Evolutionary Expansion of Structurally Complex DNA Sequences

STEVEN S. SMITH

City of Hope, 1500 E. Duarte Rd., Duarte, CA 91010, U.S.A.

Abstract. *The observed number per base pair (i.e. the frequency) of $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$ motifs has increased rapidly in the eumetazoa for which complete genomic sequences are available. This increase appears to be under positive selective pressure since it exceeds the frequency expected for a random sequence genome in every case. Since the motif is capable of forming several non-B DNA structures including quadruplexes, triplexes and hairpins, the expansion has been enabled by the presence of systems capable of suppressing non-B DNA conformations during normal replication and repair and by the emergence of proteins that promote the formation of unusual structures at these sites. Positive selection for these motifs suggests that they are not merely associated with their negative effects on genome stability, but may be useful in increasing the number of structural states in nucleic acids that are available for the elaboration of epigenetic states.*

The evolutionary progression among the eumetazoan animals is a progression that has generated increasing developmental complexity. The serial emergence of eumetazoans, bilaterians, protostomes, deuterostomes, chordates, vertebrates and mammals marks a progressive increase in developmental complexity and therefore a progressive increase in the underlying epigenetic potential of the respective genomes. The well-known expansion of genome size in this progression is consistent with evidence for combinatorial models of epigenetic complexity in which multiple inputs from regulatory proteins and regulatory RNAs in expanded genomes modulate transcription of structural and metabolic proteins to enhance epigenetic complexity (1, 2). While the evidence for gene regulatory networks and the attendant requirement for a general genome expansion and a general increase in transcription factors (3-5) is incontrovertible, additional modulating mechanisms are being recognized. For example, DNA methylation patterning in vertebrate genomes has been proposed to have important epigenetic functions in gene regulatory networks (6-8).

Non-B DNA structure potential has also been proposed as a component of epigenetic systems (9-21). Searches of genomic sequences from bacteria and partial sequences of the yeast and human genomes have suggested that the homopurine mirror repeats characteristic of the H-DNA triplex, cruciform and slipped DNA structures (Figure 1) are overrepresented in eukaryotic genomes (22).

Sequence elements defined as $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$ motifs residing on either DNA strand (14) carry the potential for G-quadruplex formation (Figure 1). In addition, they identify a subset of the imperfect homopurine mirror repeats capable of triplex formation (23) and C-strand i-motif formers (17) (Table I). Moreover, slipped and foldback structures are implicit intermediates in the formation of both quadruplex (24) and triplex (25) structures at these motifs. This report describes the prevalence of this motif in the eight eukaryotic genomes for which complete genomic sequences are currently available.

Materials and Methods

Evolutionary distance. No single bioinformatics approach that is based on nucleic acid or protein sequencing linking eukaryotic organisms in an evolutionary hierarchy has been universally adopted (26). When comparing sequenced eukaryotic genomes one is often tempted to employ a modern version of the Aristotelian concept of graded scale of existence (*i.e.* the *scala naturae* or the Great Chain of Being) as refined by Plotinus (27). Certainly, roundworms are less highly evolved than birds or mice, but quantifying the evolutionary distance between them based on genomic parameters is difficult. A variety of molecular clocks and paleonotlogical evidence (1, 28-30) tend to provide measures of evolutionary distance, and the molecular and paleontological time scales now tend to agree on the timing of the major evolutionary divergences (26, 31). In this report, the paleontological view based primarily on the fossil record was adopted as the measure of evolutionary distance (32-34).

Genomic searches, background data gathering. Genomic sequences for the different organisms were obtained through links provided by the NCBI website (http://www.ncbi.nlm.nih.gov/ sites/entrez?db=genome).

Correspondence to: Steven S. Smith, City of Hope, 1500 E. Duarte Rd., Duarte, CA 91010, U.S.A. Tel: +1 626 3598111, Fax: +1 626 3012568774, e-mail: ssmith@coh.org

Key Words: Triplex, i-motif, Quadruplex, non-B DNA stucture, epigenetics, evolution, scaling developmental complexity.

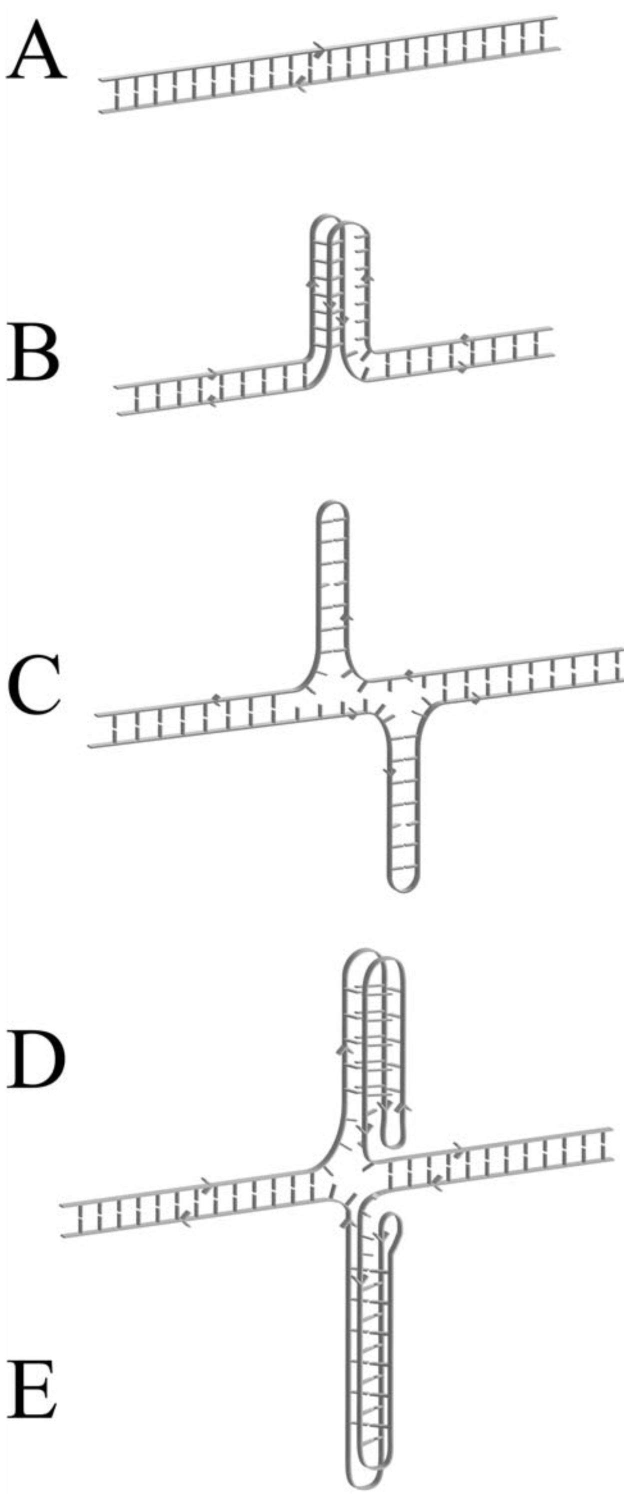

Figure 1. *Conformation space available at the $G_{3+}N_{1-7}G_{3+}$N1-7G3+$N_{1-7}G_{3+}$ motif. Ribbon drawings of canonical non-B-DNA structures formed at $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$ are depicted: (A) B-DNA, (B) an intramolecular triplex linked by base triplet formation, (C) an asymmetrical hairpin linked by base pairing, (D) an intramolecular G-quadruplex linked by Hoogsteen paired tetrads, and (E) an intramolecular C-quadruplex (i-motif) linked by intercalated C:C+ pairs.*

Human sequence (*Homo sapiens*, reference build, release 2.1, build 36.3) and the mouse sequence (*Mus musculus*, build 36.1) were obtained from the NCBI site. Chicken sequence (*Gallus gallus*, release 2.1, build 2.1) was obtained from the WUSTL website (http://genome.wustl.edu/ genomes/view/gallus_gallus/). Zebrafish (*Danio rerio*) and medaka (*Oryzias latipes*) sequences were obtained from the Ensembl FTP site (http://uswest.ensembl.org/info/ data/ftp/index.html). The Drosophila sequence (*Drosophila melanogaster*, release 5.9) was obtained from Flybase (ftp://ftp. flybase.net/genomes/). *C. elegans* (*Caenorhabditis elegans*, WS170, build 7.1) was obtained from the Sanger project website (http://www.sanger.ac.uk/Projects/C_elegans/ Genomic_Sequence. shtml). To calculate the quantity of each nucleotide, chromosomal sequences were opened using EditPad Pro (Just Great Software, Phuket, Thialand) and a search was made through each of the files. Microsoft Excel (Microsoft Corp., Redmond Washington, USA) was used to create the table with the total genome size and also used to calculate the GC fraction by dividing the number of G+C over G+C+A+T. Unknown bases, labeled N, were not included in this calculation but still counted towards the total genome size.

Data collection

Complete genome searches. All unusual structures counts were placed into Excel, and the frequencies were calculated by dividing the number of found structures by the total genome size, as determined previously. $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$ density was calculated using the Huppert algorithm (35) with the program Quadparser, available from http://www.quadruplex.org/. General quadruplex frequencies were calculated with the Maizels algorithm (36) with the program G4P, available from http:// depts.washington.edu/maizels9/G4calc.php. Similar trends were detected with both algorithms. Data from Quadparser were depicted graphically.

Quadparser was placed in the same folder as the FASTA-formatted sequence files, and the program was executed through the command prompt. Output text files were then individually examined and the total number of unique quadruplexes was counted for each file. The G4P calculator used Microsoft.NET (Microsoft) and sequence files were used as input for the algorithm. The output was the density of regions that contained a quadruplex.

Tandem repeats density was calculated using Tandem Repeats Finder (TRF) (37), available from http://tandem.bu.edu/trf/trf.html. Alignment parameters for match, mismatch, and indel were set at 2, 7, and 7 respectively. Alignment scores above 50 were reported, and the maximum period size was 500. Sequences were opened and scanned by the program. Output files were opened and counts were tallied in Excel.

Inverted repeat density was calculated using Inverted Repeats Finder (IRF) (38), available from http://tandem.bu.edu/irf/ irf.download.html. Default parameters for this program were used. Alignment parameters for match, mismatch, and indel were set at 2, 3, and 5 respectively. Matching and indel probabilities were 0.8 and 0.1, respectively. Alignment scores above 40 were reported, and the maximum period size was 2000. The program was executed through the command prompt and was placed in the same folder as the sequence files. Mirror repeat frequencies were also calculated using IRF.

Sampled genome searches. Because the algorithms for finding Z-DNA and palindromes take much longer to run on long sequences, random segments of the genome were sampled to determine the

Table I. *Biological sequences containing the $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$ and observed structures. Both quadruplex and triplex forms have been reported for this motif. The table lists the forms observed in vitro and the associated reference.*

*Gene name (ref)	Sequence	**Motif present	**DNA quadruplex	**DNA triplex	**RNA quadruplex
MYC (21, 59, 72)	5'-TGGGGAGGGTGGGGAGGGTGGGGAAGG-3'	+	+	–	+
MYC (60)	5'-GGGAGGGGCGCTTATGGGGAGGG-3'	+	–	+	ND
MYC(73)	5'-GGGGAGGGTGGGGAGGGTGGGGAGGGTGGGGAGGGT-3'	+	–	+	ND
VEGF (17, 71)	5'-GGGGCGGGCCGGGGCGGGGTCCCGGCGGGGCGGAG-3'	+	+	–	ND
IDDM2 VNTR (74)	[5'-ACAGGGGTGTGGGG-3']n	+	+	ND	ND
Human telomere (10, 24)	[5'-TTAGGG-3']n	+	–	-	ND
BCL2_mbr (63)	5'-AGGGCAGGAGGGCTCTGGGTGGGTC-3'	+	–	+	ND
PDGFA (75)	5'-GGCGGGGGGGGGGGGGCGGGGGCGGGGGCGGGGGAGGGGCG-3'	+	+	ND	ND
Test_Seq (25)	5'-AAGGGAGAAXGGGGTATAGGGGYAAGAGGGAA-3'	+	–	+	ND

*Name of gene with $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$ motif and reference to structural analysis; **type of structure reported (+/–) or not determined (ND).

density of the entire genome. Fifty sequences totaling 1.5% of the entire genome were sampled. Numbers were generated using a random number generator via the process of converting atmospheric noise to numbers (available from http://www.random.org) and then mapped to the starting locations of the sequences. To verify the accuracy of this approach, results from random samples taken from smaller genomes were compared to those of the entire genome and were found to be within around 3% of the actual results.

Z-Hunt was used to determine the density of Z-DNA in each of the genomes (39), available from http://gac-web.cgrb.oregonstate.edu/zDNA/. Segments of the genome were uploaded to the server, which then processed the data and displayed the output file listing all the potential Z-DNA sequences. Output sequences were copied to a text file and the number of Z-DNA sequences was counted.

Palindromes were counted using the palindrome function in the EMBOSS package mEMBOSS (40), available from http://emboss.sourceforge.net/. Sequences were used as input to mEMBOSS for analysis. The minimum length of the palindrome was set at 15 and the maximum at 100 base pairs, with two possible mismatches allowed in the palindrome. The maximum gap between each of the palindromes was set to be 100. The program generated a text file containing the palindromes, and the total number was tallied in Excel.

Expected density. Expected densities for quadruplex, tandem repeats, inverted repeats, Z-DNA, mirror repeats and palindromes were determined by using a randomly generated sequence (with equal probability for G, C, A, T). Each letter was assigned to a number and the random numbers were generated using the aforementioned random number generator. 10 sequences of 1 million bp were generated, as well as 1 sequence of 10 million bp, for a total of 20 million bp. The sequences were then converted to FASTA format and run through each of the algorithms to determine the density. A normal distribution between the sequences was assumed, and the 99% confidence interval for the density of random sequences was determined from their standard deviation.

Data analysis. All data was input into Mathematica (Wolfram Research Inc., Champaign, IL, U.S.A.) and processed with Mathematica subroutines. Plots were generated with the ListPlot subroutine. The subroutines LinearModelFit or NonlinearModelFit were used to model the data and to obtain fitted parameter confidence intervals and R^2 values. Parameter confidence interval shading was obtained with the MeanPredictionBands subroutine.

Results

Available genomic sequence information allowed us to calculate genome size with a high degree of accuracy. In order to detect evolutionary trends, the data for the sequenced eukaryotic genomes were plotted on a geologic time scale that links the genome to its point of divergence from the main eumetazoan lineage based on paleontological dating from the fossil record (32-34). The data are plotted in Figure 2.

The same genomic sequence information allowed calculation of the G+C content with a high degree of accuracy. The data (Figure 3) suggested a very weak tendency toward increased G+C content in the evolutionary progression of the eumetazoans. Since the frequency of the ($G_{3+}N_{1-11}G_{3+}N_{1-11}G_{3+}N_{1-11}G_{3+}$) motif is expected to increase with increasing G+C content, a baseline expectation was calculated for a random genome of 20Mb with equal frequencies for each of the four nucleotides. This baseline expectation for random sequences was calculated from a direct search of 20Mb of random sequence for the ($G_{3+}N_{1-7}$ $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$) motif. Since the random sequence was 50% G+C, the expectation exceeded that for the highest G+C content in the sequenced genomes studied here (42%).

Corresponding scans of each of the sequenced genomes yielded the data given in Figure 4. The motif was overrepresented relative to the random expectation in each organism and the frequency in number/bp increased steadily as the epigenetic potential of the genome increased.

Interestingly, inverted repeat frequency, tandem repeat frequency and Z-DNA sequence frequency did not show

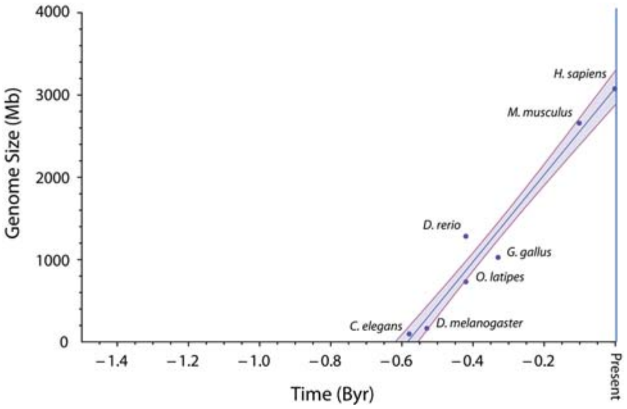

Figure 2. *Genome size for fully-sequenced organisms vs time at divergence node. The line represents the best linear fit to the data (in the least square sense). The shaded area marks the 68% confidence interval on the fitted line. The linear extrapolation to the origin of the eumetazoans is about 0.584 Byr before present, corresponding to the Ediacaran period, well after the origin of the eukaryotes. Sizes are the sum of the sequenced bases for each chromosome in each organism. Similar plots were obtained with 2N male and 2N female genome sizes. Byr: Billion years.*

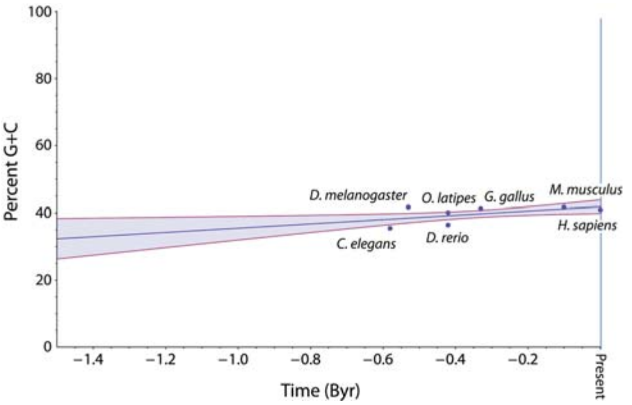

Figure 3. *Percent G+C content of fully-sequenced organisms vs. time at divergence node. The line represents the best linear fit to the data (in the least square sense). The shaded area marks the 90% confidence interval on the fitted line. Byr: Billion years.*

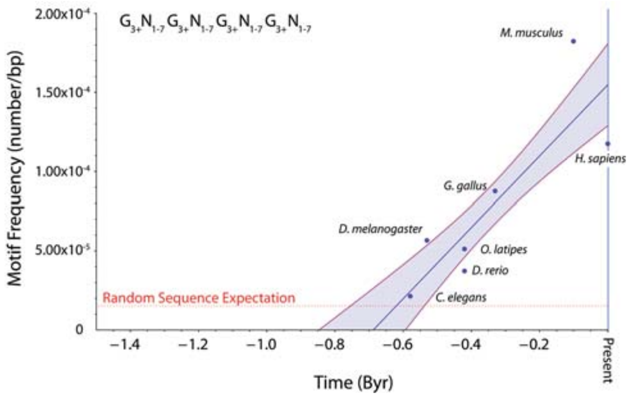

Figure 4. *$G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$ motif frequency vs. time at divergence node. The line represents the best linear fit to the data (in the least square sense). The frequency is the average over male and female genomes for each organism. The shaded area marks the 68% confidence interval on the fitted line. Random sequence expectation was calculated from random sequence scans (red). Eukaryotic genomes appear to have evolved about 1.4-2.5 Byr ago (26). The linear extrapolation for the beginning of the expansion of this motif in the eumetazoa is about 0.688 Byr before present, corresponding to the Ediacaran period, well after the origin of the eukaryotes. Byr: Billion years.*

smooth increases as a function of the time of divergence (Table II). With the exception of Z-DNA frequency, each of these motifs was present in every eukaryotic genome tested at a frequency that exceeded the expectation for a random genome as previously suggested from sampling data (22). However, only the $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$ motif approximated a linear increase as a function of time at divergence (Figure 4).

Discussion

Developmental complexity is synonymous with genetic and epigenetic complexity. Since each stage in a developmental program manipulates the same genomic sequence as new cell types and stably differentiated states are elaborated, developmental complexity results from the combined epigenetic and genetic complexity of an organism. While it is clear that the developmental complexity of a mammal is significantly greater than that of a roundworm, it is difficult to assign a measure of developmental complexity to the respective genomes. In general, the Aristotelian Chain of Being has been used to order organisms qualitatively in demonstrating that genome size (C-value paradox) (41) and gene number (G-value paradox) (42, 43) cannot account for evolutionary complexity. Current best evidence suggests that transcription factor numbers (3), or increasingly complex cis regulatory elements and multiprotein transcription complexes scale with the Chain of Being (4, 44).

Scaling developmental and epigenetic complexity. While the Chain of Being places organisms in an intuitive order, the underlying logic of this order can be seen in palenontology. Palenontology, interpreted through Darwinian evolution, offers the best approach to an unbiased scale of developmental complexity against which candidate processes can be measured. Among the animals, the date at which a given species diverged from the main evolutionary lineage can be taken as its position in the genetic and epigenetic hierarchy. This approach places the Aristotelian Chain of Being on a paleontological time scale that not only orders

Table II. *Average repeat sequence frequencies. Observed frequencies for each of the canonical repeats studied here are given for each organism in occurrences/base-pair. The expectation in occurrences/ base-pair is calculated by searching random DNA sequences for each motif.*

Species	Quadruplex -Triplex	Tandem	Inverted	z-DNA	Palindrome	Total
H. sapiens	1.18×10^{-4}	3.08×10^{-4}	1.16×10^{-3}	1.39×10^{-4}	1.09×10–3	2.82×10^{-3}
M. musculus	1.82×10^{-4}	5.87×10^{-4}	4.22×10^{-4}	2.89×10^{-4}	8.81×10–4	2.36×10^{-3}
G. gallus	8.74×10^{-5}	1.50×10^{-4}	3.40×10^{-5}	9.99×10^{-5}	3.78×10–4	7.49×10^{-4}
O. latipes	5.12×10^{-5}	1.29×10^{-4}	7.06×10^{-5}	1.88×10^{-4}	4.12×10–4	8.51×10^{-4}
D. rerio	3.72×10^{-5}	6.31×10^{-4}	1.10×10^{-3}	3.50×10^{-4}	4.20×10–3	6.31×10^{-3}
D. melanogaster	5.66×10^{-5}	2.80×10^{-4}	3.38×10^{-4}	3.48×10^{-4}	5.06×10–4	1.53×10^{-3}
C. elegans	2.14×10^{-5}	4.21×10^{-4}	6.84×10^{-4}	1.44×10^{-4}	1.67×10–3	2.94×10^{-3}
Expectation	1.56×10^{-5}	4.50×10^{-7}	0.00	5.63×10^{-4}	5.59×10^{-5}	6.35×10^{-4}

genetic and epigenetic complexity but also permits genomic analysis of the trends in the emergence of that complexity.

Based on paleontolological evidence (34) and molecular clocks (26, 28, 30) that encompass the major laboratory organisms (34), the genomes of the eukaryotic animals can now be placed on a paleontological time scale with a high degree of confidence.

When the eumetazoan genome sizes studied here are placed on this paleontological time scale, it is clear that these particular genomes scale linearly from about 584 million years ago as one might expect for a lineage thought to have originated at about the time of the Cambrian Explosion. Obviously this linear relationship for genome size holds only for the laboratory species studied here. Additional organisms plotted on this scale would obscure this linear relationship (27): nematodes can have genomes with sizes comparable to mammals. Nevertheless, these organisms can be traced to the divergence points shown in Figure 2 and, thus, provide an evolutionary series of organisms with monotonically increasing genome sizes that permit further analysis.

Given the evidence demonstrating that sequence motifs associated with non-B conformations can undermine genomic stability by promoting mutagenesis (45), dynamic mutation (46-48) and gene rearrangement (24, 49), it is reasonable to conclude that the maintenance of the larger genomes is enabled by the evolution of suppressors of non-B structure like the RecQ (50) helicase family represented in modern yeast by Sgs1 (51). By enabling the genomes to incorporate high conformation space sequence motifs (52) like those associated with the formation of quadruplex, triplex and hairpin structures, the emergence of large genomes is allowed to go forward unhindered. Given neutral selection, the aggregate frequency (namely the number per base pair) at which these motifs should occur in a random DNA sequence is expected to be a constant, independent of genome size. The present analysis showed that Z-DNA is maintained below the random expectation in every organism tested (Table II) suggesting that it is under negative selective pressure. In contrast, the observed frequency of every motif studied except that of Z-DNA fell above the associated random expectation for every organism studied (Table II), suggesting that at least some of these motifs have been the subject to positive selection. This is consistent with the results on the maintenance of cruciforms on Y (53) and the abundance of simple repeats (54) if one assumes that the positively selected palindromes, inverted repeats and tandem repeats form a class with low annealing temperatures.

Of the sequences that occured at frequencies above the respective random expectation, only the $(G_{3+}N_{1-7}G_{3+}N_{1-7}$ $G_{3+}N_{1-7}G_{3+})$ motif exhibited a monotonic increase in frequency on the palenotological scale (Figure 4). Since it is capable of forming quadruplex, triplex and hairpin structures (Figure 1 and Table I), the expanding frequency of this motif can be taken as a measure of the expansion of non-B-DNA forming potential that is linked to the expansion of developmental potential (Figure 4). Importantly, these rather large changes in sequence motif representation occurred without dramatic changes in the overall G+C content. Although the progression encompassed poikilothermic (cold blooded) eukaryotes and homeothermic (warm blooded) metazoans, none are extremophiles (requiring physically extreme conditions). The organisms studied here develop in temperatures that range from 10˚C to 44˚C, thus ranging in G+C content from about 35% to 42%. While the range from 35% to 42% suggests a trend toward a higher density of G-rich motifs in more highly evolved genomes, none of the G+C contents present in sequenced organisms approached the 50% G+C content present in a completely random genome. Clearly, the increase in the frequency of linked G-rich motifs (Figure 4) exceeded the expectation for a random sequence and cannot be due to the small increase in G+C content associated with the evolutionary progression.

Positive selection for the $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$ motif. The question of whether positive selection for this form of non-B structure potential is a reflection of the expansion of the developmental potential of the genome or is a reflection of the accumulation of junk DNA (55) sequences is an important one. While human Alu sequences lack the motif, the consensus sequence of the human L1 retrotransposon contains one copy of the motif near the polyA sequence. Thus at least some of the instances of the motif (as much as a third of the human occurrences) can be attributed to retrotransposition within the L1 family in humans. Although retrotransposition may actually be promoted by the formation of quadruplex, hairpin and triplex structures, it does not appear to be required for transposition, since P-elements in Drosophila lack the motif. Thus, the proliferation of repetitive elements does not appear to account for the observed smooth expansion of the motif frequency with developmental complexity. In short, the observed increase in the frequency of the $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$ motif was more consistent with positive selection for functionality (perhaps *via* retrotransposition) than with neutral selection associated with the accumulation of junk DNA sequences.

Positive selection is also consistent with the appearance of proteins that promote the formation of unusual structures in DNA (*e.g.* the meiotic pairing gene *Hop1* (11) and Nucleolin (56)). Moreover, positive selection is difficult to rationalize in terms of the coding capacity of these sequences for proteins given the redundancy of the genetic code and its capacity to mold codon usage (57). The analysis suggests that sequences capable of quadruplex, hairpin and triplex formation are not merely associated with harmful effects as one may expect from their association with sites of dynamic mutation (46-48, 58) and gene rearrangement (49) but also serve useful developmental functions since they appear to have been under strong positive selective pressure during the emergence of developmental complexity.

Roles for the $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$ motif. *In vitro* studies of the formation of quadruplex DNAs have generally been performed with short oligodeoxynucleotides representing regions of biologically important sequences. A key example is the nuclease hypersensitive element (NHE III_1) located between promoter P0 and promoter P1 in the human *MYC* gene (59). This element has been shown to form an intramolecular quadruplex (21) *in vitro*. However, mutagenesis studies and the presence of an imperfect homopurine mirror repeat present in the sequence suggest that it can also form an intramolecular triplex (60). Consistent with these findings, the region has also been shown to carry a canonical motif ($G_{3+}N_{1-7}G_{3+}N_{1-7}$ $G_{3+}N_{17}$ G_{3+}) characteristic of both quadruplex sequences (14) and the G-rich subset of triplex sequences formed by imperfect homopurine mirror repeats (23).

Several well-studied examples of quadruplex or triplex formation (Table 1) exhibit this motif and thus carry the capacity for the formation of either or both types of non-B DNA structure. Moreover, other similar motifs are known to form quadruplex sequences. For example, motifs such as ($G_{3+}N_{1-11}G_{3+}N_{1-11}G_{3+}N_{1-11}G_{3+}$) with extended loop size are expected to form quadruplex DNA (10, 61). However, longer loops reduce stability (62). In addition, it is well known that the $(GGC)_n$ motif in triplet repeat sequences from the *FMR1* gene has been observed to form quadruplex DNA (47), and is associated with the formation of a spontaneous slippage intermediate on the complementary C-rich strand (46, 48, 58).

Although quadruplex formation in oligodeoxynucleotides is not often studied in the presence of both complementary strands, quadruplexes have been observed to form *in vitro* in the presence of their complementary strands (58). Moreover, the extended loop size (Lamparska-Kupsik and Smith, unpublished) or the presence of the complementary strand (63) can result in the stable trimolecular triple helices. The full structural complexity of these motifs is depicted schematically in Figure 1.

A significant number of proposals for the function of these sequence motifs are consistent with the expectation that they should expand in frequency in concert with developmental complexity (35, 64). Since the intermolecular quadruplex is a molecular mimic of the synaptonemal alignment of homologues during meiosis, an active role in meiosis has been suggested (10). Support for this proposal has been adduced from the KEM1/SEP1 nuclease system (65) in yeast which can block meiosis at pachytene when mutated (66), and from the presence of quadruplex binding proteins of similar function in human cells (67). These proposals are also consistent with the expansion of the $G_{3+}N_{1-7}G_{3+}N_{1-7}$ $G_{3+}N_{1-7}G_{3+}$ motif, because homologous pairing is expected to require a higher density of interstrand links as chromosome size increases.

Gene expression. Proposals for a role in the control of gene expression through the formation of intramolecular triplex (60) or quadruplex (21) at promoter sequences in the human *c-MYC* (59), *KRAS* (68) and *c-KIT* (69) genes are consistent with the presence of these elements in the 5' UTRs of a significant number of human genes (14) and with the capacity of *Hop1* (11) and Nucleolin (56) to induce these structures in DNA.

These proposals require the generation of the structure in DNA, however, it is also possible that the RNA transcript would carry the folded structure encoded by a B-DNA sequence as discussed by Huppert *et al.* (14). For example, transcription from start site P0 at the human *MYC* gene would produce a 5' UTR capable of intramolecular quadruplex or triplex formation, while transcription from start site P1 would not. Thus, the presence of these motifs in 5' and 3' UTRs is consistent with a role in RNA processing (14) and the mechanism of action of non-coding RNAs.

The combinatorial limit. Based on Ohno's original suggestion (55) that mutation rate limits the total number of genes to between 15,000 and 20,000 genes per genome, it is tempting to speculate that once this combinatorial limit (2) is reached (apparently already at the roundworms), additional epigenetic mechanisms involving phenomena like unusual DNA structure formation and DNA methylation must come into play.

Structural complexity and epigenetic potential. A structurally complex motif like the ($G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$) motif necessarily generates a dramatic increase in the epigenetic potential of the sequence at a given site. A simple B-DNA sequence offers one DNA conformation that can be recognized by a protein or nucleic acid modulator. A ($G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$) motif can present six distinct DNA conformations that can be uniquely recognized by protein or nucleic acid modulators (Figure 1). Thus, even the pair-wise interaction potential of these motifs as sites of protein or nucleic acid recognition will be increased by fifteen times (=6!/2!(6-2)!) the number of motifs present in the genome. Thus, the increase in ($G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$) density links the positive selection for these non-B conformations with the emergence of epigenetic complexity through a clear enhancement of information content. In other words, the different intermolecular and intramolecular conformations that can be adopted or encoded in RNA by a single DNA sequence form different epigenetic signals that effectively increase the epigenetic information content of the genome (12-14, 52, 59, 70, 71) by increasing the number of protein and nucleic acid binding sites available from a single sequence. The data reported here suggest that eumetazoans have enhanced their developmental and epigenetic potential by selectively incorporating structurally complex motifs in their genomes in spite of the potential that these motifs have for chromosomal damage (50, 63).

In conclusion, similar to transcription factors and cis-regulatory elements, non-B DNA forming motifs like the $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$ motif appeared to scale with organismic complexity when placed on a paleontological scale that quantifies the Aristotelian Chain of Being. The observed steady increase in the frequency of these motifs associated with the emergence of complexity was consistent with the proposed roles for this motif in amplifying the number of structural states in nucleic acids that are available for molecular recognition during the elaboration of epigenetic states.

Acknowledgements

Yu-Loung Chang and Jarrod Clark are thanked for their help in compiling data and illustrating figures.

References

1 Britten RJ and Davidson EH: Gene Regulation for Higher Cells: A theory. Science *165*: 349-357, 1969.

2 Davidson EH and Levine MS: Properties of developmental gene regulatory networks. Proc Natl Acad Sci USA *105*: 20063-20066, 2008.

3 van Nimwegen E: Scaling laws in the functional content of genomes. Trends Genet *19*: 479-484, 2003.

4 Levine M and Tjian R: Transcription regulation and animal diversity. Nature *424*: 147-151, 2003.

5 Chen K and Rajewsky N: The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet *8*: 93-103, 2007.

6 Jones PA and Baylin SB: The epigenomics of cancer. Cell *128*: 683-692, 2007.

7 Miranda TB and Jones PA: DNA methylation: the nuts and bolts of repression. J Cell Physiol *213*: 384-390, 2007.

8 Schaefer CB, Ooi SK, Bestor TH and Bourc'his D: Epigenetic decisions in mammalian germ cells. Science *316*: 398-399, 2007.

9 Smith SS: DNA methylation in eukaryotic chromosome stability. Mol Carcinog *4*: 91-92, 1991.

10 Sen D and Gilbert W: Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. Nature *334*: 364-366, 1988.

11 Muniyappa K, Anuradha S and Byers B: Yeast meiosis-specific protein Hop1 binds to G4 DNA and promotes its formation. Mol Cell Biol *20*: 1361-1369, 2000.

12 Hershman SG, Chen Q, Lee JY, Kozak ML, Peng Y, Wang L-S, F and Brad Johnson FB: Genomic distribution and functional analyses of potential G-quadruplex-forming sequences in Saccharomyces cerevisiae. Nucleic Acids Res *36*: 144-156, 2008.

13 Goñi JP, Vaquerizas JM, Dopazo J and Orozco M: Exploring the reasons for the large density of triplex-forming oligonucleotide target sequences in the human regulatory regions. BMC Genomics *7*: 63-73, 2006.

14 Huppert JL, Bugaut A, Kumari S and Balasubramanian S: G-quadruplexes: the beginning and end of UTRs. Nucleic Acids Res *36*: 6260-6268, 2008.

15 Gonzalez V, Guo K, Hurley L and Sun D: Identification and characterization of nucleolin as a c-myc G-quadruplex-binding protein. J Biol Chem *284*: 23622-23635, 2009.

16 Grand CL, Powell TJ, Nagle RB, Bearss DJ, Tye D, Gleason-Guzman M and Hurley LH: Mutations in the G-quadruplex silencer element and their relationship to c-MYC overexpression, NM23 repression, and therapeutic rescue. Proc Natl Acad Sci USA *102*: 516, 2005.

17 Guo K, Gokhale V, Hurley LH and Sun D: Intramolecularly folded G-quadruplex and i-motif structures in the proximal promoter of the vascular endothelial growth factor gene. Nucleic Acids Res *36*: 4598-4608, 2008.

18 Palumbo SL, Ebbinghaus SW and Hurley LH: Formation of a unique end-to-end stacked pair of G-quadruplexes in the *hTERT* core promoter with implications for inhibition of telomerase by G-quadruplex-interactive ligands. J Am Chem Soc *131*: 10878-10891, 2009.

19 Palumbo SL, Memmott RM, Uribe DJ, Krotova-Khan Y, Hurley LH and Ebbinghaus SW: A novel G-quadruplex-forming GGA repeat region in the *c-myb* promoter is a critical regulator of promoter activity. Nucleic Acids Res *36*: 1755-1769, 2008.

20 Rangan A, Fedoroff OY and Hurley LH: Induction of duplex to G-quadruplex transition in the *c-myc* promoter region by a small molecule. J Biol Chem *276*: 4640-4646, 2001.

21 Yang D and Hurley LH: Structure of the biologically relevant G-quadruplex in the *c-MYC* promoter. Nucleosides Nucleotides Nucleic Acids *25*: 951-968, 2006.

22 Cox R and Mirkin SM: Characteristic enrichment of DNA repeats in different genomes. Proc Natl Acad Sci USA *94*: 5237-5242, 1997.

23 Frank-Kamenetskii MD and Mirkin SM: Triplex DNA structures. Annu Rev Biochem *64*: 65-95, 1995.

24 Sundquist WI and Klug A: Telomeric DNA dimerizes by formation of guanine tetrads between hairpin loops. Nature *342*: 825-829, 1989.

25 Mirkin SM, Lyamichev VI, Drushlyak KN, Dobrynin VN, Filippov SA and Frank-Kamenetskii MD: DNA H form requires a homopurine-homopyrimidine mirror repeat. Nature *330*: 495-497, 1987.

26 Hedges SB, Blair JE, Venturi ML and Shoe JL: A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evol Biol *4*: 2-9, 2004.

27 Gregory TR: Macroevolution, hierarchy theory, and the C-value enigma. Paleobiology Paleobiology *30*: 179-202, 2004.

28 Doolittle RF, Feng D-F, Tsang S, Cho G and Little E: Determining divergence times of the major kingdoms of living organisms with a protein clock. Science *271*: 470-477, 1996.

29 Wray GA: Dating branches on the tree of life using DNA. Genome Biol 3: REVIEWS0001, 2002.

30 Roger AJ, L.A. H. The origin and diversification of eukaryotes: problems with molecular phylogenetics and molecular clock estimation. Phil Trans R Soc B *361*: 1039-1054, 2006.

31 Benton MJ and Ayala FJ: Dating the Tree of Life. Science 300: 1698-1700, 2003.

32 Benton MJ and Donoghue PCJ: Paleontological evidence to date the Tree of Life. Mol Biol Evol *24*: 26-53, 2006.

33 Benton MJ and Donoghue PCJ: Paleontological evidence to date the Tree of Life. (*Erratum*). Mol Biol Evol *24*: 26-53, 2007.

34 Donoghue PCJ and Benton MJ: Rocks and clocks: calibrating the Tree of Life using fossils and molecules. Trends Ecol Evol *22*: 424-431, 2007.

35 Huppert JL and Balasubramanian S: Prevalence of quadruplexes in the human genome. Nucleic Acids Res *33*: 2908-2916, 2005.

36 Eddy J and Maizels N: Gene function correlates with potential for G4 DNA formation in the human genome. Nucleic Acids Res *34*: 3887-3896, 2006.

37 Benson G: Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res *27*: 573-580, 1999.

38 Warburton PE, Giordano J, Cheung F, Gelfand Y and Benson G: Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. Genome Res *14*: 1861-1869, 2004.

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

40 Rice P, Longden I and Bleasby A: EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet *16*: 276-277, 2000.

41 Thomas CA Jr: The genetic organization of chromosomes. Annu Rev Genet *5*: 237-256, 1971.

42 Hodgkin J: What does a worm want with 20,000 genes? Genome Biol 2: COMMENT2008, 2001.

43 Hahn MW and Wray GA: The G-value paradox. Evol Dev *4*: 73-75, 2002.

44 Davidson EH: The Regulatory Genome. San Deigo: Academic Press, 2006.

45 Wang G and Vasquez KM: Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells. Proc Natl Acad Sci USA *101*: 13448-13453, 2004.

46 Chen X, Mariappan SV, Catasti P, Ratliff R, Moyzis RK, Laayoun A, Smith SS, Bradbury EM and Gupta G: Hairpins are formed by the single DNA strands of the fragile X triplet repeats: structure and biological implications. Proc Natl Acad Sci USA *92*: 5199-5203, 1995.

47 Fry M, Loeb LA. The fragile X syndrome d(CGG)n nucleotide repeats form a stable tetrahelical structure. Proc Natl Acad Sci USA *91*: 4950-4954, 1994.

48 Gacy AM, Goellner G, Juranic N, Macura S and McMurray CT: Trinucleotide repeats that expand in human disease form hairpin structures *in vitro*. Cell *81*: 533-540, 1995.

49 Raghavan SC, Swanson PC, Wu X, Hsieh CL and Lieber MR: A non-B-DNA structure at the Bcl-2 major breakpoint region is cleaved by the RAG complex. Nature *428*: 88-93, 2004.

50 Johnson JE, Cao K, Ryvkin P, Wang LS and Johnson FB: Altered gene expression in the Werner and Bloom syndromes is associated with sequences having G-quadruplex forming potential. Nucleic Acids Res, *38(4)*: 1114-1122, 2010

51 Sun H, Bennett RJ and Maizels N: The *Saccharomyces cerevisiae* sgs1 helicase efficiently unwinds G-G paired DNAs. Nucleic Acids Res *27*: 1978-1984, 1999.

52 Smith SS and Crocitto L: DNA methylation in eukaryotic chromosome stability revisited: DNA methyltransferase in the management of DNA conformation space. Mol Carcinog *26*: 1-9, 1999.

53 Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK and Page DC: Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. Nature *423*: 873-876, 2003.

54 Bacolla A, Larson JE, Collins JR, Li J, Milosavljevic A, Stenson PD, Cooper DN and Wells RD: Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties. Genome Res *18*: 1545-1553, 2008.

55 Ohno S: So Much “Junk DNA” in our Genome. Brookhaven Symposia in Bioogy *23*: 366-370, 1972.

56 Hanakahi LA, Sun H and Maizels N: High affinity interactions of nucleolin with G-G-paired rDNA. J Biol Chem *274*: 15908-15912, 1999.

57 Smith SS: Species-specific differences in tumorigenesis and senescence. Trends Genet *10*: 305-306, 1994.

58 Smith SS, Laayoun A, Lingeman RG, Baker DJ and Riley J: Hypermethylation of telomere-like foldbacks at codon 12 of the human *c-Ha-ras* gene and the trinucleotide repeat of the *FMR-1* gene of fragile X. J Mol Biol *243*: 143-151, 1994.

59 Siddiqui-Jain A, Grand CL, Bearss DJ and Hurley LH: Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress *c-MYC* transcription. Proc Natl Acad Sci USA *99*: 11593-11598, 2002.

60 Belotserkovskii BP, De Silva E, Tornaletti S, Wang G, Vasquez KM and Hanawalt PC: A triplex-forming sequence from the human *c-MYC* promoter interferes with DNA transcription. J Biol Chem *282*: 32433-32441, 2007.

61 Clark J and Smith SS: Secondary structure at a hot spot for DNA methylation in DNA from human breast cancers. Cancer Genomics Proteomics *5*: 241-251, 2008.

62 Hazel P, Huppert J, Balasubramanian S and Neidle S: Loop-length-dependent folding of G-quadruplexes. J Am Chem Soc *126*: 16405-16415, 2004.

63 Raghavan SC, Chastain P, Lee JS, Hegde BG, Houston S, Langen R, Hsieh CL, Haworth IS and Lieber MR: Evidence for a triplex DNA conformation at the *bcl-2* major breakpoint region of the t(14;18) translocation. J Biol Chem *280*: 22749-22760, 2005.

64 Todd AK, Johnston M and Neidle S: Highly prevalent putative quadruplex sequence motifs in human DNA. Nucleic Acids Res 33: 2901-2907, 2005.

65 Liu Z, Frantz JD, Gilbert W and Tye BK: Identification and characterization of a nuclease activity specific for G4 tetra-stranded DNA. Proc Natl Acad Sci USA *90*: 3157-3161, 1993.

66 Tishkoff DX, Rockmill B, Roeder GS and Kolodner RD: The sep1 mutant of *Saccharomyces cerevisiae* arrests in pachytene and is deficient in meiotic recombination. Genetics *139*: 495-509, 1995.

67 Creacy SD, Routh ED, Iwamoto F, Nagamine Y, Akman SA and Vaughn JP: G4 resolvase 1 binds both DNA and RNA tetramolecular quadruplex with high affinity and is the major source of tetramolecular quadruplex G4-DNA and G4-RNA resolving activity in HeLa cell lysates. J Biol Chem *283*: 34626-34634, 2008.

68 Cogoi S, Quadrifoglio F and Xodo LE: G-rich oligonucleotide inhibits the binding of a nuclear protein to the *Ki-ras* promoter and strongly reduces cell growth in human carcinoma pancreatic cells. Biochemistry *43*: 2512-2523, 2004.

69 Bejugam M, Sewitz S, Shirude PS, Rodriguez R, Shahid R and Balasubramanian S: Trisubstituted isoalloxazines as a new class of G-quadruplex binding ligands: small molecule regulation of *c-kit* oncogene expression. J Am Chem Soc *129*: 12926-12927, 2007.

70 Catasti P, Chen X, Moyzis RK, Bradbury EM and Gupta G: Structure-function correlations of the insulin-linked polymorphic region. J Mol Biol *264*: 534-545, 1996.

71 Sun D, Guo K, Rusche JJ and Hurley LH: Facilitation of a structural transition in the polypurine/polypyrimidine tract within the proximal promoter region of the human *VEGF* gene by the presence of potassium and G-quadruplex-interactive agents. Nucleic Acids Res *33*: 6070-6080, 2005.

72 Simonsson T, Pecinka P and Kubista M: DNA tetraplex formation in the control region of c-myc. Nucleic Acids Res *26*: 1167-1172, 1998.

73 Firulli AB, Maibenco DC and Kinniburgh AJ: Triplex forming ability of a *c-myc* promoter element predicts promoter strength. Arch Biochem Biophys *310*: 236-242, 1994.

74 Lew A, Rutter WJ and Kennedy GC: Unusual DNA structure of the diabetes susceptibility locus IDDM2 and its effect on transcription by the insulin promoter factor Pur-1/MAZ. Proc Natl Acad Sci USA *97*: 12508-12512, 2000.

75 Qin Y, Rezler EM, Gokhale V, Sun D and Hurley LH: Characterization of the G-quadruplexes in the duplex nuclease hypersensitive element of the *PDGF-A* promoter and modulation of *PDGF-A* promoter activity by TMPyP4. Nucleic Acids Res *35*: 7698-7713, 2007.

Received March 20, 2010
Revised May 12, 2010
Accepted May 14, 2010