Title: Reconsidering the significance of genomic word frequencies

Short title: Genomic word frequencies

Authors: Miklós Csürös (1), Laurent Noé (2), and Gregory Kucherov (2)

Author affiliations:

(1) Department of Computer Science and Operations Research,
Université de Montréal, Québec, Canada.
CP 6128, succ. Centre-Ville, Montréal, Québec H3C 3J7, Canada;

(2) Laboratoire d'Informatique Fondamentale de Lille,
Bât. M3, 59655 Villeneuve d'Ascq Cédex, France.

Corresponding author: Miklós Csürös.
Tel: +1 (514) 343-6111 extension 1655,
Fax: +1 (514) 343-5834,
E-mail: csuros@iro.umontreal.ca.

Abstract

By conventional wisdom, a feature that occurs too often or too rarely in a genome can indicate a functional element. To infer functionality from frequency, it is crucial to precisely characterize occurrences in neutrally evolving DNA. We find that the frequency of oligonucleotides in a genomic sequence follows primarily a Pareto-lognormal distribution, which encapsulates lognormal and power-law features found across all known genomes. Such a distribution may be the result of completely random evolution by a copying process. Our characterization of the entire frequency distribution of genomic words opens a way to a more accurate reasoning about their over- and under-representation in genomic sequences.

Introduction

Determining what constitutes the surprisingly frequent and rare in a genome is a fundamental and ongoing issue in genomics [5]. Sequence motifs might be unusually rare or frequent because they belong to mobile, structural or regulatory elements, and are thus subject to selective and adaptive forces. After examining oligonucleotide occurrences in more than sixty diverse genomes, we found that a Pareto-lognormal distribution captures the crucial features of oligonucleotide frequency distributions in all the studied genomes. While prevailing random sequence models fail to produce such features, a neutral model of random duplications can. We illustrate our claim with a completely random copy-and-paste process that induces a distribution similar to those observed in real-life sequences.

Random sequence models

The simplest sequence motif is an oligonucleotide, or DNA *word*. A definite word frequency distribution that characterizes a neutrally evolving sequence is necessary to establish whether a word appears unusually often or rarely in a genome sequence. For instance, the statistical significance of the (hypothetical) overrepresentation of a word w is routinely measured by the *tail probability* $\boldsymbol{P}\{N(w) \geq n\}$, where n is the number of times w occurs in the studied sequence, and $N(w)$ is the random number of occurrences in a null model. In this context we focus on the distribution of frequencies across all words of the same length ℓ (i.e., ℓ*-mers*), called the *word frequency distribution* or *spectrum*.

Standard null models in bioinformatics are random texts [15], including Bernoulli and Markov models (see Glossary). Random text models imply a rapidly decreasing tail in the spectrum of

short words (typically, $8 \le \ell \le 16$). In particular, they imply that the number of oligonucleotides that occur the same number of times decreases exponentially with the number of occurrences. In reality, genomic word frequencies exhibit no such behavior (Figure 1). Depending on the genome and word lengths, the spectrum can show a power-law decrease on the right, a lognormal shape, or even a power-law tail on the left. Such features are at odds with random text models (see Figure 1A and Supplementary Material). As a consequence, random text models tend to underestimate the probability of frequent and rare words in long sequences.

We examined whether combining a random text model with the mosaic-like variation of cytosine-guanine content [1] explains the shape of spectra (for vertebrates, at least). Localized random shuffling, which preserves the landscape of cytosine-guanine variation, also produces a light tail (Figure 1A), and is thus not a substitute for an adequate null model.

The distribution of oligonucleotide frequencies

To date, genomic spectra have not been fully characterized, aside from the observation of power-law behavior for certain word sizes [7-9] in the right-hand tail. Here we point out that a parametric distribution describes word frequencies extremely well. The distribution in question is the so-called *double Pareto-lognormal* (DPL) distribution [14]. The DPL distribution fits many real-life size distributions, including that of personal incomes, human settlements, and files on the Internet [11]. It has four parameters: α, β, ν and τ; its density function has a power-law (Pareto) tail to the left and to the right, with slopes of $(\beta\text{-}1)$ and $(-\alpha\ -1)$ on a log-log scale, respectively; in the middle, its shape is dominated by a lognormal distribution with parameters ν and τ.

Figure 1B illustrates that the four parameters of a single DPL distribution can be adjusted to describe hundreds or thousands of word frequencies in non-vertebrate genomes. (More examples are given as Supplementary Material.) We initially sought to characterize vertebrate genomes. To our surprise, we found that in spite of considerable differences in the organization, composition and the structure of the genetic material between organisms, the DPL distribution applies to genomes from all domains of cellular life, and thus represents a universal genomic feature.

The fitted distribution's parameters reflect some idiosyncrasies of the genome at hand. For instance, the upper power-law tail, which reflects genome repetitiveness, is generally steeper in prokaryotes than in eukaryotes. The prevalence of frequent words in vertebrate genomes cannot be entirely attributed to mobile elements, as the contribution of non-repeat regions diminishes very slowly and does not vanish when moving toward higher frequencies. On human chromosome 12, for example, about 25% of very frequent 12-mers occur in non-repeat regions (see Supplementary Material for detailed analysis). Furthermore, frequent words are plentiful even in repeat-masked vertebrate and many non-vertebrate genomes (Figures 1B and D). Figures 1C and D illustrate vertebrate spectra using the example of chromosome 12, which is representative of the genome with respect to repeat element distribution and cytosine-guanine content [17].

Random evolution by duplication

Why would the DPL distribution systematically appear in genomic spectra? The answer may well lie in duplicative processes. The power-law tail of protein domain and gene family size distributions [7] can be explained by birth and death models [4,13], in which family size changes by duplication and deletion processes, and new families are introduced by a steady innovation

process. A similar model applies to genomic word frequencies. Consider a particular word's occurrences along the genome as a "family." The family size is affected by mutational events, including duplications, insertions, deletions and point mutations. The family can increase by any copying mechanism, including genomic duplication and retrotranscription. The family decreases if a mutation destroys an occurrence. Point mutations can create new words, but so can insertions (at the insertion boundaries) and deletions (by fusing two halves of a word). A neutral model equipped with constant-rate duplication, deletion, and mutation processes thus corresponds to a birth and death model of gene families. In order to illustrate how power-law features arise in a neutral duplication model, we carried out a simulation experiment in which a DNA sequence evolved solely by a "copy-and-paste" mechanism. We iteratively expanded an initial random Bernoulli sequence, by selecting a contiguous piece of a fixed length *m* in every iteration, and copying it back into the sequence at a random position. While this procedure may seem surprisingly simple (perhaps even too abstract), it is, in fact, quite effective at achieving a similar spectrum to real-life sequences (Figure 1A).

As another sign of the importance of duplications, we note the association between heavy-tail distributions and long-range autocorrelation, which are tokens of self-similarity. Long-range autocorrelation at the single nucleotide level was observed before (see [3] for a review), and it was shown that it could result from so-called expansion-randomization processes [10], which model sequence evolution by deletions, mutations and duplications of single nucleotides.

Practical implications

As we have just suggested, the birth and death model implies that some words occur often simply by chance, and not because of their functionality. Words that are abundant at an early point of

evolution tend to stay frequent in the course of random events. Therefore, even the high frequency of a particular word across many related species does not imply functionality on its own, as the word might have been frequent by chance in a common ancestor already.

The success of computational sequence analysis hinges on adequate criteria for unusual word frequencies in a wide range of applications, including identification of regulatory elements [2] and repeat families [12], whole-genome assembly [18] and homology search [6]. Random text models can cause many false signals, as they imply the statistical concentration of empirical word frequencies.

An example of underestimating the probability of frequent word occurrences is apparent in a recent study by Rigoutsos et al [16]. They reported that certain DNA words, termed *pyknons*, appear frequently in human gene-related sequences andin noncoding regions, in restricted configurations, and presented many arguments for the pyknon's functionality. By relying on a Bernoulli model, they reasoned that 16-mers should appear in a random genome sequence more than forty times with a probability $<10^{-32}$. Such a word frequency, however, is not as extraordinary if we take into account the universal shape of genomic spectra. A DPL distribution fitted to the human genome spectrum yields a *P*-value of 0.001 (see Supplementary Material).This latter translates to about four million 16-mers that are expected to occur at least forty times in a random genome-sized sequence. Strikingly, at least 460 thousand frequent words appear already in the repeat-masked sequence as accidental constituents of the fitted distribution's heavy tail.

Conclusion

Word frequencies bear witness to a long history of evolutionary tinkering: copying, deleting, and changing different parts of the genome. We argue that global features of genomic spectra arise from duplicative evolutionary processes, and not necessarily from intricate word-level selection on point mutations and deletions that are enacting adaptation and conservation, or simply obeying structural constraints. In practice, the heavy tail of word frequency distributions means that caution should be exercised when inferring functionality of motifs from frequency alone, especially if overrepresentation is related to word occurrences in random texts. Our investigations reveal the suitability of a simple Pareto-lognormal distribution for the statistical assessment of unusual word frequencies.

Acknowledgement

This project was supported by an NSERC grant.

References

[1] Bernardi, G. *et al*. (1985) The mosaic genome of warmblooded vertebrates. *Science*, 228:953-958.

[2] Brazma, A. *et al*. (1998) Predicting gene regulatory elements in silico on a genomic scale. *Genome Res*., 8:1202-1215

[3] Buldyrev, S. V. (2006) Power law correlations in DNA sequences. In *Power Laws, Scale-Free Networks and Genome Biology* (Koonin, E. V., *et al*., eds.), pp. 123-164. Landes Bioscience.

[4] Karev, G. P. *et al*. (2002) Birth and death of protein domains: a simple model of evolution explains power law behavior. *BMC Evol. Biol.*, 2:18.

[5] Karlin, S. (2005) Statistical signals in bioinformatics. *Proc. Natl. Acad. Sci. USA*, 102:13355-13362.

[6] Kent, W. J. (2002) BLAT --- the BLAST-like alignment tool. *Genome Res.*, 12:656-664.

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

[8] Mantegna, R. N. *et al*. (1995) Systematic analysis of coding and noncoding DNA sequences using methods of statistical lingustics. *Physical Review E*, 52:2939-2950.

[9] Martindale, C. and Konopka, A. K. (1996) Oligonucleotide frequencies in DNA follow a Yule distribution. *Comput. Chem.*, 20:35-38.

[10] Messer, P. W. *et al*. (2005) Universality of long-range correlations in expansion-randomization systems. *Journal of Statistical Mechanics: Theory and Experiment*, P10004. DOI: 10.1088/1742-5468/2005/10/P10004

[11] Mitzenmacher, M. (2004) Dynamic models for file sizes and double Pareto distributions. *Internet Mathematics*, 1:303-333.

[12] Morgulis, A. *et al*. (2006) WindowMasker: window-based masker for sequenced genomes. *Bioinformatics*, 22:134-141.

[13] Reed, W. J. and Hughes, B. D. (2004) A model explaining the size distribution of gene families. *Math. Biosci.*, 189:97-102.

[14] Reed, W. J. and Jorgensen, M. (2004) The double Pareto-lognormal distribution - a new parametric model for size distributions. *Communications in Statistics: Theory and Methods*, 33:1733-1753.

[15] Reinert, G. *et al*. (2000) Probabilistic and statistical properties of words: An overview. *J. Comput. Biol.*, 7:1-46.

[16] Rigoutsos, I. *et al*. (2006) Short blocks from the noncoding parts of the human genome have instances within nearly all known genes and relate to biological processes. *Proc. Natl. Acad. Sci. USA*, 103:6605-6610.

[17] Scherer, S. E. *et al*. (2006) The finished DNA sequence of human chromosome 12. *Nature*, 440:346-351.

[18] Wang, J. *et al*. (2002) RePS: A sequence assembler that masks exact repeats identified from the shotgun data. *Genome Res.*, 12:824-831.

[19] Clay, O. (2001) Standard deviations and correlations of GC levels in DNA sequences. *Gene*, 276:33-38.

[20] Messer, P. W. *et al*. (2006) Alignment statistics for long-range correlated genomic sequences. *Lecture Notes in Computer Science*, 3909: 426-440.

[21] Waterman, M.S. (1995) *Introduction to Computational Molecular Biology: Maps, Sequences and Genomes*. Chapman & Hall.

Figure legend

Genomic spectra and fitted DPL distributions. The ordinate plots the number of words that occur n times, for each n shown on the X-axis. For each spectrum, dots show the ℓ-mer frequency distribution, and a solid line traces the fitted double Pareto-lognormal (DPL) distribution.

(A) 13-mer frequencies in repeat-masked human chromosome 5 ("real"), and in random sequences of the same length (Bernoulli and a first-order Markov model). The spectrum of a shuffled sequence is also plotted, which was produced by randomly garbling the nucleotides within windows of length 1000 to preserve large-scale heterogeneity. A verisimilar word frequency distribution is achieved by random "copy-and-paste" of 33-mers. The procedure started with a Bernoulli sequence of 5000 random nucleotides with 38.5% guanine-cytosine content, matching the composition of chromosome 5.

(B) Some smaller spectra. Notice the lower power-law tail in the *B. subtilis* genome.

(C) 9-mer spectra of repeat-masked human chromosome 12. In organisms with strong dinucleotide bias, such as for CpG in vertebrates, the spectrum can be decomposed into multiple DPL distributions by dinucleotide content. By grouping the words according to the number of non-overlapping CpG dinucleotides in them, frequencies in each group follow a DPL distribution.

(D) CpG-free ℓ-mers on repeat-masked chromosome 12. Notice the transition from a lognormal to a power-law shape as the word length increases.

Glossary

Bernoulli model

The simplest random sequence model is the Bernoulli, or "coin-flip" model. Each nucleotide of the sequence is chosen independently, by the same background nucleotide probabilities p(A), p(C), p(G) and p(T). Accordingly, a DNA word $w=w_1w_2\ldots w_\ell$ occurs in a given sequence position with probability $p(w)=p(w_1)p(w_2)\ldots p(w_\ell)$. For a long random sequence, the number of occurrences $N(w)$ can be approximated [21] by a Poisson distribution with parameter $L\,p(w)$. Consequently, the ℓ-mer spectrum of a Bernoulli sequence follows a mixture of Poisson distributions, with one Poisson distribution for each possible value of $p(w)$.

Markov model

Markov models capture compositional biases present at the level of very short oligonucleotides. In this model,a random sequence is generated by a k-th order Markov chain. In other words, each random nucleotide depends on the k preceding nucleotides so that dinucleotide bias, for example, can be represented with k=1. Mathematically, the model is defined by the probabilities $p(a \mid u)$ where $a\in\{\mathrm{A,C,G,T}\}$ and u takes values in the set of k-mers. Depending on the relationship between L and ℓ, the tail probabilities may be approximated by a Poisson or Gaussian distribution [15], which imply exponentially small values for $\mathbf{P}\{N(w)\geq n\}$. Notice that the model has $3\cdot 4^k$ independent parameters $p(a \mid u)$.

Power laws and heavy tails

The term ``power law'' applies to any function $f(t)$ which is essentially polynomial, i.e., $f(t) \approx c \cdot t^k$ with some constants c and k. In the context of probabilities, an upper power-law tail means mathematically that the tail probability $p_{\geq t} = \mathbf{P}\{X \geq t\}$ is asymptotically proportional to $1/t^\alpha$ with some constant $\alpha > 1$ as $t \to \infty$. Conversely, X has a lower power-law tail if $p_{\leq t} = \mathbf{P}\{X \leq t\} \sim t^\beta$ for some $\beta > 0$ as $t \to 0$.

A power-law tail is "heavy" in the sense that $\log(p_{\geq t}) \approx -\alpha \log t$, and thus the same tail probability is reached at much larger t values than in light-tailed distributions, such as Gaussian and Poisson, where $\log(p_{\geq t}) \approx -\text{poly}(t)$ for some polynomial of t. Random quantities with light-tailed distributions have a typical magnitude, where all observations are concentrated, whereas heavy-tailed distributions span several orders of magnitude, and have no obvious ``typical'' value.

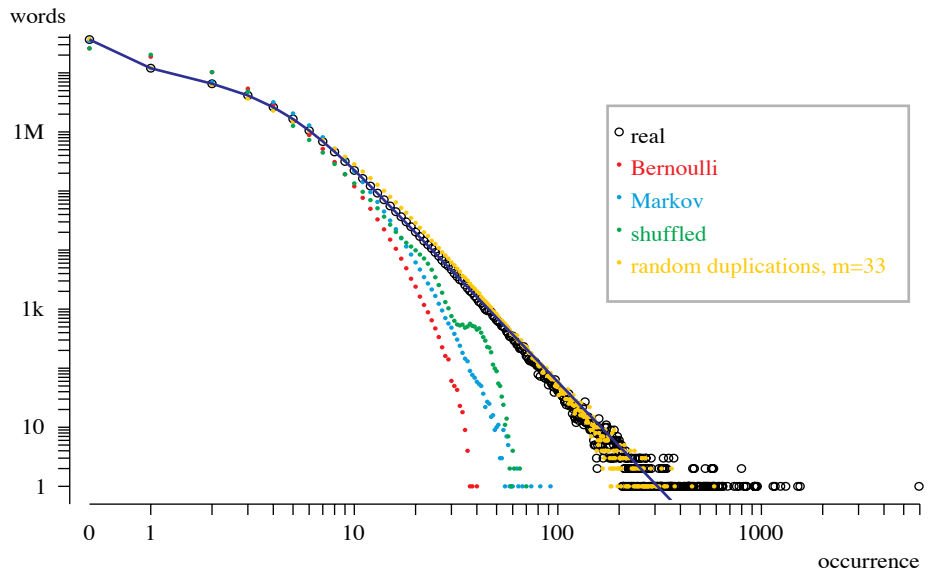
Long-range autocorrelation

Autocorrelation of a sequence is measured by the function

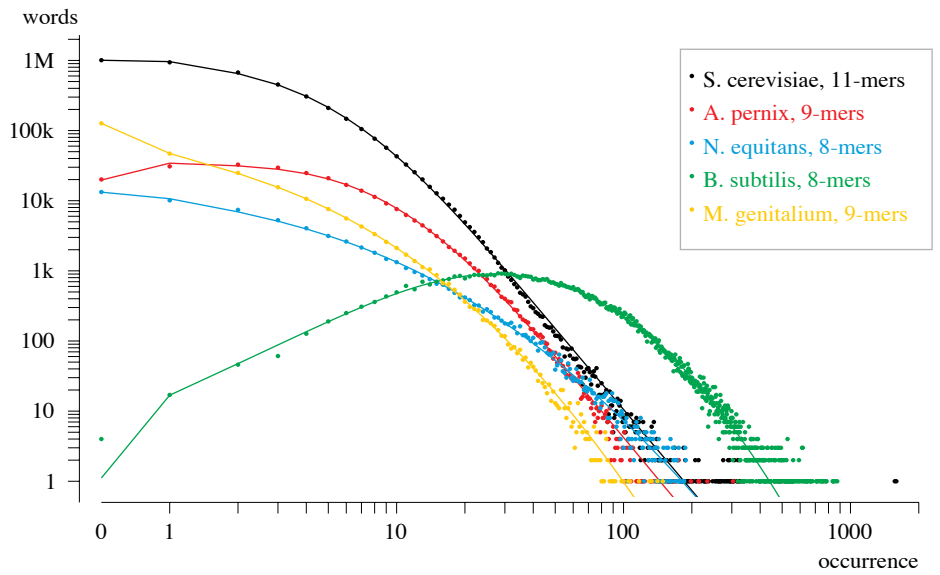
$$f(r) = \sum_{a=\text{A,C,G,T}} \left((L-r)^{-1} \sum_i g_i(a) g_{i+r}(a) - L^{-1} \sum_i (g_i(a))^2 \right)$$

where $g_i(a) = 1$ if the sequence has the nucleotide a in position i, otherwise $g_i(a) = 0$. In a Bernoulli model, the expected value of $f(r)$ is zero; in Markov models, it decays exponentially fast with r. Autocorrelation in genome sequences has a long range, as it follows a power law. It has been argued that long-range autocorrelation affects the statistical significance in homology

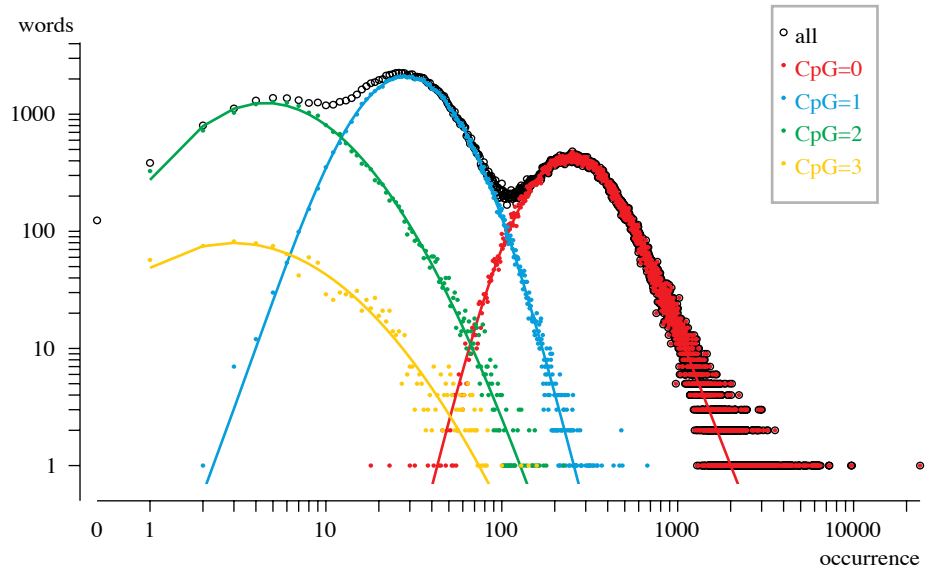
search programs such as BLAST [20], and that it should be taken into account in isochore segmentation [19].


words
1M
1k
10
1
0
1
10
100
1000
occurrence
real
Bernoulli
Markov
shuffled
random duplications, m=33

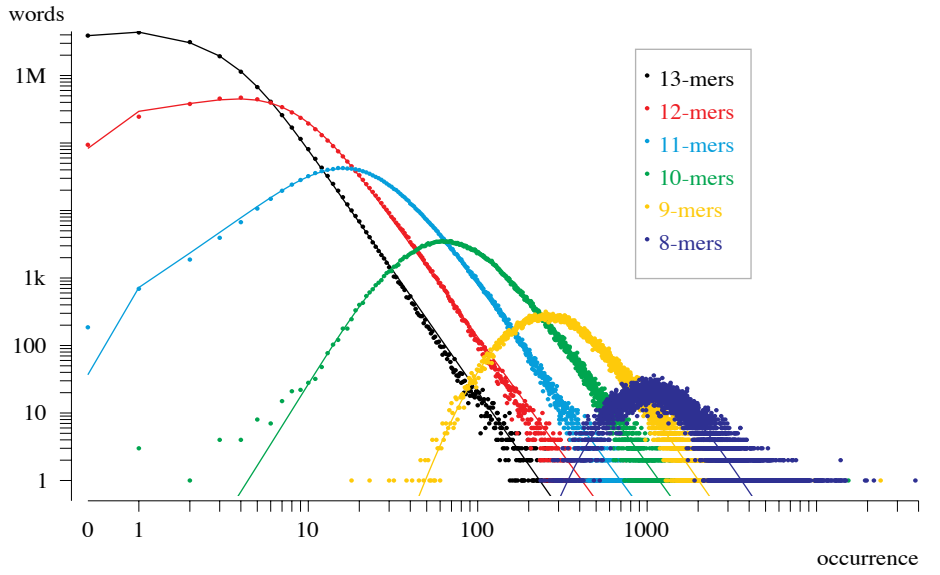
A


words
1M
100k
10k
1k
100
10
1
0
1
10
100
1000
occurrence
S. cerevisiae, 11-mers
A. pernix, 9-mers
N. equitans, 8-mers
B. subtilis, 8-mers
M. genitalium, 9-mers

B


words
1000
100
10
1
0
1
10
100
1000
10000
occurrence
all
CpG=0
CpG=1
CpG=2
CpG=3

C


words
1M
1k
100
10
1
0
1
10
100
1000
occurrence
13-mers
12-mers
11-mers
10-mers
9-mers
8-mers

D

Online Supplementary Material
Reconsidering the significance of genomic word frequencies

Miklós Csűrös*† Laurent Noé‡ Gregory Kucherov‡

1 Methods

Words were counted only on one strand of the DNA sequences (the 'plus' strand of the sequence file — counting on both strands gives similar results), with the exception of the 16-mers in the human genome, where both strands were scanned. We counted the occurrence of a word w if it appeared in a given sequence at some position $i..i+\ell-1$, without ambiguous nucleotides. The DPL distribution was fitted using its cumulative distribution function (cdf), which is

$$F(x) = \Phi\Big(\frac{\ln x - \nu}{\tau}\Big) + \frac{\alpha}{\alpha+\beta} x^{\beta} e^{-\beta\nu+\beta^2\tau^2/2} \Phi\Big(-\frac{\ln x - \nu + \beta\tau^2}{\tau}\Big) - \frac{\beta}{\alpha+\beta} x^{-\alpha} e^{\alpha\nu+\alpha^2\tau^2/2} \Phi\Big(\frac{\ln x - \nu - \alpha\tau^2}{\tau}\Big)$$

for $x > 0$ and $F(x) = 0$ for $x \le 0$, where $\Phi(\cdot)$ denotes the cdf of the standard normal distribution. The spectrum consists of the numbers $W(n)$ of ℓ-mers occurring exactly n times for all $n = 0, 1, 2, \ldots$ In order to fit

*Corresponding author.

†Department of Computer Science and Operations Research, Université de Montréal, CP 6128, succ. Centre-Ville, Montréal, Québec H3C 3J7, Canada.

‡Laboratoire d'Informatique Fondamentale de Lille, Bât. M3, 59655 Villeneuve d'Ascq Cédex, France.

the distribution's parameters, the spectrum $\Big(W(n)\colon n = 0, 1 \ldots\Big)$ was considered as a set of binned values for independently drawn samples from a continuous DPL distribution: $W(n)$ was compared to the predicted value $4^{\ell}\Big(F(n+\frac{1}{2}) - F(n-\frac{1}{2})\Big)$. We used custom-made programs to carry out the parameter fitting, using the Levenberg-Marquardt algorithm [7], a nonlinear least-squares method, for which the starting parameter values were set by likelihood maximization [8].

We defined CpG content of a word w as the number of non-overlapping CG and GC dinucleotides in w.

Human sequences (original and repeat-masked) and repeat annotations were obtained from the UCSC genome browser [2] gateway's FTP server (`ftp://hgdownload.cse.ucsc.edu/`), for version hg18 (NCBI Build 36.1). (The repeat annotations were generated by the programs RepeatMasker [11] and Tandem Repeats Finder [1].) Other sequences were downloaded from the NCBI FTP server (`ftp://ftp.ncbi.nlm.nih.gov/genomes/`).

The random sequences of Figure 1A have the same length as the repeat-masked chromosome sequence (or, more precisely, the same number of 13-mers). The k-order Markov models were constructed by counting $(k+1)$-mers in the repeat-masked sequence, and setting the transition probabilities $p(a|u_{1\ldots k}) = \frac{N(u_1u_2\cdots u_k a)}{\sum_b N(u_1u_2\cdots u_k b)}$ where $N(w)$ is the number of occurrences of the $(k+1)$-mer w. For the random shuffling, we partitioned the sequence into contiguous segments containing exactly 1000 non-ambiguous nucleotides. Non-ambiguous nucleotides were garbled in each segment by generating a uniform random permutation. The random copy-paste evolution was performed by generating an initial Bernoulli sequence of length 5000nt, with 38.5% GC-content as for the chromosome sequence. In each iteration, (1) a uniform random position was picked for the starting position of the copied sequence, and (2) an independent uniform random position was picked for the point where the copied sequence is to be inserted. The Java programs (source and bytecode) that generated the random sequences can be obtained from the corresponding author, or downloaded directly from the webpage `http://www.iro.umontreal.ca/~csuros/spectrum/`.

2 Contribution of repeats in the spectrum's tail

Figure 1 and Table 1 illustrate the contribution of repeats to the spectrum. The contribution of different annotations were computed by multiplying each $W(n)$ value in the spectrum by the fraction of occurrences within the annotated regions for words appearing n times in the entire sequence.

n	words (a)	seq (b)	nonrep (c)	SINE (d)	LINE (e)	LTR (f)	other (g)
12	19.6	69.6	47.2	16.7	21.4	8.4	6.3
25	5.3	39.7	41.5	21.3	22.6	7.6	7.0
30	3.7	33.9	39.6	23.1	22.9	7.3	7.2
50	1.3	22.6	34.0	28.8	23.2	6.5	7.5
100	0.4	14.7	27.9	36.7	22.6	5.4	7.4
200	0.1	10.6	24.6	43.0	20.7	4.7	7.0

Table 1: Composition of the 12-mer spectrum's tail in human chromosome 12: (a) fraction of words that occur at least n times; column; (b) fraction of the genome sequence covered by such words; (c–g) fraction of occurrences within non-repeat regions, short interspersed elements, long interspersed elements, long terminal repeats, and other repeat elements (including DNA transposons, simple repeats, low-complexity and tandem repeats), respectively. Fractions are expressed as percentages.

3 16-mer spectrum of the human genome

We counted 16-mers in the forward and reverse strands of the human genome sequence, NCBI version 36.1. We fitted a DPL distribution to word occurrences between 0 and 100 for the unmasked sequence, and ignored the shape of the upper tail consisting of words occurring more than 100 times, since it is determined by the mixture of repeat elements. The DPL curve has parameters $\alpha = 1.988, \beta = 0.209, \nu = 1.08, \tau = 0.528$, which corresponds to a tail probability $\mathbb{P}\{N(w) \geq 40\} = 9.2 \cdot 10^{-4}$. We found that other parameter settings also provide a reasonable fit to the spectrum, corresponding to tail probabilities between $8.6 \cdot 10^{-4}$ and $1.0 \cdot 10^{-3}$. Figure 2 shows the 16-mer spectrum of the repeat-masked sequence. The fitted DPL distribution has parameters $\alpha = 2.625, \beta = 0.191, \nu = 0.828, \tau = 0.556$, giving

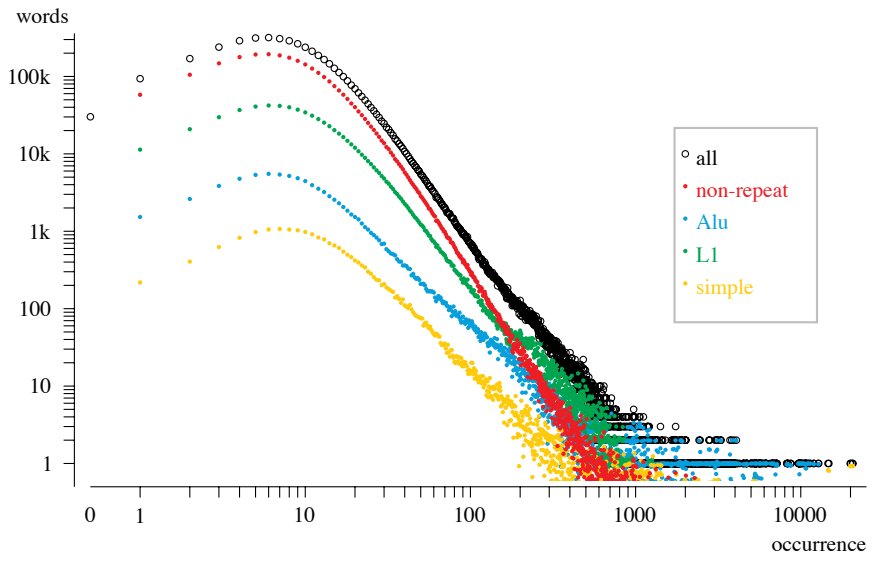


Figure 1: Contribution of different repeat families to the full spectrum of CpG-free 12-mers along chromosome 12. Abundant repeat elements may cause deviations from the distribution, which may be the basis of their identification using a DPL null model, but they are often absorbed in the fundamental curve.

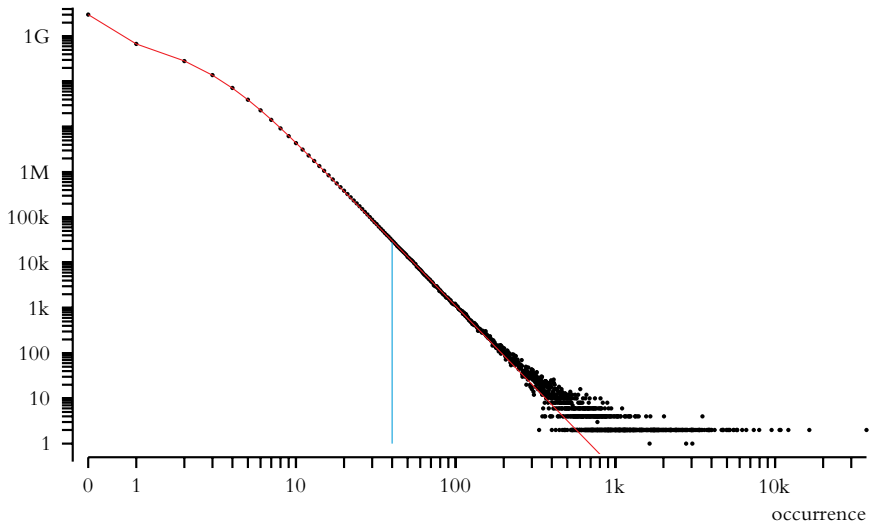


Figure 2: 16-mer spectrum of the repeat-masked human genome. The solid curve shows the fitted DPL distribution. The vertical line indicates the cutoff of forty occurrences chosen by Rigoutsos et al.

$\mathbb{P}\{N(w) \geq 40\} = 1.08 \cdot 10^{-4}$. Since there are 4^{16} 16-mers, the expected number of words occurring at least 40 times is $4^{16} \cdot 1.08 \cdot 10^{-4} \approx 464000$.

4 Fit of Markov models

Markov models of random sequences are often used to predict word frequencies. Markov and Bernoulli models capture the word distribution for all word lengths simultaneously, but the models are not necessarily more compact in practice than a few DPL distributions, if the goal is to recognize unusual word occurrences. A third-order Markov model has $3 \cdot 4^3 = 192$ independent parameters, and, thus, uses almost 20 parameters for each practically interesting word length (say, between six and fifteen). A DPL model is then five times simpler with four parameters per word length distribution. A second-order Markov model routinely used to capture the codon distribution in a prokaryotic genome uses $3 \cdot 4^2 = 48$ parameters, which is equivalent to twelve DPL distributions in complexity. Even third- to seventh-order Markov models are employed in practice [9, 3, 12, 10, 4, 6, 5] in order to provide P-values

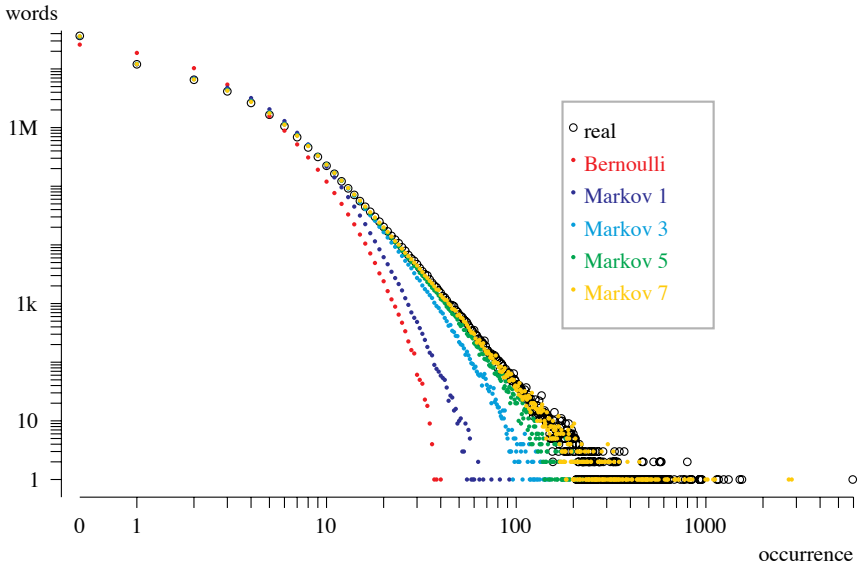


Figure 3: Spectra of Markov models. The models were estimated from the repeat-masked sequence of human chromosome 5. The plots compare the 13-mer spectra of the Markov models to that of the true sequence.

for word overrepresentation.

Figure 3 illustrates how much Markov models fail to capture the word frequency distributions for lengths above their order, despite a substantial number of parameters. Notice that it is necessary to use a seventh-order Markov model (with 49152 independent parameters) to predict 13-mer frequencies.

References

[1] G. Benson. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, 27(2):573–580, 1999.

[2] A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, and W. J. Kent. The

UCSC genome browser database: update 2006. *Nucleic Acids Res.*, 34:D590–598, 2006.

[3] X. Liu, D. B. Brutlag, and J. S. Liu. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pacific Symposium on Biocomputing*, 6:127–138, 2001. PMID: 11262934.

[4] C. Narasimhan, P. LoCascio, and E. Uberbacher. Background rareness-based iterative multiple alignment algorithm for regulatory element detection. *Bioinformatics*, 19(15):1952–1963, 2003.

[5] G. Pavesi, F. Zambelli, and G. Pesole. WeederH: an algorithm for finding conserved regulatory motifs and regions in homologous sequences. *BMC Bioinformatics*, 8:46, 2007. doi:10.1186/1471-2105-8-46.

[6] A. A. Pilippakis, G. S. He, and M. L. Bulyk. ModuleFinder: a tool for computational discovery of *cis* regulatory modules. *Pacific Symposium on Biocomputing*, 10:519–530, 2005.

[7] W. H. Press, S. A. Teukolsky, W. V. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge UniversIty Press, second edition, 1997.

[8] W. J. Reed and M. Jorgensen. The double Pareto-lognormal distribution — a new parametric model for size distributions. *Communications in Statistics: Theory and Methods*, 33(8):1733–1753, 2004.

[9] S. Scherer, M. S. McPeek, and T. P. Speed. Atypical regions in large genomic sequences. *Proc. Natl. Acad. Sci. USA*, 91:7134–7138, 1994.

[10] S. Sinha and M. Tompa. Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.*, 30(24):5549–5560, 2002.

[11] A. F. A. Smit, R. Hubley, and P. Green. Repeatmasker open-3.0, 1996–2004. `http://www.repeatmasker.org`.

[12] G. Thijs, M. Lescot, K. Marchal, S. Rombauts, B. De Moor, P. Rouzé, and Y. Moreau. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, 17(12):1113–1122, 2001.