Chapter 3

An Intricate Mosaic of Genomic Patterns at Mid-range Scale

Alexei Fedorov and Larisa Fedorova*

1. 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," are up to 17-nucleotide-long sequences which are over-abundant in the exons and introns of humans and other mammals (Rigoutsos *et al.*, 2006; Tsirigos and Rigoutsos, 2008). 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-nucleotide-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 ($\sim 3 \times 10^9$ 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 other chapters of this book. Here we consider genomic patterns longer than 30 and up to several thousands of nucleotides to be called mid-range scale. At this 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 one

*Department of Medicine, The University of Toledo, Health Science Campus, Ohio, USA

represented by genomic isochores reviewed elsewhere (Bernardi, 2007). 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 by ten thousand bases. The main focus of this chapter 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 waiting for thorough exploration and recognition.

2. Results and Discussion

2.1. *DNA repeats — important elements at genomic mid-range scale*

All eukaryotic genomes contain several extra-large "words" recurring many times — so called DNA repetitive elements, the size of which are generally within mid-range scale. DNA repeats are classified into three major classes based on the molecular mechanisms of their origin and propagation: transposons, retrotransposons, and tandemly organized repeats. There is a large variety of transposons and retrotransposons that can be specific for narrow taxons (like the Alu-elements within primates), or that have a much broader representation (like the L1-repeats found in all vertebrates). We will not consider these DNA repetitive elements, but only refer a reader to several excellent, detailed reviews on their genomic organization and evolution (Jurka *et al.*, 2007; Eickbush and Jamburuthugoda, 2008; Richard *et al.*, 2008). Here we concentrate only on the simple tandem repeats that exist in almost every eukaryotic organism. Our examples well illustrate the common trend for mid-range scale sequence patterns to associate with DNA conformation abnormalities or alternative 3D structures.

We begin the examination of a simple tandem repeat from a well-characterized type composed of a reiterating pentamer sequence AATGG. Our computer analysis of completely sequenced mammalian genomes demonstrates that there were 162 different loci inside the euchromatic genomic regions of humans, 24 in mouse, 14 in rat, 21 in cow, and 58 in dog that contain $(\mathrm{AATGG})_N$ perfect repeats, where $N \geq 4$. These

sequences are proportionally distributed between intergenic regions and introns and, often, there are up to several dozens of tandem AATGG pentamers in one locus. The location of these repeats inside genes is not evolutionarily conserved, since we have not detected their presence in the same intron of the named species. In addition, the same tandemly repeated pentamer AATGG is one of the most evolutionarily conserved parts of a centromere, where it exists in thousands of copies and serves as an attachment point for the two sister chromosomes during mitosis (Grady *et al.*, 1992; Lee *et al.*, 1997). Interestingly, under physiological conditions, this DNA-repeat comprised of at least four pentamers could exist not only as B-form Watson–Crick duplex but also in an unusual form with highly asymmetrical conformations of AATGG-strand and its complementary CCATT-strand (Jaishree and Wang, 1994; Catasti *et al.*, 1999). Its transition from Watson–Crick duplex to single-stranded structures is facilitated by acidic pH conditions. Figure 1 demonstrates the NMR solution structure of the anti-parallel stranded non-B-form DNA duplex 5′-TGGAATGGAA:TGGAATGGAA-3′ created by two repetitive pentamers published by Chou *et al.* (1994). This particular structure is also known as "interdigitated" or zipper-like stacking (Chou *et al.*, 2003). The scheme of the unusual 3D structure of AATGG repeats is illustrated in Figs. 8–10 of Catasti and co-authors (Catasti *et al.*, 1999) while a slightly different variant of spatial organization of the same repeat is illustrated in Fig. 1 of Jaishree and Wang (1994). These two papers demonstrate that the AATGG strand of the repeat forms stable doubly folded hairpins with Watson–Crick A-T and non-Watson–Crick A-G and G-G base pairs in the stems. The stability of these stems is reached partially due to stacking of the three purines shown by arrows in Fig. 1. Moreover, the same authors demonstrated that a greater number of the AATGG pentamers might form higher-order structure in which doubly folded hairpins are compactly organized in a helical array of $(AATGG)_4$ units. At the same time, the complementary strand formed by CCATT pentamers is unstable under physiological conditions and likely represents loose structures. Under acidic conditions, this CCATT tandem repeat might also form unusual structures known as i-motif with intercalated cytosine bases shown in Fig. 2 (Catasti *et al.*, 1999; Nonin-Lecomte and Leroy, 2001).

Other tandemly organized simple repeats could also have conformations far distinct from Watson–Crick double helices. One of the prominent noncanonical structures, known as G-quadruplex, G-quartet, or

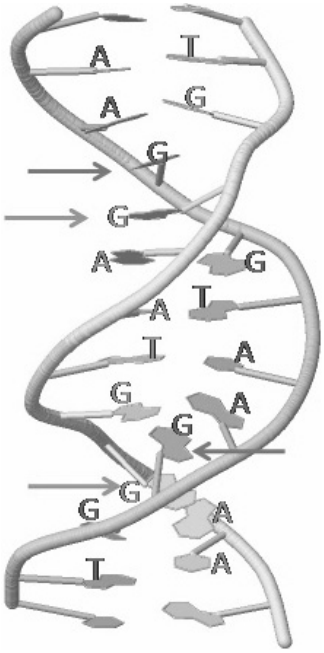

Fig. 1. Cartoon of 3D-structure of anti-parallel DNA duplex 5′TGGAATGGAA: TGGAATGGAA3′ formed by two copies of TGGAA pentamer repeat. This picture is a snapshot of the structure with the identifier 103D obtained from the Protein Data Bank. The structure was resolved via solution NMR approach by Chow and co-authors (Chou *et al.*, 1994). Four arrows point to unpaired guanosine residues stacked between Hoogsteen G-A pairs.

G4, is formed by guanine-rich strands of the repeats. Quadruplexes are arranged in four-stranded structures with stands connected to each other via Hoogsteen hydrogen bonding. G-quadruplex has been well characterized in human telomeric and related sequences with the core repetitive element TTAGGGG and also within promoters and 5'-untranslated regions of human genes whose sequences have a loose consensus of $G_{3-5}N_{L1}G_{3-5}N_{L2}G_{3-5}N_{L3}G_{3-5}$, where N_{L1}, N_{L2}, and N_{L3} are loops with the length from 1 to 7 nucleotides and variable nucleotide composition (Neidle, 2009). There are several alternative conformations of G-quadruplexes due to the organization of the strands relative to each other. Among them are anti-parallel, parallel, and parallel/anti-parallel hybrids (Oganesian and Bryan, 2007; Huppert, 2008). One example of a

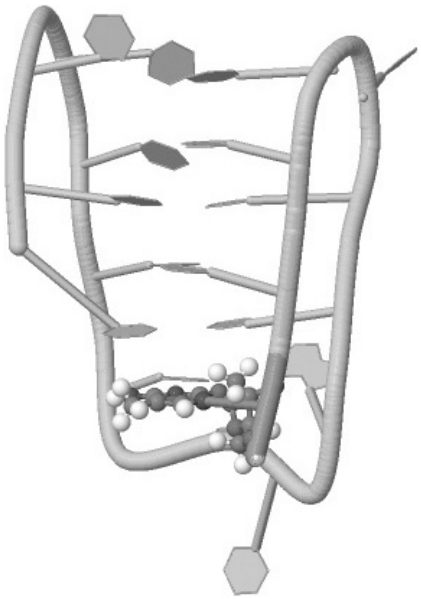

Fig. 2. Cartoon of 3D-structure of a C-rich strand fragment of the human centromeric satellite III d(CCATTCCATTCCTTTCC) that forms intramolecurlar i-motif structure with C.C(+) pairs from parallel strands intercalated head-to-tail. This picture is a snapshot of the structure with the identifier 1G22 obtained from the Protein Data Bank. The structure was resolved via solution NMR approach by Nonin-Lecomte and Leroy (2001) for uridine derivative methylated on the first cytidine base, d(5mCCATTCCAUTCCUTTCC), whose proton spectrum is better resolved. Modified residue 5-METHYL-2'-DEOXYCYTIDINE is demonstrated with white, blue, red, and grey spheres.

parallel-stranded G4 NMR structure is illustrated in Fig. 3. G-quadruplex structure has been demonstrated for several G-rich short tandem repeats. Among them are GGA and CGG triplet repeats ((Matsugami *et al.*, 2003; Nakagama *et al.*, 2006) and the GGCAG mouse minisatellite Pc-1 (Katahira *et al.*, 1999). Four elements of GGA-repeat can form intramolecular parallel quadruplex, while the neighboring quadruplexes can form a dimer stabilized through the stacking interaction between the heptads of the two quadruplexes (Matsugami *et al.*, 2003). While G-quadruplexes are formed by a guanine-rich strand, their complementary strand being C-rich may also form a completely different four-stranded structure known as i-motif or intercalated cytosine tetraplex. This structure is only stable

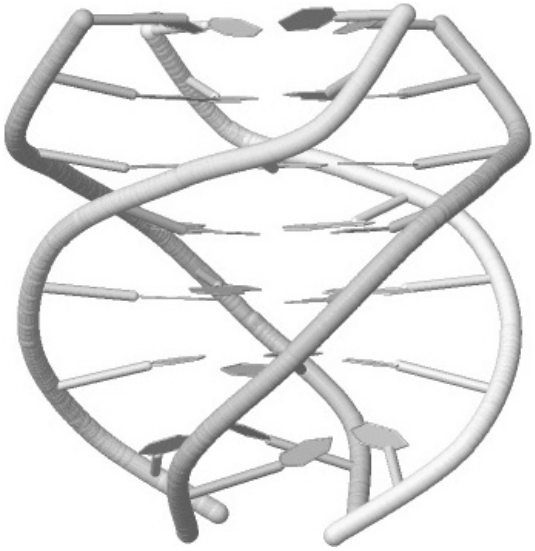

Fig. 3. Cartoon of 3D-structure of a parallel-stranded G-quadruplex DNA formed by the Tetrahymena telomeric sequence d(T-T-G-G-G-G-T). This picture is a snapshot of the structure with the identifier 139D obtained from the Protein Data Bank. The structure was resolved by combined NMR-computational approach by (Wang and Patel, 1993).

under acidic conditions (Huppert, 2008). Intercalated cytosines have been found in several unusual conformations, one of which is shown in Fig. 2.

There are up to 10 different non-B-from DNA conformations associated with the simple repeats listed and well illustrated by Wells (Wells, 2007). Among them are slipped structure formed by CNG repeats; triplex DNA formed by purine (R)-rich or pyrimidine (Y)-rich mirror repeats (described below in details); sticky DNA formed by G + A-rich tracts like $(GAA)_N$; and DNA unwinding elements formed by A + T-rich regions. In addition, Wells describes cruciforms created by inverted repeats; and left-handed Z-DNA formed by alternated R/Y bases in $(RY)_N$ repeats (Wells, 2007). There is evidence that CGG triplet repeats could form a non-B-type higher order structure (Nakagama *et al.*, 2006). This particular CGG repeat is widespread in animal genomes and it expands inside the first exon of the FMR1 gene that causes Fragile X syndrome (Mandel, 1993; Crawford *et al.*, 2001). NMR and X-ray crystallography studies of DNA oligonucleotides strongly suggest that $(GA)_N$

and $(A)_N$ repeats could form parallel-stranded homeoduplexes (Kypr *et al.*, 2007; Chakraborty *et al.*, 2009). However it is questionable whether such structures might exist *in vivo*.

Tandem repeats with longer units could also form non-B structures. For instance, the $(ACAGGGGTGTGGGG)_N$ insulin minisatellite has a complex loop-folding conformation (Catasti *et al.*, 1999). The listed non-B DNA conformations likely do not exist permanently, but only under specific conditions. Their formation can be facilitated by negative supercoiling during transcription or by binding with transcription factors (Mirkin, 2008). The very specific pattern of mutagenesis within simple repeats associated with particular bases and particular sites strongly suggests the existence of non-B structures *in vivo* (Wells, 2007). Also, several non-B structures have been confirmed in various experiments including *in vivo* studies (Wells, 2007; Fernando *et al.*, 2009; Kypr *et al.*, 2009). Currently, more than 70 human genetic disorders have been associated with changes in simple repeats (Lupski, 1998; Wells, 2007).

In summary, simple repeats are abundant in the genomes of diverse animals and plants. In rodents, 2.4% of the euchromatic part of their genome is represented by simple repeats, which is two times bigger than the length of all protein-coding sequences (Gibbs *et al.*, 2004). Additionally, tandemly organized short sequences are abundant and are key components of telomeres and centromeric regions. Many simple repeats, whose total length reaches 20 bases and above, under certain conditions can exist in a variety of non-B DNA conformations *in vivo* associated with specific genomic functions. Computationally, simple repeats can be detected by the RepeatMasker program (Smit AFA). However, the default parameters of this program could skip recognition of simple repeat loci whose copy numbers are low or where the sequences have accumulated mutations (fuzzy repeats). In this case the best choice is the stand-alone Tandem Repeat Finder with advanced search parameters (Benson, 1999).

2.2. *Genomic Mid-Range Inhomogeneity (MRI): Nucleotide compositional extremes and sequence nonrandomness*

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 base pairs but less than 10 000 base pairs. To characterize genomic MRI patterns, a public computational resource (*Genomic MRI*) has been created that allows detecting sequence regions with any type of extreme composition (Bechtel *et al.*, 2008). 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 (Prakash *et al.*, 2009).

2.2.1. *Genomic MRI toolkit*

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 studying 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 nucleotides 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 (Bechtel *et al.*, 2008). 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. Firstly, 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 1–9 nucleotide long "word." Then, a second program, SRI-*generator*, creates a random sequence with a short-range inhomogeneity that approximates 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 ones, e.g. G + C-poor are A + T-rich). An example of *MRI analyzer* graphical output is shown in Fig. 4 that illustrates the

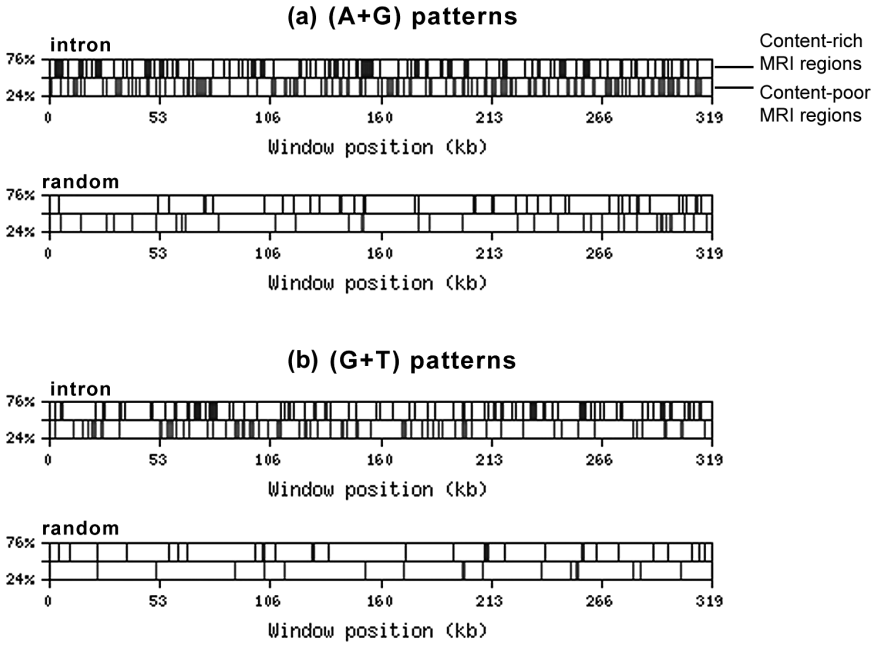

Fig. 4. shows the graphical output of the *MRI-analyzer* program for the first intron of the Dystrophin gene (marked as "intron") and also for the SRI-generated random sequence based on the tetramer oligonucleotide frequency table of the intron (marked as "random"). The entire sequence of the 319 kb intron and the random sequence is displayed on the *x*-axis. Shaded bars on each top row represent positions of content-rich MRI regions on the sequence. Bars on the bottom row 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.

MRI-patterns for an extra-large human intron of the dystrophin gene from chromosome X.

Two scales of MRI regions should be considered. First, regions from 30 to 1000 bp, whose properties have been investigated in detail and for which several periodicities have been reported (Trifonov, 1991; Herzel *et al.*, 1999; Ioshikhes *et al.*, 1999). Second, 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-nucleotide-long window, there are 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.

Extended regions with compositional extremes satisfying G + C- or A + T-richness are abundant in vertebrates and can be as long as several million bases (known as genomic isochores). Other composition extremes, such as R-, Y-, G + T-, A + C-richness, that extend over long chromosomal regions are not as abundant as C + G- and A + T-rich genomic areas. Nonetheless, for more than 2100 human chromosomal regions with lengths exceeding 10 kb, we have detected frequencies of more than 60% for G + A-, T + C-, A + C-, or G + T-nucleotides. As for extremes, our computations have shown that there are 22 regions in the human genome where R or Y composition exceeds 70% within a sequence longer than 10 kb.

Recently, by studying the distribution of more than four 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 (Prakash *et al.*, 2009). 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. All in all, the observed strong fixation bias for mutations helps to preserve MRI regions during evolution, indicating their involvement in genomic operations.

2.2.2. *G + C-rich and A + T-rich MRI regions are associated with several unusual DNA structures*

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 thirty to several thousand nucleotides are 4–20 times over-represented in the mammalian genomes compared to random expectation (Bechtel, 2008; Bechtel *et al.*, 2008). 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 (Hackenberg *et al.*, 2006). CpG-islands are found in nearly 60% of human genes including almost all of the housekeeping ones (Hackenberg *et al.*, 2006). According to two different definitions of these islands, their length must be at least 200 or 500 bp long; 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 (Gardiner-Garden and Frommer, 1987; Takai and Jones, 2003; respectively). 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 (Suzuki and Bird, 2008). It has been shown recently that CpG dinucleotide 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 (Greenbaum, Pang, and Tullius, 2007; Greenbaum, Parker, and Tullius, 2007) and during the sonication of double-stranded DNA (Grokhovsky *et al.*, 2008). The crucial involvement of cytosine methylation in the regulation of gene expression is well described in a number of reviews including some recent ones (Prokhortchouk and Defossez, 2008; Suzuki and Bird, 2008; Illingworth and Bird, 2009). Thus, here we concentrate on the other physicochemical properties of G + C-rich and A + T-rich regions.

It is well known that A-form of 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 (Warne and deHaseth, 1993; Stefl *et al.*, 2001; Kypr *et al.*, 2009). A special form of DNA which is an intermediate between A- and B-forms, has been characterized in G + C-rich sequences with methylated cytosines (Vargason *et al.*, 2000). In addition, short $(CpG)_n$ repeats could adopt Z-DNA as

recently reviewed by P.S. Ho (Ho, 2009). This Z-DNA is proposed to serve as transcriptional co-activator (Liu *et al.*, 2001).

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 (Kowalski *et al.*, 1988). These elements are often associated with the origins of replication in eukaryotes and prokaryotes (Umek *et al.*, 1989). 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 anti-parallel double helical duplex in which the base pairing is of the Hoogsteen type (Abrescia *et al.*, 2004). 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 base pairs per turn. Theoretical analysis has demonstrated that energies of the Hoogsteen form and B-form of DNA are practically identical (Cubero *et al.*, 2003). Most recently, Chakraborty and co-authors 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 (Chakraborty *et al.*, 2009).

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 (Liebich *et al.*, 2002). Second, centromere DNA of diverse animals, plants and fungi always contain A + T-rich regions (Choo, 1997; Abrescia *et al.*, 2004).

2.2.3. *R-rich/Y-rich MRI regions are associated with 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- and Y-rich sequences and also T + G- and A + C-rich ones have physically the same loci, yet representing 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 (Felsenfeld and Rich, 1957). 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. 5. There are four kinds of H-DNA depending on strand type and orientation (Jain *et al.*, 2008). 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

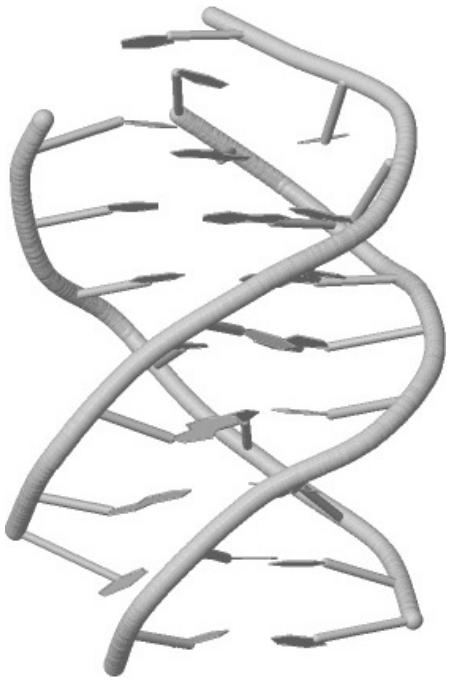

Fig. 5. Cartoon of 3D-structure of a purine.purine.pyrimidine DNA triplex containing G.GC and T.AT triples. This picture is a snapshot of the structure with the identifier 134D obtained from the Protein Data Bank. The structure was resolved using a combined NMR and molecular dynamics approach by Radhakrishnan and Patel (1993).

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 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 (Van Raay *et al.*, 1996; Blaszak *et al.*, 1999). Another kind of intramolecular triplex, can be formed at neutral pH and requires bivalent cations for stability. It is formed by the interaction of 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 (Malkov *et al.*, 1993).

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 recently (Jain *et al.*, 2008; Wang, Zhao, and Vasquez 2009). 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 homo-purine 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 and others (Jain *et al.*, 2008). Bacolla with co-authors characterized nearly 3000 homo-purine tracks in the human genome longer than 100 nucleotides (Bacolla *et al.*, 2006). 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 co-authors surveyed the evidence for how H-DNA influences the activity of DNA and RNA polymerases. Finally, Goni and others (Goni *et al.*, 2006) 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; likely they are not direct targets for transcription factors. Goni and co-authors have suggested that these sequences act as pacing fragments in promoter regions and help in the correct positioning of transcription factors.

2.2.4. *DNA and RNA properties of GT-rich/AC-rich MRI regions*

Recall that the complementary strands of G + T-rich regions are naturally A + C-rich regions. They co-exist with each other and we consider them interchangeably 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 *aMino* or M (Moss). Thus, sometimes these regions are referred to as K.M-tracks or motifs (Yagil, 2004). Bechtel and co-authors demonstrated that G + T-regions are about five times more abundant in the mammalian genomes compared to random expectation (Bechtel *et al.*, 2008). Moreover, these regions practically do not intersect with interspersed DNA repeats at all. In 2004 Yagil 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 *D. melanogaster* genome, where they outnumber other motifs such as R/Y-rich motifs (Yagil, 2004). 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. Firstly, $(CA)_N$ simple repeats are one of the most profuse tandem repeats in mammalian genomes (Waterston *et al.*, 2002). They also should be considered as alternating R/Y sequence, and, due to this property, associated with a Z-DNA conformation (Vogt *et al.*, 1988), 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 (Berger *et al.*, 1996). We mentioned such structures (i-motifs) above in the Simple Repeat section and present an example of it in Fig. 2. Third, telomeres of various eukaryotic species are represented by G + T-rich regions which form G-quadruplexes (see above). Fourth, short G + T-rich regions could represent transcription factor binding sites such as for factor Sp1 (Wang *et al.*, 2009). Intriguingly, G + T-rich oligonucleotides possess antiviral

activities. For example, $T_2(G_4T_2)_3$ sequences are virucidal against herpes simplex virus (Shogan *et al.*, 2006). At the RNA level, C + A-rich sequences within intronic segments could regulate alternative splicing by being binding sites for the hnRNP L protein (Hui *et al.*, 2005). The presence of C + A-rich sequences at the 3'-UTR of mRNA could regulate gene expression at the level of translation (Hamilton *et al.*, 2008). 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 (Giraudeau *et al.*, 1999). 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.

2.2.5. *Alternated R/Y MRI regions adopt Z-DNA conformation*

Left-handed anti-parallel Z-DNA double helix conformation has been first characterized in 1979 by Wang and co-authors for $(GC)_3$ repeats (Wang *et al.*, 1979). Detailed Z-DNA structure has been considered elsewhere (Rich and Zhang, 2003; Ho, 2009). 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 (Ho, 2009). Thus, Z-DNA, which is characterized by alternating pattern of *anti-syn* conformations, is formed by alternating R/Y sequences (Johnston, 1992). The latest version of the *Genomic MRI* package has a new feature allowing 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 over-abundance of alternating R/Y stretching over 50–100 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 human heparanase-2 gene highly enriched with alternated R and Y bases: 5′AAATGGATGTGTGTATATATATGAAGTCGATACACACACATATA CACATA3′. We showed that such alternating R/Y sequences are plentiful throughout the mammalian genomes either inside introns or within intergenic regions.

In 1986 Ho and others developed a ZHUNT program for detection of genomic sequences with high propensity to form Z-DNA (Ho *et al.*, 1986). They found a high concentration of these sequences near the transcription start sites (Schroth *et al.*, 1992; Rich and Zhang, 2003). Most recently, human genomic Z-DNA segments have been detected experimentally using a Z-DNA binding protein domain as a probe (Li *et al.*, 2009). These authors found an abundance of Z-DNA hotspots 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 (Wang *et al.*, 2006). 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 (Rich and Zhang, 2003).

2.3. *Weak periodicities and loose patterns*

In addition to MRI patterns, there are several weak genomic periodicities and specific signals at the mid-range scale. Many of them are described in "The codes of life" (Barbieri, 2008). Wherein, Trifonov reviewed different codes that exist in the genomes at DNA, RNA, and protein levels. He emphasized a special property of genomic sequences to make superposition (overlapping) of the codes they carry. The overlapping is possibly due to degeneracy of the codes and might be useful for organism survivability (Peleg *et al.*, 2004). Here we consider three types of such patterns in eukaryotic genomes.

2.3.1. *Chromatin periodicities*

There exists a nonrandom positioning of nucleosomes along genomic DNA of eukaryotes (Salih *et al.*, 2007). Nucleosome binding preferences are achieved via sequence-dependent deformational anisotropy of DNA (Barbieri, 2008). On average, one nucleosome occupies 200 bp including 145 nucleotides that contact its core particle while the rest corresponds to linkers between nucleosomes. Due to this specificity in nucleosome

positioning, Trifonov and co-authors described sequence features that are repeated with 200- and 400-base periodicities (Trifonov 1998; Cohanim *et al.*, 2006).

2.3.2. *Periodicities in protein-coding sequences*

There are well-known short-range periodicities in coding sequences that exist due to nonsymmetry in the genetic code, nonrandom amino acid appearance and association of neighboring amino acids within protein sequences, and also regularities in codon bias and context-dependent codon bias (Fedorov *et al.*, 2002). In addition, there exist longer periodicities in the coding sequences that correspond to modular organization of globular proteins. They extend over 20–30 codons and represent initial protein folding modules (Aharonovsky and Trifonov, 2005; Barbieri, 2008).

2.3.3. *Transcription-associated mutational asymmetry in mammals*

In 2003 Green *et al.* demonstrated that the transcribed strands of mammalian DNA have an excess of G + T over A + C due to the difference of particular mutation frequencies (Green *et al.*, 2003). Specifically, the A → G transition occurs at a 28% higher rate than the complementary transition T → C on the transcribed strand in most human genes. This transcription-associated mutational bias exists for both the exonic and intronic parts of genes. Thus, if we look at the nucleotide frequencies in the combined sequences of all human introns (T = 30.7%; A = 28.0%; G = 21.1%; C = 20.2%) there is a 3.6% excess of G + T over A + C. (These calculations were obtained on our nonredundant set of 11 315 human genes containing 96 931 introns (Bechtel *et al.*, 2008). For each intron we removed the first 10 and the last 30 bases). We detected the same preference of G + T over A + C in introns of other mammals, a smaller preference for sea urchin (2.2%), and the highest preference in Arabidopsis (11.8%). For the mouse-ear cress, the nucleotide bias in introns is mainly due to significant excess of T (39.6%) over A (28.2%). Such strong transcriptional asymmetry in the preference of G + T over A + C is typical for other plants. No difference for G + T versus A + C composition has been detected for fruit fly and worm introns.

2.4. *A complex mosaic of MRI patterns and their fundamental importance*

2.4.1. *Intricate arrangement of genomic MRI patterns*

Different MRI regions are not randomly arranged relative to each other (Bechtel, 2008). For example, Fig. 6 illustrates that G + C-rich regions

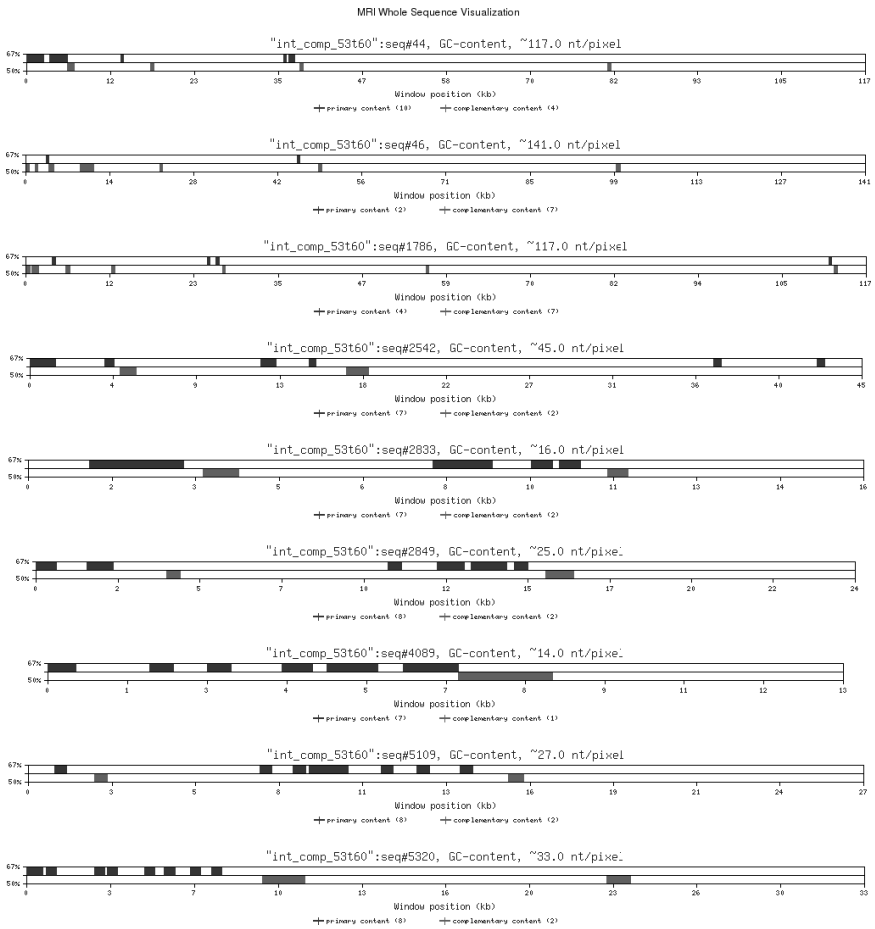

Fig. 6. Visualization of G + C-rich (top row) and A + T-rich (bottom row) MRI features in human introns using a 400-nt base window size. The scale for each sequence is independent and is given in its subheading in nucleotides per pixel. The figure represents a fragment of Fig. 17 in Bechtel (2008).

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 (Bechtel, 2008). So far, investigators have examined only individual genomic patterns. The mutual arrangement of various genomic mid-range patterns has never been thoroughly investigated yet. 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 *MRI-generator* program from the *Genomic MRI* package has proven not to be a trivial computational task (Bechtel *et al.*, 2008). This has made us appreciate that the reconstruction of the entire set of MRI patterns in modeling DNA sequences is an extremely challenging mission due to a complex multi-layer nonrandomness in genomic sequences. In addition, genomic sequences have an intricate organization of nested patterns and also with respect to the clustering of particular patterns. Some features of this complex organization were described as genomic fractals in several publications (Havlin *et al.*, 1995; Cheng *et al.*, 2007; Pellionisz, 2008). This arrangement has been studied by methods such as "detrended fluctuation analysis" and a "Brownian walk" 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 (Peng *et al.*, 1992; Havlin *et al.*, 1995; Peng *et al.*, 1995; Haring and Kypr, 2001; Nicolay *et al.*, 2004; Cheng and Zhang, 2005; Cheng *et al.*, 2007).

2.4.2. *The purpose 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 principles is the Principle of Recursive Genome Function (PRGF) illuminated by Pellionisz (Pellionisz, 2008). 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 John Sanford (Sanford, 2005). 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.

3. 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 multi-cellular

eukaryotes. Understanding of genomic MRI functions is critical for the newly emerged field of personal genomics and also for drug discovery.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. 0643542.

References

Abrescia NG, Gonzalez C, Gouyette C, Subirana JA. (2004) X-ray and NMR studies of the DNA oligomer d(ATATAT): Hoogsteen base pairing in duplex DNA. *Biochemistry* **43**: 4092–4100.

Aharonovsky E, Trifonov EN. (2005) Protein sequence modules. *J Biomol Struct Dyn* **23**: 237–242.

Bacolla A, Collins JR, Gold B *et al.* (2006) Long homopurine*homopyrimidine sequences are characteristic of genes expressed in brain and the pseudoautosomal region. *Nucleic Acids Res* **34**: 2663–2675.

Barbieri M. (2008) *The Codes of Life. The rules of Macroevolution. Biosemiontics.* Springer.

Bechtel JM. (2008) *Characterization of Genomic Mid-Range Inhomogenity.* PP. 97. Health Science Campus. University of Toledo, Toledo.

Bechtel JM, Wittenschlaeger T, Dwyer T *et al.* (2008) Genomic mid-range inhomogeneity correlates with an abundance of RNA secondary structures. *BMC Genomics* **9**: 284.

Benson G. (1999) Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573–580.

Berger I, Egli M, 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 USA* **93**: 12116–12121.

Bernardi G. (2007) The neoselectionist theory of genome evolution. *Proc Natl Acad Sci USA* **104**: 8385–8390.

Blaszak RT, Potaman V, Sinden RR, Bissler JJ. (1999) DNA structural transitions within the PKD1 gene. *Nucleic Acids Res* **27**: 2610–2617.

Catasti P, Chen X, Mariappan SV, Bradbury EM, Gupta G. (1999) DNA repeats in the human genome. *Genetica* **106**: 15–36.

Chakraborty S, Sharma S, Maiti PK, Krishnan Y. (2009) The poly dA helix: A new structural motif for high performance DNA-based molecular switches. *Nucleic Acids Res* **37**: 2810–2817.

Cheng J, Tong ZS, Zhang LX. (2007) Scaling behavior of nucleotide cluster in DNA sequences. *J Zhejiang Univ Sci B* **8**: 359–364.

Cheng J, Zhang LX. (2005) Statistical properties of nucleotide clusters in DNA sequences. *J Zhejiang Univ Sci B* **6**: 408–412.

Choo KH. (1997) *The Centromere*. Oxford Univ Press, Oxford, UK.

Chou SH, Cheng JW, Fedoroff O, Reid BR. (1994) DNA sequence GCGAATGAGC containing the human centromere core sequence GAAT forms a self-complementary duplex with sheared G.A pairs in solution. *J Mol Biol* **241**: 467–479.

Chou SH, Chin KH, Wang AH. (2003) Unusual DNA duplex and hairpin motifs. *Nucleic Acids Res* **31**: 2461–2474.

Chou SH, Zhu L, Reid BR. (1994) The unusual structure of the human centromere (GGA)2 motif. Unpaired guanosine residues stacked between sheared G.A pairs. *J Mol Biol* **244**: 259–268.

Cohanim AB, Kashi Y, Trifonov EN. (2006) Three sequence rules for chromatin. *J Biomol Struct Dyn* **23**: 559–566.

Crawford DC, Acuna JM, Sherman SL. (2001) FMR1 and the fragile X syndrome: Human genome epidemiology review. *Genet Med* **3**: 359–371.

Cubero E, Abrescia NG, Subirana JA, Luque FJ, Orozco M. (2003) Theoretical study of a new DNA structure: The antiparallel Hoogsteen duplex. *J Am Chem Soc* **125**: 14603–14612.

Eickbush TH, Jamburuthugoda VK. (2008) The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Res* **134**: 221–234.

Fedorov A, Saxonov S, Gilbert W. (2002) Regularities of context-dependent codon bias in eukaryotic genes. *Nucleic Acids Res* **30**: 1192–1197.

Felsenfeld G, Rich A. (1957) Studies on the formation of two- and three-stranded polyribonucleotides. *Biochim Biophys Acta* **26**: 457–468.

Fernando H, Sewitz S, Darot J, Tavare S, Huppert JL, Balasubramanian S. (2009) Genome-wide analysis of a G-quadruplex-specific single-chain antibody that regulates gene expression. *Nucleic Acids Res* **37**: 6716–6722.

Gardiner-Garden M, Frommer M. (1987) CpG islands in vertebrate genomes. *J Mol Biol* **196**: 261–282.

Gibbs RA, Weinstock GM, Metzker ML *et al.* (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**: 493–521.

Giraudeau F, Petit E, Avet-Loiseau H, Hauck Y, Vergnaud G, Amarger V. (1999) Finding new human minisatellite sequences in the vicinity of long CA-rich sequences. *Genome Res* **9**: 647–653.

Goni JR, Vaquerizas JM, Dopazo J, 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.

Grady DL, Ratliff RL, Robinson DL, McCanlies EC, Meyne J, Moyzis RK. (1992) Highly conserved repetitive DNA sequences are present at human centromeres. *Proc Natl Acad Sci USA* **89**: 1695–1699.

Green P, Ewing B, Miller W, Thomas PJ, Green ED. (2003) Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet* **33**: 514–517.

Greenbaum JA, Pang B, Tullius TD. (2007) Construction of a genome-scale structural map at single-nucleotide resolution. *Genome Res* **17**: 947–953.

Greenbaum JA, Parker SC, Tullius TD. (2007) Detection of DNA structural motifs in functional genomic elements. *Genome Res* **17**: 940–946.

Grokhovsky SL, Il'icheva IA, Nechipurenko DY, Panchenko LA, Polozov RL, Nechipurenko YD. (2008) Heterogeneity of DNA local structure and dynamics: ultrasound studies. *Biofizika* **53**: 417–425.

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

Hamilton BJ, Wang XW, Collins J *et al.* (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.

Haring D, Kypr J. (2001) Mosaic structure of the DNA molecules of the human chromosomes 21 and 22. *Mol Biol Rep* **28**: 9–17.

Havlin S, Buldyrev SV, Goldberger AL *et al.* (1995) Statistical and linguistic features of DNA sequences. *Fractals* **3**: 269–284.

Herzel H, Weiss O, Trifonov EN. (1999) 10–11 bp periodicities in complete genomes reflect protein structure and DNA folding. *Bioinformatics* **15**: 187–193.

Ho PS. (2009) Methods to study nucleic acid structure. *Methods* **47**: 141.

Ho PS, Ellison MJ, Quigley GJ, Rich A. (1986) A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences. *EMBO J* **5**: 2737–2744.

Hui J, Hung LH, Heiner M *et al.* (2005) Intronic CA-repeat and CA-rich elements: A new class of regulators of mammalian alternative splicing. *EMBO J* **24**: 1988–1998.

Huppert JL. (2008) Four-stranded nucleic acids: Structure, function and targeting of G-quadruplexes. *Chem Soc Rev* **37**: 1375–1384.

Illingworth RS, Bird AP. (2009) CpG islands–'a rough guide'. *FEBS Lett* **583**: 1713–1720.

Ioshikhes I, Trifonov EN, Zhang MQ. (1999) Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. *Proc Natl Acad Sci USA* **96**: 2891–2895.

Jain A, Wang G, Vasquez KM. (2008) DNA triple helices: Biological consequences and therapeutic potential. *Biochimie* **90**: 1117–1130.

Jaishree TN, Wang AH. (1994) Human chromosomal centromere (AATGG)n sequence forms stable structures with unusual base pairs. *FEBS Lett* **347**: 99–103.

Johnston BH. (1992) Generation and detection of Z-DNA. *Methods Enzymol* **211**: 127–158.

Jurka J, Kapitonov VV, Kohany O, Jurka MV. (2007) Repetitive sequences in complex genomes: Structure and evolution. *Annu Rev Genomics Hum Genet* **8**: 241–259.

Katahira M, Fukuda H, Kawasumi H, Sugimura T, Nakagama H, Nagao M. (1999) Intramolecular quadruplex formation of the G-rich strand of the mouse hypervariable minisatellite Pc-1. *Biochem Biophys Res Commun* **264**: 327–333.

Kowalski D, Natale DA, Eddy MJ. (1988) Stable DNA unwinding, not "breathing," accounts for single-strand-specific nuclease hypersensitivity of specific A + T-rich sequences. *Proc Natl Acad Sci USA* **85**: 9464–9468.

Kypr J, Kejnovska I, Renciuk D, Vorlickova M. (2009) Circular dichroism and conformational polymorphism of DNA. *Nucleic Acids Res* **37**: 1713–1725.

Kypr J, Kejnovska I, Vorlickova M. (2007) Conformations of DNA strands containing GAGT, GACA, or GAGC tetranucleotide repeats. *Biopolymers* **87**: 218–224.

Lee C, Wevrick R, Fisher RB, Ferguson-Smith MA, Lin CC. (1997) Human centromeric DNAs. *Hum Genet* **100**: 291–304.

Li H, Xiao J, Li J, Lu L, Feng S, Droge P. (2009) Human genomic Z-DNA segments probed by the Z alpha domain of ADAR1. *Nucleic Acids Res* **37**: 2737–2746.

Liebich I, Bode J, Reuter I, Wingender E. (2002) Evaluation of sequence motifs found in scaffold/matrix-attached regions (S/MARs). *Nucleic Acids Res* **30**: 3433–3442.

Liu R, Liu H, Chen X, Kirby M, Brown PO, Zhao K. (2001) Regulation of CSF1 promoter by the SWI/SNF-like BAF complex. *Cell* **106**: 309–318.

Lupski JR. (1998) Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet* **14**: 417–422.

Malkov VA, Voloshin ON, Veselkov AG *et al.* (1993) Protonated pyrimidine-purine-purine triplex. *Nucleic Acids Res* **21**: 105–111.

Mandel JL. (1993) Questions of expansion. *Nat Genet* **4**: 8–9.

Matsugami A, Okuizumi T, Uesugi S, Katahira M. (2003) Intramolecular higher order packing of parallel quadruplexes comprising a G:G:G:G tetrad and a G(:A):G(:A):G(:A):G heptad of GGA triplet repeat DNA. *J Biol Chem* **278**: 28147–28153.

Mirkin SM. (2008) Discovery of alternative DNA structures: A heroic decade (1979–1989). *Front Biosci* **13**: 1064–1071.

Moss GP. Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences.

Nakagama H, Higuchi K, Tanaka E *et al.* (2006) Molecular mechanisms for maintenance of G-rich short tandem repeats capable of adopting G4 DNA structures. *Mutat Res* **598**: 120–131.

Neidle S. (2009) The structures of quadruplex nucleic acids and their drug complexes. *Curr Opin Struct Biol* **19**: 239–250.

Nicolay S, Argoul F, Touchon M, d'Aubenton-Carafa Y, Thermes C, 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.

Nonin-Lecomte S, Leroy JL. (2001) Structure of a C-rich strand fragment of the human centromeric satellite III: A pH-dependent intercalation topology. *J Mol Biol* **309**: 491–506.

Oganesian L, Bryan TM. (2007) Physiological relevance of telomeric G-quadruplex formation: A potential drug target. *Bioessays* **29**: 155–165.

Peleg O, Kirzhner V, Trifonov E, Bolshoy A. (2004) Overlapping messages and survivability. *J Mol Evol* **59**: 520–527.

Pellionisz AJ. (2008) The principle of recursive genome function. *Cerebellum* **7**: 348–359.

Peng CK, Buldyrev SV, Goldberger AL *et al.* (1995) Statistical properties of DNA sequences. *Physica A* **221**: 180–192.

Peng CK, Buldyrev SV, Goldberger AL *et al.* (1992) Long-range correlations in nucleotide sequences. *Nature* **356**: 168–170.

Prakash A, Shepard SS, Mileyeva-Biebesheimer O *et al.* (2009) Evolution of genomic sequence inhomogeneity at mid-range scales. *BMC Genomics* **10**: 513.

Prokhortchouk E, Defossez PA. (2008) The cell biology of DNA methylation in mammals. *Biochim Biophys Acta* **1783**: 2167–2173.

Radhakrishnan I, Patel DJ. (1993) Solution structure of a purine.purine.pyrimidine DNA triplex containing G.GC and T.AT triples. *Structure* **1**: 135–152.

Rich A, Zhang S. (2003) Timeline: Z-DNA: the long road to biological function. *Nat Rev Genet* **4**: 566–572.

Richard GF, Kerrest A, Dujon B. (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol Mol Biol Rev* **72**: 686–727.

Rigoutsos I, Huynh T, Miranda K, Tsirigos A, McHardy A, 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 USA* **103**: 6605–6610.

Salih F, Salih B, Trifonov EN. (2007) Sequence-directed mapping of nucleosome positions. *J Biomol Struct Dyn* **24**: 489–493.

Sanford JC. (2005) Genetic Entropy & the Mystery of the Genome. Elim Publishing.

Schroth GP, Chou PJ, Ho PS. (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.

Shogan B, Kruse L, Mulamba GB, Hu A, Coen DM. (2006) Virucidal activity of a GT-rich oligonucleotide against herpes simplex virus mediated by glycoprotein B. *J Virol* **80**: 4740–4747.

Smit AFA, Hubley R, Green P. (1996–2008) RepeatMasker Open-3.1.8 <http://www.repeatmasker.org>.

Stefl R, Trantirek L, Vorlickova M, Koca J, Sklenar V, Kypr J. (2001) A-like guanine-guanine stacking in the aqueous DNA duplex of d(GGGGCCCC). *J Mol Biol* **307**: 513–524.

Suzuki MM, Bird A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* **9**: 465–476.

Takai D, Jones PA. (2003) The CpG island searcher: a new WWW resource. *In Silico Biol* **3**: 235–240.

Trifonov EN. (1998) 3-, 10.5-, 200- and 400-base periodicities in genome sequences. *Physica a-Statistical Mechanics and Its Applications* **249**: 511–516.

Trifonov EN. (1991) DNA in profile. *Trends Biochem Sci* **16**: 467–470.

Tsirigos A, 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.

Umek RM, Linskens MH, Kowalski D, Huberman JA. (1989) New beginnings in studies of eukaryotic DNA replication origins. *Biochim Biophys Acta* **1007**: 1–14.

Van Raay TJ, Burn TC, Connors TD *et al.* (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.

Vargason JM, Eichman BF, Ho PS. (2000) The extended and eccentric E-DNA structure induced by cytosine methylation or bromination. *Nat Struct Biol* **7**: 758–761.

Vogt N, Rousseau N, Leng M, 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.

Wang AH, Quigley GJ, Kolpak FJ *et al.* (1979) Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature* **282**: 680–686.

Wang G, Christensen LA, Vasquez KM. (2006) Z-DNA-forming sequences generate large-scale deletions in mammalian cells. *Proc Natl Acad Sci USA* **103**: 2677–2682.

Wang G, Zhao J, Vasquez KM. (2009) Methods to determine DNA structural alterations and genetic instability. *Methods* **48**: 54–62.

Wang L, Sommer M, Rajamani J, Arvin AM. (2009) Regulation of the ORF61 promoter and ORF61 functions in varicella-zoster virus replication and pathogenesis. *J Virol* **83**: 7560–7572.

Wang Y, Patel DJ. (1993) Solution structure of a parallel-stranded G-quadruplex DNA. *J Mol Biol* **234**: 1171–1183.

Warne SE, deHaseth PL. (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.

Waterston RH, Lindblad-Toh K, Birney E *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520–562.

Wells RD. (2007) Non-B DNA conformations, mutagenesis and disease. *Trends Biochem Sci* **32**: 271–278.

Yagil G. (2004) The over-representation of binary DNA tracts in seven sequenced chromosomes. *BMC Genomics* **5**: 19.