Long range clustering of oligonucleotides containing the CG signal

P. Katsaloulis, T. Theoharis, A. Provata

www.elsevier.com/locate/yjtbi

Long range clustering of oligonucleotides containing the CG signal

P. Katsaloulis, T. Theoharis, A. Provata

PII: S0022-5193(09)00022-8
DOI: doi:10.1016/j.jtbi.2009.01.014
Reference: YJTBI 5433

To appear in: *Journal of Theoretical Biology*

Received date: 15 February 2008
Revised date: 14 January 2009
Accepted date: 14 January 2009

Cite this article as: P. Katsaloulis, T. Theoharis and A. Provata, Long range clustering of oligonucleotides containing the CG signal, *Journal of Theoretical Biology* (2009), doi:10.1016/j.jtbi.2009.01.014

Long range clustering of oligonucleotides containing the CG signal

P. Katsaloulis[1,2], T. Theoharis[1] and A. Provata[2] *

[1] Department of Informatics and Telecommunications, University of Athens,15784 Athens, Greece
[2] Institute of Physical Chemistry, National Center for Scientific Research "Demokritos", 15310 Athens, Greece

January 12, 2009

Abstract

The distance distributions between successive occurrences of the same oligonucleotides in chromosomal DNA are studied, in different classes of higher eucaryotic organisms. A two-parameter modeling is undertaken and applied on the distance distribution of quintuplets (sequences of size five bps) and hexaplets (sequences of size six bps); the first parameter k refers to the short range exponential decay of the distributions, whereas the second parameter m refers to the power law behavior. A 2-dimensional scatter plot representing the model equation demonstrates that the points corresponding to the distance distribution of oligonucleotides containing the `CG` consensus sequence (promoter of the RNA polymerase II) cluster together (group α), apart from all other oligonucleotides (group β). This is shown for the available chordata *Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Rattus norvegicus*, *Gallus gallus* and *Danio rerio*. This clustering is less evident in lower Animalia and plants, such as *Drosophila melanogaster*, *Caenorhabditis elegans* and *Arabidopsis thaliana*. Moreover, in all organisms the oligonucleotides which contain any consensus sequence are found to be described by long range distributions, whereas all others have a stronger influence of short range decay.

Various measures are introduced and evaluated, to numerically characterize the clustering of the two groups. The one which most clearly discriminates the two classes is shown to be the Proximity Factor.

*Corresponding author. E-mails: P. Katsaloulis pkatsaloulis@chem.demokritos.gr; A. Provata aprovata@chem.demokritos.gr

1 Introduction

In recent years there has been great improvement in biological DNA decoding tools, in terms of speed and efficiency (Altschul et al. 1990, 1997; Pearson and Lipman 1988; Higgins et al. 1992; Thompson et al. 1994). This growth produced a spectacular increase of biological data, especially in decoding chromosomes from different organisms. The analysis has been extended from *Homo sapiens* (Venter and al 2001) to various mammals, birds, insects and plants. This rapid growth made obvious the need of management and evaluation tools, while the main objective remains to estimate the functional role of the DNA. To address this question approaches have been proposed, which combine computational tools with statistical methods.

The use of computation tools in parallel with methods originating from equilibrium and non-equilibrium statistical mechanics have revealed a number of unexpected features in the primary genomic structure. The most striking one was the observation of long range correlations discovered in 1992 in the noncoding DNA of higher organisms (Peng et al. 1992; Li and Kaneko 1992; Voss 1992). Later-on, other researchers verified and further explored the existence of nontrivial structural correlations in genomic sequences (Karlin and Brendel 1992; Ebeling and Nicolis 1992; Buldyrev et al. 1993; Li et al. 1994; Czirok et al. 1995; Arneodo et al. 1996; Hao 2000; Yu et al. 2000; Li and Holste 2005; Li and Miramontes 2006; Bernaola-Galvan et al. 2002; Carpena et al. 2007; Scafetta et al. 2002; Allegrini et al. 1995). Some of these nontrivial characteristics were used for identification of functional units in DNA sequences of unknown origin (Hackenberg et al. 2006; Carpena et al. 2002).

In previous studies in the primary structure of DNA, scaling behavior was observed within the noncoding DNA areas, whereas short range behavior is found in the coding ones (Almirantis and Provata 1997, 1999; Provata and Almirantis 1997, 2000; Katsaloulis et al. 2002, 2005, 2006). In these studies, the form of the size distribution of noncoding DNA segments was shown to include a major long range contribution of the form:

$$P(S_{non-cod}) \sim S_{non-cod}^{-1-\mu} \tag{1}$$

where $S_{non-cod}$ is the length of a noncoding region, $P(S_{non-cod})$ the length distribution of noncoding regions and μ is the critical exponent of the distribution.

It is noted here that with the term "noncoding DNA segments" we mean collectively all introns and intragenic regions(Provata and Oikonomou 2007). Both introns and intragenic regions can suffer modifications during evolution which can be moderate, mostly in the case of introns, and/or extensive, mostly in the case of intragenic regions. This "open-to-environmental-influence" structure of the introns and intragenic regions is in the origin of the long range, power law behavior, which is more prominent in the length scales characterizing the intragenic regions.

The size distribution of introns in particular, has been the subject of earlier studies and short range features, analogous to the ones of the coding regions,

have been reported (Alberts et al. 1994; Hawkins 1988; Deutsch and Long 1999; Lander and et al. 2001; Lim and Burge 2001; Sakharkar et al. 2002; Carpena et al. 2007). Introns may also present some long range characteristics, due to the fact that their structure is open to environmental influence (Provata and Oikonomou 2007).

Coding regions, or exons, mostly conserve their structure through evolution and thus are compatible with short range characteristics as has been extensively reported in the literature. Gaussian-type distributions, skew distributions, log-normal and exponential type, this is not an exhaustive list of functions used to describe the exon size distributions (Hawkins 1988; Lander and et al. 2001; Sakharkar et al. 2002; Carpena et al. 2007). As in most cases of short range behavior, the tails of the coding regions can be approximated by an exponential decay, of the form:

$$P(S_{cod}) \sim e^{-aS_{cod}} \tag{2}$$

where a is the positive parameter characterizing the decay law.

Eucaryotic organisms, were considered to obey the distributions 1 and 2 for their noncoding and coding parts, respectively. Since the RNA polymerase promoter marks the beginning of a coding region, and the size of a coding region is much smaller compared to a noncoding region (Alberts et al. 1994; Etienn-Decant 1988; Vinogradov 1999), the hypothesis was formulated that the distance distribution between two consecutive consensus sequences of the promoter should follow the distance distribution of the corresponding noncoding regions. This hypothesis was elaborated in references (Katsaloulis et al. 2005, 2006), where it was shown that oligonucleotides containing the signature of a consensus sequence follow long range distributions, while all other oligonucleotides follow short range distributions. The consensus sequences with the major contributions are the `CG` box for all animals and plants, and the `TATA` box for the plant *Arabidopsis thaliana* (Blake et al. 1990; Hoffman et al. 1987). The distance distribution of consensus sequences includes a strong long range contribution in the tails of the form:

$$P(S_p) \sim S_p^{-1-\mu} \tag{3}$$

where S_p is the distance between two consecutive appearances of the oligonucleotides containing the promoter consensus sequence, $P(S_p)$ is the distance distribution of the oligonucleotides and μ is the critical exponent of the distribution. As in the case of the size distributions of noncoding regions, the value of the critical exponent μ is calculated based on the form of the tails of the distribution. When the tails in the noncoding area are characterized by values of μ in the range $0 \leq \mu \leq 2$ the corresponding size distributions are classified as long range.

This approach does not sufficiently cover the majority of the cases. Some distributions of oligonucleotides, especially in evolutionary newer organisms, do not have a well developed, observable tail, and have a mixed behavior characterized both by long range tails and short range decay. This mixed behavior

may be explained by the many different evolutionary mechanisms which have acted upon the genome during evolution.

Summing up the behaviour of the size distribution between different types of oligonucleotides we note the following:
A) Oligonucleotides which do not have any specificity in the genome can appear in any part part of the genome (coding or noncoding), are not expected to have any peculiar characteristics in their distribution throughout the chromosomes and can thus be modeled by short range (exponential like) distributions in all scales.
B)Specific oligonucleotides containing the CG or other promoter signatures maybe encountered in the chromosomes in the following two cases:

- They may appear randomly (such as in coding areas where almost all combinations are found equiprobably). These will enter with a short range type of distribution. Because they can be found within the coding areas which are relatively short in sizes, or maybe separated by introns (also relatively short), they will be always separated by short distances. This indicates that in the short scales, short range behaviour is expected.

- They may designate the presence of a promoter and then their distributions would carry the power law correlations (this property is connected with the fact that they are separated by at least one noncoding region, as noted earlier).

Because in the very-very long scales, finite size effects are expected to cover up the long range behaviour, the power law property is expected to be detectable mostly in the intermediate scales. Thus for the oligonucleotides which carry the promoter signature we expect i) short range (exponential decay type) behaviour at the short scales, ii) mixed short and long range behaviour at the intermediate scales and iii) short range decay, due to finite size effects in the very long scales. The mixed behaviour, at the intermediate scales, need then to be modeled, using a mixed law containing long-range (power like) and short range (exponential like) parts. In this case the oligonucleotides which contain the promoter signature are expected to present most important power law regions in the intermediate scales.

To address this problem it is unavoidable to take into account more than one parameters. In particular, we introduce here a phenomenological mixed description, which includes long range power law terms together with exponential decay ones. The results of this method are presented in this study, where we have examined and compared the forms of the distributions $P(S_p)$ in various chromosomes and in different classes of organisms.

Using this 2-parameter approach, we have shown certain tendencies for clustering in the parameters characterizing the distributions of the oligonucleotides. In most cases oligonucleotides tend to accumulate in two distinct areas, which correspond to different parameters. We examined this phenomenon in terms of evolution, by taking into account organisms varying from mammals and chordata in general, to insects, nematodes and plants. We have found that this

clustering is static, although in evolutionary newer organisms is more evident. For the quantitative study of the clustering property various clustering measures were tested and we ended up with selecting the Proximity Factor (PF) as the one which could distinguish between classes of organisms. We have shown that the measure PF identifies clusters and is able to discriminate between newer and older organisms.

This work deals mostly with the qualitative and quantitative study of the clustering tendency of the oligonucleotide distribution parameters and focuses on the mathematical properties of the distributions, for understanding the structure and functionality of DNA. In this respect, this work distinguishes interesting sequences based only on their distribution on the genome - a pure mathematical approach. We have also tried to verify these observations from a biological point of view, by noting that the prominent sequences are actually consensus sequences of the RNA polymerase promoter. Although these findings could be used to annotate oligonucleotides, and thus specific DNA areas, we do not opt to present here a prediction tool. Our aim is to present a comparative study of the clustering across organisms and to show how this property fades away in evolutionary older organisms.

In the next section the methodology involving the 2-parameter modeling of the oligonucleotide distribution is described. In section 3 the analysis of the chromosomes of various organisms is presented. In section 4 quantitative measures of the oligonucleotide clustering are proposed and validated. In section 5 the main conclusions are recapitulated and open problems are discussed.

2 Two-parameter phenomenological description of clustering

Since the value of the critical exponent μ is not adequate to describe the behavior of the distance size distribution of the oligonucleotides, we need to enrich formula 3 which describes the from of the distributions. Many oligonucleotides do not demonstrate a clear linear region in their histogram (in double logarithmic scale) but a curved line, indicating mixed short as well as long range features.

In a previous study (Katsaloulis et al. 2006) a two-parameter model was used to describe the behavior of the distribution . Since both long range and short range distribution tendencies were detected, it appeared natural to experiment with a phenomenological formula which includes both. The proposed equation takes the following form:

$$P(S_o) = AS_o^{-1-m}e^{-kS_o} \tag{4}$$

where S_o is the distance between two consecutive appearances of the same oligonucleotide o, $P(S_o)$ is the distance distribution of the oligonucleotides o and A, m and k are the parameters used to describe this distribution. Eq. 4 contains both a power law expression (S_o^{-1-m}) and an exponential term ($exp[-kS_o]$). We have to note that the parameter m corresponds to the critical exponent μ of

Eq. 3. Parameter A is mostly used for normalization. This parameter does not appear to affect the results, and thus we will take into account only the other

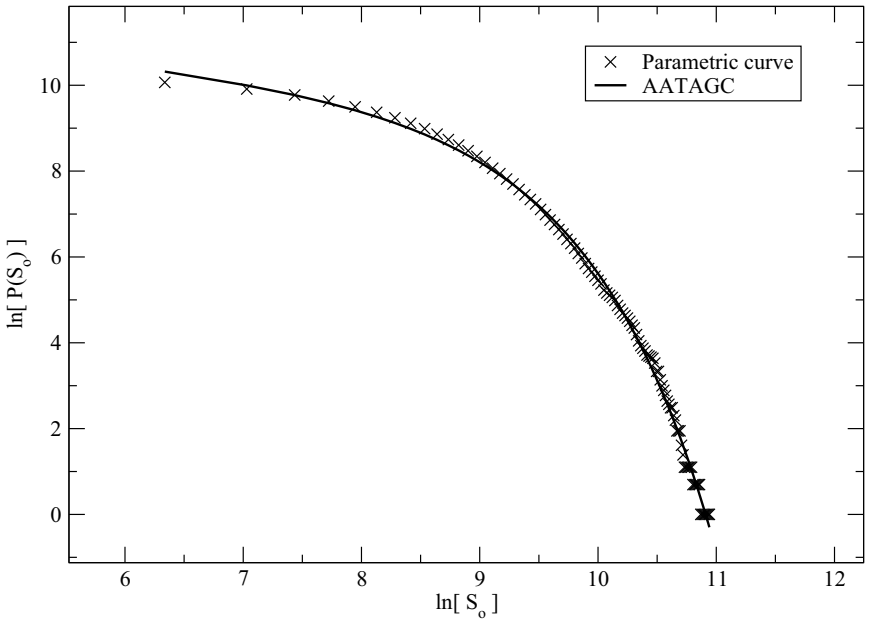

Figure 1: Cumulative distance distribution of the *Mus musculus* sequence `AATAGC`, chromosome 16 (symbol X). The solid line represents the corresponding fitted curve. The value of $(-1-m)$ and $-k$ of this sequence is -3.956 and -2.267×10^{-05}. The value of correlation coefficient is 0.975

form of the size distributions are used(Provata and Oikonomou 2007). Figure 1 corresponds to the oligonucleotide sequence `AATAGC`. The value of $m = 2.956 > 2$ and $k = 2.267 \times 10^{-5}$ indicate clearly that the behavior is of exponential decay type, purely short ranged (the correlation coefficient of this fit is $r = 0.975$). The form of the figure also does not include any noticeable power law region. On the other hand, Figure 2 corresponds to the distribution of the oligonucleotide `CGATCG`, a sequence which contain twice the CG complex. This figure clearly indicates a power law region. The solid line is drawn for comparison and has slope -1.6. The calculated parameter values $m = 0.532 < 2$ and $k = 9.507 \times 10^{-7} \sim 0$ indicate clearly that the behavior is purely long ranged (the correlation coefficient of this fit is $r = 0.984$). Similar features are observed in all other oligonucleotide sequences.

For each chromosome of the selected set we calculate the cumulative distance

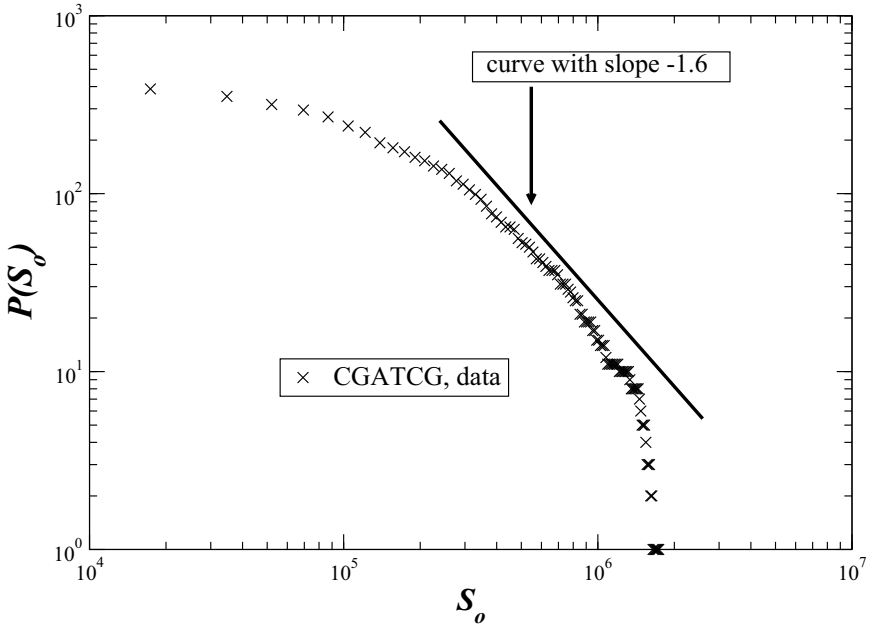

Figure 2: Cumulative distance distribution of the *Mus musculus* sequence `CGATCG` (chromosome 15). Values of (-1-m), -k and correlation coefficient are -1.315, $5.256 \cdot 10^{-7}$ and 0.98 respectively.

distributions between all oligonucleotides o of fixed length(Katsaloulis et al. 2006). The curve of the distribution is used to calculate the values of parameters A, m and k, using a combined Levenberg-Marquardt with Gauss-Newton method (Levenberg 1944; Marquardt 1963). Finally, we form a 2-dimensional scatter plot of all oligonucleotides of the same length for each chromosome, taking into account the exponent values of $(-1-m)$ versus $(-k)$. In the current study we computed all oligonucleotides with fixed length of five (quintuplets) and six (hexaplets) base pairs.

In order to quantify the quality of the curve fitting, we have also calculated the value of correlation coefficient (Cc)(see relative discussion of Figs. 1 and 2). As an example, in chromosome 15 of *Mus musculus* the value of Cc is between 0.955 and 1 for all oligonucleotides, with average value of 0.995. Similarly, in chromosome 16 Cc takes values between 0.94 and 1, with average of 0.979.

Although in some cases the exponential parameter k takes negative values, these values are very-very small and are expected to be rounding errors of the curve fitting algorithm. Due to computer precision limitations, these values are *almost zero* and thus have minimum or non existing contribution to the size distribution of oligonucleotides. In the presented graphs we retain this information though, for completeness reasons.

For evolutionary newer organisms it is shown that the parameters appear to cluster, as will be discussed in the next section. In all plots presented hereafter, the sequences which include the sub-sequence CG (Bernaola-Galvan et al. 2004) will be marked with the symbol X (group α), while all other sequences will be marked with the symbol O (group β), to make the clustering property more obvious.

3 Comparative clustering in exponent plots

Tests have been performed in the chromosomes of various organisms for which long decoded chromosome sequences already exist. An effort was undertaken to analyze organisms of as many different classes as possible. The organisms that were studied are the following: *Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Rattus norvegicus*, *Gallus gallus*, *Danio rerio*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*. The complete chromosome data was obtained from the NCBI server (`ftp://ftp.ncbi.nih.gov/genomes`).

3.1 Mammals

The first group of organisms that was studied was the Mammals and especially chromosomes 19, 21 and 22 of *Homo sapiens* (Katsaloulis et al. 2006), chromosomes 1 and 2 of *Pan troglodytes*, chromosomes 15 and 18 of *Mus musculus* and chromosome 1 of *Rattus norvegicus* (see corresponding Figs. 3 , 4 , 5 and 6). The analysis was performed solely for the quintuplets and hexaplets of these chromosomes.

We have found that sequences which include the binucleotide CG (and belong to group α) are completely separated from all other oligonucleotides (group β). The situation on quintuplets and hexaplets is exactly the same; the two groups in both cases are clearly separated from one another (see Figs. 3-6).

This behavior persists also across organisms and chromosomes. In all cases we have studied, the separation between the two groups is always evident. Although the distance between these two groups may vary, the separation is always quite obvious.

For the study of the separation of the two groups in a straightforward way, it is possible to consider the two axes separately. By calculating the average value of $< k_i >$ for the two groups ($i = \{CG\}, \{nonCG\}$) and the corresponding standard errors σ_{k_i}, the clustering in the $k-$direction (y-axis) would be significant if $< k_i >><< k_j > +1.96 \times \sigma_{k_j}$ or $< k_i ><< k_j > -1.96 \times \sigma_{k_j}$, for $i \neq j$. In the same way one can approach the x-axis ($|-1-m|$-direction). A distinct example is the case of Chromosome 15 of Mus musculus (see Fig. 5) where $< k_{\{CG\}} >= 0.000018$ and $\sigma_{k_{\{nonCG\}}} = 1.02 \times 10^{-5}$, while $< k_{\{nonCG\}} >= 0.0002$ and $\sigma_{k_{\{nonCG\}}} = 0.9 \times 10^{-5}$. In this case the $< k >$-value of the oligonucleotides containing CG is outside the critical interval of the oligonucleotides not containing CG and thus clustering is evident on the y-axis. Other cases might not be so evident and composite clustering in both axes must then be taken into account.

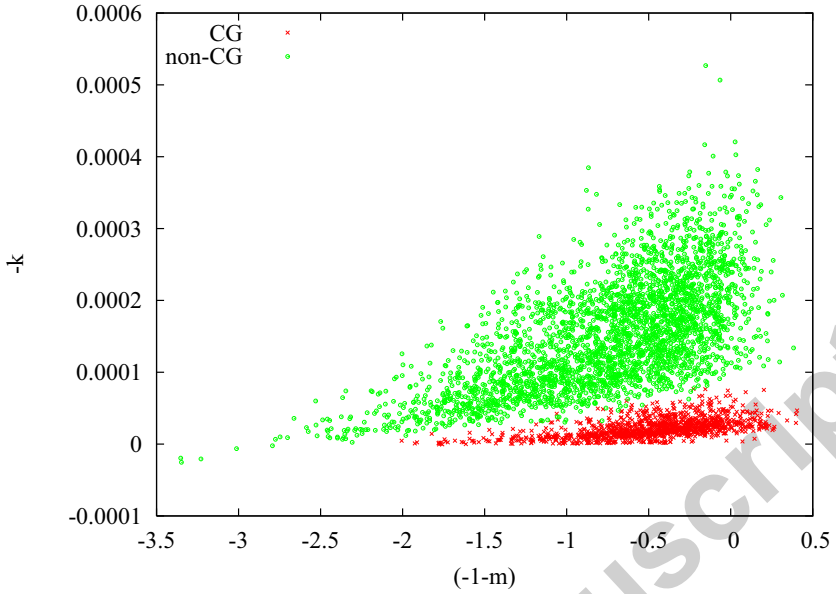

Figure 3: 2-dimensional exponent plot of organism *Homo sapiens*, chromosome 21, hexaplets

Quantitative cluster analysis taking into account both axes will be undertaken in section 4.

Another interesting observation is the positioning of the two groups, which mirrors the values of the parameters $-1-m$ and $-k$. The value of parameter $|-1-m|$ in group α is between the limits [0,2] and we have a direct correspondence with the critical exponent of μ in Eq. 1. This is in accordance with earlier observations that the subsequence `CG` marks oligonucleotides with the smaller values of $|\mu|$ (Katsaloulis et al. 2005, 2006), since they are marked with a consensus sequence of the promoter. On the other hand, parameter $|k|$ is very close to zero, which displays the minor role of the exponential term e^{-kS_o}. In other words group α seems to be influenced almost completely from the power law factor with minor contribution from the exponential one.

The situation in group β is different. The values of $|-1-m|$ are larger than in group α, while the value of $|k|$ has broader spectrum and is non zero. Thus the key role here is played by the exponential factor. This behavior differs from that of group α and reflects the fact that the distributions are different for these oligonucleotides. The oligonucleotides which do not contain the `CG` sequence follow short range distributions, such as exponential decay, as expected (Peng et al. 1992; Li and Kaneko 1992; Almirantis and Provata 1997, 1999; Katsaloulis et al. 2006).

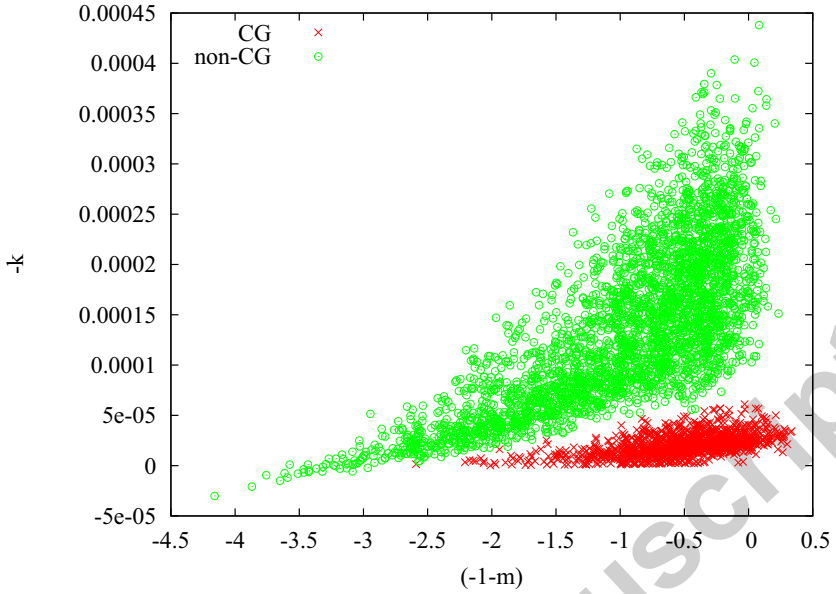

Figure 4: 2-dimensional exponent plot of organism *Pan troglodytes*, chromosome 1, hexaplets

3.2 Other chordata

Studies have been performed on the organisms *Gallus gallus* (bird) and *Danio rerio* (fish), which are the only chordata with fully decoded chromosomes up to now. Fig. 7 presents the results for quintuplets of chromosome 1 of *Gallus g.* and Fig. 8 presents the results for *Danio r.*. The observation for both organisms is the same, as in mammals. Two distinct areas are presented again, group α, which consists solely of oligonucleotides with the signature `CG` of the consensus sequence of the RNA polymerase promoter, and group β of all other oligonucleotides. Group α again appears to have the smallest values of $|-1-m|$, while the $|k|$ parameter takes values around zero, which means it is more affected by the power law factor of Eq. 4. Group β has again a similar behavior as in mammals. Still, the value of the $-1-m$ parameter is larger than in group α and thus we assume that these sequences follow short range distribution.

All these observations show that in the evolutionary recent organisms we have studied, all oligonucleotides which have the consensus sequence `CG` seem to follow long range distributions. On the other hand, the sequences which do not posses CG follow short range distributions, such as the exponential one. This observation persists across organisms, for mammals, birds and fish, and we expect it will appear in all evolutionary newer organisms.

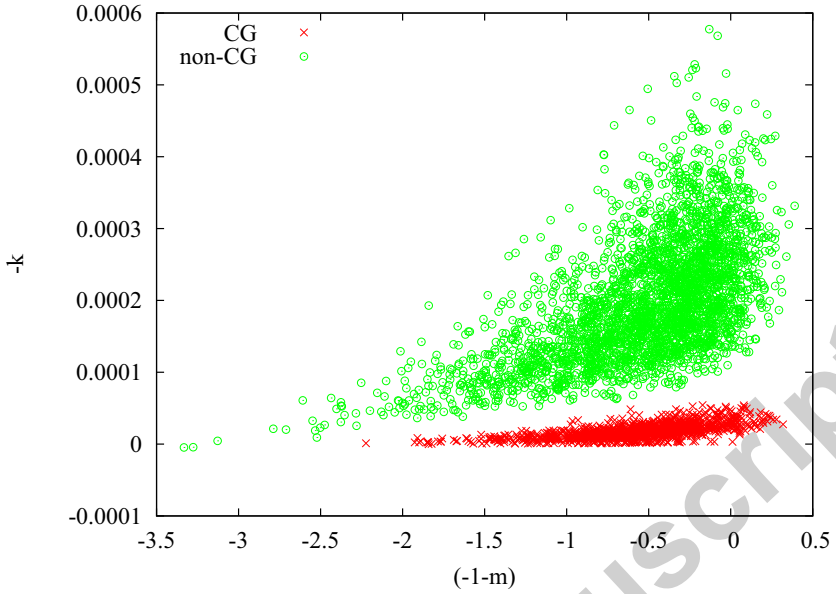

Figure 5: 2-dimensional exponent plot of organism *Mus musculus*, chromosome 15, hexaplets

3.3 Evolutionary older organisms and plants

Evolutionary older organisms were analyzed, such as *Drosophila melanogaster*, *Caenorhabditis elegans* and the plant *Arabidopsis thaliana*. We have tested quintuplets and hexaplets of the regions NT_033777 and NT_037436 from chromosome 3 of *Drosophila m.*, chromosome I (area NC_003279) of *Caenorhabditis el.* and chromosome 1 (area NC_003070) of *Arabidopsis th.* (see Figs. 9 , 10 and 11).

The results from these organisms differ from those of mammals and birds. The separation of the two groups α and β is no longer evident and the two regions appear close to each other. In hexaplets of *Arabidopsis th.* these two groups seem to form different structures. Since *Arabidopsis th.* is evolutionary newer than the other two animals, we would expect that this behavior is due to the evolutionary age difference of these organisms.

Although the cluster separation in the above organisms is not so obvious, the values of the equation parameters appear to be in agreement with those of mammals, birds and fish. Oligonucleotides that include the consensus sequence `CG` are found in the region where the value of the parameter $|k|$ is near zero, while the value of parameter $|-1-m|$ is between zero and two. On the other hand, oligonucleotides that do not include the `CG` sequence, have larger values for the parameter $|-1-m|$. In other words, oligonucleotides with the signature of the

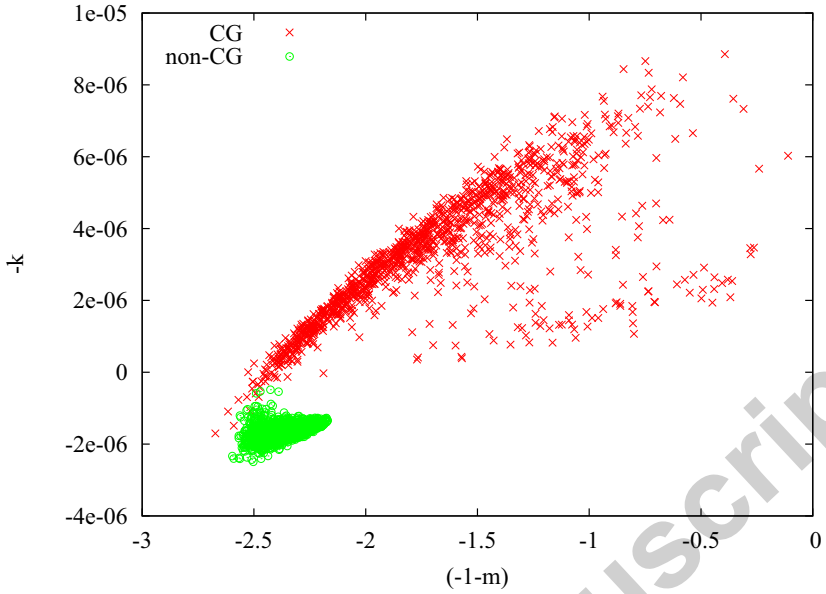

Figure 6: 2-dimensional exponent plot of organism *Rattus norvegicus*, chromosome 1, hexaplets

consensus sequences of RNA polymerase seem to follow long range distributions, while all other sequences appear to be affected mainly by the exponential term.

Another example for the study of the separation of the two groups, is also presented here, considering both axes separately. In the case of area NC_003070 of Arabidopsis thaliana (see Fig. 11), the averages and standard errors in the $k-$direction, for the two groups, are calculated as $< k_{\{CG\}} >= 0.000028$ and $\sigma_{k_{\{nonCG\}}} = 2.3 \times 10^{-5}$, while $< k_{\{nonCG\}} >= 0.000030$ and $\sigma_{k_{\{nonCG\}}} = 3.3 \times 10^{-5}$. In this case the clustering is not prominent, the average value of one cluster falls inside the critical area of the other cluster, thus the two clusters seem to overlap. More detailed, quantitative cluster analysis taking into account both axes will be undertaken in section 4.

We have to note that, in these organisms, there is a difference concerning consensus sequences, compared with evolutionary recent organisms. It is known that, in *Arabidopsis th.* for example, important role as consensus sequences is played by other sequences as well, such as `TATA` and `CAT` (Campbell and Gowri 1990). This property might be mirrored on the position of the oligonucleotides, since now the promoting is not basically attributed to the `CG` sequence but also to other competitive sequences. This phenomenon might influence the presented graphs, since oligonucleotides which do not have the `CG` sequence might have values of $|-1-m|$ between [0,2] and parameter $|k| \sim 0$. Still, the importance of `CG` is high and thus we still see oligonucleotides with the `CG` sequence to appear

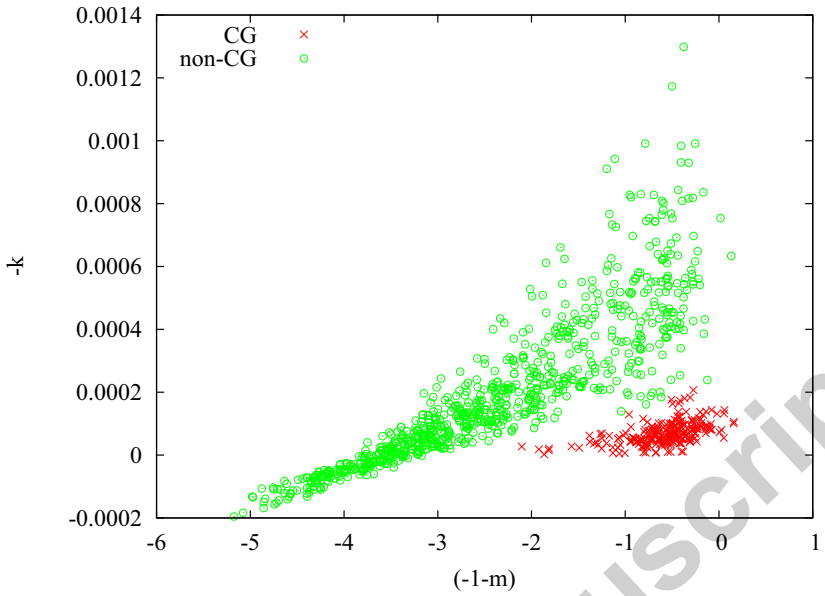

Figure 7: 2-dimensional exponent plot of the organism *Gallus gallus*, chromosome 1, quintuplets

in the same position in the diagram, as in mammals, birds and fish.

4 Cluster quantitative analysis

In an attempt to comprehend the degree of overlapping between the two cluster areas α and β, we introduce some distance measures between the two clusters. Calculating a distance between areas α and β can be used as a measure of proximity and furthermore as a way to quantify the coverage between the two groups.

We have computed the positions in two dimensional space of every point, and the group to which this point belongs. Points of group α are defined as $\vec{p_i} = (-1 - m_i^a, -k_i^a)$, with i belonging to group α and $1 \leq i \leq n_a$, where n_a is the total number of oligonucleotides in group α. Similarly, the points of group β are defined as $\vec{q_j} = (-1 - m_j^b, -k_j^b)$ with j belonging to group β and $1 \leq j \leq n_b$, where n_b is the total number of oligonucleotides in group β. The basic distance used between two points is Euclidean.

It is noted that the scale of the two axes differs by many orders of magnitude which can introduce proximity errors. For example, two points can be distant on the axis with the smaller scaling and relatively proximate on the other. The distance between these points will be smaller than that for two other points,

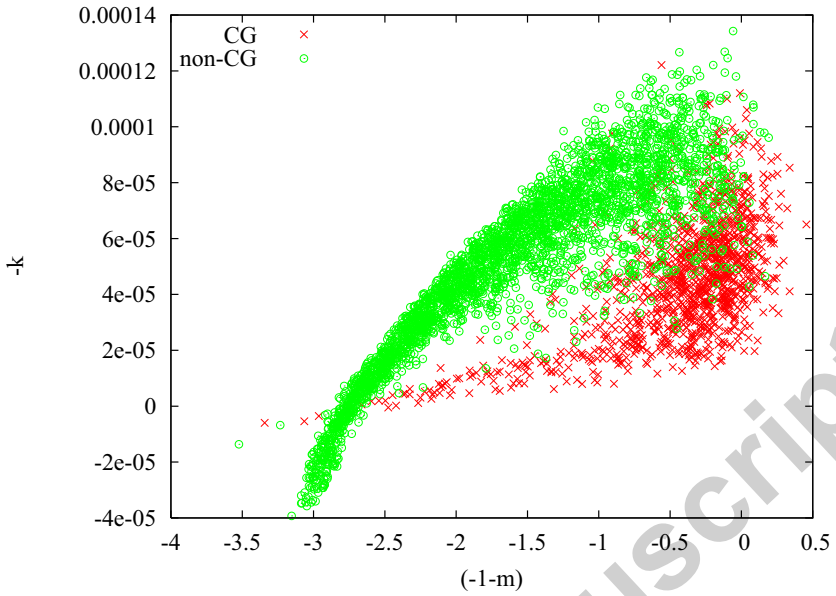

Figure 8: 2-dimensional exponent plot of the organism *Danio rerio*, chromosome 1, hexaplets

which are very close in the small-scale axis and not so proximate on the large-scale axis. In order to minimize scaling errors, the data has been normalized in both axes, thus $0 \leq \tilde{m}_i^a \leq 1$ and $0 \leq \tilde{m}_j^b \leq 1$, where $\tilde{m}_i^a$ and $\tilde{m}_j^b$ is the normalized value of $-1 - m_i^a$ and $-1 - m_j^b$ respectively. Similarly, we define $0 \leq \tilde{k}_i^a \leq 1$ and $0 \leq \tilde{k}_j^b \leq 1$, where $\tilde{k}_i^a$ and $\tilde{k}_j^b$ is the normalized value of $-k_i^a$ and $-k_j^b$ respectively. The position of the points is then defined as

$$\vec{p_i} = (\tilde{m}_i^a, \tilde{k}_i^a) \quad and \quad \vec{q_j} = (\tilde{m}_j^b, \tilde{k}_j^b) \tag{5}$$

The Euclidean distance between two points p_i, q_j (where $1 \leq i \leq n_a$ and $1 \leq j \leq n_b$) is defined as:

$$d(\vec{p_i}, \vec{q_j}) = \sqrt{(\tilde{m}_i^a - \tilde{m}_j^b)^2 + (\tilde{k}_i^a - \tilde{k}_j^b)^2} \tag{6}$$

Various distance measures have been used and computed here, to represent the distance of the two groups:

- Average Distance (AD):

$$AD = \sum_{i=1}^{n_a} \sum_{j=1}^{n_b} \frac{d(\vec{p_i}, \vec{q_j})}{n_a n_b} \tag{7}$$

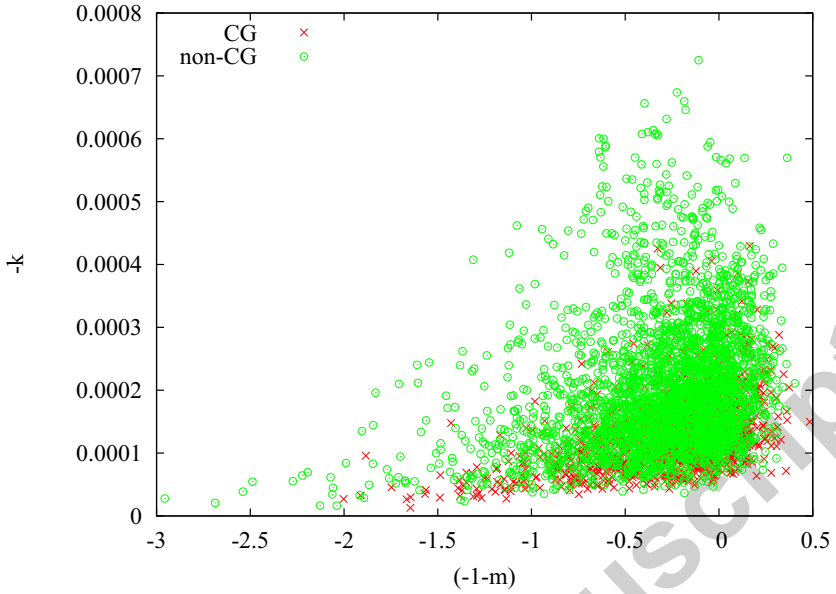

Figure 9: 2-dimensional exponent plot of the organism *Drosophila melanogaster*, chromosome 3, area NT_033777, hexaplets

- Average Minimum Distance (AMD):

$$AMD_a = \frac{1}{n_a} \sum_{i=1}^{n_a} min_i[d(\vec{p_i}, \vec{q_j})]|_{j=1}^{n_b} \tag{8}$$

where $min_i[d(\vec{p_i}, \vec{q_j})]|_{j=1}^{n_b}$ is the minimum of all distances between one point $\vec{p_i}$ in group α and all points $\vec{q_j}$ of group β. Likewise :

$$AMD_b = \frac{1}{n_b} \sum_{j=1}^{n_b} min_j[d(\vec{p_i}, \vec{q_j})]|_{i=1}^{n_a} \tag{9}$$

To compare clustering distances between different data-sets and organisms, we take into account the minimum ($AMD_{min} = min(AMD_a, AMD_b)$) and maximum ($AMD_{max} = max(AMD_a, AMD_b)$) of these two numbers.

- Proximity Factor (PF):

$$PF = \frac{1}{n_a + n_b} \left(\sum_{i=1}^{n_a} S_i^a + \sum_{j=1}^{n_b} S_j^b \right) \tag{10}$$

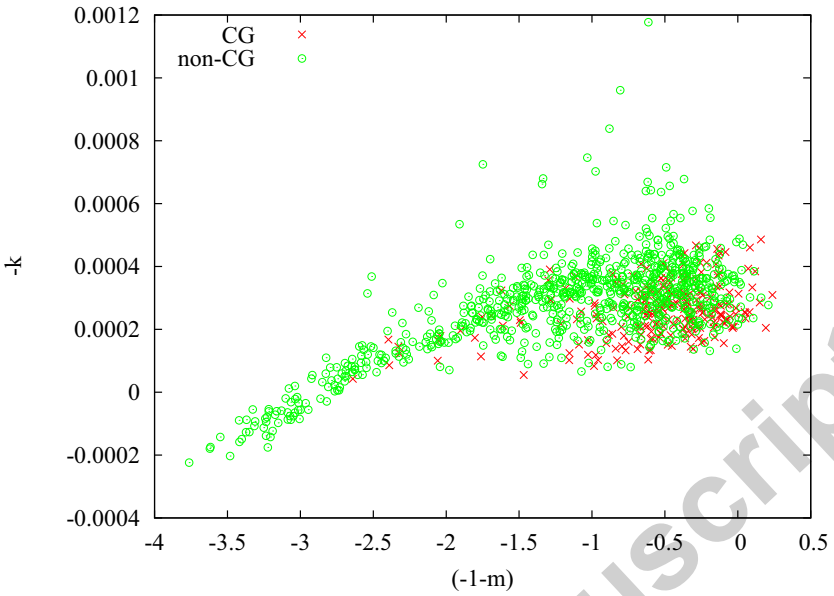

Figure 10: 2-dimensional exponent plot of the organism *Caenorhabditis elegans*, chromosome I, area NC_003279, quintuplets

where

$$S_i^a = \begin{cases} 1, & \text{if } min[d(\vec{p_i}, \vec{p_l})]|_{l=1,l\neq i}^{n_a} \leq min[d(\vec{p_i}, \vec{q_j})]|_{j=1}^{n_b} \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

$$S_j^b = \begin{cases} 1, & \text{if } min[d(\vec{q_j}, \vec{q_g})]|_{g=1,g\neq j}^{n_b} \leq min[d(\vec{q_j}, \vec{p_i})]|_{i=1}^{n_a} \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

$\vec{p_l}$ is a point of group α excluding self, $\vec{q_g}$ is a point of group β excluding self, $min[d(\vec{p_i}, \vec{p_l})]$ is the minimum of all distances between point $\vec{p_i}$ and all other points of group α. The other $min[\ldots]$ functions are likewise defined.

The function S_i^a takes the value 1 if the minimum distance between a specific point of group α and all points of group α is smaller than the minimum distance between this point and all points of group β. Likewise the value S_j^b is computed. These measures are representative of the relative distances between the two groups. Larger values indicate that points are more proximate to the same group. Maximum value is 1, which means that the closest neighbor to any point belongs to the same group as itself, while 0.5 indicates that the points are randomly distributed.

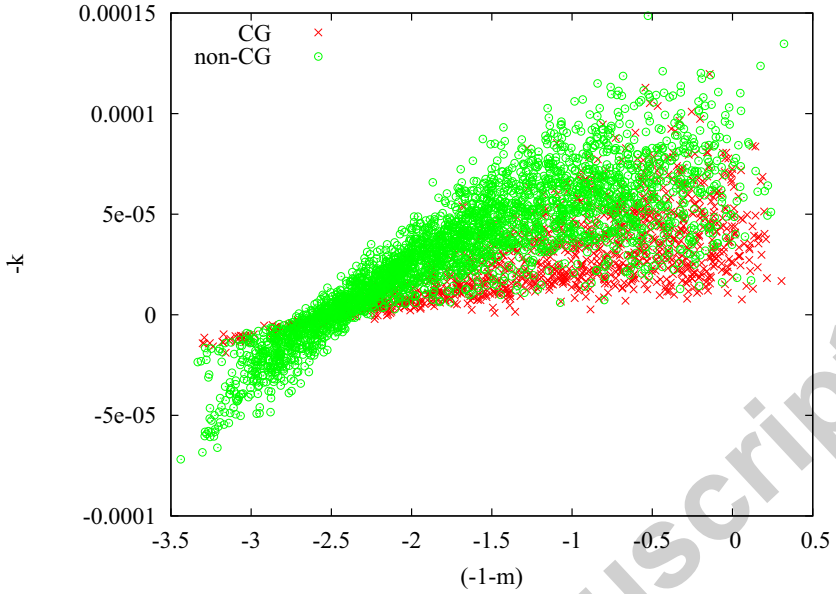

Figure 11: 2-dimensional exponent plot of the organism *Arabidopsis thaliana*, chromosome 1, area NC_003070, hexaplets

- Average Distance Factor (ADF):

$$ADF_a = \frac{\sum_{i=1}^{n_a} \sum_{j=1}^{n_b} \frac{d(\vec{p_i},\vec{q_j})}{n_a n_b}}{\sum_{i=1}^{n_a} \sum_{l=1,l\neq i}^{n_a} \frac{d(\vec{p_i},\vec{p_l})}{(n_a-1)n_a}} \tag{13}$$

and

$$ADF_b = \frac{\sum_{i=1}^{n_a} \sum_{j=1}^{n_b} \frac{d(\vec{p_i},\vec{q_j})}{n_a n_b}}{\sum_{j=1}^{n_b} \sum_{g=1,g\neq j}^{n_b} \frac{d(\vec{q_j},\vec{q_g})}{(n_b-1)n_b}} \tag{14}$$

where $\vec{p_l}$ is a point of group α excluding self and $\vec{q_g}$ is a point of group β excluding self. For comparison we take into account again the minimum ($ADF_{min} = min(ADF_a, ADF_b)$) and maximum ($ADF_{max} = max(ADF_a, ADF_b)$) values among these numbers.

The nominators in Eqs. 13 and 14 represent the average distance between the two groups, as described in Eq. 7. The denominator is the average distance between the points of the same group. If this factor is near 1, means that the average distance between the two groups is the same. If clustering appears and clusters are apart from each other, then this factor should have values greater than 1, since the distances between the two groups should be greater than the distances within the same group.

The results for each of these measures are presented in Table 1 for chromosomes of different organisms.

We observe that some measures better describe the relative position and separation of the two groups than others. The average distance (AD) is unable to distinguish the two groups in organisms *Drosophila m.* and *Arabidopsis th.* The reason for this is the geometry of the clusters. The distances between points belonging to the same group might be greater or smaller than the distances between points of different groups.

The AMD_{min} and AMD_{max} measures demonstrate a clear tendency for larger values in chordata, but still the separation is not absolute. The reason for this is that, in general, the distance between two points in different groups is smaller when the groups intermix and larger when they are separate. We note that we also calculate the average maximum distance between the two groups, which is defined similarly to the AMD, but taking the maximum instead of the minimum. Again, the data produced fails to distinguish the two groups, due to their geometry.

The metric which better describes the clustering of the oligonucleotides is PF. This factor represents the percentage of the points, for which their nearest neighbor belongs to the same group. It is clear that when we have obvious clustering, as in chordata this factor is near 1. In *Arabidopsis th.* and *Drosophila m.* this factor decreases. Even in these last two organisms, the PF factor is away from the value of 0.5, which defines the case of random mixing between the two groups α and β, thus clustering is present in all cases.

Factors ADF_{min} and ADF_{max} also distinguish the two groups. When these factors are equal to 1, the average distance between the points of one group is the same as the average distance between the points from different groups. In other words, all distances between points are the same, on average, for all points. If this factor is greater than one then the average distance is larger for points belonging to different groups than for points belonging to the same group. The more obvious the clustering is, the larger the value of ADF_{min}. In rare cases, as in *Drosophila m.,* we also find that the ADF_{min} factor is less than 1, which could be due to the geometry of the graph, in which large distances appear even within the same group.

As a result we propose the measures PF, ADF_{min} and ADF_{max} to characterize the clustering of the two groups. These factors are able to demonstrate the clustering and provide a quantitative approach for describing the group geometry.

5 Discussion

In this study the distance size distribution between oligonucleotides is studied in many model organisms, using a two-parameter model to take into account short and long tail tendencies (see Eq. 4). The characteristic parameters are depicted in a 2-dimensional scatter plot where clustering appears. It is shown that oligonucleotides tend to cluster depending on their composition, in evolutionary

newer organisms and particularly in mammals, birds and fish. Oligonucleotides which include the consensus sequence of the promoter CG (group α) have the smallest value of the $|-1-m|$ parameter. These values correspond to the smallest value of the critical exponent $|\mu|$ and they follow long range distributions. All other oligonucleotides (group β) seem to follow exponentially decaying distributions or, in general, short range distributions. These two groups of oligonucleotides appear to be completely separated in evolutionary newer organisms.

In evolutionary older organisms and in plants, the clustering of the oligonucleotides is not so evident. Although there appears to be some difference between these two groups, they tend to overlap. The statistical behavior though remains similar. Oligonucleotides of group α still demonstrate values of $|-1-m|$ between [0,2], while the value of $|k| \sim 0$. We recall that in these organisms other prominent consensus sequences coexist (such as TATA) apart from the CG sequence.

This behavior becomes clear when we numerically calculate the distance between groups α and β. By using the Proximity Factor (PF) metric we can clearly demonstrate the clustering. The situation is also mirrored in older organisms, where the clustering is not so obvious, but there is still a small structure present. These results are also supported by other distance metrics introduced here, although not so clearly.

Although the clustering of the distribution parameters is more evident in evolutionary newer organisms, we would like to stress that this behaviour is present in all organisms studied. It seems that the distribution of oligonucleotides containing the CG sequence is conserved and is evolutionary stable, since all organisms inherited and kept it. The distribution of the CG-containing sequences differs from all other sequences, across organisms, and follows power law statistics.

We also propose that the difference between chordata on the one hand and insects and plants on the other might have evolutionary roots. Organisms which appear later than insects seem to have different distribution of small sequences. It seems that the importance of the CG sequence has been "upgraded" by evolutionary forces. This hypothesis is corroborated by the various quantitative analysis measures, which demonstrate not only the clustering described above, but also quantify the differences between evolutionary newer and older organisms.

Unfortunately, DNA decoding for chromosomes of organisms between insects and chordata has not yet begun. For this reason we are unable to draw a line and propose the evolutionary stage in which this separation appeared for the first time. This question remains open and is a future objective for investigation.

It would also be interesting to investigate the oligonucleotide statistical behavior of evolutionary old organisms, such as procaryotes, in which multifractal characteristics have been found (Yu et al. 2001). These organisms have different DNA constitution, in terms of coding / noncoding DNA, as well as different consensus sequences. It is unlikely though to find extreme differences from the eucaryotes, since the general transcription mechanism remains the same.

The current study is also limited by the availability of data characterising many organisms within one class. Usually one or two organisms have been decoded within each class, especially in the case of higher eucaryotes which have many long chromosomes. Still, all the available organisms are used in Modern Biology as "model organisms", representing their class. Certainly, our goal is to improve this study with more organisms, as soon as they become publicly available in the genomic databases and to compare the distribution details between organisms belonging to the same class.

Finally, even though the proposed model is a good description of the distribution of oligonucleotides, there are cases (especially *Arabidopsis th.*) where it is not adequate. Sometimes the distance size distribution appears to have one (or more) plateau. Although this type of distribution is rather rare, it would be interesting to further investigate this phenomenon. A possible solution would be to use more parameters in our model, in an effort to distinguish more clearly oligonucleotides according to their functionality.

6 Acknowledgments

The authors would like to thank Drs. Th. Oikonomou and D. Verganelakis for helpful discussions.

References

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. 1994. *Molecular Biology of the Cell.* Garland Publishing Inc., New York.

Allegrini, P., Barbi, M., Grigolini, P., and West, B. J. 1995. Dynamical model for dna sequences. *PHYS REV E* 52, 5281.

Almirantis, Y. and Provata, A. 1997. The 'clustered structure' of the purines/pyrimidines distribution in DNA distinguishes systematically between coding and non-coding sequences. *Bull. Math. Biol.* 59, 975.

Almirantis, Y. and Provata, A. 1999. Long- and short-range correlations in genome organization. *J. Stat. Phys.* 97, 233.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990. Basic local alignment search tool. *Mol Biol* 215, 403.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. 1997. Gapped blast and pslblast: a new generation of protein database search programs. *Nucleic Acids Research* 25, 3389.

Arneodo, A., d'Aubenton Carafa, Y., Bacry, E., Graves, P. V., Muzy, J. F., and Thermes, C. 1996. Wavelet based fractal analysis of DNA sequences. *Physica D* 96, 291.

Bernaola-Galvan, P., Carpena, P., Oman-Roldan, R., and Oliver, J. 2002. Study of statistical correlations in DNA sequences. *Gene* 200, 105.

Bernaola-Galvan, P., Oliver, J. L., Carpena, P., Clay, O., and Bernardi, G. 2004. Quantifying intrachromosomal GC heterogeneity in prokaryotic genomes. *Gene* 333, 121.

Blake, M. C., Jambou, R. C., Swick, A. G., Kahn, J. W., and Azizkhan, J. C. 1990. Transcriptional initiation is controlled by upstream GC-box interactions in a TATAA-less promoter. *Mol Cell Biol* 10, 6632.

Buldyrev, S. V., Goldberger, A. L., Peng, C. K., Simons, M., and Stanley, H. E. 1993. Generalized levy-walk model for DNA nucleotide sequences. *Phys. Rev. E* 47, 4514.

Campbell, W. H. and Gowri, G. 1990. Codon usage in higher plant, green algae and cyanobacteria. *Plant Physiol* 92, 1.

Carpena, P., Bernaola-Galvan, P., Coronado, A., Hackenberg, M., and Oliver, J. 2007. Identifying characteristic scales in the human genome. *PHYS REV E* 75, 032903.

Carpena, P., Bernaola-Galvan, P., Roman-Rodlan, P., and Oliver, J. 2002. A simple and species-independent coding measure. *Gene* 300, 97.

Czirok, A., Mantegna, R. N., Havlin, S., and Stanley, H. E. 1995. Correlations in binary sequences and a generalized Zipf analysis. *Physical Review E* 52, 446.

Deutsch, M. and Long, M. 1999. Exon-intron structures of eukaryotic model organisms. *Nucleic Acids Research* 27, 3219.

Ebeling, W. and Nicolis, G. 1992. Word frequency and entropy of symbolic sequences: a dynamical perspective. *Chaos, Solitons and Fractals* 2, 635.

Etienn-Decant, J. 1988. *Genetic Biochemistry - From gene to protein.* Ellis Horwood Limited, N.Y.

Hackenberg, M., Previti, C., Luque-Escamilla, P., Carpena, P., Martinez-Aroza, J., and Oliver, J. 2006. CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics* 7, 446.

Hao, B. L. 2000. Fractals from genomes - exact solutions of a biology-inspired problem. *Physica A* 282, 225.

Hawkins, J. D. 1988. A survey on intron and exon lengths. *Nucleic Acids Research* 16, 9893.

Higgins, D. G., Bleasby, A. J., and Fuchs, R. 1992. Clustal V: improved software for multiple sequence alignment. *CABIOS* 8, 189.

Hoffman, E. P., Gerring, S. L., and Corces, V. G. 1987. The ovarian, ecdysterone, and heat-shock-responsive promoters of the Drosophila melanogaster hsp27 gene react very differently to perturbations of DNA sequence. *Mol Cell Biol* 7, 973.

Karlin, S. and Brendel, V. 1992. Chance and statistical significance in protein and DNA sequence analysis. *Science* 257, 39.

Katsaloulis, P., Theoharis, T., and Provata, A. 2002. Statistical distributions of oligonucleotide combinations: applications in human chromosomes 21 and 22. *Physica A* 316, 380.

Katsaloulis, P., Theoharis, T., and Provata, A. 2005. Statistical algorithms for long dna sequences: Oligonucleotide distributions and homogeneity maps. *Scientific Programming* 13, 177.

Katsaloulis, P., Theoharis, T., Zheng, W. M., Hao, B. L., Bountis, A., Almirantis, Y., and Provata, A. 2006. Long-range correlations of rna polymerase ii promoter sequences across organisms. *Physica A* 366, 308.

Lander, E. S. and et al., L. M. L. 2001. Initial sequencing and analysis of the human genome. *Nature (London)* 409, 860.

Levenberg, K. 1944. A method for the solution of certain problems in least squares. *Quart. Appl. Math.* 2, 164.

Li, W. and Kaneko, K. 1992. Long-range correlations and partial 1/F-Alpha spectrum in a noncoding DNA sequence. *Europhys. Lett.* 17, 655.

Li, W. T. and Holste, D. 2005. Universal 1/f noise, crossovers of scaling exponents, and chromosome-specific patterns of guanine-cytosine content in dna sequences of the human genome. *Phys. Rev. E* 71, 041910.

Li, W. T., Marr, T. G., and Kaneko, K. 1994. Understanding long-arnge correlations in dna sequences. *Physica D* 75, 392.

Li, W. T. and Miramontes, P. 2006. Large-scale oscillation of structure-related dna sequence features in human chromosome 21. *Phys. Rev. E* 74, 021912.

Lim, L. P. and Burge, C. B. 2001. A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl. Acad. Sci. USA* 98, 11193.

Marquardt, D. 1963. An algorithm for least-squares estimation of nonlinear parameters. *SIAM J. Appl. Math.* 11, 431.

Pearson, W. R. and Lipman, D. J. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* 85, 2444.

Peng, C. K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M., and Stanley, H. E. 1992. Long-range correlations in nucleotide sequences. *Nature* 356, 168.

Provata, A. and Almirantis, Y. 1997. Scaling properties of coding and non-coding DNA sequences. *Physica A* 247, 482.

Provata, A. and Almirantis, Y. 2000. Fractal cantor patterns in the sequence structure of DNA. *Fractals* 8, 15.

Provata, A. and Oikonomou, T. 2007. Power law exponents characterising the human dna. *Phys. Rev. E* 75, 056102.

Sakharkar, M., Passetti, F., de Souza, J. E., Long, M., and de Souza, J. S. 2002. Exint: an exon intron datbase. *Nucleic Acids Research* 30, 191.

Scafetta, N., Latora, V., and Grigolini, P. 2002. Levy scaling: the diffusion entropy analysis applied to dna sequences. *PHYS REV E* 66, 031906.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Computational Biology* 22, 4673.

Venter, J. C. and al. 2001. The sequence of the human genome. *Science* 291, 5507.

Vinogradov, A. E. 1999. Intron-genome size relationship on a large evolutionary scale. *J Mol Evol* 49(3), 376.

Voss, R. 1992. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phys. Rev. Lett.* 68, 3805.

Yu, Z.-G., Anh, V., and Lau, K.-S. 2001. Multifractal characterisation of length sequences of coding and noncoding segments in a complete genome. *Physica A* 301, 356.

Yu, Z.-G., Anh, V., and Wang, B. 2000. Correlation property of length sequences based on the global structure of complete genomes. *Physical Review E* 63, 11903.

Chromosome	AD	AMD_{min}	AMD_{max}	PF	ADF_{min}	ADF_{max}
Dros. m.	0.328223	0.0099115	0.0311811	0.733398	0.88209	1.39532
Arab. th.	0.332287	0.0083081	0.0171636	0.758545	1.08311	1.16287
Dan. r. 1,5	0.701532	0.0903898	0.125794	0.946289	2.42046	4.98622
Dan. r. 1,6	0.398682	0.0238059	0.0631456	0.9104	1.21719	1.92106
Dan. r. 25,5	0.440494	0.0497386	0.102181	0.923828	1.81863	4.15382
Gal. g.	0.459963	0.106491	0.309101	0.99707	1.42238	4.42571
Mus m.	0.418231	0.1424	0.317655	1	1.5331	2.79288
Pan t.	0.511827	0.0686806	0.434286	0.99707	1.60466	10.6364
Homo s. 21,5	0.363733	0.0960787	0.184398	0.998047	1.20283	3.58647
Homo s. 21,6	0.313492	0.0613419	0.178214	0.994385	1.28021	2.56637
Homo s. 22,5	0.752239	0.160707	0.186834	0.976562	3.25357	3.35892
Homo s. 22,6	0.387476	0.0698528	0.0929383	0.953857	1.17362	2.2189

Table 1: Numerical results of the various clustering measures. The chromosomes taken into account are quintuplets of area NT_037436 of *Drosophila m.*, chromosome 1 hexaplets of *Arabidopsis th.*, chromosome 1 quintuplets of *Danio r.*, chromosome 1 hexaplets of *Danio r.*, chromosome 25 quintuplets of *Danio r.*, chromosome 1 quintuplets of *Gallus g.*, chromosome 19 quintuplets of *Mus m.*, chromosome 1 quintuplets of *Pan t.* and chromosomes 21 and 22 quintuplets and hexaplets of *Homo s.*