Mini-Review
TheScientificWorldJOURNAL (2011) 11, 842–854
ISSN 1537-744X; DOI 10.1100/tsw.2011.82

TheScientificWorldJOURNAL
www.thescientificworld.com

Mid-Range Inhomogeneity of Eukaryotic Genomes

Larisa Fedorova and Alexei Fedorov*

Department of Medicine, University of Toledo, Health Science Campus, Toledo, Ohio

E-mail: Alexei.Fedorov@utoledo.edu

Received August 5, 2010; Revised January 18, 2011; Accepted March 9, 2011; Published April 5, 2011

Multicellular eukaryotic genomes are replete with nonprotein coding sequences, both within genes (introns) and between them (intergenic regions). Excluding the well-recognized functional elements within these sequences (ncRNAs, transcription factor binding sites, intronic enhancers/silencers, etc.), the remaining portion is made up of so-called "dark" DNA, which still occupies the majority of the genome. This dark DNA has a profound nonrandomness in its sequence composition seen at different scales, from a few nucleotides to regions that span over hundreds of thousands of nucleotides. At the mid-range scale (from 30 up to 10,000 nt), this nonrandomness is manifested in base compositional extremes detected for each of four nucleotides (A, G, T, or C) or any of their combinations. Examples of such compositional nonrandomness are A-rich, purine-rich, or G+T-rich regions. Almost every combination of nucleotides has such enriched regions. We refer to these regions as being "inhomogeneous". These regions are associated with unusual DNA conformations and/or particular DNA properties. In particular, mid-range inhomogeneous regions have complex arrangements relative to each other and to specific genomic sites, such as centromeres, telomeres, and promoters, pointing to their important role in genomic functioning and organization.

KEYWORDS: pattern, DNA structure, CpG islands, isochore, triplex, Z-DNA, A-DNA, H-DNA

INTRODUCTION

Genomic patterns on short-range scales represent various "words" composed from nucleotide "letters". Each of these words occurs many times within DNA sequences. The longest words, also known as "pyknons", comprise sequences up to 17 nucleotides (nt) long that are overabundant in the exons and introns of humans and other mammals[1,2]. The vast majority of sequences, only a little bit longer than pyknons, are unique even for the large genomes of animals and plants. For example, the complete theoretical set of 20-nt-long sequences is comprised of 4^{20} different words of length 20, which is just over one trillion. More than 99% of these 20-mer oligonucleotides never occur in the entire human genome (~$3*10^9$ bp). Therefore, biologists frequently use 20-mer oligonucleotides as PCR primers or hybridization probes for experimental characterization of particular genomic segments. The genomic arrangement of short sequences (<20 bp) is covered in insightful papers[3,4]. Here we consider genomic patterns longer than 30 and up to several thousands of nucleotides to be called the mid-range scale. At this

*Corresponding author.

Published by TheScientificWorld; www.thescientificworld.com

mid-range, most of the sequences are unique, i.e., occur only once in the entire genome; hence, it is more appropriate to characterize or group them not by their exact sequence of nucleotides, but rather by their overall nucleotide composition, such as G+C richness, purine richness, etc. We also distinguish mid-range genomic scales from the long-range scale represented by genomic isochores, reviewed elsewhere[5]. Traditionally, G+C-rich and G+C-poor isochores are considered to be from 100 kb and longer. Recently, scientists have started to describe ultra-short isochores in the range of tens of thousands of nucleotides. In order not to interfere with isochores, we limit the length of mid-range patterns to 10,000 bases. The main focus of this paper is to show that at mid-range scales, genomes of complex eukaryotes consist of a number of different patterns and are associated with unusual DNA conformations. Some of these patterns are scarcely investigated and still wait for thorough exploration and recognition.

G+C-RICH AND A+T-RICH REGIONS

We start considering mid-range genomic compositional patterns from the most-studied case: G+C-rich and A+T-rich regions. These G+C-rich and A+T-rich regions of various lengths from 30 to several thousand nucleotides are four to 20 times over-represented in the mammalian genomes compared to random expectation[6,7]. Among G+C-rich genomic segments, CpG islands have drawn the most public attention, due to their functional properties and involvement in gene expression regulation[8]. CpG islands are found in nearly 60% of human genes, including almost all of the housekeeping genes[8]. According to two different definitions of these islands, their length must be at least 200 or 500 bp, G+C content more than 50 or 55%, and the number of CpG dinucleotides in the islands should exceed more than twice their occurrence in other genomic regions[9,10]. CpG dinucleotides are important sites for cytosine methylation in all vertebrates and some invertebrates and plants. However, inside CpG islands, CpG dinucleotides are predominantly nonmethylated[11]. It has been shown recently that CpG dinucleotides without methylation exhibit structural abnormalities in the DNA helix. Particularly, they are one of the most frequent sites for DNA backbone cleavage by hydroxyl radicals[12,13] and during the sonication of double-stranded DNA[14]. The crucial involvement of cytosine methylation in the regulation of gene expression is well described in a number of reviews, including some recent ones[11,15,16]. Thus, we concentrate here on the other physicochemical properties of G+C-rich and A+T-rich regions.

It is well known that the A form of the DNA helix exists in high salt concentrations and in ethanol-containing solutions. However, G+C-rich regions may be present in A-form DNA even in aqueous solutions[17,18,19]. A special form of DNA that is an intermediate between A and B forms has been characterized in G+C-rich sequences with methylated cytosines[20]. In addition, short $(CpG)_n$ repeats could adopt Z-DNA (reviewed by Ho[21]). This Z-DNA is proposed to serve as a transcriptional coactivator[22].

A+T-rich regions, on the other hand, are also associated with special DNA conformations. Some of these sequences with specific distributions of A and T bases form an unusual structure known as the DNA unwinding element[23]. These elements are often associated with the origins of replication in eukaryotes and prokaryotes[24]. There are several A+T-rich simple repeats widespread in eukaryotes. Among them, $(AT)_n$ is one of the most common in animals. X-ray and NMR studies of the DNA oligomer d(ATATAT) have shown that, in addition to B-DNA, it could form an antiparallel, double-helical duplex in which the base pairing is of the Hoogsteen type[25]. The adenines in this duplex are flipped over, making the minor groove narrow and hydrophobic. This structure is very similar to the standard B-form helix with about 10 bp per turn. Theoretical analysis has demonstrated that energies of the Hoogsteen form and B form of DNA are practically identical[26]. Most recently, Chakraborty et al.[27] demonstrated that poly-dA oligonucleotides (dA_{15}) under acidic pH conditions could allow the formation of a double-helical parallel-stranded duplex held together by reversed Hoogsteen-type AH^+-H^+A base pairs.

A+T-rich regions presumably have several important cellular functions. First, the most indicative compositional characteristic of scaffold/matrix-attached regions is that they are A+T rich [28]. Second, centromere DNA of diverse animals, plants, and fungi always contain A+T-rich regions[25,29].

PURINE/PYRIMIDINE-RICH REGIONS AND H-DNA TRIPLEX

All combinations of nucleotide pairs, except G+C and A+T, have strand asymmetry. For example, if one strand is enriched by purines (R), the complementary strand is enriched by pyrimidines (Y). Therefore, R-rich and Y-rich sequences and T+G-rich and A+C-rich sequences are physically the same loci, yet represent complementary strands. From here on, we will consider them together and refer to them as R/Y-rich and T+G/A+C-rich, respectively.

Since 1957, it has been shown that complementary DNA strands, one of which is R-rich and another Y-rich, can form three-stranded helical structures or triplexes[30]. Intramolecular triplexes, known also as H-DNA, materialize under certain conditions, like supercoiling, when half of the DNA duplex may dissociate into single strands and one of the stand-alone strands can interact via Hoogsteen base pairing with the remaining Watson-Crick DNA duplex along its major groove, forming a triplex structure. The remaining stand-alone strand stays unpaired. An example of a DNA triplex is shown in Fig. 1. There are four kinds of H-DNA depending on strand type and orientation[31]. One type of H-DNA forms under acidic conditions when the stand-alone Y-rich strand interacts with the R-rich strand of the remaining duplex. Particularly, thymines of the stand-alone strand interact with adenosines of the A-T Watson-Crick pairs of the duplex via Hoogsteen hydrogen bonding, while cytosines of the stand-alone strand interact with guanines of G-C Watson-Crick pairs. Due to this base match requirement for the assembly of this kind of triplex, the sequences of the Y-rich stand-alone strand and the Y-rich strand in the duplex should have sequence mirror symmetry. (Here is an example of two sequences with mirror symmetry: 5-TAGTTCC-3 and 5-CCTTGAT-3.) In many R/Y-rich regions of the genomes, such mirror symmetry has been observed. For example, a 2.5-kb R-rich sequence of the 21st intron of the human PKD1 gene has 23 mirror repeats that form H-DNA[32,33]. Another kind of intramolecular triplex can be formed at neutral pH and requires bivalent cations for stability. It is formed by the interaction of the R-rich stand-alone strand with the remaining duplex via Hoogsteen bonding. It does not require strong mirror symmetry within its sequences, since the adenines of the stand-alone R-rich strand could interact with the A-T pair of the duplex or with the G-C pair[34].

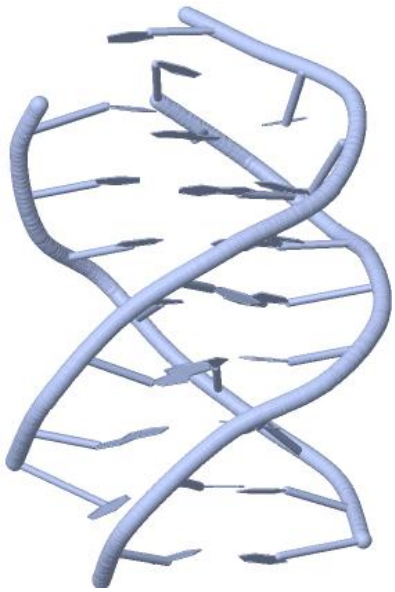

FIGURE 1. Cartoon of 3D structure of a purine-purine-pyrimidine DNA triplex containing G-GC and T-AT triples. This picture is a snapshot of the structure with the identifier 134D obtained from the Protein Data Bank. The structure was resolved using a combined NMR and molecular dynamics approach by Radhakrishnan and Patel[35].

There are several documented functions of H-DNA. It is well established that H-DNA could exist *in vivo* under certain conditions. Various experimental methods for the characterization of H-DNA have been reviewed by Jain et al.[31] and Wang et al.[36]. Single-stranded DNA not participating in the triplex is accessible to S1-nuclease cleavage. Eukaryotic genomes contain many S1-nuclease–sensitive sites within runs of homopurine sequences. These segments of single-stranded DNA are frequently involved in the recombination of homologous DNA and, thus, are sites for genetic instability. Different schemes of recombination involving H-DNA have been described by Jain et al.[31]. Bacolla et al.[37] characterized nearly 3,000 homopurine tracks in the human genome longer than 100 nt. They supported evidence for these tracks in promoting recombination and association with higher rates of mutations. In addition, stable H-DNA structures are able to block transcription and replication. Jain and coauthors surveyed the evidence for how H-DNA influences the activity of DNA and RNA polymerases. Finally, Goni and others[38] performed a large-scale bioinformatic analysis of the distribution of short R-rich sequences in the human genome. They demonstrated that short R-rich sequences are several times more abundant in the downstream promoter regions compared to other regions and to random expectation models. These short R-rich sequences hold evolutionary conservation between human and mouse, yet they likely are not direct targets for transcription factors. Goni and coauthors have suggested that these sequences act as pacing fragments in promoter regions and help in the correct positioning of transcription factors.

G+T-RICH/A+C-RICH REGIONS

Recall that the complementary strands of G+T-rich regions are naturally A+C-rich regions. They coexist with each other and we consider them interchangeable with respect to their description in the literature. According to nucleic acid nomenclature, G or T nucleotides are also known as *Keto* or K, while A or C are known as *Imino* or M. Thus, sometimes these regions are referred to as K.M tracks or motifs[39]. Bechtel and coauthors demonstrated that G+T regions are about five times more abundant in the mammalian genomes compared to random expectation[7]. Moreover, these regions practically do not intersect with interspersed DNA repeats at all. In 2004, Yagil[39] demonstrated that K.M motifs are significantly over-represented in the genomes of diverse animals, plants, and fungi. Specifically, K.M motifs are predominant in the *Drosophila melanogaster* genome, where they outnumber other motifs such as R/Y-rich motifs. Despite their abundance, G+T-rich motifs are much less investigated than other regions with extremes in base compositions. Possible functions that could be associated to G+T-rich regions are the following. First, $(CA)_N$ simple repeats are one of the most profuse tandem repeats in mammalian genomes[40]. They also should be considered as an alternating R/Y sequence and, due to this property, associated with a Z-DNA conformation[41], which is considered in the next section. Second, C-rich regions, which could be a component of CA-rich regions, are capable of forming four-stranded intercalated molecules[42]. Third, short G+T-rich regions could represent transcription factor binding sites, such as for factor Sp1[43]. Fourth, telomeres of various eukaryotic species are represented by G+T-rich regions that form G-quadruplexes, also known as G-quartets or G-4. Quadruplexes are arranged in four-stranded structures with strands connected to each other via Hoogsteen hydrogen bonding between guanines. The G-quadruplex has been well characterized in human telomeric and related sequences with the core repetitive element TTAGGGG, and within promoters and 5´-untranslated regions of human genes whose sequences have a loose consensus of $G_{3\text{-}5}N_{L1}G_{3\text{-}5}N_{L2}G_{3\text{-}5}N_{L3}G_{3\text{-}5}$, where N_{L1}, N_{L2}, and N_{L3} are loops with the length from 1 to 7 nt and variable nucleotide composition[44]. Intriguingly, G+T-rich oligonucleotides possess antiviral activities. For example, the $T_2(G_4T_2)_3$ sequence is virucidal against the herpes simplex virus[45]. At the RNA level, C+A-rich sequences within intronic segments could regulate alternative splicing by being binding sites for the hnRNP L protein[46]. The presence of C+A-rich sequences at the 3´-UTR of mRNA could regulate gene expression at the level of translation[47]. The distribution of C+A-rich sequences enriched by $(CA)_N$ imperfect repeats is highly skewed towards telomeres and minisatellites can usually be found in the vicinity as well[48]. Despite the listed properties

associated with G+T-rich regions, they seem significantly underinvestigated and may yet reveal unknown important functional properties in the near future.

ALTERNATED PURINE/PYRIMIDINE REGIONS AND Z-DNA

Left-handed antiparallel Z-DNA double-helix conformation has been first characterized in 1979 by Wang and coauthors for $(GC)_3$ repeats[49]. Detailed Z-DNA structure has been considered elsewhere[21,50]. This particular conformation is characterized by rotation of R bases that adopt *syn* form and stack over the deoxyribose ring, while Y bases do not adopt unfavorable *syn* form[21]. Thus, Z-DNA, which is characterized by an alternating pattern of *anti-syn* conformations, is formed by alternated R/Y sequences[51].

In 1986, Ho and others[52] developed a ZHUNT program for detection of genomic sequences with high propensity to form Z-DNA. They found a high concentration of these sequences near the transcription start sites[50,53]. Most recently, human genomic Z-DNA segments have been detected experimentally using a Z-DNA binding protein domain as a probe[54]. The authors found an abundance of Z-DNA hot spots located in centromeres of 13 human chromosomes. Z-DNA–forming sequences induce high levels of genetic instability in both mammalian and bacterial cells. These sequences could be causative factors for gene translocations found in leukemias and lymphomas[55]. The discovery of certain classes of proteins bound to Z-DNA with high affinity and specificity indicated a biological role of this structure. Yet, it is a common view that Z-DNA is an unstable conformation that is formed and disappears during particular physiological activities, such as transcription[50].

The latest version of the *Genomic MRI* package (described in detail in the next section) has a new feature that allows the detection of excesses and shortages of alternating bases, including R/Y patterns. It reveals that in mammalian genomes, there is more than 40 times the overabundance of alternating R/Y stretching over 50- to 100-bp genomic segments, where RY plus YR comprise more than 80% of all dinucleotides. A considerable portion of these alternating R/Y patterns are represented by short $(GC)_n$, $(AC)_n$, $(AT)_n$, and $(TG)_n$ repeats that can alternate with each other and be accompanied by alternating R/Y bases without strong periodic sequence pattern. For example, here is a sequence of a 50-bp segment from the third intron of a human heparanase-2 gene highly enriched with alternated R and Y bases: 5′AAATGGATGTGTGTATATATATGAAGTCGATACACACACATATACACATA3′. We showed that such alternating R/Y sequences are plentiful throughout the mammalian genomes, either inside introns or within intergenic regions.

Genomic MRI Program Package

In thousands of genomic regions, the composition of A, T, C, or G content or different combinations of these bases exist at extremes far different from the average base composition. We call such compositional extremes genomic mid-range inhomogeneity (or MRI) if they stretch at least 30 bp, but <10,000 bp. To characterize genomic MRI patterns, a public computational resource (*Genomic MRI*) has been created that allows us to detect sequence regions with any type of extreme composition[7]. Using this resource, it was demonstrated that various MRI regions occupy up to a quarter of the human genome and their existence is maintained via strong fixation bias[56]. For examining mid-range sequence patterns, *Genomic MRI* programs do not characterize particular "words", but only the overall compositional content of particular base(s) that we refer to as X (X could be a single nucleotide A, G, C, or T, or any of their combinations like A+C or G+T+C, etc.). *Genomic MRI* allows us to study the distribution of X-rich regions in any sequence of interest. These X-rich MRI regions are highly over-represented in mammalian genomes for all kinds of X contexts. For instance, in the human genome, G+C-rich sequences with lengths from 100 to 200 nt are 20 times over-represented; A+T-rich sequences in the same length range are about 12 times over-represented; A+G-rich and T+C-rich sequences 10 times; and G+T-rich and A+C-rich sequences up

to six times over-represented[7]. In order to measure the abundance of X-rich regions in the sequences under analysis, *Genomic MRI* compares their presence inside a specifically generated random sequence that has the same oligonucleotide distribution as the real one. This evaluation is achieved by the following computational steps. First, the short-range inhomogeneity (SRI) of a given sequence is analyzed by the SRI-*analyzer* program from the *Genomic MRI* package to create an oligonucleotide frequency table for each possible 1- to 9-nt-long "word". Then, a second program, SRI-*generator*, creates a random sequence with SRI approximating the oligonucleotide frequency table of the natural sequence. This random sequence is used further for comparison with the natural one. Finally, the third program, MRI-*analyzer*, scans a sequence under analysis and the random sequence with a window of a specified size, and checks whether the nucleotide composition of the sequence in the current window is X rich or X poor for a particular chosen combination of nucleotides (X), e.g., A, T, C, G, G+C, A+G, G+T, etc. A window is rich for the X content if its X composition is above a user-specified threshold-X_1, while a window is X poor if it is below another user-specified threshold-X_2. (Note that X-poor regions can be referred to as non–X-rich regions, e.g., G+C poor are A+T rich.) An example of MRI-*analyzer* graphical output is shown in Fig. 2, which illustrates the MRI patterns for an extra-large human intron of the dystrophin gene from chromosome X.

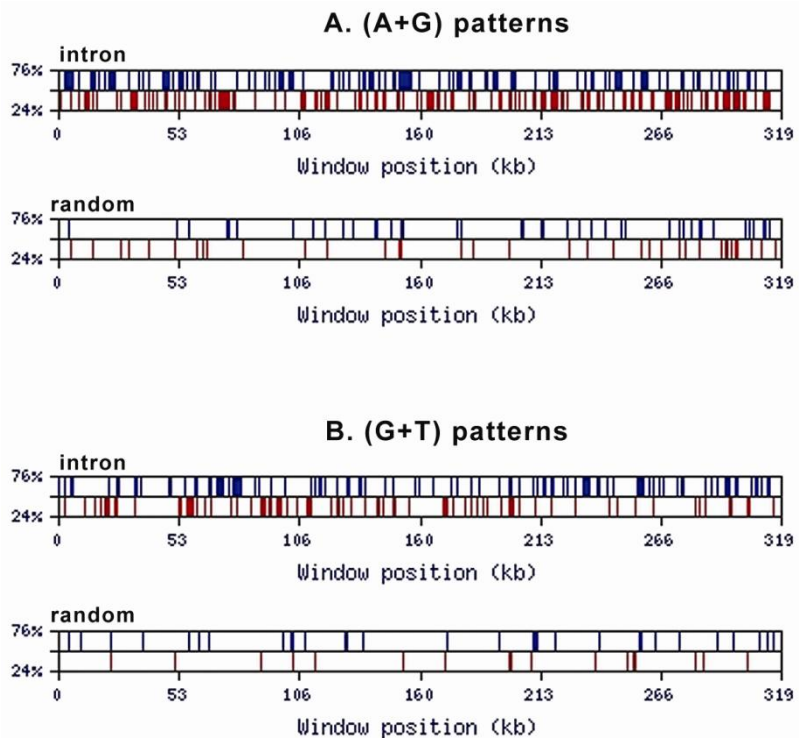

FIGURE 2. The graphical output of the MRI-*analyzer* program for the first intron of the dystrophin gene (marked as "intron") and also for the SRI-*generator* random sequence based on the tetramer oligonucleotide frequency table of the intron (marked as "random"). The entire sequence of the 319-kb intron and the random sequence is displayed on the x axis. Blue bars represent content-rich MRI regions on the sequence. Red bars represent content-poor MRI regions. The y axis contains upper and lower thresholds for the given content type. (A) *Genomic MRI* analysis of A+G-rich and A+G-poor (or T+C-rich) regions; (B) *Genomic MRI* analysis of G+T-rich and G+T-poor (or C+A-rich) regions.

Two scales of MRI regions should be considered: (1) regions (from 30 to 1000 bp) whose properties have been investigated in detail and for which several periodicities have been reported[57,58,59]; (2) larger regions (from 1 to 10 kb), which are one of the least-studied areas in genomic composition and where as-yet-unknown biological properties may be found. Such subdivisions are important for the proper choice of parameters for the MRI thresholds. For instance, for a 100-nt-long window, there is a vast number of regions in mammals where G+C composition is 85% or higher. However, for studying regions with a window size of around 5 kb, the upper threshold for G+C content should not be more than 65% to find the areas satisfying the criterion.

A COMPLEX MOSAIC OF MRI PATTERNS

Different MRI regions are not randomly arranged relative to each other[6]. For example, Fig. 3 illustrates that G+C-rich regions tend to be associated in clusters. On the other hand, the distribution of A+T-rich regions is much more close to a random distribution with the exception that A+T-rich regions avoid very close proximity to each other[6]. So far, investigators have examined only individual genomic patterns. The mutual arrangement of various genomic mid-range patterns has never been thoroughly investigated. Our preliminary results suggest that within mammalian genomes, there is a complex mosaic picture of MRI regions. Modeling sequences only with one particular type of MRI compositional bias using the MRI-*generator* program from the *Genomic MRI* package has proven to not be a trivial computational task[7]. This has given us an appreciation that the reconstruction of the entire set of MRI patterns in modeling DNA sequences is an extremely challenging mission due to a complex multilayer nonrandomness in genomic sequences. In addition, genomic sequences have an intricate organization of nested patterns with respect to the clustering of particular patterns. Some features of this complex organization were described as genomic fractals in several publications[60,61,62]. This arrangement has been studied by methods such as “detrended fluctuation analysis” and a “Brownian walk” in order to uncover relationships such as power law correlations and exponential decays, which assess the scaling behavior of a system. This scaling behavior is related to fractal geometry and deals with “self-similarity”, defined as the property of resembling a subset of oneself. Earlier investigations of this kind generally confined themselves to clusters of purines and pyrimidines, but later studies have shifted to examining G+C and A+T clusters for the thermodynamic implications of their pair binding[60,61,63,64,65,66,67].

Recently, by studying the distribution of more than 4 million SNPs in the human genome and by taking into account their frequencies in the population, the influence of mutations on different MRI regions has been examined[56]. The authors demonstrated that MRI regions have comparable levels of *de novo* mutations to the control genomic sequences with average base composition. *De novo* substitutions rapidly erode MRI regions, bringing their nucleotide composition toward genome-average levels. However, those substitutions that favor the maintenance of MRI properties have a higher chance to spread through the entire population. The observed strong fixation bias for mutations helps to preserve MRI regions during evolution, indicating their potential significance to genomic operations.

On the other hand, a large portion of MRI regions could have a mechanistic origin due to the bias in frequencies of different types of mutations as well as fixation bias of these mutations. Indeed, the rate of transition mutations is generated at higher frequency than transversions, even though there are twice as many possible transversions. Moreover, the rate of particular types of mutation (e.g., A->C) is influenced by the surrounding nucleotides (context). For example, CpG dinucleotides are a hot spot for C->T and G->A changes due to methylation of the cytosines within this context. The charts for all possible human substitution frequencies within the context of a single 5′ nucleotide are presented by Zhang and Gerstein[68]. Among transversions, the highest frequency was observed for tA->tC substitutions, which is about three to four times higher than those for cT->cG, cC->cG, and tT->tA substitutions having the lowest frequencies[68]. Many sequence patterns may arise in accordance to the widely accepted neutral theory of molecular evolution without involvement of negative (purifying) and/or positive selection[69]. The

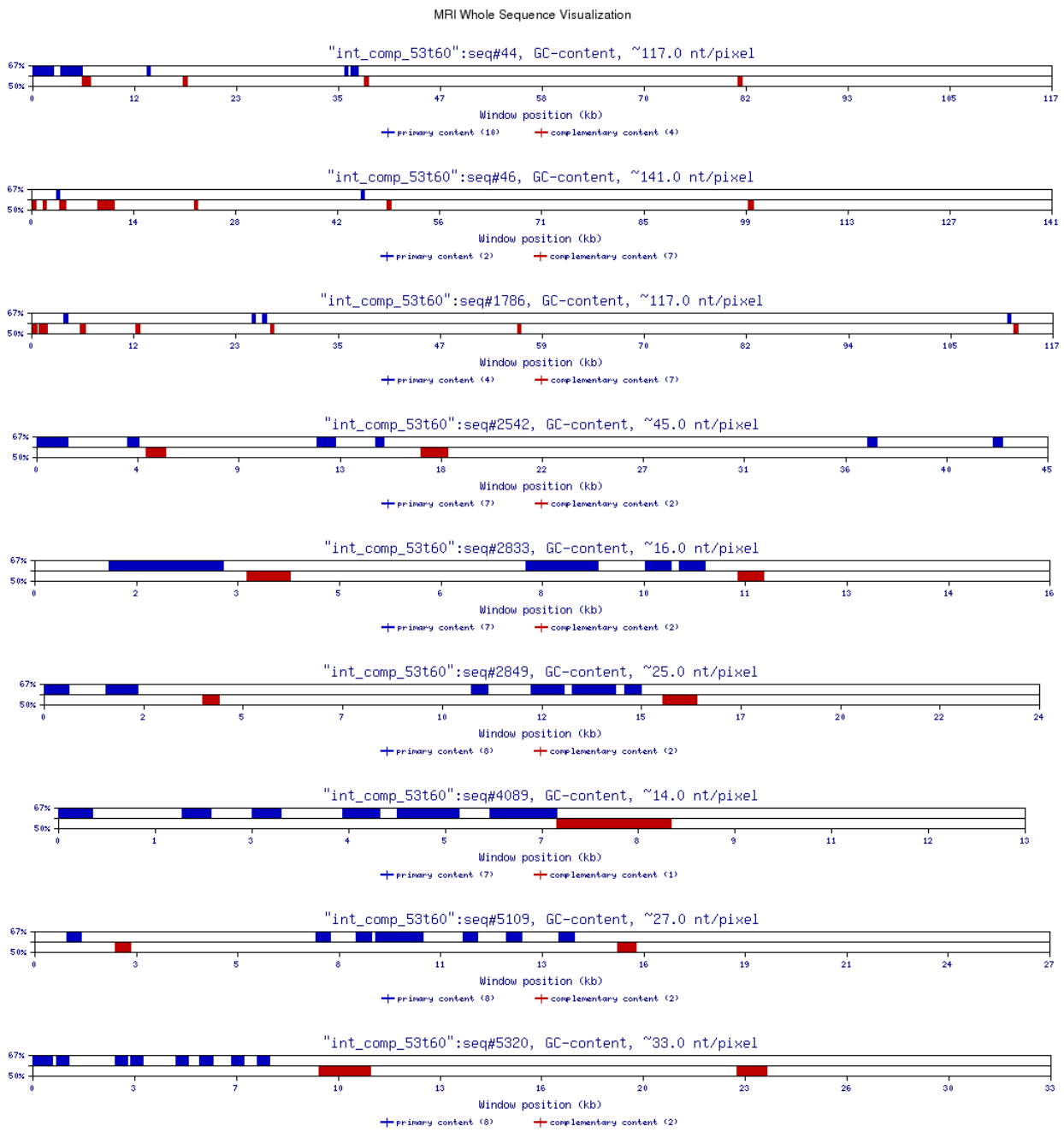

FIGURE 3. Visualization of G+C-rich (blue) and A+T-rich (red) MRI features in human introns using a 400-nt base window size. The scale for each sequence is independent and is given in its subheading in nucleotides per pixel. The figure represents a fragment of Fig. 17 in Bechtel[6].

mechanistic origin of genomic compositional inhomogeneity is the basis for the Biased Gene Conversion (BGC) theory for the origin of GC-rich isochores, lately detailed by Duret and Galtier[70]. Particularly, BGC theory explains GC-rich isochores due to fixation bias in favor of AT->GC mutations that occur without positive Darwinian selection. Currently, a popular view is that both selective and neutral processes drive GC content evolution in the human genome[70,71].

It is important to note that various simple repeats make up one of the widespread components of MRI regions, e.g., $(AT)_n$ repeats for A+T-rich MRI regions and $(AGG)_n$ for purine-rich regions. During evolution, simple repeats are subjected to growth via replication slippage and interchromosomal exchange[72], and hence their existence may lack functional importance. At the same time, there are up to 10 different non–B-form DNA conformations connected with simple repeats, as are listed and well illustrated by Wells[73]. Interestingly, more than 70 human genetic disorders have been associated with changes in simple repeats[73,74]. In rodents, 2.4% of their euchromatin is represented by simple repeats, which is two times bigger than the length of all their protein-coding sequences[75]. An important public toolkit is available online for the characterization of simple repeats as well as the analysis of DNA sequence complexity. Programs include Compexity, LZcomposer, OligoRep, and more[76,77].

THE ROLE OF MRI REGIONS

Often, in the popular literature, genomes are presented as a set of texts or instructions. Such a representation implies that there should be an intelligent creature somewhere inside a cell interpreting these DNA texts. Thus, it is more appropriate to compare genomes with self-realization programs that autonomously fulfill their tasks and are able to respond to environment signals and conditions. Such programs must be extremely complicated for complex organisms, like humans, which are built from trillions of cells of hundreds of different kinds, yet sharing the same genomic sequence. There must be fundamental principles for construction and functioning of genomic programs. One of the most important is the Principle of Recursive Genome Function (PRGF) illuminated by Pellionisz[62]. The author considers the genome as an unsupervised operating system. The well-known examples of such a system are neural networks for which mathematical models describing their behavior have been developed. According to Pellionisz, "the recursive genome function is a process when at every step of development already-built proteins iteratively access sets of primary and ensuing auxiliary information packets of DNA to build constantly developing hierarchies of protein structures." In other words, there is a crucial flow of information from proteins back to the genomic DNA. According to Pellionisz, this principle converts a genome from a *closed* to an *open* physical system and resolves the paradox of genomic entropy posed by Sanford[78]. This perspective elucidates the importance of MRI regions as specific sites for changing genomic information by proteins. Indeed, MRI regions are intricately associated with unusual DNA conformations, which in turn are binding sites for a number of proteins. These proteins could stabilize and/or initiate DNA conformation transformation and propagate the signal along neighboring DNA segments. For instance, Z-DNA binding proteins could initiate this transformation from right-handed B-DNA to the left-handed Z form. This structural transition changes the DNA supercoiling for the regional DNA landscape and additionally creates specific B-Z boundaries with flipped-over bases. Such transformation could modify, open, and/or hide some information on the genomic DNA, not only at the protein binding site, but within neighboring regions.

CONCLUSIONS

Overall, within vast areas of previously thought "junk DNA", represented by introns and intergenic sequences, there exists an intricate mosaic of various MRI regions with extreme base compositions. Various genomic MRI regions are tightly associated with unusual DNA conformations and must be of crucial importance for proper functioning of multicellular eukaryotes.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (Career MCB-0643542).

A modified version of this paper is to be published as a chapter for a book: "Advances in Genome Sequence Analysis and Pattern Discovery" 2011.

REFERENCES

1. Rigoutsos, I., Huynh, T., Miranda, K., Tsirigos, A., McHardy, A., and Platt, D. (2006) Short blocks from the noncoding parts of the human genome have instances within nearly all known genes and relate to biological processes. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 6605–6610.
2. Tsirigos, A. and Rigoutsos, I. (2008) Human and mouse introns are linked to the same processes and functions through each genome's most frequent non-conserved motifs. *Nucleic Acids Res*. **36**, 3484–3493.
3. Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., Makeev, V.J., Mironov, A.A., Noble, W.S., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol*. **23**, 137–144.
4. Lichtenberg, J., Yilmaz, A., Welch, J.D., Kurz, K., Liang, X., Drews, F., Ecker, K., Lee, S.S., Geisler, M., Grotewold, E., and Welch, L.R. (2009) The word landscape of the non-coding segments of the Arabidopsis thaliana genome. *BMC Genomics* **10**, 463.
5. Bernardi, G. (2007) The neoselectionist theory of genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 8385–8390.
6. Bechtel, J.M. (2008) Characterization of Genomic Mid-Range Inhomogenity [Thesis]. Health Science Campus. University of Toledo, Toledo, OH. p. 97.
7. Bechtel, J.M., Wittenschlaeger, T., Dwyer, T., Song, J., Arunachalam, S., Ramakrishnan, S.K., Shepard, S., and Fedorov, A. (2008) Genomic mid-range inhomogeneity correlates with an abundance of RNA secondary structures. *BMC Genomics* **9**, 284.
8. Hackenberg, M., Previti, C., Luque-Escamilla, P.L., Carpena, P., Martinez-Aroza, J., and Oliver, J.L. (2006) CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics* **7**, 446.
9. Gardiner-Garden, M. and Frommer, M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282.
10. Takai, D. and Jones, P.A. (2003) The CpG island searcher: a new WWW resource. *In Silico Biol.* **3**, 235–240.
11. Suzuki, M.M. and Bird, A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet*. **9**, 465–476.
12. Greenbaum, J.A., Pang, B., and Tullius, T.D. (2007) Construction of a genome-scale structural map at single-nucleotide resolution. *Genome Res*. **17**, 947–953.
13. Greenbaum, J.A., Parker, S.C., and Tullius, T.D. (2007) Detection of DNA structural motifs in functional genomic elements. *Genome Res.* **17**, 940–946.
14. Grokhovsky, S.L., Il'icheva, I.A., Nechipurenko, D.Y., Panchenko, L.A., Polozov, R.L., and Nechipurenko, Y.D. (2008) Heterogeneity of DNA local structure and dynamics: ultrasound studies. *Biofizika* **53**, 417–425.
15. Prokhortchouk, E. and Defossez, P.A. (2008) The cell biology of DNA methylation in mammals. *Biochim. Biophys. Acta* **1783**, 2167–2173.
16. Illingworth, R.S. and Bird, A.P. (2009) CpG islands--'a rough guide'. *FEBS Lett.* **583**, 1713–1720.
17. Warne, S.E. and deHaseth, P.L. (1993) Promoter recognition by Escherichia coli RNA polymerase. Effects of single base pair deletions and insertions in the spacer DNA separating the -10 and -35 regions are dependent on spacer DNA sequence. *Biochemistry* **32**, 6134–6140.
18. Stefl, R., Trantirek, L., Vorlickova, M., Koca, J., Sklenar, V., and Kypr, J. (2001) A-like guanine-guanine stacking in the aqueous DNA duplex of d(GGGGCCCC). *J. Mol. Biol*. **307**, 513–524.
19. Kypr, J., Kejnovska, I., Renciuk, D., and Vorlickova, M. (2009) Circular dichroism and conformational polymorphism of DNA. *Nucleic Acids Res*. **37**, 1713–1725.
20. Vargason, J.M., Eichman, B.F., and Ho, P.S. (2000) The extended and eccentric E-DNA structure induced by cytosine methylation or bromination. *Nat. Struct. Biol*. **7**, 758–761.
21. Ho, P.S. (2009) Methods to study nucleic acid structure. *Methods* **47**, 141.
22. Liu, R., Liu, H., Chen, X., Kirby, M., Brown, P.O., and Zhao, K. (2001) Regulation of CSF1 promoter by the SWI/SNF-like BAF complex. *Cell* **106**, 309–318.
23. Kowalski, D., Natale, D.A., and Eddy, M.J. (1988) Stable DNA unwinding, not "breathing," accounts for single-strand-specific nuclease hypersensitivity of specific A+T-rich sequences. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 9464–9468.
24. Umek, R.M., Linskens, M.H., Kowalski, D., and Huberman, J.A. (1989) New beginnings in studies of eukaryotic DNA replication origins. *Biochim. Biophys. Acta* **1007**, 1–14.
25. Abrescia, N.G., Gonzalez, C., Gouyette, C., and Subirana, J.A. (2004) X-ray and NMR studies of the DNA oligomer d(ATATAT): Hoogsteen base pairing in duplex DNA. *Biochemistry* **43**, 4092–4100.
26. Cubero, E., Abrescia, N.G., Subirana, J.A., Luque, F.J., and Orozco, M. (2003) Theoretical study of a new DNA structure: the antiparallel Hoogsteen duplex. *J. Am. Chem. Soc.* **125**, 14603–14612.

27. Chakraborty, S., Sharma, S., Maiti, P.K., and Krishnan, Y. (2009) The poly dA helix: a new structural motif for high performance DNA-based molecular switches. *Nucleic Acids Res.* **37**, 2810–2817.
28. Liebich, I., Bode, J., Reuter, I., and Wingender, E. (2002) Evaluation of sequence motifs found in scaffold/matrix-attached regions (S/MARs). *Nucleic Acids Res.* **30**, 3433–3442.
29. Choo, K.H. (1997) *The Centromere*. Oxford University Press, Oxford, U.K.
30. Felsenfeld, G. and Rich, A. (1957) Studies on the formation of two- and three-stranded polyribonucleotides. *Biochim. Biophys. Acta* **26**, 457–468.
31. Jain, A., Wang, G., and Vasquez, K.M. (2008) DNA triple helices: biological consequences and therapeutic potential. *Biochimie* **90**, 1117–1130.
32. Van Raay, T.J., Burn, T.C., Connors, T.D., Petry, L.R., Germino, G.G., Klinger, K.W., and Landes, G.M. (1996) A 2.5 kb polypyrimidine tract in the PKD1 gene contains at least 23 H-DNA-forming sequences. *Microb. Comp. Genomics* **1**, 317–327.
33. Blaszak, R.T., Potaman, V., Sinden, R.R., and Bissler, J.J. (1999) DNA structural transitions within the PKD1 gene. *Nucleic Acids Res.* **27**, 2610–2617.
34. Malkov, V.A., Voloshin, O.N., Veselkov, A.G., Rostapshov, V.M., Jansen, I., Soyfer, V.N., and Frank-Kamenetskii, M.D. (1993) Protonated pyrimidine-purine-purine triplex. *Nucleic Acids Res.* **21**, 105–111.
35. Radhakrishnan, I. and Patel, D.J. (1993) Solution structure of a purine.purine.pyrimidine DNA triplex containing G.GC and T.AT triples. *Structure* **1**, 135–152.
36. Wang, G., Zhao, J., and Vasquez, K.M. (2009) Methods to determine DNA structural alterations and genetic instability. *Methods* **48**, 54–62.
37. Bacolla, A., Collins, J.R., Gold, B., Chuzhanova, N., Yi, M., Stephens, R.M., Stefanov, S., Olsh, A., Jakupciak, J.P., Dean, M., Lempicki, R.A., Cooper, D.N., and Wells, R.D. (2006) Long homopurine*homopyrimidine sequences are characteristic of genes expressed in brain and the pseudoautosomal region. *Nucleic Acids Res.* **34**, 2663–2675.
38. Goni, J.R., Vaquerizas, J.M., Dopazo, J., and Orozco, M. (2006) Exploring the reasons for the large density of triplex-forming oligonucleotide target sequences in the human regulatory regions. *BMC Genomics* **7**, 63.
39. Yagil, G. (2004) The over-representation of binary DNA tracts in seven sequenced chromosomes. *BMC Genomics* **5**, 19.
40. Waterston, R., Lindblad-Toh, H.K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S.E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M.R., Brown, D.G., Brown, S.D., Bult, C., Burton, J., Butler, J., Campbell, R.D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A.T., Church, D.M., Clamp, M., Clee, C., Collins, F.S., Cook, L.L., Copley, R.R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K.D., Deri, J., Dermitzakis, E.T., Dewey, C., Dickens, N.J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D.M., Eddy, S.R., Elnitski, L., Emes, R.D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G.A., Flicek, P., Foley, K., Frankel, W.N., Fulton, L.A., Fulton, R.S., Furey, T.S., Gage, D., Gibbs, R.A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T.A., Green, E.D., Gregory, S., Guigo, R., Guyer, M., Hardison, R.C., Haussler, D., Hayashizaki, Y., Hillier, L.W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D.B., Johnson, L.S., Jones, M., Jones, T.A., Joy, A., Kamal, M., Karlsson, E.K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W.J., Kirby, A., Kolbe, D.L., Korf, I., Kucherlapati, R.S., Kulbokas, E.J., Kulp, D., Landers, T., Leger, J.P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D.R., Mardis, E.R., Matthews, L., Mauceli, E., Mayer, J.H., McCarthy, M., McCombie, W.R., McLaren, S., McLay, K., McPherson, J.D., Meldrim, J., Meredith, B., Mesirov, J.P., Miller, W., Miner, T.L., Mongin, E., Montgomery, K.T., Morgan, M., Mott, R., Mullikin, J.C., Muzny, D.M., Nash, W.E., Nelson, J.O., Nhan, M.N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M.J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K.H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C.S., Poliakov, A., Ponce, T.C., Ponting, C.P., Potter, S., Quail, M., Reymond, A., Roe, B.A., Roskin, K.M., Rubin, E.M., Rust, A.G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M.S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J.B., Slater, G., Smit, A., Smith, D.R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J.P., Von Niederhausern, A.C., Wade, C.M., Wall, M., Weber, R.J., Weiss, R.B., Wendl, M.C., West, A.P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R.K., Winter, E., Worley, K.C., Wyman, D., Yang, S., Yang, S.P., Zdobnov, E.M., Zody, M.C., and Lander, E.S. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562.
41. Vogt, N., Rousseau, N., Leng, M., and Malfoy, B. (1988) A study of the B-Z transition of the AC-rich region of the repeat unit of a satellite DNA from Cebus by means of chemical probes. *J. Biol. Chem*. **263**, 11826–11832.
42. Berger, I., Egli, M., and Rich, A. (1996) Inter-strand C-H...O hydrogen bonds stabilizing four-stranded intercalated molecules: stereoelectronic effects of O4' in cytosine-rich DNA. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 12116–12121.
43. Wang, L., Sommer, M., Rajamani, J., and Arvin, A.M. (2009) Regulation of the ORF61 promoter and ORF61 functions in varicella-zoster virus replication and pathogenesis. *J. Virol.* **83**, 7560–7572.
44. Neidle, S. (2009) The structures of quadruplex nucleic acids and their drug complexes. *Curr. Opin. Struct. Biol.* **19**, 239–250.
45. Shogan, B., Kruse, L., Mulamba, G.B., Hu, A., and Coen, D.M. (2006) Virucidal activity of a GT-rich oligonucleotide against herpes simplex virus mediated by glycoprotein B. *J. Virol.* **80**, 4740–4747.

46. Hui, J., Hung, L.H., Heiner, M., Schreiner, S., Neumuller, N., Reither, G., Haas, S.A., and Bindereif, A. (2005) Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. *EMBO J.* **24**, 1988–1998.
47. Hamilton, B.J., Wang, X.W., Collins, J., Bloch, D., Bergeron, A., Henry, B., Terry, B.M., Zan, M., Mouland, A.J., and Rigby, W.F. (2008) Separate cis-trans pathways post-transcriptionally regulate murine CD154 (CD40 ligand) expression: a novel function for CA repeats in the 3'-untranslated region. *J. Biol. Chem*. **283**, 25606–25616.
48. Giraudeau, F., Petit, E., Avet-Loiseau, H., Hauck, Y., Vergnaud, G., and Amarger, V. (1999) Finding new human minisatellite sequences in the vicinity of long CA-rich sequences. *Genome Res.* **9**, 647–653.
49. Wang, A.H., Quigley, G.J., Kolpak, F.J., Crawford, J.L., van Boom, J.H., van der Marel, G., and Rich, A. (1979) Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature* **282**, 680–686.
50. Rich, A. and Zhang, S. (2003) Timeline: Z-DNA: the long road to biological function. *Nat. Rev. Genet.* **4**, 566–572.
51. Johnston, B.H. (1992) Generation and detection of Z-DNA. *Methods Enzymol*. **211**, 127–158.
52. Ho, P.S., Ellison, M.J., Quigley, G.J., and Rich, A. (1986) A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences. *EMBO J.* **5**, 2737–2744.
53. Schroth, G.P., Chou, P.J., and Ho, P.S. (1992) Mapping Z-DNA in the human genome. Computer-aided mapping reveals a nonrandom distribution of potential Z-DNA-forming sequences in human genes. *J. Biol. Chem.* **267**, 11846–11855.
54. Li, H., Xiao, J., Li, J., Lu, L., Feng, S., and Droge, P. (2009) Human genomic Z-DNA segments probed by the Z alpha domain of ADAR1. *Nucleic Acids Res.* **37**, 2737–2746.
55. Wang, G., Christensen, L.A., and Vasquez, K.M. (2006) Z-DNA-forming sequences generate large-scale deletions in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 2677–2682.
56. Prakash, A., Shepard, S.S., Mileyeva-Biebesheimer, O., He, J., Hart, B., Chen, M., Amarachintha, S.P., Bechtel, J., and Fedorov, A. (2009) Evolution of genomic sequence inhomogeneity at mid-range scales. *BMC Genomics* **10**, 513.
57. Trifonov, E.N. (1991) DNA in profile. *Trends Biochem. Sci.* **16**, 467–470.
58. Herzel, H., Weiss, O., and Trifonov, E.N. (1999) 10-11 bp periodicities in complete genomes reflect protein structure and DNA folding. *Bioinformatics* **15**, 187–193.
59. Ioshikhes, I., Trifonov, E.N., and Zhang, M.Q. (1999) Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 2891–2895.
60. Havlin, S., Buldyrev, S.V., Goldberger, A.L., Mantegna, R.N., Peng, C.K., Simons, M., and Stanley, H.E. (1995) Statistical and linguistic features of DNA sequences. *Fractals* **3**, 269–284.
61. Cheng, J., Tong, Z.S., and Zhang, L.X. (2007) Scaling behavior of nucleotide cluster in DNA sequences. *J. Zhejiang Univ. Sci. B* **8**, 359–364.
62. Pellionisz, A.J. (2008) The principle of recursive genome function. *Cerebellum* **7**, 348–359.
63. 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**, 168–170.
64. Peng, C.K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.N., Simons, M., and Stanley, H.E. (1995) Statistical properties of DNA sequences. *Physica A* **221**, 180–192.
65. Haring, D. and Kypr, J. (2001) Mosaic structure of the DNA molecules of the human chromosomes 21 and 22. *Mol. Biol. Rep*. **28**, 9–17.
66. Nicolay, S., Argoul, F., Touchon, M., d'Aubenton-Carafa, Y., Thermes, C., and Arneodo, A. (2004) Low frequency rhythms in human DNA sequences: a key to the organization of gene location and orientation? *Phys. Rev. Lett.* **93**, 108101.
67. Cheng, J. and Zhang, L.X. (2005) Statistical properties of nucleotide clusters in DNA sequences. *J. Zhejiang Univ. Sci. B* **6**, 408–412.
68. Zhang, Z. and Gerstein, M. (2003) Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res.* **31**, 5338–5348.
69. Nei, M., Suzuki, Y., and Nozawa, M. (2010) The neutral theory of molecular evolution in the genomic era. *Annu. Rev. Genomics Hum. Genet*. **11,** 265–289.
70. Duret, L. and Galtier, N. (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genomics Hum. Genet*. **10**, 285–311.
71. Pozzoli, U., Menozzi, G., Fumagalli, M., Cereda, M., Comi, G.P., Cagliani, R., Bresolin, N., and Sironi, M. (2008) Both selective and neutral processes drive GC content evolution in the human genome. *BMC Evol. Biol.* **8**, 99.
72. Buschiazzo, E. and Gemmell, N.J. (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. *Bioessays* **28**, 1040–1050.
73. Wells, R.D. (2007) Non-B DNA conformations, mutagenesis and disease. *Trends Biochem. Sci.* **32**, 271–278.
74. Lupski, J.R. (1998) Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* **14**, 417–422.

75. Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., Okwuonu, G., Hines, S., Lewis, L., DeRamo, C., Delgado, O., Dugan-Rocha, S., Miner, G., Morgan, M., Hawes, A., Gill, R., Celera, Holt, R.A., Adams, M.D., Amanatides, P.G., Baden-Tillson, H., Barnstead, M., Chin, S., Evans, C.A., Ferriera, S., Fosler, C., Glodek, A., Gu, Z., Jennings, D., Kraft, C.L., Nguyen. T., Pfannkoch, C.M., Sitter, C., Sutton, G.G., Venter, J.C., Woodage, T., Smith, D., Lee, H.M., Gustafson, E., Cahill, P., Kana, A., Doucette-Stamm, L., Weinstock, K., Fechtel, K., Weiss, R.B., Dunn, D.M., Green, E.D., Blakesley, R.W., Bouffard, G.G., De Jong, P.J., Osoegawa, K., Zhu, B., Marra, M., Schein, J., Bosdet, I., Fjell, C., Jones, S., Krzywinski, M., Mathewson, C., Siddiqui, A., Wye, N., McPherson, J., Zhao, S., Fraser, C.M., Shetty, J., Shatsman, S., Geer, K., Chen, Y., Abramzon, S., Nierman, W.C., Havlak, P.H., Chen, R., Durbin, K.J., Egan, A., Ren, Y., Song, X.Z., Li, B., Liu, Y., Qin, X., Cawley, S., Worley, K.C., Cooney, A.J., D'Souza, L.M., Martin, K., Wu, J.Q., Gonzalez-Garay, M.L., Jackson, A.R., Kalafus, K.J., McLeod, M.P., Milosavljevic, A., Virk, D., Volkov, A., Wheeler, D.A., Zhang, Z., Bailey, J.A., Eichler, E.E., Tuzun, E., Birney, E., Mongin, E., Ureta-Vidal, A., Woodwark, C., Zdobnov, E., Bork, P., Suyama, M., Torrents, D., Alexandersson, M., Trask, B.J., Young, J.M., Huang, H., Wang, H., Xing, H., Daniels, S., Gietzen, D., Schmidt, J., Stevens, K., Vitt, U., Wingrove, J., Camara, F., Mar Albà, M., Abril, J.F., Guigo, R., Smit, A., Dubchak, I., Rubin, E.M., Couronne, O., Poliakov, A., Hübner, N., Ganten, D., Goesele, C., Hummel, O., Kreitler, T., Lee, Y.A., Monti, J., Schulz, H., Zimdahl, H., Himmelbauer, H., Lehrach, H., Jacob, H.J., Bromberg, S., Gullings-Handley, J., Jensen-Seaman, M.I., Kwitek, A.E., Lazar, J., Pasko, D., Tonellato, P.J., Twigger, S., Ponting, C.P., Duarte, J.M., Rice, S., Goodstadt, L., Beatson, S.A., Emes, R.D., Winter, E.E., Webber, C., Brandt, P., Nyakatura, G., Adetobi, M., Chiaromonte, F., Elnitski, L., Eswara, P., Hardison, R.C., Hou, M., Kolbe, D., Makova, K., Miller, W., Nekrutenko, A., Riemer, C., Schwartz, S., Taylor, J., Yang, S., Zhang, Y., Lindpaintner, K., Andrews, T.D., Caccamo, M., Clamp, M., Clarke, L., Curwen, V., Durbin, R., Eyras, E., Searle, S.M., Cooper, G.M., Batzoglou, S., Brudno, M., Sidow, A., Stone, E.A., Venter, J.C., Payseur, B.A., Bourque, G., López-Otín, C., Puente, X.S., Chakrabarti, K., Chatterji, S., Dewey, C., Pachter, L., Bray, N., Yap, V.B., Caspi, A., Tesler, G., Pevzner, P.A., Haussler, D., Roskin, K.M., Baertsch, R., Clawson, H., Furey, T.S., Hinrichs, A.S., Karolchik, D., Kent, W.J., Rosenbloom, K.R., Trumbower, H., Weirauch, M., Cooper, D.N., Stenson, P.D., Ma, B., Brent, M., Arumugam, M., Shteynberg, D., Copley, R.R., Taylor, M.S., Riethman, H., Mudunuri, U., Peterson, J., Guyer, M., Felsenfeld, A., Old, S., Mockrin, S., Collins, F.; Rat Genome Sequencing Project Consortium (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521.
76. Orlov, Y.L. and Potapov, V.N. (2004) Complexity: an internet resource for analysis of DNA sequence complexity. *Nucleic Acids Res.* **32**, W628–633.
77. Orlov, Y.L., Te Boekhorst, R., and Abnizova, I.I. (2006) Statistical measures of the structure of genomic sequences: entropy, complexity, and position information. *J. Bioinform. Comput. Biol.* **4**, 523–536.
78. Sanford, J.C. (2005) *Genetic Entropy and the Mystery of the Genome*. ILN, London.

This article should be cited as follows:

Fedorova, L. and Fedorov, A. (2011) Mid-range inhomogeneity of eukaryotic genomes. *TheScientificWorldJOURNAL* **11**, 842–854. DOI 10.1100/tsw.2011.82.