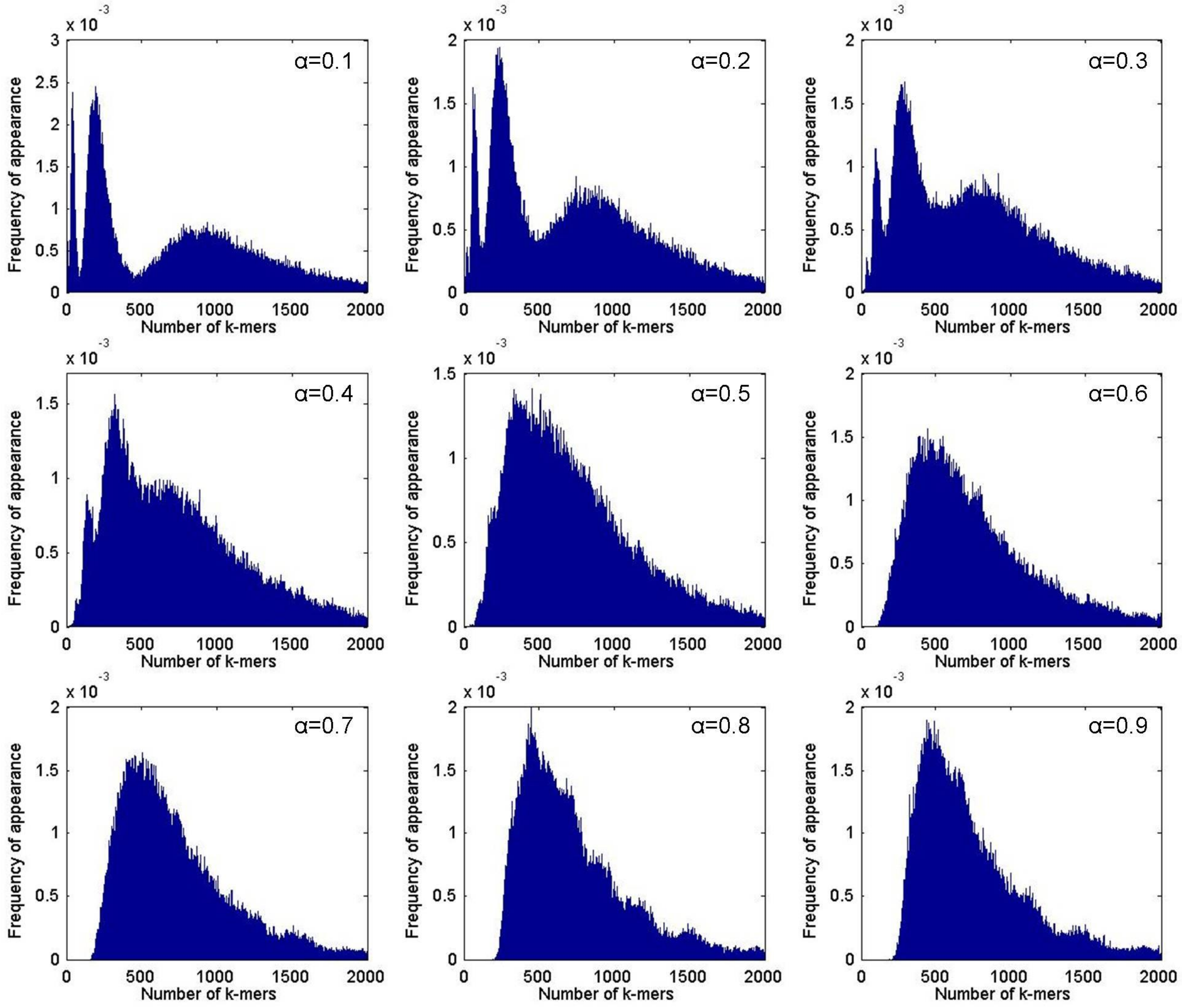

Figure 3.18: Simulated estimate of the border of the multi-modal distribution. Using a convex combination of the Markov model transition probabilities from one species to another. In this example we start at the human distribution ($\alpha = 0$) and finish at the mosquito distribution ($\alpha = 1$). We show the distributions for the varying values of α from $\alpha = 0.1$ to $\alpha = 0.9$ (k=9). In this case $\alpha_s = 0.3$, as the multi-modal distribution is clearly evident for $\alpha \leq 0.3$.

The convex combination was obtained using the following:

Let D capture the probabilities of all 16 possible dimers, with respect to species s.

$$\mathrm{D_s} = \mathrm{Pr}(\mathtt{ab}) \quad \mathtt{a}, \mathtt{b} \in \mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}$$

So now for each species s, and every value of $0 \leq \alpha \leq 1$, define the convex combination

$$\mathrm{D}'_{\alpha,\mathrm{s}} = (1-\alpha)\mathrm{D}_{\mathrm{human}} + \alpha\mathrm{D}_{\mathrm{s}}$$

We then plug these new dimer distribution values in to our Markov model simulator as the new transition probabilities, and run the simulation. We then visually decide whether the graph is "uni-modal" or "multi-modal", and thus define α_s as the border where for $\alpha \leq \alpha_s$ we will get a "multi-modal" distribution.

Figure 3.18 shows an example of the k-mer distribution graphs of the Markov model simulation from human ($\alpha = 0$) to mosquito ($\alpha = 1$). We used $k = 9$, and the length of each simulation run was $50,000,000$ bases.

3.9 Extended Chargaff Phenomenon

When analyzing single strand sequences, we saw a close similarity between the number of appearances of k-mers and their reverse-complements (this is obvious for two strands, but not for one). We note Chargaff's rules [2], and specifically the second rule.

Chargaff's 2^{nd} rule: $\%A \approx \%T$ and $\%C \approx \%G$ on single stranded DNA.

We observed that this rule can be extended to larger k-mers:

For almost all k-mers: # of occurrences of a k-mer $\approx$ # of occurrences of its reverse-complement-k-mer.

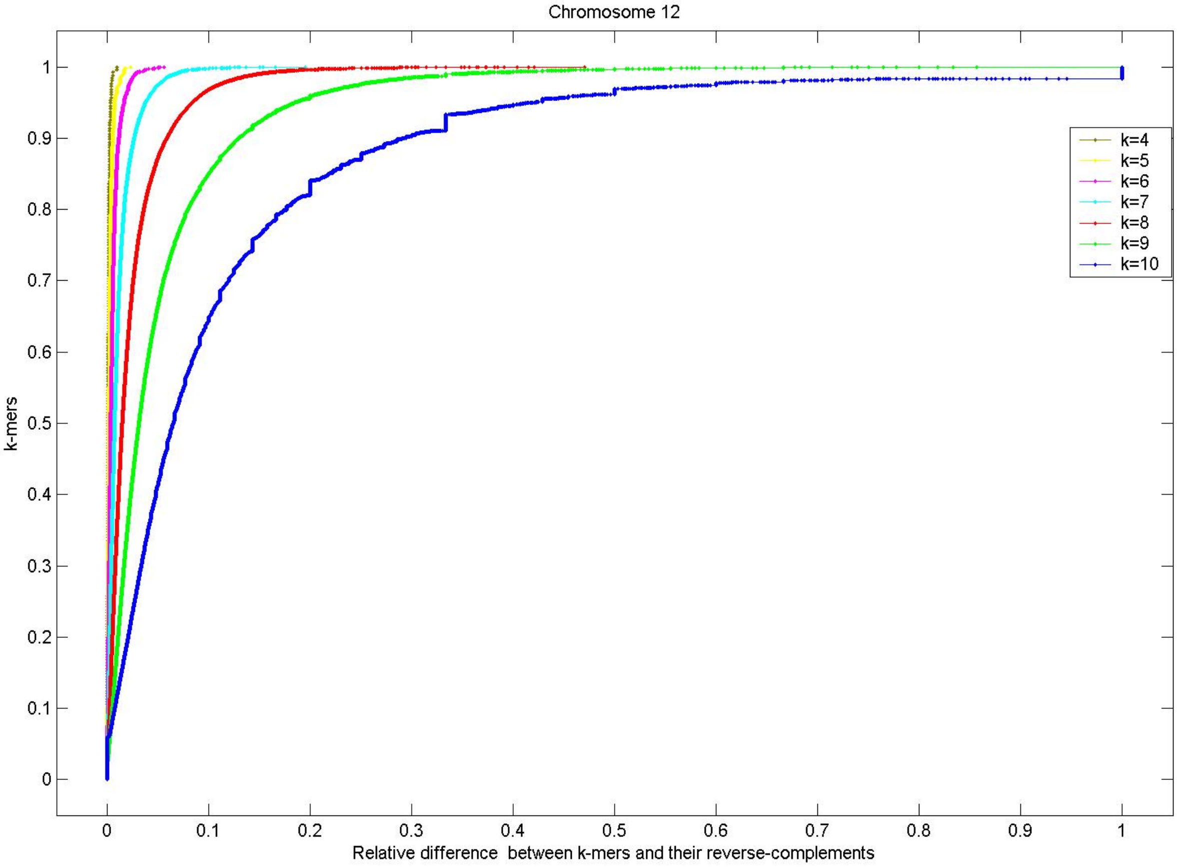

Figure 3.19: Extending Chargaff's 2^{nd} rule for k-mers, $4 \leq k \leq 10$.
k-mers and their reverse complements on a single strand, human chromosome 12, $k = 10$.

We define a simple score d_i for the relative difference between the number of occurrences of a k-mer i (N_i) and the number of occurrences of its reverse complement (N_i^{rc}):

$$\mathrm{d_i} = \left| \frac{\mathrm{N_i} - \mathrm{N_i^{rc}}}{\mathrm{N_i} + \mathrm{N_i^{rc}}} \right|$$

It is easy to see that $d = 0$ for k-mers which appear the exact same number of times as their reverse complements, and d increases with the relative difference between these occurrences.

Figure 3.19 shows this observation on the human chromosome 12, for $4 \leq k \leq 10$. The y-axis represents the k-mers sorted by their S score and normalized on a $[0..1]$ scale. The x-axis represents the score d between the k-mer and its reverse complement. It can be clearly seen that the vast majority of k-mers exhibit a very low relative difference score.

% of k-mers	k=4	k=5	k=6	k=7	k=8	k=9	k=10
10	0.000	0.000	0.001	0.001	0.002	0.005	0.007
30	0.000	0.001	0.002	0.004	0.008	0.017	0.034
50	0.000	0.002	0.003	0.007	0.015	0.031	0.065
70	0.001	0.003	0.006	0.012	0.026	0.056	0.120
85	0.002	0.004	0.009	0.020	0.045	0.100	0.217
90	0.002	0.006	0.011	0.027	0.059	0.130	0.294
95	0.003	0.007	0.017	0.037	0.084	0.185	0.417
99	0.006	0.016	0.027	0.068	0.148	0.333	0.999

Table 5: The relative difference score d at different percentiles of k-mers.

Table 5 provides some actual values as depicted in the figure. For example, if we look at $k = 8$, we can see that 90% of k-mers have a relative difference score of less than 0.006, and 99% have a relative difference score of 0.148 or less.

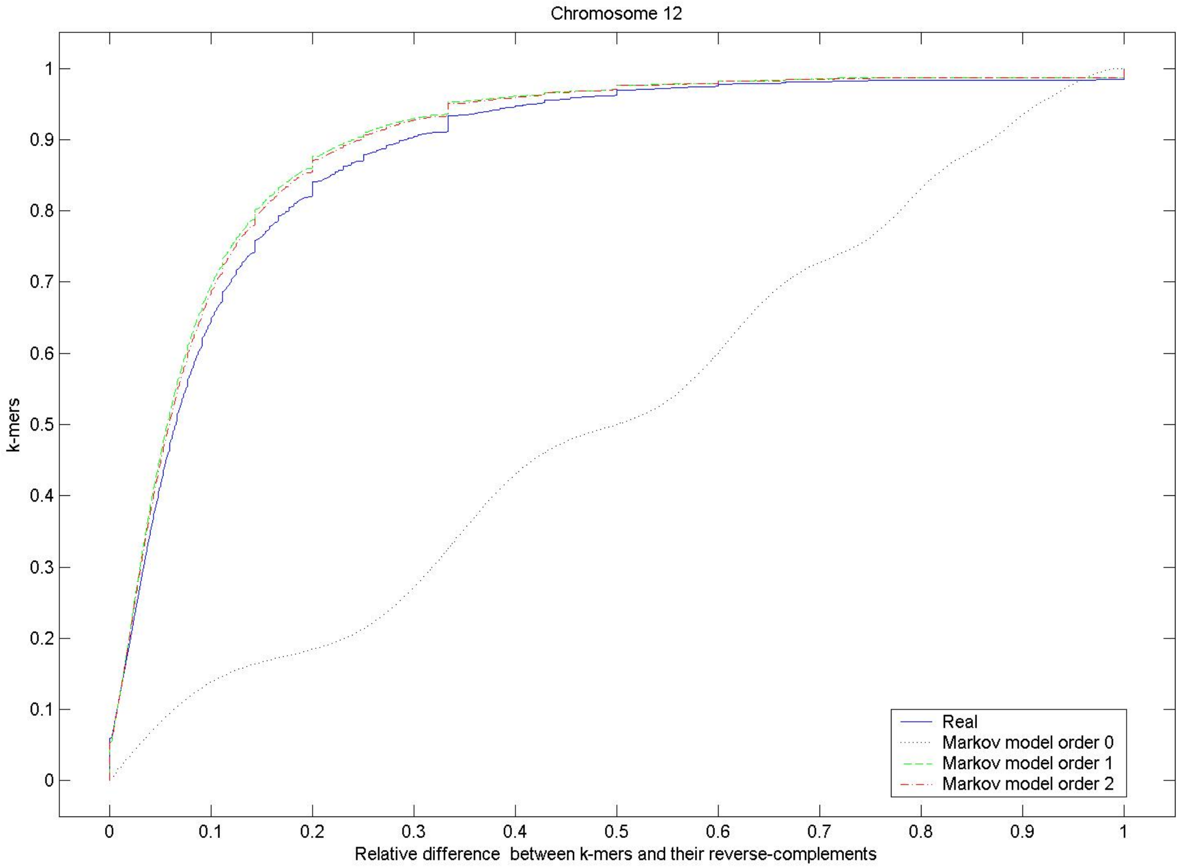

Figure 3.20: Low order Markov models preserve properties of extended Chargaff rule phenomena.
k-mers and their reverse complements on a single strand, human chromosome 12, $k = 10$. Real data (blue) versus the simulated models of order 0 (black dots), order 1 (green dashes) and order 2 (red dot-dashes).

Figure 3.20 shows that the low order (non-zero) Markov models preserve the properties of the extended Chargaff rule phenomena as well. The models of orders 0, 1 and 2 were based on the human chromosome 12.

4 Discussion

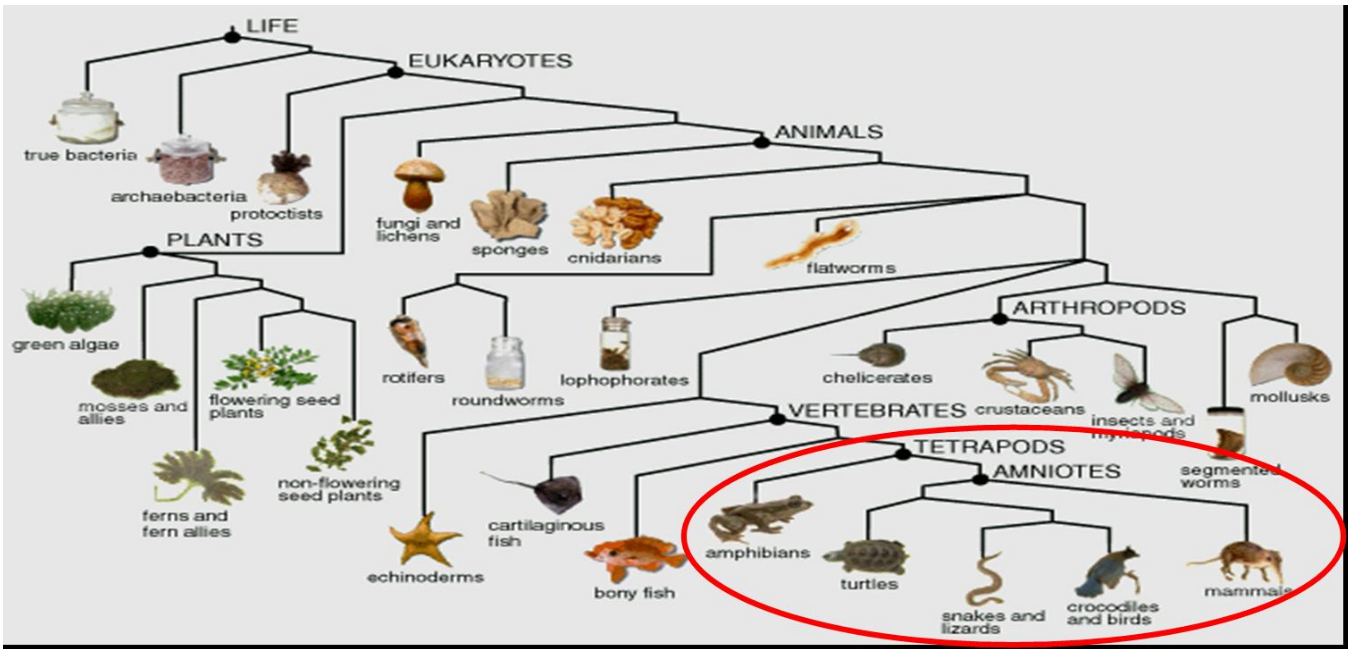

Figure 4.1: The tetrapod clade on a symbolic "tree of life".
Figure taken from http://biology.unm.edu/ccouncil/Biology_203/Summaries/Phylogeny.htm

When placing the species with multi-modal k-mer histograms on the "tree of life" [22], we observe that they define *exactly* the tetrapod clade (Figure 4.1). We note that mammals, birds, amphibians and reptiles (all analyzed here) compose the majority of the tetrapod groups. (Tetrapods are terrestrial vertebrates, and other vertebrates that bear limbs with digits rather than fins.) Another family of species within the tetrapods are the turtles, whose genomes are not yet sequenced, but we expect they will also exhibit the same multi-modal distribution. We note that bony fish are the outgroup to tetrapods within the vertebrates. As pointed out in Section 3, the five bony fish analyzed here exhibit unimodal k-mer distributions.

As mentioned earlier, Reinert *et al.* have shown that the distribution for the number of occurrences of a particular k-mer has two distinct large sample regimes: a normal distribution for abundant k-mers, and a Poisson (or compound Poisson) distribution for extremely rare k-mers [18]. Comparing their results to our empirical findings, we observe that the tetrapods' genomes exhibit k-mer distributions where a superposition of *both* such regimes are evident. The non-tetrapods exhibit just one of the two distributions.

A major cause of the low frequency of appearance of the `CpG` dimer (termed "`CpG` suppression" [13]) in tetrapods is methylation [3]. Methylation has far ranging effects, from embryonic development to neuron apoptosis and cancer. In humans, for example, more than 70% of the `C`s in `CpG`s are methylated, and these `CpG`s tends to mutate at greater rates, specifically to `TG` and to `CA`. The dimers `TG` and `CA` together account for most of the `CpG` loss: With no suppression, `CpG` in tetrapods is expected to appear at approximately 4% of the dimers, but in actuality it appears at approximately 1%, so approximately 3% are missing; TG and CA each appear about 1.3% more than their “fair share”. Together this accounts for about 2.6% of the dimers, covering about 90% of the CpG loss.

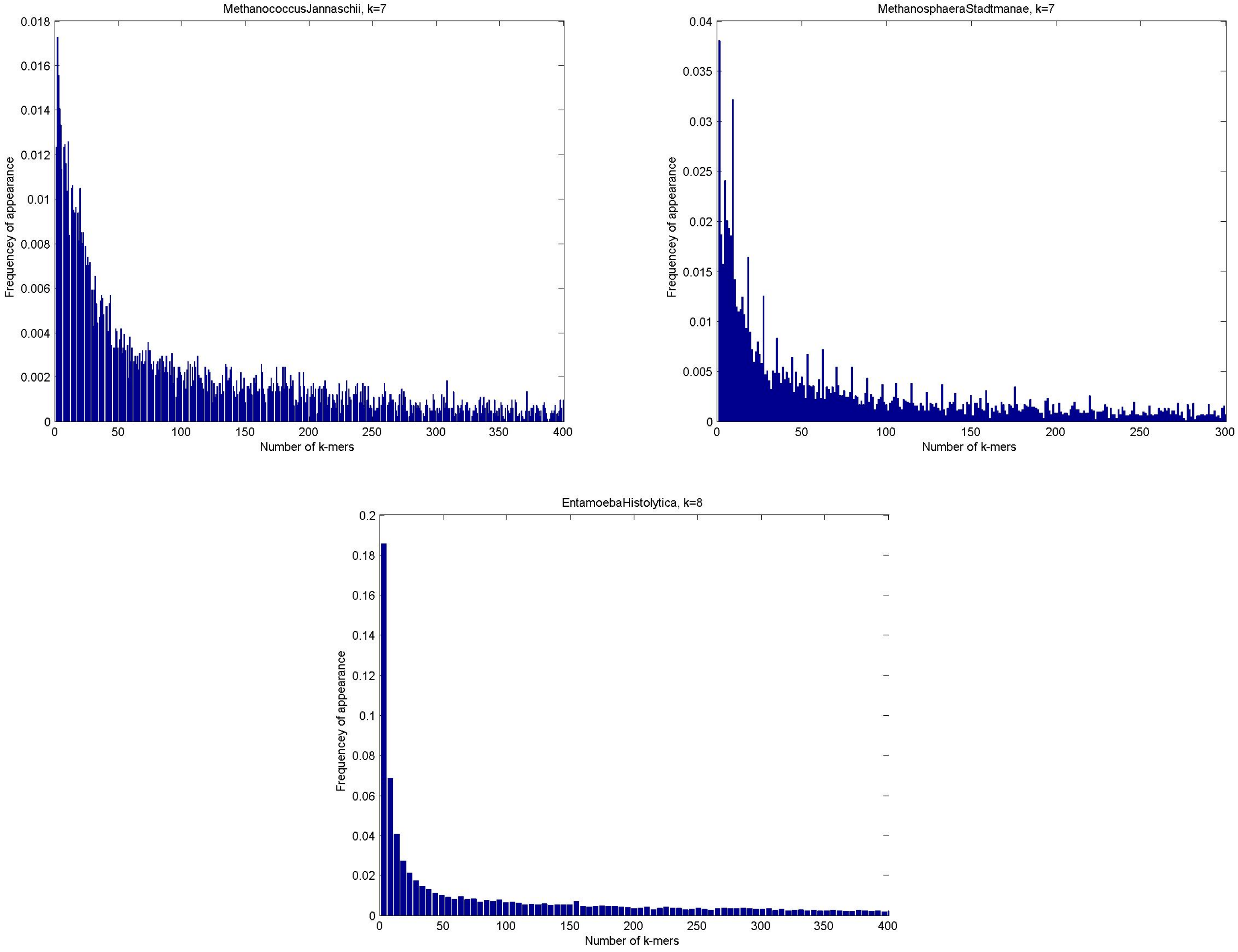

Figure 4.2: Species with unimodal distributions and low $\rho_{\mathtt{CG}}$ values.
k-mer distribution for Methanococcus Jannaschii (k=7, top left), Methanosphaera Stadtmanae ($k = 7$, top right), and Entamoeba Histolytica ($k = 8$, bottom).

Since all tetrapods exhibit a significant level of CpG suppression, one may wonder if their multi-modal k-mer distributions are not simply a direct consequence of low $\rho_{\mathtt{CG}}$ values. Clearly, our findings indicate that low $\rho_{\mathtt{CG}}$ values are a necessary ingredient in having multi-modal k-mer distributions. However, low $\rho_{\mathtt{CG}}$ values by themselves are not enough. Figure 4.2, reveals three non-tetrapods (two archea and one bacteria, with $\rho_{\mathtt{CG}}$ values smaller than 0.33, yet unimodal k-mer distributions. These three species have a substantially lower `C+G` contents than the tetrapods, and this also is likely to be an important factor.

More generally, even the first order Markov model has 20 parameters, and it is not clear how to determine which regimes give rise to each type of distributions, but clearly $\rho_{\texttt{CG}}$ and the `C+G` contents are important. Despite their limitations, low order models do reveal interesting features. For example, by allowing to measure how surprising a k-mer is, we get a quantitative measure that is more significant than just “missing” or “highly abundant”.

We believe that our study of empirical DNA distributions is only an initial step in this direction, and that additional insights are yet to be found.

5 Appendix A

5.1 Algorithms

The algorithms used in this research were fairly simple and straightforward.

5.1.1 Extracting k-mer data from a genome

This algorithm was used to extract all data for all k's in range $4 \leq k \leq 13$ from a genome data file, *in one pass.* The general scheme was to create for each k an array of length 4^k. Each entry in the array holds a counter for the number of appearances for a specific k-mer. The genome would then be read letter by letter, and the entry for the relevant k-mer encountered would be counted. Even though the procedure to do this is quite straightforward, there were some issues that needed some closer attention:

1. Files can reach significant sizes (the complete human genome, for example, is larger that 2GB). Loading a file like this to the RAM can fail, or at least slow down the machine. The solution used to this problem was to load the file to memory in a "chunks" of 100MB each, which essentially act as a buffer. This allowed to keep the RAM manageable, thus allowing the calculative parts of the algorithm to work faster.

2. How to efficiently map a k-mer to its designated place in the array. The solution to this was done by mapping each nucleotide to a 2-bit number ($\mathtt{A} \rightarrow 00, \mathtt{C} \rightarrow 01, \mathtt{G} \rightarrow 10, \mathtt{T} \rightarrow 11$), and thus mapping a k-letter string to an integer in the range 0 to $4^k - 1$. When a new letter is encountered, all that needs to be done is a 2-bit left shift of the previous number, and OR-ing the new letter's 2-bit value as the least significant bits (LSB). This also enabled easy calculation of double-strand (reverse-complement) data by switching the direction of the 2-bit shift to a right-shift, and OR-ing the complement 2-bit value as the most significant bits (MSB). This allows all operations on a single letter be done in O(1), enabling the total time be O(l), where l is the genome length. Total memory usage is O($\max(4^k, b)$), where b is the buffer size used to read the genome file.

3. Output files can reach large sizes. Assuming we want all extracted data written to a file, that file can reach large sizes. For example, for $k = 13$, the raw output file is almost 200MB. This is not too large, but loading and manipulating that kind of data for many species is not recommended. The solution was to partition the data into histogram-like bins, thus controlling the file sizes with the number of bins. Since we show histograms anyway, this was an easy solution to choose.

5.2 Genomic data acquiring

Genomic data of species was downloaded from the following sites:

- **Ensembl** ftp://ftp.ensembl.org
- **NCBI** ftp://ftp.ncbi.nih.gov
- **TIGR** ftp://ftp.tigr.org
- **UCSC** ftp://hgdownload.cse.ucsc.edu
- **HGSC** ftp://ftp.hgsc.bcm.tmc.edu
- **Genoscope** http://www.genoscope.cns.fr

Data regarding human genomic regions was downloaded from:

- **Exons, Introns** http://hsc.utoledo.edu/bioinfo/eid/index.html
- **3'UTRs, 5'UTRs** ftp://bighost.ba.itb.cnr.it/pub/Embnet/Database/UTR/data
- **Gene Promoters** http://www.epd.isb-sib.ch/seq_download.html
- **miRNA Data** ftp://ftp.sanger.ac.uk/

For some of the species the whole genomic data was not fully assembled, so we took what was available in the form of scaffolds and chromosomes. We assume that the final assemblies will probably not differ significantly from these preliminary files in terms of the k-mer statistics.

5.3 Species list

Species (Archea)	**Classification**	**Length**	**%G+C**	ρ_{CG}	**dist. type**
AeropyrumPernix	Archea	1,509,911	56.5	0.70	unimodal
ArchaeoglobusFulgidus	Archea	2,076,061	48.6	0.78	unimodal
HaloarculaMarismortui	Archea	3,208,489	62.0	1.32	unimodal
HalobacteriumSp	Archea	1,887,389	68.0	1.36	unimodal
HaloquadratumWalsbyi	Archea	2,925,353	47.9	1.11	unimodal
HyperthermusButylicus	Archea	1,509,911	53.8	0.77	unimodal
MethanobacteriumThermoautotrophicum	Archea	1,604,247	49.6	0.51	unimodal
MethanococcoidesBurtonii	Archea	2,453,539	40.8	0.72	unimodal
MethanococcusJannaschii	Archea	1,604,238	31.4	0.32	unimodal
MethanococcusMaripaludis	Archea	1,509,911	33.1	0.89	unimodal
MethanocorpusculumLabreanum	Archea	1,698,650	50.1	1.19	unimodal
MethanopyrusKandleri	Archea	1,604,247	61.2	1.18	unimodal
MethanosaetaThermophila	Archea	1,792,986	53.5	0.86	unimodal
MethanosarcinaAcetivorans	Archea	5,473,296	42.7	0.79	unimodal
MethanosarcinaBarkeri	Archea	4,624,004	39.3	0.72	unimodal
MethanosarcinaMazei	Archea	3,869,048	41.5	0.72	unimodal
MethanosphaeraStadtmanae	Archea	1,604,247	27.6	0.27	unimodal
MethanospirillumHungatei	Archea	3,302,831	45.1	0.77	unimodal
NanoarchaeumEquitans	Archea	377,477	31.1	0.61	unimodal
NatronomonasPharaonis	Archea	2,453,539	63.4	1.39	unimodal
PicrophilusTorridus	Archea	1,415,508	36.0	0.76	unimodal
PyrobaculumAerophilum	Archea	2,076,061	51.3	0.97	unimodal
PyrobaculumIslandicum	Archea	1,698,650	49.5	0.93	unimodal
PyrococcusAbyssi	Archea	1,604,247	44.7	0.71	unimodal
PyrococcusFuriosus	Archea	1,792,986	40.8	0.50	unimodal
PyrococcusHorikoshii	Archea	1,604,247	41.9	0.61	unimodal
SulfolobusAcidocaldarius	Archea	2,076,061	36.7	0.55	unimodal
SulfolobusSolfataricus	Archea	2,831,017	35.8	0.67	unimodal
SulfolobusTokodaii	Archea	2,547,942	32.8	0.55	unimodal
ThermococcusKodakaraensis	Archea	1,981,725	52.0	0.88	unimodal
ThermofilumPendens	Archea	1,698,650	57.7	1.00	unimodal
ThermoplasmaAcidophilum	Archea	1,415,508	45.9	0.91	unimodal
ThermoplasmaVolcanium	Archea	1,509,911	39.9	0.83	unimodal

Species (Bacteria)	**Classification**	**Length**	**%G+C**	ρ_{CG}	**dist. type**
AcidobacteriaBacteriumEllin345	Bacteria	5,378,893	58.4	1.27	unimodal
BacillusSubtilis	Bacteria	3,963,384	43.5	1.04	unimodal
BrucellaMelitensis	Bacteria	3,114,055	57.2	1.20	unimodal
BurkholderiaXenovoransLB400	Bacteria	9,247,928	62.6	1.38	unimodal
ChlamydophilaPneumoniaeAR39	Bacteria	1,132,366	40.6	0.73	unimodal
ChlorobiumTepidumTLS	Bacteria	1,981,718	56.5	1.21	unimodal
ChromobacteriumViolaceum	Bacteria	4,529,601	64.8	1.12	unimodal
CyanobacteriaBacteriumYellowstoneA-Prime	Bacteria	2,736,681	60.2	0.74	unimodal
EscherichiaColi536	Bacteria	4,718,340	50.5	1.14	unimodal
FrancisellaTularensisHolarctica	Bacteria	1,792,986	32.2	0.54	unimodal
GeobacterSulfurreducens	Bacteria	3,585,973	60.9	1.00	unimodal
HelicobacterHepaticus	Bacteria	1,698,650	35.9	0.70	unimodal
IdiomarinaLoihiensisL2TR	Bacteria	2,642,278	47.1	1.06	unimodal
LactobacillusPlantarum	Bacteria	3,114,092	44.5	1.12	unimodal
LegionellaPneumophilaLens	Bacteria	3,208,092	38.4	0.73	unimodal
MagnetococcusMC-1	Bacteria	4,435,265	54.2	0.80	unimodal
MarinobacterAquaeoleiVT8	Bacteria	4,529,611	56.9	0.94	unimodal
MycobacteriumTuberculosis	Bacteria	4,151,874	65.6	1.18	unimodal
NeisseriaMeningitidisFAM18	Bacteria	2,076,061	51.6	1.31	unimodal
NitrobacterWinogradskyiNb-255	Bacteria	3,208,495	62.1	1.33	unimodal
NostocSp	Bacteria	6,888,767	41.3	0.78	unimodal
PhotobacteriumProfundumSS9	Bacteria	6,039,402	41.8	0.99	unimodal
ProchlorococcusMarinusNATL2A	Bacteria	1,698,650	35.1	0.57	unimodal
RalstoniaEutrophaH16	Bacteria	6,605,672	66.6	1.13	unimodal
SaccharophagusDegradans2-40	Bacteria	4,812,743	45.8	1.08	unimodal
SalmonellaTyphi	Bacteria	4,907,083	51.9	1.23	unimodal
ShewanellaOneidensis	Bacteria	4,907,081	45.9	1.00	unimodal
ShigellaDysenteriae	Bacteria	4,340,901	51.0	1.13	unimodal
StreptococcusMutans	Bacteria	1,887,389	36.9	0.71	unimodal
ThermoanaerobacterTengcongensis	Bacteria	2,547,942	37.6	0.52	unimodal
VibrioCholerae	Bacteria	3,774,608	47.5	1.04	unimodal
ZymomonasMobilisZM4	Bacteria	1,887,389	46.2	1.10	unimodal
PlasmodiumFalciparum (malaria)	Protozoa	21,798,142	19.4	0.76	unimodal
TetrahymenaThermophila	Protozoa	98,669,430	22.3	0.44	unimodal
LeishmaniaMajor	Protozoa	31,116,157	59.7	1.02	unimodal
EntamoebaHistolytica	Protozoa	12,298,665	33.6	0.30	unimodal

Species (Eukaryotes)	**Classification**	**Length**	**%G+C**	ρ_{CG}	**dist. type**
CaenorhabditisElegans (worm)	Nematode	95,970,454	35.4	0.99	unimodal
SaccharomycesCerevisiae (yeast)	Fungi	11,512,682	38.3	0.80	unimodal
CandidaGlabrata (haploid yeast)	Fungi	11,700,869	38.6	0.66	unimodal
BiomphalariaGlabrata (mollusca)	Mollusk	48,845,378	39.1	0.75	unimodal
CionaIntestinalis (sea squirt)	Tunicate	83,180,064	35.6	0.86	unimodal
ArabidopsisThaliana (arabidopsis)	Plant	113,813,113	36.0	0.72	unimodal
OryzaSativa (rice)	Plant	385,822,552	43.5	0.87	unimodal
VitisVinifera (grape)	Plant	468,739,222	34.6	0.43	unimodal
DrosophilaMelanogaster (fly)	Insect	115,134,437	42.4	0.93	unimodal
AnophelesGambiae (mosquito)	Insect	213,534,127	44.6	1.07	unimodal
ApisMellifera (bee)	Insect	175,355,887	34.9	1.64	unimodal
TriboliumCastaneum (beetle)	Insect	145,486,340	33.9	1.15	unimodal
DanioRerio (zebrafish)	Bony Fish	996,230,784	36.3	0.52	unimodal
TakifuguRubripes (fugu)	Bony Fish	329,961,080	45.5	0.57	unimodal
TetraodonNigroviridis (pufferfish)	Bony Fish	177,300,831	45.9	0.60	unimodal
GasterosteusAculeatus (stickleback)	Bony Fish	424,233,346	44.6	0.66	unimodal
OryziasLatipes (Japanese Medaka)	Bony Fish	552,716,066	40.1	0.48	unimodal
AnolisCarolinensis (lizard)	Reptile	1,676,035,836	40.4	0.30	multi-modal
XenopusTropicalis (frog)	Amphibian	1,288,438,558	40.0	0.34	multi-modal
GallusGallus (chicken)	Bird	942,474,046	41.3	0.24	multi-modal
OrnithorhynchusAnatinus (platypus)	Mammal	388,840,627	43.3	0.30	multi-modal
BosTaurus (cow)	Mammal	1,408,036,795	42.8	0.25	multi-modal
CanisFamiliaris (dog)	Mammal	2,187,364,344	41.1	0.26	multi-modal
RattusNorvegicus (rat)	Mammal	2,327,934,313	42.1	0.22	multi-modal
MacacaMulatta (rhesus monkey)	Mammal	2,503,126,668	40.9	0.25	multi-modal
MonodelphisDomestica (opossum)	Mammal	3,287,333,976	37.6	0.13	multi-modal
MusMusculus (mouse)	Mammal	2,387,461,979	41.9	0.19	multi-modal
PanTroglodytes (chimpanzee)	Mammal	2,596,334,645	40.8	0.23	multi-modal
HomoSapiens (human)	Mammal	2,735,501,651	40.9	0.24	multi-modal
HumanIntrons (human regions)		1,368,981,774	41.5	0.24	multi-modal
Human3’UTR (human regions)		30,577,457	44.4	0.29	multi-modal
HumanExons (human regions)		108,962,293	49.9	0.44	unimodal
Human5’UTR (human regions)		10,160,999	55.4	0.60	unimodal
HumanPromoters600 (human regions)		1,025,512	58.8	0.74	unimodal
HumanPromoters1000 (human regions)		1,762,896	53.7	0.64	unimodal
HumanPromoters5000 (human regions)		8,753,016	47.3	0.39	multi-modal
HumanMiRNA1000 (human regions)		494,332	48.3	0.38	multi-modal
HumanMiRNA5000 (human regions)		2,594,039	47.2	0.35	multi-modal
HumanMiRNA10000 (human regions)		5,193,844	46.5	0.33	multi-modal
HumanChr12 (human regions)		108,585,772	40.8	0.24	multi-modal
HumanChr5RepeatMasked (human regions)		74,495,291	38.6	0.21	multi-modal

References

[1] K. P. Burnham, and D. R. Anderson (1998) *Model selection and inference: a practical information-theoretic approach.* Springer, New York.

[2] E. Chargaff (1951), Structure and function of nucleic acids as cell constituents. *Federal Proceedings*, Sep; 10(3):654-9

[3] D. N. Cooper and M. Krawczak (1989), Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. *Human Genetics*, Vol. 83, No. 2, September 1989.

[4] M. Csuros, L. Noe, and G. Kucherov (2007). Reconsidering the significance of genomic word frequencies. *Trends in Genetics*, 23(11):543-546

[5] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison (1998) *Biological sequence analysis: probabilistic models of proteins and nucleotides.* Cambridge University Press, Cambridge.

[6] Ensembl: http://www.ensembl.org/index.html

[7] Y. Fofanov *et al.* (2004), How independent are the appearances of n-mers in different genomes?. *Bioinformatics*, 20(15):2421-2428.

[8] G. Hampikian and T. Andersen (2007), Absent sequences: Nullomers and primes. *Pacific Symposium on Biocomputing*, 12:355-366.

[9] C. Hunte, E. Screpanti, M. Venturi, A. Rimon, E. Padan, H. Michel (2005), Structure of a Na+/H+ antiporter and insights into mechanism of action and regulation by pH. *Nature*, 435:1197-202.

[10] S. Karlin and J. Mra'zek (1997). Compositional differences within and between eukaryotic genomes. *Proc Natl Acad Sci, USA*, 94(19): 10227-10232.

[11] J. Lee, Z. Li, R. Brower-Sinning, B. John (2007), Regulatory Circuit of Human MicroRNA Biogenesis. *PLoS Computational Biology*, Apr 20; 3(4):e67

[12] X. Liu, D. L. Brutlag, and J. S. Liu (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.* 6:127–38.

[13] A.L. Mazin and B.F. Vanyushin (1987), CpG-suppression in DNA. 1. Methylated and nonmethylated compartments of genome in eukaryotes with different 5-methylcytosine content in DNA. *Mol. Biol.*, 1987, vol.21, no.2, pp. 543–551.

[14] C. Narasimhan, P. LoCascio, and E. Uberbacher (2003) Background rareness-based iterative multiple sequence alignment algorithm for regulatory element detection. *Bioinformatics* 19:1952–63.

[15] J. R. Norris (1997) *Markov chains.* Cambridge University Press, Cambridge.

[16] J. Otaki, S. Ienaka, T. Gotoh, and H. Yamamoto (2005), Availability of short amino acid sequences in proteins. *Protein Science*, 14:617-625.

[17] W. J. Reed (2004), The Double Pareto-Lognormal Distribution A New Parametric Model for Size Distributions. *Com.Stats Theory & Method* Vol. 33, No. 8, 1733–1753

[18] G. Reinert, S. Schbath, and M. Waterman (2000) Probabilistic and statistical properties of words: an overview. *J. Comp. Biol.* 7:1–46.

[19] S. Robin and S. Schbath (2001) Comparison of Several Approximations of the Word Count Distribution in Random Sequences. *J. Comp. Biol.* 8:349–359.

[20] A. F. A. Smit, R. Hubley, and P. Green. (1996-2004) RepeatMasker Open-3.0. http://www.repeatmasker.org

[21] R. R. Sokal and F. J. Rohlf (1995), Biometry (3rd Ed.), W. H. Freeman and Company, New York.

[22] Tree of Life: http://tolweb.org/tree/

[23] T. Tuller, B. Chor, and N. Nelson (2007). Forbidden Penta-Peptides. Accepted to *Protein Science.*