Short subsequences in genomes: How random are they?

Yuriy Fofanov,[1] *Yi Luo,*[1] *Charles Katili,*[1] *Jim Wang,*[1] *Yuri Y. Belosludtsev,*[3] *Thomas F. Powdrill,*[3] *Viacheslav Fofanov,*[1] *Tong-Bin Li,*[1] *Sergey Chumakov,*[1,4] *and B. Montgomery Pettitt*[1,2]

[1]Department of Computer Science University of Houston, Houston, Texas
[3]Vitruvius Biosciences, The Woodlands, Texas
[2]Department of Chemistry University of Houston, Houston, Texas
[4]Department of Physics, University of Guadalajara, Guadalajara, Mexico

Corresponding author: Dr. Y. Fofanov
Department of Computer Science
The University of Houston
4800 Calhoun Road
Houston, Texas 77204-3010
Tel: 713-743-8553
Fax: 713-743-1250
Email: yfofanov@uh.edu

Running title: Randomness of short subsequences in genomes
Keywords: genomes, subsequence, motifs, n-mers, statistical properties

Abstract

A comparative statistical analysis of the presence of all possible short subsequences of length 5 to 20 nucleotides in the genomes of more than 250 microbial, viral and multicellular organisms was performed. A remarkable similarity of the presence/absence distributions for different *n*-mers in all genomes was found. The same analysis applied analytically and numerically to random sequences also shows a similar shape of the distribution, yielding the random boundary, with differences that correlate with biology. We hypothesize that the presence/absence distribution of *n*-mers in all genomes considered (provided that the condition $M<<4^n$ holds, where M is the total genome sequence length) can be treated as nearly random. The relative deviation of the frequency of presence of *n*-mers from the purely random distribution can be used as a measure of "non-randomness" or self-similarity of a genome. Our results indicate that larger genomes are often less random than shorter ones.

There is supplementary material. Accession number requested.

Introduction

Statistical analysis of the appearance of short subsequences in different DNA sequences, from individual genes to full genomes is important for various reasons. Applications include PCR primer (Fislage 1998; Fislage et al. 1997) and microarray probe design (Southern 2001). Several attempts (Deschavanne et al. 1999; Karlin and Ladunga 1994; Karlin and Mrazek 1997; Nakashima et al. 1997; Nakashima et al. 1998; Nussinov 1984; Sandberg et al. 2001) have been made to employ the frequency distribution of short subsequences (*n*-mers) to identify species with relatively short genome sizes (microbial). In such an approach, the shape of the frequency distribution for certain short subsequences: 2-4-mers (Deschavanne et al. 1999; Karlin and Ladunga 1994; Karlin and Mrazek 1997; Nakashima et al. 1997; Nakashima et al. 1998; Nussinov 1984) and 8-9-mers (Deschavanne et al. 1999; Sandberg et al. 2001) have been used to decide what microbial genome one is dealing with, based on a given piece of genome or a whole genome.

Many sequencing projects are in progress and more full genomes have recently become available. The several hundred projects completed so far provide sufficient material to consider them from a statistical viewpoint. Yet, we are still far from having a complete or even reasonable statistical picture. There are simply too many species and variations yet to be sequenced.

Here we present the results of the comparative statistical analysis of the presence/absence of all possible *n*-mers (n=5-20) for all genomes available (before May 2002) in the NCBI [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome], including microbial (76 genomes), viral (176 genomes), and five genomes of multicellular organisms. Let us stress that we do not consider the number of appearances of *n*-mers in a genome

(frequency of appearance), but just the information whether the given *n*-mer is present or absent (frequency of presence) in a given genome.

It is well-known that when genome size $M > 4^n$, the appearance of *n*-mers in various genomes are not random (Karlin and Ladunga 1994; Karlin and Mrazek 1997; Nakashima et al. 1997; Nakashima et al. 1998; Nussinov 1984). The basic motivation of our analysis is to explore the statistical properties of the presence of longer *n*-mers if the condition $M<<4^n$ is held. There are several reasons by which one could expect that the distributions of presence of longer *n*-mers are also not random. First, genomes (especially large ones) contain structural repeats. Second, since the occurrence statistics for short oligonucleotides (2- and 3-mers) is not random, this affects the occurrence distributions for longer *n*-mers, since they contain 2- and 3-mers as structural elements. However, our analysis of more than 250 genomes of microbial, viral and multicellular organisms shows that the distributions of presence in the range $M<<4^n$ remains nearly random or at least contain a strong random component.

Results

Microbial and viral genomes.

We have calculated the number of all distinct 7 - 15 -mers present in each of the viral and microbial genomes. Tables 1 and 2 contain representative results for some of the analyzed genomes (microbial and viral), for n = 8 and 12. Complete tables including all of the 252 genomes can be found on a supplementary data website (http://www.bioinfo.uh.edu/publications/how_random_are_genomes/). It is worth mentioning that as n increases, the total number of possible n-mers, 4^n, strongly exceeds the total sequence length M and most of the possible n-mers do not appear at all because the maximum number of n-mers contained in this sequence is M-n+1 ≈ M . Moreover, for a reasonably high ratio, $4^n/M$ **,** most of the n-mers which appear tend to appear only once, in accordance with the fact that the number of present n-mers becomes very close to M (see Tables 1,2 and supplementary data). That is why we have chosen to use the statistics for "present/absent" (frequency of presence) in our analysis instead of the usual "frequency of appearance", which is reasonable for short n-mers (total sequence length $M > 4^n$). We give precise definitions of these quantities in the Appendix.

We now consider the results obtained for different n-mers in the various genomes. We plot the frequency of presence, f, of n-mers in genomes (the number of different n-mers present in a given genome over the total number of n-mers, 4^n) against the ratio $4^n/M$. Figures 1-3 correspond to the microbial, RNA containing viruses and DNA containing viruses, respectively. The analytical distribution that corresponds to the frequency of presence of n-mers in a purely random "genome" (see Appendix) is also shown for comparison in all figures. Note the extraordinary similarity between these plots. All of the

different genomes form a well-defined pattern, when plotted against the ratio $4^n/M$ and not against the size of the genome or the length of the *n*-mer separately.

Multicellular organisms.

For much longer genomes of multicellular organisms practically all *n*-mers for $n < 12$ are present. Therefore, we have calculated the number of distinct 13 - 20 -mers present in each genome. The results are shown in Figure 4 and Table 3. In addition to that, we performed the same calculation for each human chromosome separately (see Figure 5 and Table 4). Note that the well-pronounced pattern can be observed in all these figures. It is noteworthy that multicellular organisms, especially rice and human, demonstrate much higher systematic deviation from the random boundary.

Discussion

A very similar rough shape of the dependences in Figures 1-5 can be observed. This remarkable similarity leads us to the hypothesis that the frequency f of presence/absence of relatively long n-mers ($M < 4^n$) can be treated as a result of a random process, or at least may contain a strong random component. This assumption motivated us to perform the following Monte Carlo simulation and analytical analysis.

We generated 100,000 random sequences of varying length M (from M=1Kb to M=10Mb), and applied to them the same analysis as for real genomes. We considered two cases: First, we used equal probabilities, p_i, of appearance of every nucleotide ($p_a = p_c = p_t = p_g = 0.25$) to generate random sequences. Second, to make our random sequences closer to real genomes, we calculated probabilities for each nucleotide in the three groups (see supplementary data) of genomes mentioned above (microbial, DNA viruses and RNA viruses) and also used them for our simulations. It turns out that the difference between these two simulations is negligibly small. This is, in fact, natural for actual probabilities that are close to 0.25; namely, for all cases, $0.22 < p_i < 0.29$. The results of the simulations fit the real data remarkably well.

In fact, the frequencies of presence of n-mers, f, in various genomes nearly belong to the same universal curve representing the random boundary (always being below it). The analytical derivation for this curve can be found in the Appendix. Assuming equal probabilities of appearance of every nucleotide, we have (in full agreement with the Monte Carlo simulations),

$$f_0 = 1 - \exp\left(-\frac{1}{x}\right), \qquad x = \frac{4^n}{M}, \qquad 1)$$

where, f_0 is the frequency of presence of n-mer in a random sequence of length M, x is the ratio of the total number of possible n-mers to the number of n-mers in the sequence in consideration, M-n+1 ≈ M. Equation 1 defines the analytical form of the above-mentioned "random boundary". It is shown in all our Figures 1-5 as a solid line.

The relative deviation

$$D = 1 - \frac{f}{f_0} \qquad 2)$$

of real results from the random boundary can be used as a definition of "non-randomness", or "self-similarity" of a given genome. Corresponding data for several genomes are given in Tables 1-4 and in supplementary data. It can be observed that shorter genomes are more random (based on this definition) than long ones. To quantify this statement we may compare the deviation from the random boundary, D, for n-mers corresponding to certain reasonable range of x=$4^n/M$. Indeed, when x <<1 (i.e. 4^n<<M) practically all of the n-mers are present in genomes. On the other hand, when x>>1 (M<<4^n) non-random processes may play an important role. For instance, some repetitions of intermediate length subsequences may appear. For example, if we want to see f_0 between 15-40%, the appropriate value of x would be between 2 and 6. It corresponds to 7-8-mers for RNA viruses, 8-10-mers for DNA viruses, and 11-12-mers for microbials. For large genomes this range of n can vary in accordance to the different sizes: *Homo sapiens* (16-17), *Drosophila melanogaster* (14-15), *Oryza sativa* (rice) (15-16), *Schizosaccharomyces pombe*(12-13), *Caenorhabditis elegans* (14-15).

The average value of the relative deviation, D, for 128 RNA viruses calculated for 7-8 mers is 7.6%, for 48 DNA viruses for 8-10 -mers it is 12.6%, and for 76 microbial genomes for 11-12 -mers it is 29.2%. It is worth mentioning that a few genomes show unusually high

self-similarity, such as *Simian Human immunodeficiency virus* (RNA): 47.8% for 7-mers, *Melanoplus sanguinipes entomopoxvirus* (DNA): 66.9% for 9-mers, *Mycoplasma pulmonis* (Bacteria): 56.0% for 10-mers. For five large genomes considered, the most "non-random" behavior (the largest self-similarity) is demonstrated by human genome: 50.2% for 16-mers and rice genome: 45.9% for 15-mers. The next is *C. elegans*: 40.5% for 14-mers. The most "random" is *Drosophila*: 24.8% for 14-mers.

We also considered the deviation from the random boundary for all of the 24 human chromosomes separately, which have average *D*-values of 40.4% for 13 -14 mers. The least random is the Y chromosome at 50.5% for 13-mers. (For complete details see supplementary data.)

Human and rice genomes are located especially far from the random boundary, which is in agreement with the presence of a significant number of structural repeats in these genomes. All three classes of *n*-mers (all *n*-mers, *n*-mers present once and n-mers present more that once) for human genome are shown in Table 5. One can observe that, when *n* grows, the fraction of over-represented *n*-mers rapidly decreases comparing to the fraction of *n*-mers present only once. For instance, 4.19% of the total number of 18-mers are present once and only 1.00% on the total number of 18-mers are present more than once.

We have not found an example when the frequency of presence is different from the random boundary by more than 67%. Theoretically, there may exist genomes with the frequency of presence curve for intermediate length sequences above the random boundary, however we have never observed them. It is worth mentioning that another possible explanation of high self-similarity of large genomes could be the occurrence of errors during the sequence assembly procedure used for obtaining the "complete" genomes.

Conclusion

Comparative statistical analysis of presence of all possible short subsequences (*n*-mers) for more than 250 complete microbial, viruses and multicellular organism genomes has been performed. To the best of our knowledge, no such analysis has been carried out before for $n>11$. Unlike the previous studies, we concentrated on the distribution of the frequency of presence/absence of all possible *n*-mers disregarding the information of how many times a given *n*-mer appears in a given genome. Beforehand, one could expect that the frequency of presence of all possible *n*-mers is a significantly non-random characteristic. However, our results point to the conclusion that the presence of *n*-mers in all genomes considered (in the range of *n*, when the condition $M<<4^n$ holds) can be treated as a nearly random process.

We find remarkable similarity of presence/absence distributions for different *n*-mers in all genomes studied so far. These distributions are found to be near the "random boundary" defined analytically and numerically. This universal behavior is intriguing in a variety of biological contexts.

Such a unique property of genomes leads to several practical applications. For example, relatively small random subset on *n*-mers of particular size can be placed on the DNA microarray and used for fast estimation of the genome size of unknown organisms. Furthermore, if future research reveals similarity between the presence/absence statistics if *n*-mers in coding and noncoding regions of genomes, such DNA microarray can be employed to estimate the size of transcriptome (the expressed part of the genome) under different circumstances.

The self-similarity (D) was found to be between 0 and 0.67 for all 250+ genomes examined, and closer to 0 for shorter genomes. This indicates that larger genomes are less

random than shorter ones. In our opinion this has interesting implications in terms of evolution and the complexity of the genomes.

Methods

For our analysis we have picked 76 complete microbial genome sequences with sizes ranging from 0.58 Mb to 7.04 Mb and 176 viral genomes (128 RNA containing viruses with genome sizes from 0.32 Kb to 130.76 Kb and 48 DNA containing viruses with genome sizes from 2.0 Kb to 671.19 Kb). We also used the genomes of five multicellular organisms: *Caenorhabditis elegans* (99.99 Mb), *Drosophila melanogaster* (119.98 Mb), *Oryza sativa* (Rice, 255.87 Mb), *Schizosaccharomyces pombe* (12.49 Mb), and *Homo sapiens* (human, 2.875 Gb) genomes. See supplementary data available from http://www.bioinfo.uh.edu/publications/how_random_are_genomes/.

We compared the frequencies of presence/absence of each n-mer in each of the genomes for $5 \le n \le 20$. To our knowledge, no such studies have been performed for $n>11$ due to the rapid growth of computational complexity with traditional algorithms.

To be able to perform calculations for longer ($n>11$) n-mers new algorithms and specific data structures (such as *counting arrays* (Fofanov et al. 2002a) and *incomplete search trees* (Fofanov et al. 2002b)) were utilized. The principal advantage of our approach is its time and memory efficiency, since only n-mers that are present in a genome under consideration (but not all possible 4^n n-mers) are taken into account. This approach also provides an efficient way to store sequences for later use. See http://bioinfo.uh.edu/publications/ for more details.

For our computations with multicellular genomes, microbial genomes and viral genomes, we used both complementary sequences (concatenating the sequence with its inverted complementary sequence). This apparent redundancy does not affect the statistical outcome and allows us to simplify the analysis.

References

Deschavanne, P.J., A. Giron, J. Vilain, G. Fagot, and B. Fertil. 1999. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol* **16:** 1391-1399.

Fislage, R. 1998. Differential display approach to quantitation of environmental stimuli on bacterial gene expression. *Electrophoresis* **19:** 613-616.

Fislage, R., M. Berceanu, Y. Humboldt, M. Wendt, and H. Oberender. 1997. Primer design for a prokaryotic differential display RT-PCR. *Nucleic Acids Res* **25:** 1830-1835.

Fofanov, V., Y. Fofanov, and B.M. Pettitt. 2002a. Counting array algorithms for the problem of finding appearances of all possible patterns of size n in a sequence. In *The 2002 Bioinformatics Symposium, Keck/GCC Bioinformatics Consortium*, pp. 14.

Fofanov, V., Y. Fofanov, and B.M. Pettitt. 2002b. Fast subsequence search using imcomplete search trees. In *The seventh Structural Biology Symposium of Seely Center for Structural Biology*, pp. 51, Galveston, Texas.

Karlin, S. and I. Ladunga. 1994. Comparisons of eukaryotic genomic sequences. *Proc Natl Acad Sci U S A* **91:** 12832-12836.

Karlin, S. and J. Mrazek. 1997. Compositional differences within and between eukaryotic genomes. *Proc Natl Acad Sci U S A* **94:** 10227-10232.

Nakashima, H., K. Nishikawa, and T. Ooi. 1997. Differences in dinucleotide frequencies of human, yeast, and Escherichia coli genes. *DNA Res* **4:** 185-192.

Nakashima, H., M. Ota, K. Nishikawa, and T. Ooi. 1998. Genes from nine genomes are separated into their organisms in the dinucleotide composition space. *DNA Res* **5:** 251-259.

Nussinov, R. 1984. Doublet frequencies in evolutionary distinct groups. *Nucleic Acids Res* **12:** 1749-1763.

Sandberg, R., G. Winberg, C.I. Branden, A. Kaske, I. Ernberg, and J. Coster. 2001. Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. *Genome Res* **11:** 1404-1409.

Southern, E.M. 2001. DNA microarrays - history and overview. *Methods of Molecular Biology* **170:** 1-15.

Acknowledgments

S.C. is grateful to the Department of Computer Science, University of Houston, Texas for hospitality. B.M.P, acknowledges the NIH for partial support and NPACI for computational support. M. Hogan is thanked for many stimulating conversations and encouragement.

Appendix

Here we will analytically find the frequency of presence of *n*-mers in random sequences. We will use the following definitions. Let *G* be a random sequence of length *M* of four characters $\{a,c,g,t\}$, and *S* be one of the 4^n possible subsequences of length n ("*n*-mer"). We will enumerate them, so that $\sum_{S=1}^{4^n}$ will stand for the sum with respect to all *n*-mers.

Let $F^{M,n}(S,k)$ be the probability that *S* appears exactly *k* times in *G*. We will refer to this also as "frequency of appearance" of *S* in *G*. To define this probability one can imagine a random statistical set of *N* sequences of the same length. If in this set there are N_k sequences that contain *S* exactly *k* times, then

$$F^{M,n}(S,k) = \lim_{N\to\infty} \frac{N_k}{N} \qquad 3)$$

Now let $f^{M,n}(S)$ be the probability that *S* is present in *G* (the *frequency of presence*). It is clear that

$$f^{M,n}(S) = \sum_{k=1}^{M} F^{M,n}(S,k) = 1 - F^{M,n}(S,0)\,. \qquad 4)$$

We now consider $P(\{k_S\})$ the probability distribution of appearance of *n*-mers.

Let p_S be the probability to find the *n*-mer *S* in *G*. For $n = 1$ they are reduced to the "elementary probabilities", p_l to find the character *l* in *G*, $l = \{a,c,t,g\}$. If the p_l are given, and we assume that *n*-mers *S* for $n > 1$ are composed in a random manner (*i.e.* the characters in *S* are not correlated), then

$$p_S = p_a p_t ... p_c, \qquad S = [at...c]. \tag{5}$$

For instance, equal probabilities $p_l = 1/4$ lead to homogeneous distribution,

$$p_S = const = \frac{1}{4^n}. \tag{6}$$

We are interested in the characteristics of appearance of *n*-mers in *G*. All of the related statistical information is contained in the distribution of probabilities such that

$$\{n\text{-mer } S \text{ appears } k_S \text{ times, } S=1,2,\ldots 4^n \equiv n_1\}, \quad \sum_{S=1}^{n_1} k_S = M - n + 1 \equiv M_n, \tag{7}$$

where M_n is a total number of *n*-mers in *G*. We will denote this distribution by

$$P(k_1, k_2, ..., k_S, ..., k_{n_1}) = P(\{k_S\}). \tag{8}$$

This distribution has a multinomial form,

$$P(\{k_S\}) = M_n! \prod_{S=1}^{n_1} \frac{{p_S}^{k_S}}{k_S!} \tag{9}$$

Here the product is taken over all configurations, such that $\sum_{S=1}^{n_1} k_S = M_n$.

One finds immediately the frequency of appearance of *S* in *G*,

$$F^{M,n}(S,k) = \sum_{\{k_T, T \neq S\}} P(\{k_T\}) = \frac{M_n! p_S^k q_S^{M_n - k}}{k!(M_n - k)!}, \qquad q_S = 1 - p_S \tag{10}$$

The mean number of appearance, $\bar{k}_S$, the variance $\sigma_S^2 = \overline{k_S^2} - (\bar{k}_S)^2$, covariance,

$\sigma_{ST}^2 = \overline{k_S k_T} - \bar{k}_S \bar{k}_T$ and the correlation coefficient, $C_{ST} = \sigma_{ST}^2 / \sigma_S \sigma_T$ are given as

$$\bar{k}_S = M_n p_S, \quad \sigma_{ST}^2 = -M_n p_S q_S, \quad \sigma_{ST}^2 = -M_n p_S p_T, \quad C_{ST} = -\sqrt{\frac{p_S p_T}{q_S q_T}} \qquad 11)$$

Therefore, if $p_S = 1/4^n$, $q_S = 1 - 1/4^n$, the correlation coefficient takes on the value $C_{ST} = -1/(4^n - 1)$.

Let us find the probability of presence, $f^{M,n}(S) = 1 - F^{M,n}(S,0) = 1 - (1 - p_S)^{M_n}$. In the homogeneous case, all probabilities are equal $p_S = 1/4^n$. It is convenient to introduce the variable, $y = M/4^n$, and consider the common Poisson limit of the Bernoulli distribution:

$$f^{M,n}(S) = 1 - (1 - \frac{y}{M_n})^{M_n} \to 1 - e^{-y}. \qquad 12)$$

Introducing another variable, $x = 4^n / M_N = 1/y$ we come to the formula,

$$f^{M,n}(S) = 1 - e^{-1/x}. \qquad 13)$$

Accession	Genome	Total Sequence length (bp)	Number of presence 8-mers	Frequency of presence 8-mers	Random boundary	Self-similarity
NC_001436	*Human T-cell lymphotropic virus type* 1	17,014	13,739	20.96%	22.86%	8.31%
NC_001707	*Hepatitis B virus*	6,430	5,963	9.10%	9.35%	2.64%
NC_001503	*Mouse mammary tumor virus*	17,610	14,307	21.83%	23.56%	7.35%
NC_001547	*Sindbis Virus*	11,703	10,431	15.92%	16.35%	2.67%
NC_001434	*Hepatitis E virus*	7,176	6,517	9.94%	10.37%	4.12%
NC_003312	*Swine hepatitis E virus*	7,257	6,608	10.08%	10.48%	3.81%
NC_001489	*Hepatitis A virus*	7,478	6,543	9.98%	10.78%	7.42%
NC_001433	*Hepatitis C virus*	9,413	8,480	12.94%	13.38%	3.29%
NC_001653	*Hepatitis D virus*	1,682	1,608	2.45%	2.53%	3.17%
NC_001802	*Human immunodeficiency virus type 1*	9,181	7,725	11.79%	13.07%	9.83%
NC_003461	*Human parainfluenza virus 1*	15,600	12,242	18.68%	21.18%	11.82%
NC_001796	*Human parainfluenza virus 3*	15,462	11,506	17.56%	21.02%	16.46%
NC_003443	*Human parainfluenza virus 2*	15,646	12,702	19.38%	21.24%	8.74%

Table 1. Frequency of presence of 8-mers and self-similarity (see the definition in the text) for several viral genomes.

Accession	Genome	Total Sequence length (bp)	Number of present 12-mers	Frequency of present 12-mers	Random boundary	Self-similarity
NC_000964	*Bacillus subtilis*	8,429,628	5,346,103	31.87%	39.50%	19.32%
NC_002696	*Caulobacter crescentus*	8,033,894	3,399,234	20.26%	38.05%	46.75%
NC_000913	*Escherichia coli K*12	9,278,442	5,695,881	33.95%	42.48%	20.08%
NC_000916	*Methanobacterium thermoautotrophicum*	3,502,754	2,658,450	15.85%	18.84%	15.91%
NC_003197	*Salmonella typhimurium LT*2	9,714,864	5,821,910	34.70%	43.96%	21.06%
NC_002758	*Staphylococcus aureus Mu*50	5,756,080	3,398,622	20.26%	29.04%	30.25%
NC_003098	*Streptococcus pneumoniae R*6	4,077,230	2,992,091	17.83%	21.57%	17.34%
NC_002737	*Streptococcus pyogenes*	3,704,882	2,778,223	16.56%	19.81%	16.43%
NC_002578	*Thermoplasma acidophilum*	3,129,812	2,602,761	15.51%	17.02%	8.84%
NC_002689	*Thermoplasma volcanium*	3,169,608	2,590,718	15.44%	17.22%	10.30%
NC_000919	*Treponema pallidum*	2,275,888	1,978,453	11.79%	12.69%	7.04%
NC_000853	*Thermotoga maritima*	3,721,450	2,755,886	16.43%	19.89%	17.43%
NC_002162	*Ureaplasma urealyticum*	1,503,438	948,274	5.65%	8.57%	34.06%
NC_002505	*Vibrio cholerae* chromosome I, chromosome II	8,066,854	5,383,520	32.09%	38.17%	15.94%
NC_002488	*Xylella fastidiosa* 9*a*5*c*	5,358,610	3,996,398	23.82%	27.34%	12.88%

Table 2. Frequency of presence of 12-mers and self-similarity for several microbial genomes.

Genome	Total Sequence length (bp)	Number of present n-mers	Percent of present n-mers	Random boundary: (1-exp(-1/x))	Self-similarity
Caenorhabditis elegans (14-mers)	199,980,344	83,915,577	31.26%	52.53%	40.5%
Drosophila melanogaster (14-mers)	239,963,692	119,253,045	44.43%	59.10%	24.8%
Oryza sativa (15-mers)	511,742,384	220,383,196	20.52%	37.91%	45.9%
Schizosaccharomyces pombe (12-mers)	24,980,160	9,256,101	55.17%	31.08%	28.8%
Homo Sipiens 16-mers	5,749,472,188	1,577,086,225	36.72%	73.78%	50.2%

Table 3. Frequency of presence of *n*-mers and self-similarity for several genomes of multicellular organisms (*n* is different for every genome).

Chromosome	Total Sequence length (bp)	Number of present 14-mers	Percent of present 14-mers	Random boundary	Self-similarity
1	447,066,010	120,482,569	45%	0.624829	28%
2	483,605,166	123,530,238	46%	0.643057	28%
3	386,994,240	111,160,119	41%	0.590444	30%
4	384,625,114	107,927,999	40%	0.588958	32%
5	368,501,246	108,267,069	40%	0.578552	30%
6	360,746,676	106,403,089	40%	0.573358	31%
7	316,339,548	102,552,382	38%	0.540959	29%
8	286,437,774	98,412,806	37%	0.516222	29%
9	226,712,656	88,055,274	33%	0.457868	28%
10	266,605,056	96,655,967	36%	0.498289	28%
11	260,847,876	95,175,227	35%	0.492832	28%
12	252,370,980	92,449,874	34%	0.484577	29%
13	195,223,308	80,365,520	30%	0.42105	29%
14	177,154,566	77,225,666	29%	0.397573	28%
15	164,503,442	74,009,598	28%	0.379969	27%
16	157,882,630	71,951,060	27%	0.37034	28%
17	158,162,596	71,571,114	27%	0.370753	28%
18	156,510,924	71,908,450	27%	0.368307	27%
19	113,275,840	52,794,505	20%	0.296758	34%
20	118,774,342	62,812,716	23%	0.306744	24%
21	67,658,204	41,829,261	16%	0.201308	23%
22	67,643,376	39,930,274	15%	0.201272	26%
X	286,494,168	93,586,369	35%	0.516271	32%
Y	45,336,450	22,663,720	8%	0.144489	42%

Table 4. Frequency of presence of 14-mers and self-similarity for every chromosome of the human genome.

***n*-mer size**	**Number of different *n*-mers 4^n**	**Number of absent *n*-mers**		**Number of *n*-mers present only once**		**Number of *n*-mers present more than once**	
7	16,384	0	0.00%	0	0.00%	16,384	100.00%
8	65,536	0	0.00%	0	0.00%	65,536	100.00%
9	262,144	0	0.00%	0	0.00%	262,144	100.00%
10	1,048,576	0	0.00%	0	0.00%	1,048,576	100.00%
11	4,194,304	42	0.00%	324	0.01%	4,193,938	99.99%
12	16,777,216	42,501	0.25%	91,146	0.54%	16,643,569	99.20%
13	67,108,864	2,382,096	3.55%	2,642,582	3.94%	62,084,186	92.51%
14	268,435,456	41,634,971	15.51%	30,411,367	11.33%	196,389,118	73.16%
15	1,073,741,824	410,828,287	38.26%	166,998,278	15.55%	495,915,259	46.19%
16	4,294,967,296	2,717,880,983	63.28%	671,192,253	15.63%	905,894,060	21.09%
17	17,179,869,184	14,452,040,667	84.12%	1,790,043,813	10.42%	937,784,704	5.46%
18	68,719,476,736	65,147,397,575	94.80%	2,881,849,256	4.19%	690,229,905	1.00%
19	274,877,906,944	270,850,664,602	98.53%	3,538,156,028	1.29%	489,086,314	0.18%
20	1,099,511,627,776	1,095,257,688,530	99.61%	3,866,031,543	0.35%	387,907,703	0.04%

Table 5. Presence/absence statistics for human genome of size 2,874,736,094 base pairs. Calculation was performed using both (original and complementary) DNA strand sequences.

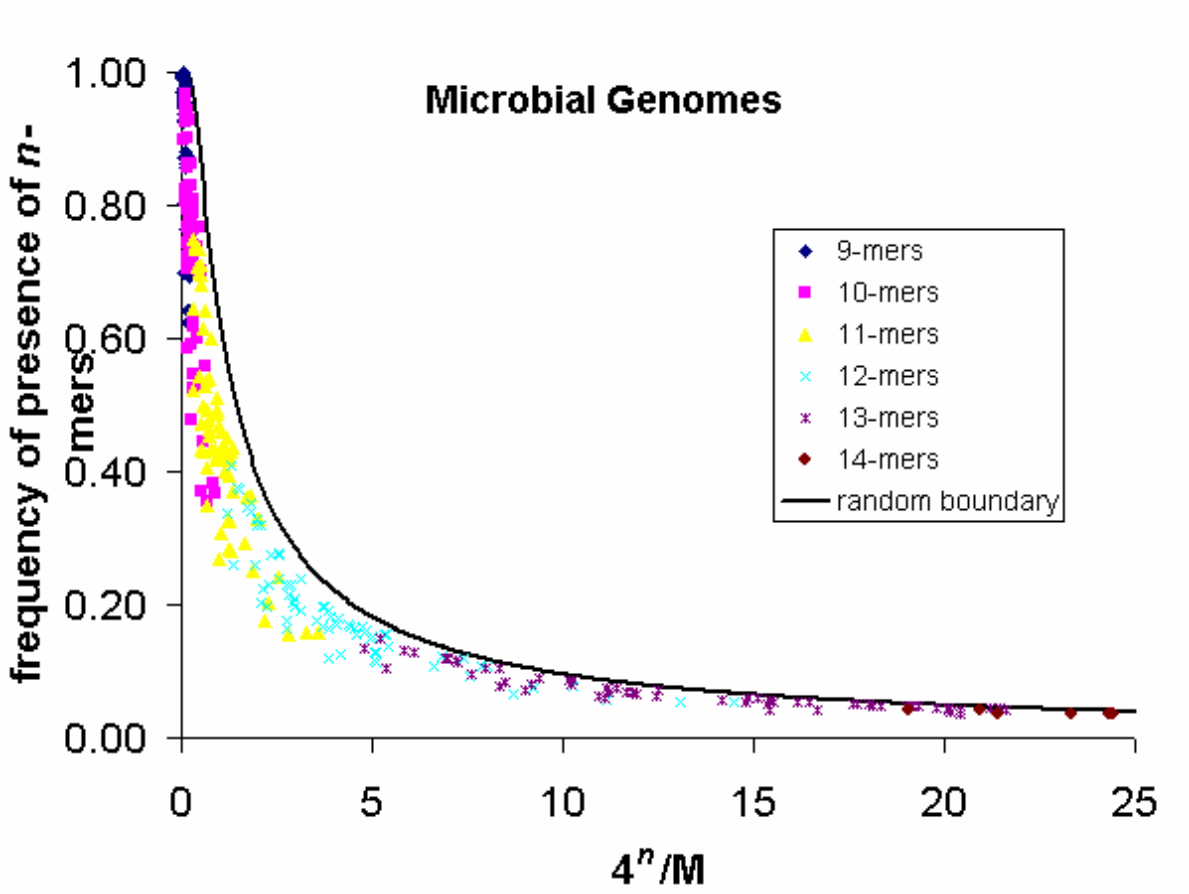


Figure 1. Frequency of presence of 9-14-mers in 76 microbial genomes.

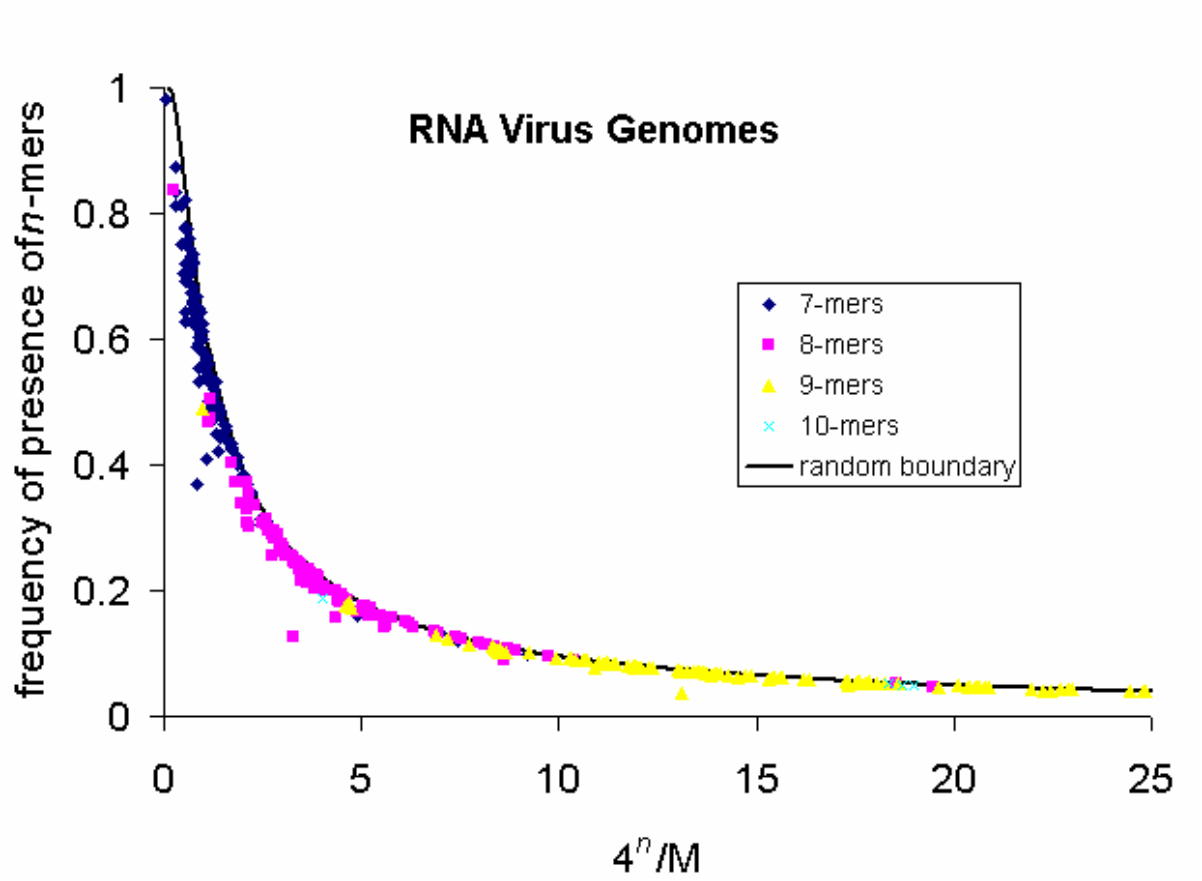


Figure 2. Frequency of presence of 7-10-mers in 129 RNA viral genomes.

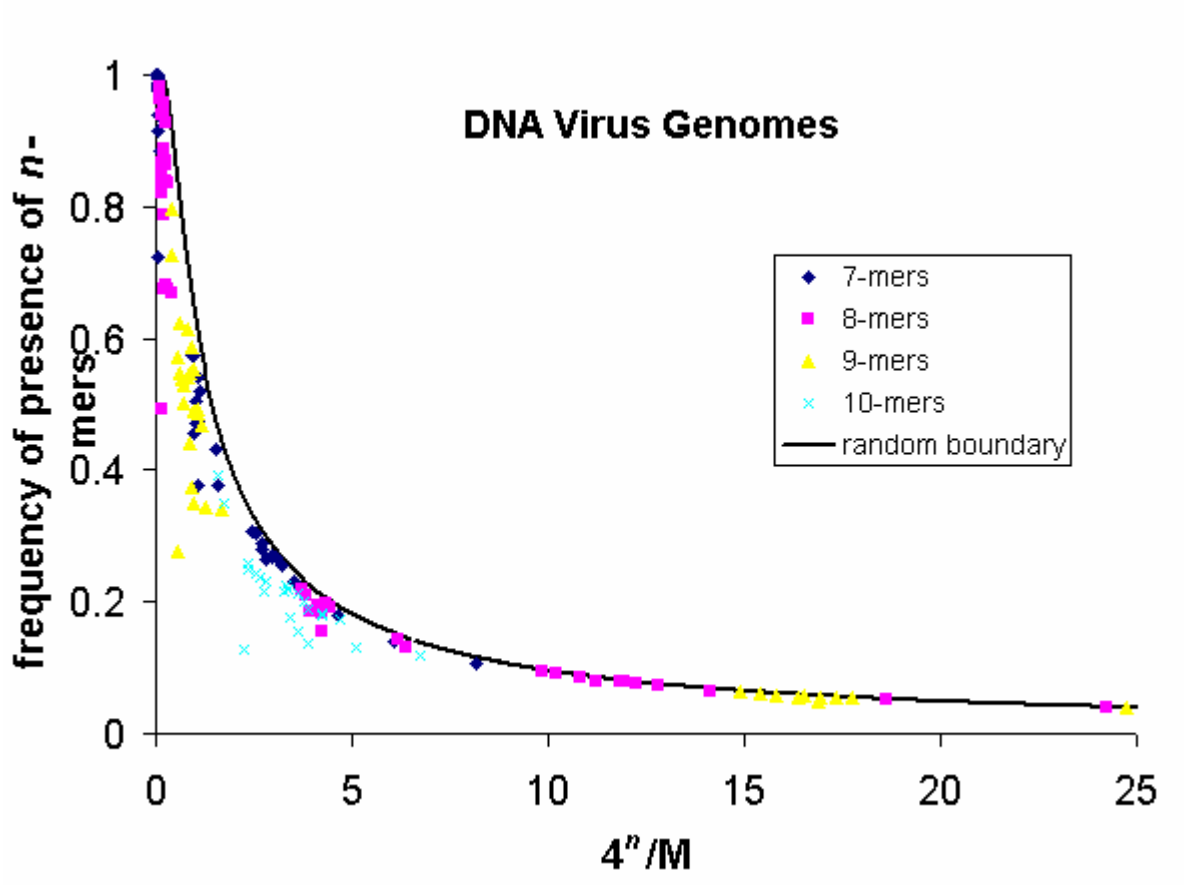
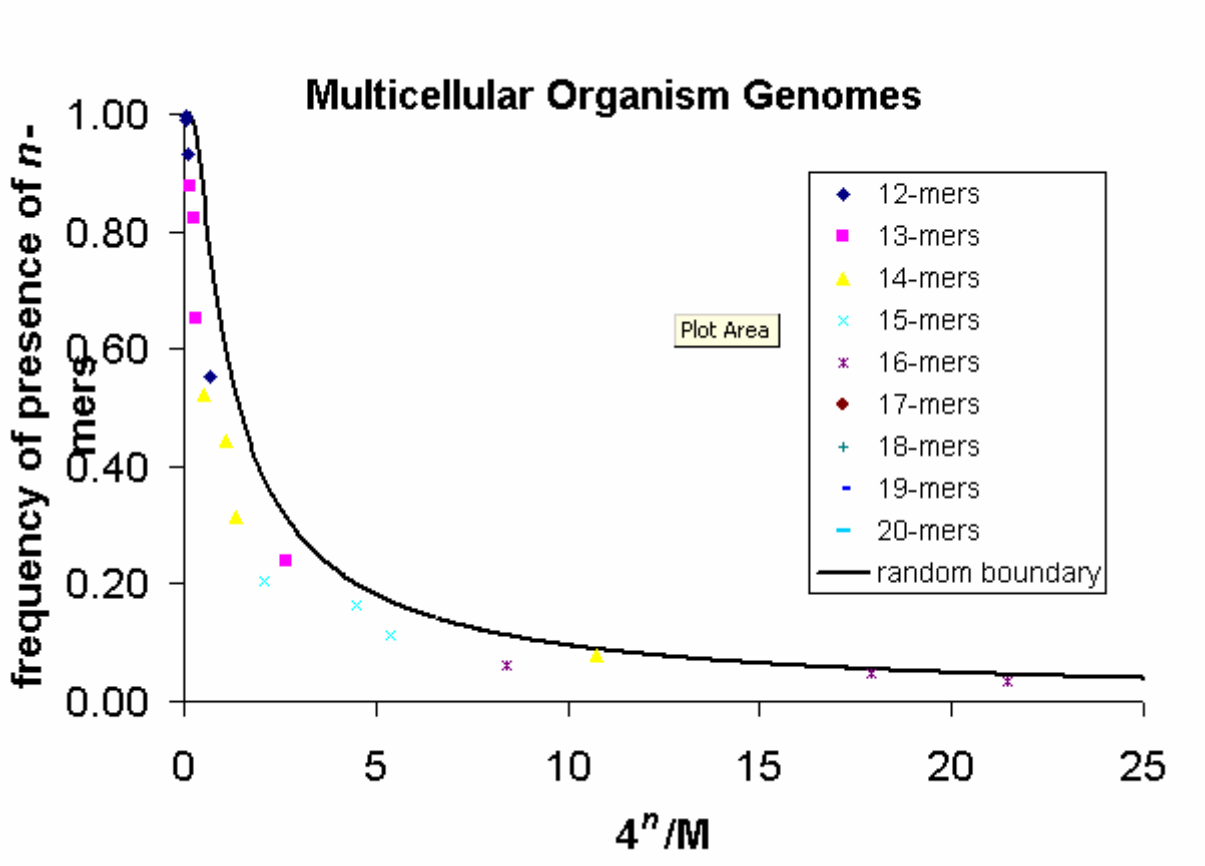
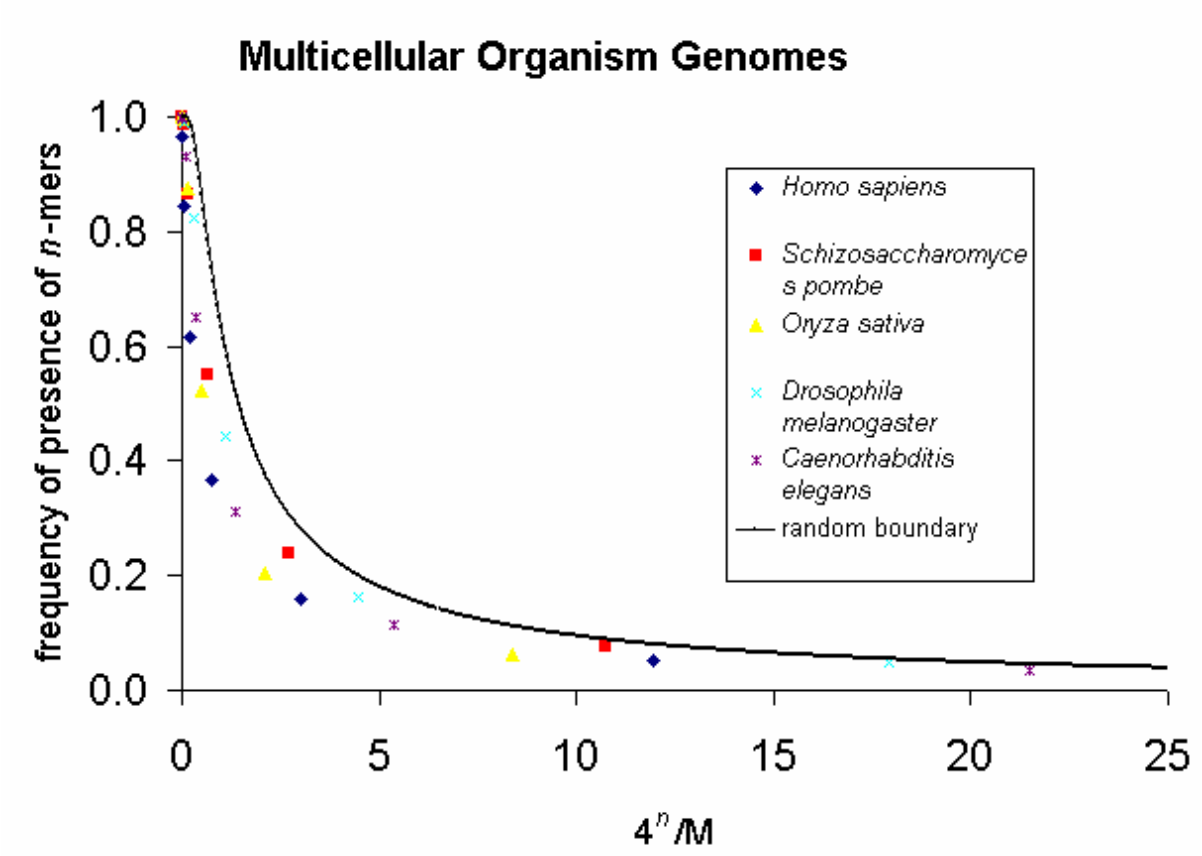


Figure 3. Frequency of presence of 7-10-mers in 48 DNA viral genomes


Multicellular Organism Genomes
frequency of presence of n-mers
1.00
0.80
0.60
0.40
0.20
0.00
0
5
10
15
20
25
4^n/M
Plot Area
12-mers
13-mers
14-mers
15-mers
16-mers
17-mers
18-mers
19-mers
20-mers
random boundary

a)



b)

Figure 4 (a). Frequency of presence of 12-20-mers in the genomes of multicellular organisms, (b) The same as in (a), but different genomes are indicated differently.

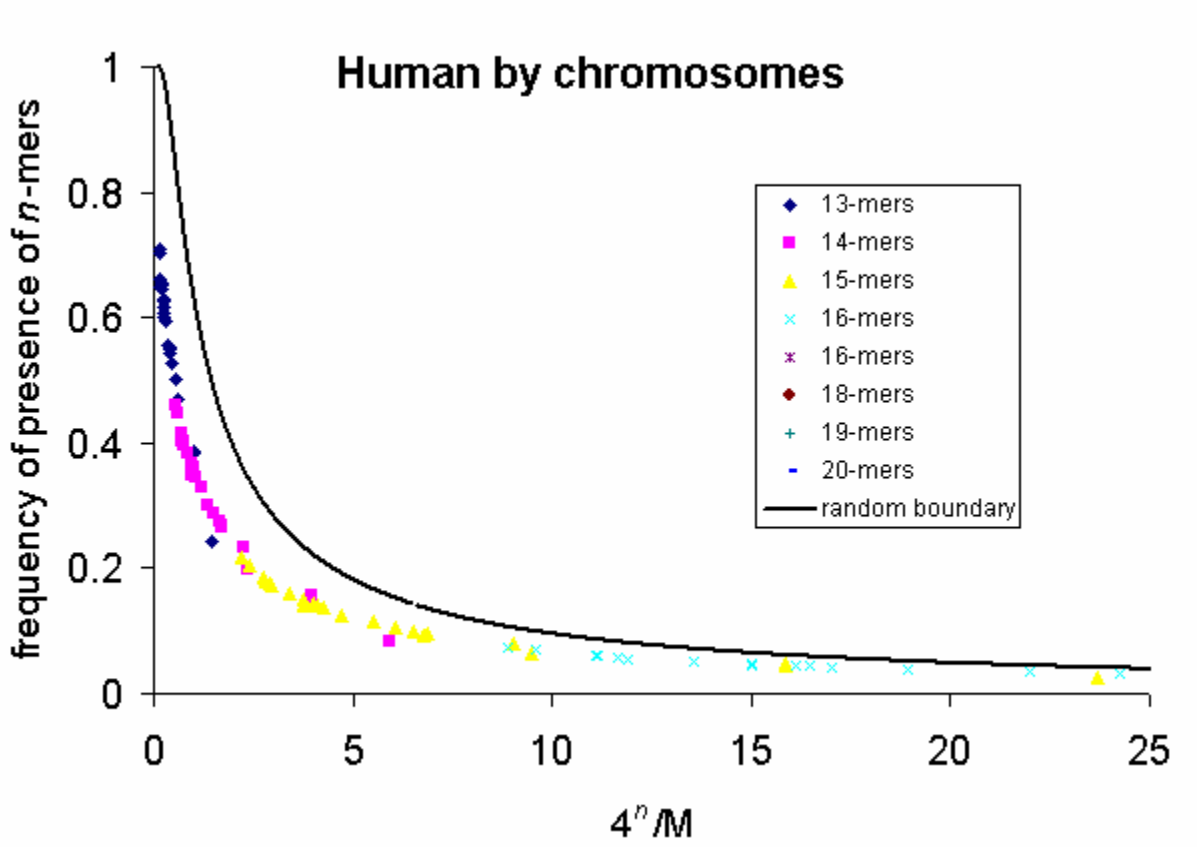


Figure 5. Frequency of presence of 13-20-mers in the different chromosomes of the human genome.

.