ARTICLE IN PRESS

Available online at www.sciencedirect.com
SCIENCE DIRECT®

Physica A ▮ (▮▮▮▮) ▮▮▮–▮▮▮

PHYSICA A

www.elsevier.com/locate/physa

Statistical analysis of simple repeats in the human genome

F. Piazza[a,*], P. Liò[b]

[a]*Laboratoire de Biophysique Statistique, ITP-FSB, École Politechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland*
[b]*Computer Laboratory, University of Cambridge, Cambridge CB3 0FD, UK*

Received 4 May 2004; received in revised form 14 July 2004

Abstract

The human genome contains repetitive DNA at different level of sequence length, number and dispersion. Highly repetitive DNA is particularly rich in homo- and di-nucleotide repeats, while middle repetitive DNA is rich of families of interspersed, mobile elements hundreds of base pairs (bp) long, among which belong the Alu families. A link between homo- and di-polymeric tracts and mobile elements has been recently highlighted. In particular, the mobility of Alu repeats, which form 10% of the human genome, has been correlated with the length of poly(A) tracts located at one end of the Alu. These tracts have a rigid and non-bendable structure and have an inhibitory effect on nucleosomes, which normally compact the DNA. We performed a statistical analysis of the genome-wide distribution of lengths and inter-tract separations of poly(X) and poly(XY) tracts in the human genome. Our study shows that in humans the length distributions of these sequences reflect the dynamics of their expansion and DNA replication. By means of general tools from linguistics, we show that the latter play the role of highly-significant content-bearing terms in the DNA text. Furthermore, we find that such tracts are positioned in a non-random fashion, with an apparent periodicity of 150 bases. This allows us to extend the link between repetitive, highly mobile elements such as Alus and low-complexity words in human DNA. More precisely, we show that Alus are sources of

*Corresponding author.
E-mail addresses: francesco.piazza@epfl.ch (F. Piazza), pietro.lio@cl.cam.ac.uk (P. Liò).

0378-4371/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2004.08.038

poly(X) tracts, which in turn affect in a subtle way the combination and diversification of gene expression and the fixation of multigene families.
© 2004 Elsevier B.V. All rights reserved.

PACS: 87.10.+e; 02.50.Cw

Keywords: DNA; Homopolymeric repeats (HR); Alu repeats; Mobile elements; Human genome

1. Introduction

Experiments on kinetics of DNA denaturation and renaturation and the analysis of DNA sequences have revealed that most of our genome is populated by DNA repeats of different length, number and degree of dispersion [1]. Long repeats in few copies are usually orthologous genes, which may contain hidden repeats in the form of runs of amino acids, and retroviruses inserted in the genome. For example, the human genome contains more than 50 chemokine receptor genes which have high sequence similarity [2] and almost one thousands olfactory receptor genes and pseudogenes [3]. Short repetitive DNA sequences may be categorized in high and middle repetitive. The first is formed by tandemly clustered DNA of variable length motifs (5–100 bp) and is present in large islands of up to 100 Mb. The middle-repetitive can be either short islands of tandemly repeated microsatellites/minisatellites ('CA repeats', tri- and tetra-nucleotide repeats) or mobile genetic elements. Mobile elements include DNA transposons, short and long interspersed elements (SINEs and LINEs), and processed pseudogenes [4,5].

Why should we be interested in repetitive DNA? Tandem repeats with 1–3 base motif can differ in repeat number among individuals; therefore, they are used as genetic markers for assessing genetic differences in plants and animals and in forensic testing. It is known that trinucleotide repeats are involved in several human neurodegenerative diseases (e.g., fragile X and Huntington's disease), and instability of short tandemly repeated DNA has been associated with cancer [6–8]. DNA repeats increase DNA recombination events and have the potential to destroy (by insertional mutagenesis), to create (by generating functional retropseudogenes), and to empower (by giving old genes new promoters or regulatory signals). Despite their importance in genome dynamics and for medical diagnosis, and despite the advances in the understanding of their role in several prokaryotic and eukaryotic genomes [9–13], a robust, genome-wide, statistical analysis of interspersed repetitive elements in human genome is still lacking. In particular, the analysis of short repetitive DNA needs to fully exploit the relationship between simple repeats and mobile elements.

Interestingly, all mobile elements such as SINEs, LINEs, and processed pseudogenes, contain A-rich regions of different length [14]. In particular, the Alu elements, present exclusively in the primates, are the most abundant repeat elements in terms of copy number ($>10^6$ in the human genome) and account for more than 10% of the human genome [1]. They are typically 300 nucleotides in length, often

form clusters, and are mainly present in non-coding regions. Higher Alu densities were observed in chromosomes with a greater number of genes and vice versa. Alus have a dimeric structure and are ancestrally derived from the gene 7SL RNA. They amplify in the genome by using a RNA polymerase III-derived transcript as template, in a process termed retroposition [15,16]. The mobility is facilitated by a variable-length stretch of an A-rich region located at the 3' end [17,18].

Although all Alus elements have poly(A) stretches, only a very few are able to retropose [14]. Therefore, the mere presence of a poly(A) stretch is not sufficient to confer on an Alu element the ability to retropose efficiently. However, the length of the A stretch correlates positively with the mobility of the Alu [19].

The Alu repeats are divided into three subfamilies on the basis of their evolutionary age: Alu J (oldest), S (intermediate age) and Y (youngest) [20]. There is an inverse correlation between the age of the Alu subfamily and the proportion of the members with long A-tails in the genome, indicating that loss of A stretches may be a primary, though not the only, inactivating feature in the older subfamilies [19].

In this study, we first investigate exhaustively the distribution and characteristic length size of all homopolymeric repeats (HR) of the kind poly(X) in the complete human genome, where X∈[A,C,G,T]. By means of simple tools drawn from linguistics, we show that stretches of homopolymeric repeats play a highly specialized role in the DNA text. In addition, we show that the former are more specific words within the human genome with respect to other repeats coded from different alphabets (see Table 1 for a list of alphabets considered here).

We quantify this effect by studying the characteristic positioning patterns of stretches of given composition and length. We then focus on long A stretches in human chromosomes 20, 21 and 22. The latter chromosomes differ substantially in both Alu density and gene density. Chromosomes 21 and 22, for example, are of similar size (together about 1.6% of the human genome), even though chromosome 22 has four times as many genes and twice as many Alu repeats [21]. The comparative analysis of genome-wide distribution of poly(A) and other homo-dinucleotide polymers allows us to examine the mechanisms and constraints of poly(A) elongation or shortening, and how the elongation dynamics is related to the evolutionary instabilities [4,22], DNA bendability [23], and nucleosome inhibition [24].

Table 1
List of different alphabets used to code simple repeat stretches

Alphabet	*A*	*C*	*G*	*T*
S/W	W	S	S	W
R/Y	R	Y	R	Y

W (weak) and S (strong) code with respect to the strength of the inter-strand H–bond between pairs of complementary nucleotides. R (purine), Y (pyrimidine).

2. Length distributions of poly(X) repeats in human genome

For our study of homopolymeric tracts in the human genome, we used all the finished sequences of the 24 chromosomes, among which were the published sequences of chromosomes 21 and 22, as well as a set of compiled sequences together covering about 3 Giga bases (Gb). DNA sequences were obtained from the Genbank directory of the web site of the National Center for Biotechnology Information (ftp://ncbi.nlm.nih.gov/).

We find that homopolymeric tracts of the type poly(X) are substantially over-represented in all human chromosomes. A first simple quantitative measure of over-abundance of poly(X) strings can be obtained by calculating the conditional probabilities $p(X_{i+1}|X_i)$, i.e., the probabilities of finding nucleotide X at position $i+1$, given that position i is also occupied by X. In the totally uncorrelated case (sometimes referred to as the Bernoulli case) $p(X|X) = p(X)$, whereas $p(X|X) \neq p(X)$ denotes the presence of spatial correlations along the sequence. This effect in the human genome is depicted in Fig. 1, where we show as an example data from chromosomes 1, 3, 5, and 7. The presence of positive correlations (here measured at the di-nucleotide level) is very clear.

A more quantitative analysis of the correlations underlying the structure of poly(X) words can be performed by studying the length distributions of such repetitive sequences. For this purpose, it is better to work with the cumulative (integrated) distributions, in order to gain some statistical weight in the low-frequency regions. These can be easily obtained from the data by noting that they are nothing but rank-size plots with the axes inverted.

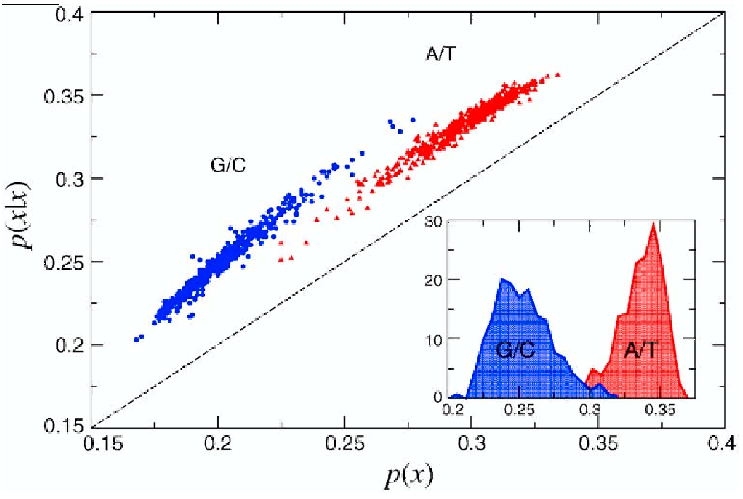

Fig. 1. Human Chromosomes 1, 3, 5 and 7. Plot of the conditional probabilities $p(X|X)$ vs. the single-nucleotide probabilities $p(X)$ for X=A,T and X=C,G. The dashed line marks the totally uncorrelated (Bernoulli) case $p(X|X) = p(X)$. Inset: normalized histograms of the conditional probabilities. The probabilities are measured in windows of length 22 Kbp along the chromosomes.

We find that each chromosome displays approximately the same statistical properties as for the length distributions of poly(X) strings, in both coding and non-coding regions. In particular, the main features of such distributions may be summarized as follows.

In non-coding regions there exists a clear difference between the distributions of lengths of W-type words (poly(A) and poly(T)) and S-type ones (poly(C) and poly(G)). Poly(A) and poly(T) words are much more represented and with much longer words than poly(C) and poly(G) words (see Fig. 2). We do not observe the same difference in coding regions (Fig. 3). In particular, the characteristic trends of such distributions in non-coding regions are the same for inter-genic tracts and for introns.

In Fig. 4, we compare the cumulative length distributions of poly(A) tracts in coding, non-coding inter-genic and intronic regions in a representative case (chromosome 1, about 8% of the genome) with the same distribution from a fictitious sequence of the same length which has been randomly reshuffled according to the same single-nucleotide probabilities and coding/non-coding pattern. We note that the latter reshuffling reproduces in the fictitious sequence the same set of exons as from the true sequence. It is clear that non-coding DNA displays a marked over-abundance of poly(A) words, whereas the length distribution found in coding tracts nicely fits within the random uncorrelated scenario.

The distributions of all poly(X) sequences in all other human chromosomes display the same statistical properties (data not shown).

In general, the peculiar behaviour of length distributions of poly(X) words in non-coding regions are well described by a sum of two exponential laws, with different length constants and weights. It is clear from the data shown in Fig. 4 that the first exponential law describes a trend which is to a good extent common to coding and non-coding regions. The second exponential law takes over from a characteristic length onward. The latter behaviour is very clear for poly(A) and poly(T) words, while it is less marked for poly(C) and poly(G) words. This makes the identification

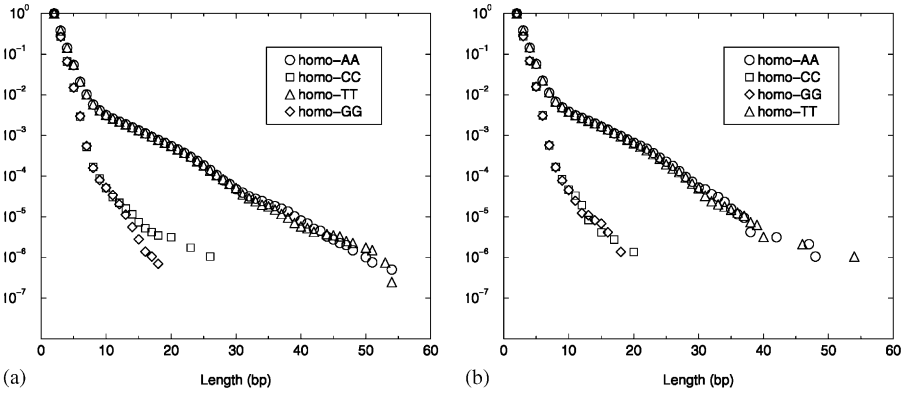

Fig. 2. Cumulative length distributions of poly(X) words (X=[A,C,T,G]) in non-coding regions in a representative case (Chromosome 1): (a) Intergenic tracts; (b) Introns.

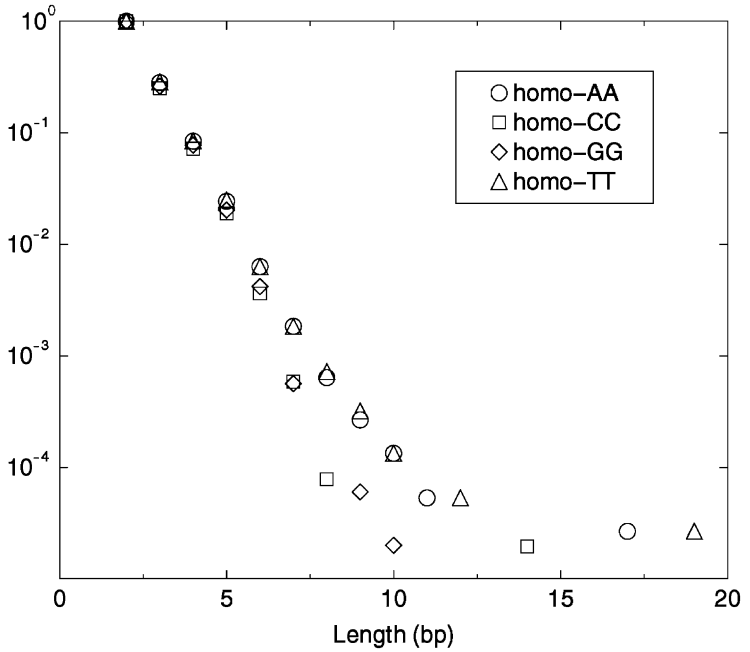

Fig. 3. Cumulative length distributions of poly(X) words (X=[A,C,T,G]) in coding regions in a representative case (Chromosome 1).

of a crossover in the length distributions of S-type words as ambiguous. In the case of poly(A) and poly(T) words, the crossover length can be calculated by separately fitting the first and second portions of the length distributions. We call L_0 the crossover length of the cumulative distributions $\mathscr{C}_h$:

$$\mathscr{C}_h(L) = \begin{cases} A_1 \exp[\log(p_1)L] & \text{for } L < L_0\,, \\ A_2 \exp[\log(p_2)L] & \text{for } L > L_0\,. \end{cases}$$

Hence, we can calculate L_0 as

$$L_0 = -\frac{\log(A_1/A_2)}{\log(p_1/p_2)}\,. \tag{1}$$

The crossover length l_0 in the probability density will then be given by

$$l_0 = L_0 + \Delta l\,, \tag{2}$$

where

$$\Delta l = -\frac{\log[\log(p_1)/\log(p_2)]}{\log(p_1/p_2)}\,. \tag{3}$$

The best-fit values of the parameters A_i and p_i $(i = 1, 2)$ are reported for chromosome 1 in Table 2 along with the corresponding values of L_0 and l_0. We found similar results in the other chromosomes. It is clear from the measured

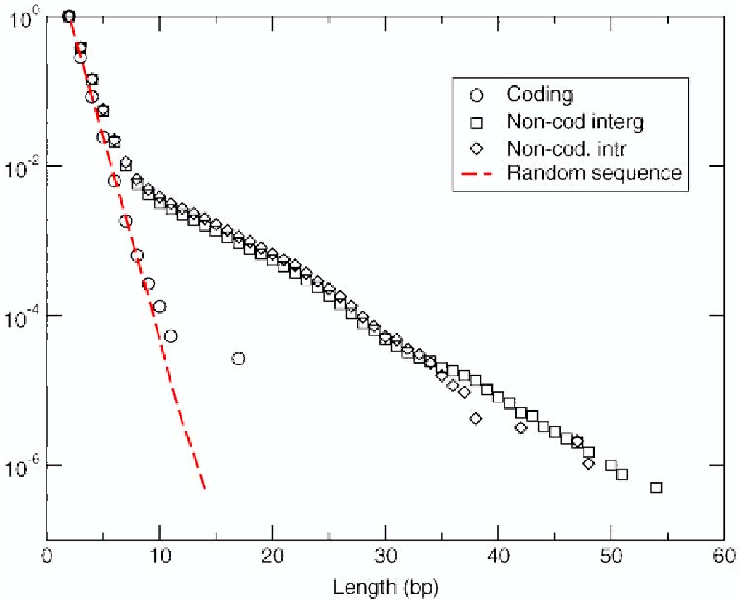

Fig. 4. Comparison of the cumulative distributions of poly(A) tracts in coding and non-coding regions with a simple Bernoulli model (representative case of chromosome 1).

Table 2
Results of the fits performed with the function (1) on the poly(X) length distributions in a representative case (chromosome 1)

	A_1	p_1	A_2	p_2	L_0 (bp)	l_0 (bp)
Poly(A)						
Coding	12.7	0.28	—	—	—	—
Non–cod.	6.09	0.39	0.029	0.81	7.4	9.45
Poly(T)						
Coding	12.8	0.28	—	—	—	—
Non–cod.	6.09	0.39	0.026	0.81	7.6	9.65
Poly(C)						
Coding	21.85	0.23	—	—	—	—
Non–cod. interg.	24.0	0.22	2.5×10^{-4}	0.8	8.9	10.4
Non–cod. intron.	23.5	0.22	1.4×10^{-3}	0.7	8.4	9.65
Poly(G)						
Coding	22.6	0.23	—	—	—	—
Non–cod. interg.	23.4	0.22	0.014	0.57	7.8	8.84
Non–cod. intron.	23.3	0.22	1.95×10^{-3}	0.67	8.4	9.6

Note that no appreciable difference is found in the best-fit values of the floating parameters between intergenic and intronic regions for the distributions of poly(A) and poly(T) words.

single-nucleotide and di-nucleotide probabilities (see Fig. 1) that p_1 matches to a good extent the single-nucleotide probabilities for poly(C) and poly(G) words, while it corresponds to the di-nucleotide probabilities $p(A|A)$ and $p(T|T)$ for poly(A) and poly(T) words. This means that in the first region ($L<l_0$) the length distribution of poly(X) words can be adequately reproduced by treating words as uncorrelated or short-range correlated sequences. This conclusion holds for both coding and non-coding regions. On the contrary, it is apparent that the quantity p_2 cannot be associated with any of the single- or di-nucleotide probabilities, nor with any m-nucleotide probability, with $m>2$. This finding is consistent with the general acceptance of the failure of zeroth- and first-order Markovian models of simple repeats to fully account for "linguistic" features of non-coding DNA [13].

Such intrinsic long-range effect is a signature of the peculiarities of DNA replication dynamics. In particular, it is also indicative of a particular mutual-help relationships between A-tracts and Alu repeats. The change in slope of length distribution of poly(A) tracts reflects the characteristics of DNA replication process. At a certain length of the repeats the replicative process is more prone to errors. The poly(A) tracts elongate during DNA replications by a slippage mutation mechanism, in such a way that the longer the tract the more likely it is to change (elongate or shrink) [25,26]. The dynamical process explains qualitatively the slope change in the length distribution of poly(A) tracts at a specific length, which is the same in all chromosomes. Our estimates of such characteristics length $l_0 \approx 10$ correlates well with the size of the open complex of DNA during the replication (10–12 bp). On the other hand, Alu elements multiply within the genome through RNA polymerase III-derived transcripts in a process termed retroposition; thus contributing with their poly(A) tails to the spreading of A-tracts through the genome. This sort of mutual-help relationship between Alus and A-tracts is the major finding reported in this paper.

In the next section, we shall establish in a clear quantitative way such link between poly(A) tracts and Alus in the human genome.

2.1. Statistics of separations between consecutive poly(X) words

In this section, we analyze the positioning patterns of poly(X) words along the sequence of human chromosomes. Very generally, the statistics of separations between consecutive words is a very useful tool in linguistics to isolate content-bearing terms from generic ones. In general, the former words will tend to cluster themselves as a consequence of their high specificity (attraction or repulsion), while the latter ones will have a tendency to be evenly distributed across the whole text. In order to eliminate the dependency on frequency for different words, it is convenient to analyze the sequences of normalized separations between consecutive words of length $L, s = x(L)/\langle x(L)\rangle$. If homogeneous tracts were distributed at random in the genome, the inter-tract distribution $P_L(s)$ for words of length L would be a Poissonian

$$P_L(s) = \mathrm{e}^{-s}\,.$$

As a consequence, we expect that non-specific words will run close to a Poisson law, while larger deviations should occur for highly specific content-bearing words. Such analysis may be implemented systematically in a quantitative fashion by studying the standard deviations $\sigma_L = \sqrt{\bar{s}^2 - \bar{s}^2}$ of the distributions $P_L(s)$, which is the simplest variable for characterizing a normalized distribution and its fluctuations. For a Poisson distribution $\sigma_L = 1$, while if there is attraction, $\sigma_L > 1$. In the case of repulsion among words one should expect $\sigma_L < 1$.

In Fig. 5, we compare the results of our analysis on separations among tracts of direct repeats of given length in the whole human genome with the result of the same analysis performed on short repeats of the type poly(XY), coded according to the alphabets reported in Table 1. The normalized histograms refer to distances between HRs of length $L \geqslant 2$. The figure clearly evidences that direct repeats are more highly specialized like higher-content words with respect to words coded in the other alphabets, namely Purine/Pyrimidine (poly(AG) and poly(CT)) and weak/strong (poly(AT) and poly(GC)). Interestingly, we find that the regions where there is attraction ($\sigma_L > 1$) or repulsion ($\sigma_L < 1$) they are systematically associated with strings shorter or longer than a characteristic length of about 25 bp, respectively. This effect is shown in a representative case in Fig. 6. We remark that we systematically obtain analogous curves for all human chromosomes.

It is important to remark that the parameters σ_L are estimated from finite series of spacings. We may calculate the uncertainty $\Delta\sigma_L$ associated with such estimates in the hypothesis of random positioning of HRs. In general, one has $\Delta\sigma_L = t_\beta[\mu_4/N_L - (N_L - 3)/N_L(N_L - 1)\mu_2]$, where N_L is the number of tracts of length L in the series, μ_n is the nth order moment, and t_β specifies the required confidence level (e.g., $t_\beta = 1$

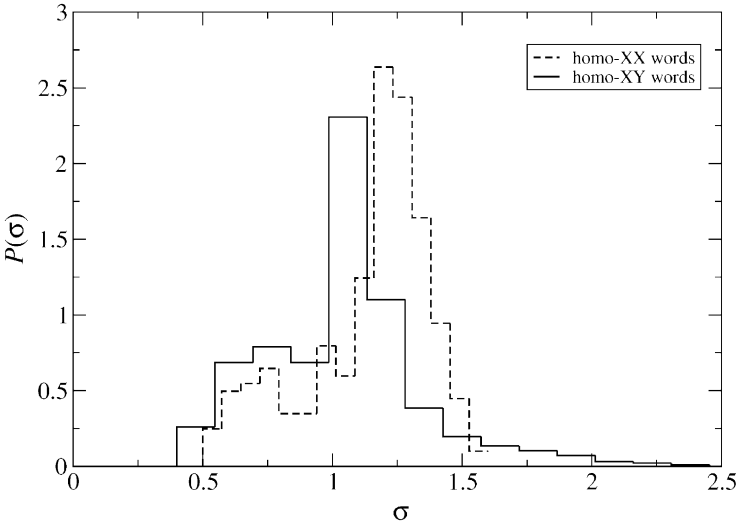

Fig. 5. Normalized histograms of all standard deviations of the distributions $P_L(s)$ for $L \geqslant 2$ in the whole human genome. Comparison of sequences of direct repeats of the kind poly(X) and repetitive sequences of the kind poly(XY), coded according to the alphabets reported in Table 1.

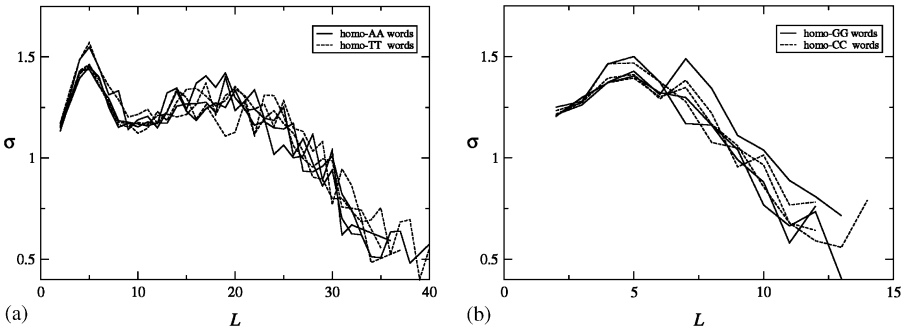

Fig. 6. Plots of the standard deviation of the distributions $P_L(s)$ vs. L in a representative case (Chromosomes 1, 3, 5): (a) poly(A) and poly(T) words; (b) poly(C) and poly(G) words.

for a confidence of one standard deviation, $\beta = 0.68$). Under the hypothesis of random positioning of HRs, we get $\mu_4 = 9$, $\mu_2 = 1$. Hence, at 68% confidence, we have

$$\Delta\sigma_L = \frac{8N_L - 6}{N_L(N_L - 1)} \,. \tag{4}$$

We have employed formula (4) to check the statistical significance of the values of σ_L deduced from our series of spacings. As an example, for $L = 25$, we have in the worst case, $N_{25} = 45, \sigma_{25} = 1.05$ (chr. Y), and hence, a relative error $\Delta\sigma_{25}/\sigma_{25} \approx 0.17$. More generally, the number of tracts N_L grows exponentially for $L<25$. This means that we rapidly get very small errors on the estimates of σ_L in the region of HR clustering. By the same token, in the region $L>25$ we may still state that the observed repulsion among tracts is statistically significative for nearly all chromosomes, chromosomes smaller than the 20th being at the limit of statistical significance.

We interpret the clustering of poly(A) words shorter than about 25 bp in terms of the correlation between Alu repeats and their flanking motifs of direct repeats of A stretches [17]. Our results are consistent with recent finding by Holste et al., who show that histograms of distances between adjacent Alu repeats show significant deviations from an exponential decay, expected from random chromosomal positions of repeats [10]. In other words, the content-bearing clustering of poly(A) oligomers shorter than about 25 bp may be interpreted in terms of clustering of Alu repeats. This conclusion establishes an important connection between the dynamics of A stretches and that of Alu repeats in the human genome.

2.2. *Distribution of separations in the* 100–400 *bp region*

In order to further investigate the relation between the positions of poly(A) sequences and of Alu repeats, we computed the distances between poly(A)–poly(T) tracts considering only distances smaller than 800 bases. We took that precaution in order to neglect other chromosome structures that may be present on larger length

scales and that may be subject of separate studies in the future. For this purpose, we made use of kernel density plots to analyze distance distribution. Histograms depend on the starting point on the grid of bins and the differences between histograms realized with different choices of the bin grid can be surprisingly large. Another way to look at the result is to compute the histograms averaging over a large number of shifts of starting points and having very small bins [27]. Kernel density estimators are smoother than histograms and converge faster to the true density. The density function, which has unit total area is computed through the following formula:

$$d(s) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{b}K\left(\frac{s-s_j}{b}\right). \tag{5}$$

for a sample of distance values $s_1, s_2, ..., s_n$, a given kernel function $K(x)$ and a bandwidth parameter b. It can be shown theoretically that the choice of the kernel is not crucial [28], whereas the choice of the proper bandwidth is the important issue. The correct choice is a compromise between smoothing enough to remove insignificant bumps and not smoothing too much to smear real peaks away. We used a Gaussian kernel and we selected data-dependent bandwidths, using the formula $b = 0.9\ \min(\sigma, R/1.34)n^{-1/5}$, where n is the sample size, σ is the standard deviation, and R is the inter-quantile range [29]. Applying this formula, we determined the bandwidths for the human genome to be $b = 38.01$ bp. Bandwidths for different genomes resulted in very similar values.

In Fig. 7, we show the density plots of the distances between 12 bp poly(A)–poly(T) tracts in the human genome (a) and for chromosomes 21 and 22 (b). Many of the 12 bp tracts in the human genome sequences were found to be positioned at intervals of either ≈ 150 or ≈ 300 bp. We found similar patterns in all human chromosomes. Although chromosome 22 has the highest tract density of the two (0.329 tracts/Kbp), its 300 bp band is very pronounced and much bigger than that for chromosome 21, which has a much lower tract density (0.217/Kbp). Using the chromosome 22 and 21 human sequences, we also analyzed poly(A) and poly(T) tracts >8 bp in length, and also this yielded distinct peaks at ≈ 150 and ≈ 300 bp (data not shown). Other types of repeat sequences which are not thought to be rigid [23,30,31], such as poly(AT) repeats of 12 bases, did not show this distinct periodicity. Next, we investigated distances between longer poly(A)–poly(T) tracts and found that tracts longer than 25 bases are located at longer and non-periodical distances. This finding is consistent with the general tendency of stretches of direct repeats longer than about 25 bp to repel each other, which has been established in the previous section. This might reflect the tendency of very long rigid homopolymeric tracts to be accommodated in the linker regions between nucleosomes so as to be as more scattered as possible in order to favor the tightest packing of chromatine. Although in human chromosomes, the densities of poly(A)–poly(T) tracts of 12 bases or more range from 0.188/Kbp (chromosomes Y) to 0.374/Kbp (chromosome 16), there is in all a clear 150 bp periodicity, and in most of the chromosomes the largest peak is the one at ≈ 300 bp. The different intensities of the 150 and 300 bp periodicities in human chromosomes reveal the clustering of Alus of 150 and 300 bp.

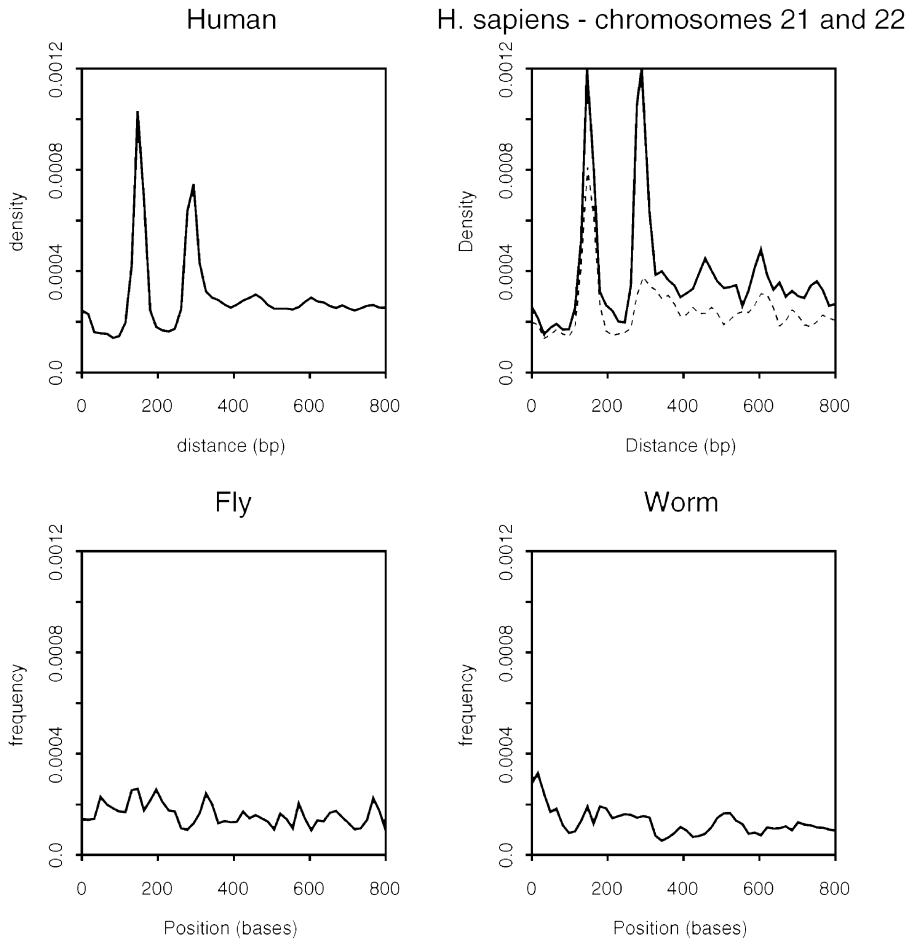

Fig. 7. Distribution of poly(A) and poly(T) tracts in different eukaryotes. Density plots of distances between poly(A) and poly(T) tracts of 12 bp (line) derived from (from top to bottom and left to right): (a) *Homo Sapiens* entire genome; (b) *Homo Sapiens* chromosomes 21 (dotted line) and 22 (solid line); (c) the *Drosophila melanogaster* genome sequence (113.5 Mb); and (d) the *Caenorabditis elegans* genome sequence (87.6 Mb).

The 150 bp reflects the dimeric structure of Alus and shows that the central, short poly(A)–poly(T) tract often elongates [17].

In Fig. 8, we plot the density of poly(A) tracts longer than 12 bp (i.e., the number of tracts divided by the length of the chromosome) versus the density of Alus for each chromosome. The plot shows that poly(A) longer than 12 bp are less abundant than Alus. Moreover, it is clear that a linear correlation exists between the density of poly(A) and the density of Alus over the whole genome. If Alus were completely clustered, we would observe 100% of the area under three peaks at 150, 300 and 450 bp. The absence of the peak at 450 bp means that Alu clusters contain elements

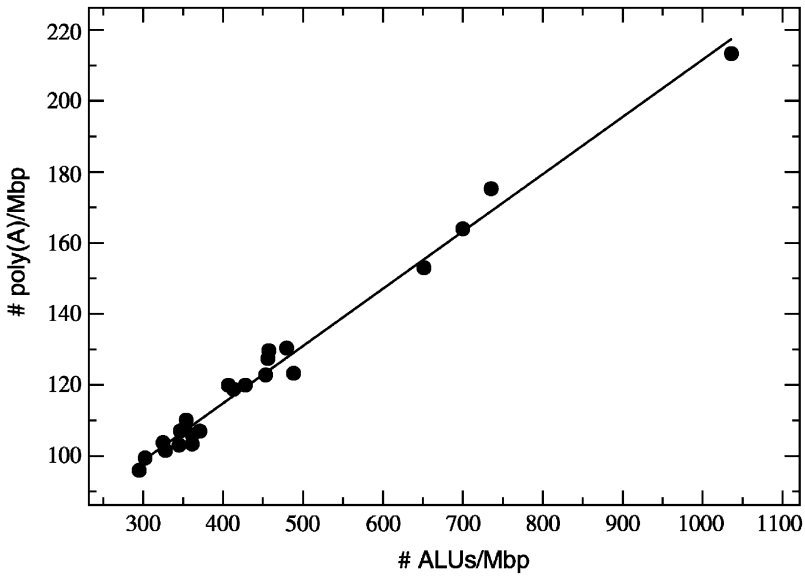

Fig. 8. Density of poly(A) tracts longer than 12 bp vs. density of Alus for the 24 human chromosomes (symbols) and linear fit (correlation coefficient $r = 0.993$).

of the same length (either 150 or 300 bp). We found that Alus are tandemly clustered in small groups along each chromosome and that the different spacing between consecutive clusters explains the peaks at 150 and 300 bp. Note that the genomes of certain other eukaryotes such as *C. elegans* (worm) and *D. melanogaster* (fly) that do not have Alu-like sequences, do not show clear peaks and have much lower density of poly(A) (see Fig. 7).

3. Discussion

The sequence of the human genome is highly repetitious at different sequence length-scales and the coding sequences comprise less than 5% of it. Many of human genome repeats can be found in mature mRNA and total cellular RNA. RNAs containing repetitive elements include Alu-containing mRNAs which amount to 5% of all known mRNAs [1,32].

Patterns of homo- and di-nucleotide expansion in human genome suggest an explanation as to why, contrary to vertebrate, low eukaryotes and bacteria avoid the genome-wide accumulation and expansion of tandem repeats. Selection acting on all of the repeats in a bacterial genome would generate a very high mutational load (loss of many individuals due to selection) and would have to act against very small incremental increases in genome size or repetitivity, many of which would be expected to have minimal phenotypic effects. This high cost of selection is tolerated in bacteria by the large population size and the short cellular

division time. The spreading of repeats across vertebrate genomes occurred as non-adaptive when the organism size increased and the population-size decreased. In vertebrates, the selection against HRs may be neutral or weakly negative because these regions are localised in the periphery of the nucleus, where they are replicated lately [33,34]. Moreover, even if they are inserted in the regulatory or coding regions of genes, there is a high chance that metabolic or genetic redundancy would buffer the effect.

3.1. Why As strings are more abundant than other strings?

Strings of As and Ts have several peculiar properties not shared by strings of Gs and Cs. They are very rigid, straight, show high stability with respect to the mutation erosion and found in Alu repeats a perfect alliance for spreading within the genome. First, in vitro studies have shown that such poly(A) and poly(T) sequences cannot be readily wound around the nucleosome [35]. Therefore, they remain exposed and not affected by the silencing mechanisms [36]. Since poly(A)–poly(T) tracts are scarcely compatible with nucleosome formation, very long tracts may affect chromatin organization. In humans, most of the large genome is organized in gene-poor and densely compacted chromatin. This might involve a relatively tight positioning of nucleosomes and thereby be responsible for the observed spreading of Alus and, consequently, for the observed inter-tract statistics.

3.2. The link between homopolymeric repeats and organismal complexity

The abundance of HRs show an apparent correlation with the organismal complexity. For example, simple repeats are absent in viruses, rather rare in bacteria and low eukaryotes and very abundant in high vertebrate genomes. A strong association has been found between organismal complexity and the complexity of regulatory regions upstream and downstream the genes and the complexity of the coding region.

In a typical mammalian gene, HRs and other repeats or mobile genetic elements can have different effects on gene functioning. In particular, HRs can be found in both the regulatory and the coding regions. The multiple regulatory regions that upstream the gene are binding sites for transcription factors and represent subfunctions that might finely tune, positively or negatively, the level of expression in a specific tissue and developmental stage. The coding region generally codes for proteins with several structural/functional domains that may interact with different ligands/proteins.

Phenotype and the acquisition of new gene functions has been associated with gene duplication [37,38]. We hypothesize that it can be also affected by the presence of HRs. The probability of preservation of both gene duplicates increases with the number of independent subfunctions in the regulatory or coding region due to a greater number of ways that gene duplicates can differently evolve, by means of keeping or loosing some of these subfunctions. The insertion of repeats, mobile elements and the elongation of homo or dimer strings in the regulatory region in one

duplicate, may change the binding affinities between transcription factors and the basal transcription machinery and increase the probability of recombination, loss, acquisition, shuffling, duplication of regulatory sites with respect to the other gene copy. The same may occur in the coding regions. These events will increase rather than reduce the probability of duplicated gene fixation because each gene can now perform a function the other gene cannot, for example the two genes being expressed in different tissues or developmental stage [38–43]. Moreover, several hundred genes use fragments of mobile elements in the regulatory sequences that control expression and transcription termination [14,17]. This suggests that, at least in part, mobile elements such as Alus are retained because they confer some advantages [44]. SINEs and LINEs appear to be subject to RNA interference (RNAi) that is a form of post-transcriptional gene silencing triggered by double-stranded RNA [14].

The probability of fixation by differential subfunctionalization approaches zero in large populations because the long time fixation magnifies the chances that secondary mutations will completely incapacitate one copy before joint preservation is complete; therefore, subfunctionalization is a more important factor in high eukaryotes than in bacteria and low eukaryotes, where neofunctionalization, i.e., arising of completely new genes is a more frequent event.

Several examples of subfunctionalization of regulatory and coding regions are reported in literature [42]. We hypothesize that HRs affect the mobility of Alus and increase in a subtle way the combination and diversification of gene expression and the fixation of multigene families. Since poly(A)–poly(T) tracts are scarcely compatible with nucleosome formation, and very long tracts may affect chromatin higher structure, we aim at investigating how their distribution can be used as an indirect means to obtain insights into the structural organization of DNA, on both genome-wide scale and on individual chromosomes.

4. Conclusions

Despite the availability of several high eukaryote genomes, the evolutionary dynamics of the simplest repeats are not yet fully understood. Since microsatellite slippage mutation rates depend on many factors, among which, repeat motif-length, here, we have studied the genome-wide base composition of the microsatellites and we have particularly focused on the relationships between poly(A) and Alus in the human genome.

We have shown by means of standard linguistic analysis that HRs are highly specific, content-bearing terms within the DNA sequence of humans. More specifically, we have provided evidence that a quantitative analysis of length and inter-tract distributions of HRs may provide insight into the mobility of these elements within the genome.

The clear 150 and 300 bp periodical patterns of poly(A)–poly(T) tracts, revealed with the aid of a kernel density estimator, show that these tracts are almost entirely associated to Alus and to mobile elements of similar length. Since most Alus are 300 bp long, the fact that the signal at 150 bp is higher than the one at 300 bp in

almost all distributions of separations between A-tracts and T-tracts from all chromosomes suggests that the central, short poly(A)–poly(T) tracts often elongates.

We have shown that a quantitative analysis of the link between poly(A) repeats and Alus has important consequences on the understanding of the joint dynamics of Alus and simple repeats in the human genome. On one hand, the intrinsic rigidity of sequences such as poly(A) tracts helps Alus mobilization, hindering packing of Alu-containing tracts into the nucleosome structure. On the other hand, middle and 3' end regions of Alus are source of longer poly(A)–poly(T) tracts. Furthermore, these tracts may mutate and populate also other poly(X) sequences. Work in progress focuses on reconstructing the dynamics of Alu clusters formation in the human genome. In conclusion, we wish to stress that the genome-wide statistical analysis of low complexity is a thriving field of research. In particular, it may have two important benefits: improve the understanding of the processes that shaped the genome organization and improve the ability to correlate phenotype complexity with genome organization.

Acknowledgements

F.P. acknowledges funding from the Italian Institute for Condensed Matter Physics, under the Forum project STADYBIS.

References

[1] E.S. Lander et al., Nature 409 (2001) 860–921.
[2] P. Lio, M. Vannucci, Gene 317 (2003) 29–37.
[3] G. Glusman, I. Yanai, I. Rubin, D. Lancet, Genome Res. 11 (2001) 685–702.
[4] A.M. Weiner, Curr. Op. Cell Biology 14 (2002) 343–350.
[5] W.H. Li, Molecular Evolution, Sinauer Associates, Sunderland.
[6] A. Umar, J.I. Risinger, E.T. Hawk, J.C. Barret, Nat. Rev. Cancer 4 (2004) 153–158.
[7] P. Bois, A.J. Jeffreys, Cell. Mol. Life Sci. 55 (1999) 1636–1648.
[8] G. Gifford, R. Brown, Methods Mol. Med. (2004) 237–250.
[9] C. Acquisti et al., Chaos Sol. Fract. 20 (2004) 127–137.
[10] D. Holste, I. Grosse, S. Beirer, P. Schieg, H. Herzel, Phys. Rev. E 67 (2003) 061913.
[11] F. Lillo, S. Basile, R.N. Mantegna, Bioinformatics 18 (7) (2002) 971–979.
[12] H. Herzel, E.N. Trivonov, O. Weiss, I. Grosse, Physica A 249 (1–4) (1998) 449–459.
[13] R.N. Mantegna, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.K. Peng, M. Simons, H.E. Stanley, Phys. Rev. E 52 (3) (1995) 2939–2950.
[14] D. Dieringer, C. Schlotterer, Genome Res. 13 (2003) 2242–2251.
[15] H.H. Kazazian Jr., Science 303 (2004) 1626–1632.
[16] P.L. Deininger, M.A. Batzer, Genome Res. 12 (2002) 1455–1465.
[17] M.A. Batzer, P.L. Deininger, Nat. Rev. Genet. 3 (2002) 370–379.
[18] P.L. Deininger, M.A. Batzer, Mol. Genet. Metab. 67 (1999) 183–193.
[19] A.M. Roy-Engel, A.-H. Salem, O.O. Oyeniran, L. Deininger, D.J. Hedges, G.E. Kilroy, M.A. Batzer, P.L. Deininger, Genome Res. 12 (2002) 1333–1344.
[20] K. Johanning, C.A. Stevenson, O.O. Oyeniran, Y.M. Gozal, A.M. Roy-Engel, J. Jurka, P.L. Deininger, J. Mol. Evol. 56 (2003) 658–664.
[21] C. Chingfer, A.J. Gentles, J. Jurka, S. Karlin, PNAS 99 (2002) 2930–2935.

[22] K.J. Fryxell, E. Zuckerkandl, Mol. Biol. Evol. 17 (2000) 1371–1383.
[23] I. Brukner, R. Sanchez, D. Suck, S. Pongor, EMBO J. 14 (1995) 1812–1818.
[24] E.W. Englander, B.H. Howard, J. Biol. Chem. 271 (1996) 5819–5823.
[25] Y. Lai, F. Sun, Mol. Biol. Evol. 20 (2003) 2123–2131.
[26] S. Deepali, L. Yinglei, S. Fengzhu, N. Arnheim, Nucl. Ac. Res. 31 (2003) 974–998.
[27] D.W. Scott, Multivariate Density Estimation, Theory Practice and Visualization, Wiley, New York, 1992.
[28] L. Wassermann, All of Statistics, Springer, Berlin, 2003.
[29] B.W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman & Hall, London, 1986, pp. 45–47.
[30] M.L. Coté, M. Pflomm, M.M. Georgiadis, J. Mol. Biol. 330 (2003) 57–74.
[31] D. Bhattacharyya, S. Kundu, A.R. Thakur, R. Majumbar, J. Biomol. Struct. Dyn. 17 (1999) 289–300.
[32] W.H. Li, Z. Gu, H. Wang, A. Nekrutenko, Nature 409 (2001) 847–849.
[33] A. Eyre-Walker, L.D. Hurst, Nat. Rev. Genet. 2 (2001) 549–555.
[34] P. Fransz, J.H. De Jong, M. Lysak, M.R. Castiglione, I. Schubert, Proc. Natl. Acad. Sci. USA 99 (2002) 14584–14589.
[35] H.C. Nelson, J.T. Finch, B.F. Luisi, A. Klug, Nature 330 (1987) 221–226.
[36] V. Iyer, K. Struhl, EMBO J. 14 (1995) 2570–2579.
[37] V. Lynch, J.S. Conery, Science 290 (2000) 1151–1155.
[38] A. Wagner, BioEssays 20 (1998) 785–788.
[39] E.T. Dermitazakis, A.G. Clark, Mol. Biol. Evol. 18 (4) (2001) 557–562.
[40] M. Lynch, J.S. Conery, Science 302 (2001) 1401–1405.
[41] A. Wagner, Trends Genet. 17 (2001) 237–239.
[42] M. Lynch, A. Force, Genetics 154 (2000) 459–473.
[43] A. Force, M. Lynch, F.B. Pickett, A. Amores, Y. Yan, J. Postlethwait, Genetics 151 (1999) 1531–1545.
[44] J.S. Han, S.T. Szak, J.D. Boeke, Nature 429 (2004) 268–274.