Occurrence of potential cruciform and H-DNA forming sequences in genomic DNA

Gary P. Schroth* and P. Shing Ho

Department of Biochemistry and Biophysics, ALS 2011, Oregon State University, Corvallis, OR 97331, USA

Received February 1, 1995; Revised and Accepted April 3, 1995

ABSTRACT

We have used computer-assisted methods to search large amounts of the human, yeast and *Escherichia coli* genomes for inverted repeat (IR) and mirror repeat (MR) DNA sequence patterns. In highly supercoiled DNA some IRs can form cruciforms, while some MRs can form intramolecular triplexes, or H-DNA. We find that total IR and MR sequences are highly enriched in both eukaryotic genomes. In *E.coli*, however, only total IRs are enriched, while total MRs only occur as frequently as in random sequence DNA. We then used a set of experimentally derived criteria to predict which of the total IRs and MRs are most likely to form cruciforms or H-DNA in supercoiled DNA. We show that strong cruciform forming sequences occur at a relatively high frequency in yeast (1/19 700 bp) and humans (1/41 800 bp), but that H-DNA forming sequences are abundant only in humans (1/49 400 bp). Strong cruciform and H-DNA forming sequences are not abundant in the *E.coli* genome. These results suggest that cruciforms and H-DNA may have a functional role in eukaryotes, but probably not prokaryotes.

INTRODUCTION

The polymorphic nature of DNA structure has been intensely studied during the last 15 years, resulting in a vast literature on the biochemical properties of many non-B-DNA conformations, including left-handed Z-DNA, cruciforms and triple-stranded H-DNA (reviewed in 1–6). As a result of this large body of work we have developed a strong fundamental understanding of the requirements for non-B-DNA structure formation. In general biochemical terms, non-B-DNA conformations are stabilized by high levels of negative supercoiling and sometimes by the presence of multivalent cations (6). In genetic terms non-B-DNA conformations are most favored in specific DNA sequence patterns or motifs, some of which are very simple repeating sequences (1). In this paper we study two classes of DNA sequence patterns, inverted repeats (IRs) and mirror repeats (MRs), and their occurrence in genomic DNA. These sequence patterns are interesting because they are prerequisites for the formation of two non-B-DNA structures: the formation of cruciforms is most favored in DNA sequences with IR symmetry (4,7,8) and the formation of intramolecular triplex DNA structures, or H-DNA, is most favored in homopurine sequences with MR symmetry (5,9–12). The symmetry of IR and MR DNA sequences and their relationship to the structures of cruciforms and H-DNA are introduced in Figure 1.

Many of the simple repetitive sequences found in genomic DNA have the potential to form non-B-DNA structures. The sequence poly(dT–dC/dA–dG), which is abundant in eukaryotic genomes (13,19), will readily form triple-stranded H-DNA structures *in vitro* (10,12). Other abundant simple sequences in eukaryotic genomes include poly(dC–dA/dT–dG) (14), which can form Z-DNA (15), and poly(dT–dA) repeats (16,17) which can easily convert to cruciforms (18). Many previous studies have shown these repeating sequences occur relatively frequently in eukaryotic genomes, but are virtually absent from prokaryotic genomes (13,14,16,17,19). Although a direct functional role for simple repeating sequences in any cellular process has yet to be conclusively established, many reports have noted that they are found near functionally interesting regions of the genome, such as promoters or sites of recombination (reviewed in 1–6,20,21).

In order to gain a better understanding of the possible biological significance of non-B-DNA structures, it is important to compare the frequency of occurrence of these structures in the genomes of various organisms. Since the formation of non-B-DNA structures is a dynamic process, sensitive to the levels of negative supercoiling, they have been difficult to detect directly in genomic DNA (6). However, we can quantitate the number of naturally occurring DNA sequences which seem to best fit the current biochemical rules for forming such structures.

In this paper the occurrence of IR and MR sequences are compared across the human, yeast and *Escherichia coli* genomes. Although IRs or MRs are not generally classified as repetitive DNA sequence elements *per se*, we find that they tend to follow the same phylogenetic distributions as other repeated sequences. This is especially true for MRs, which are not enriched at all in *E.coli* DNA, but are highly enriched in yeast and even more abundant in human DNA. In contrast, we find that total inverted repeats are abundant in all three genomes. We then use a set of empirical rules to aid us in predicting which of the IRs and MRs have the highest potential to form H-DNA or cruciforms. We show that cruciform forming sequences occur frequently in yeast and humans and that potential H-DNA forming sequences occur at higher levels than expected only in human DNA. These results suggest that non-B-DNA forming sequences are relatively rare in prokaryotes but are abundant in eukaryotic genomes.

* To whom correspondence should be addressed

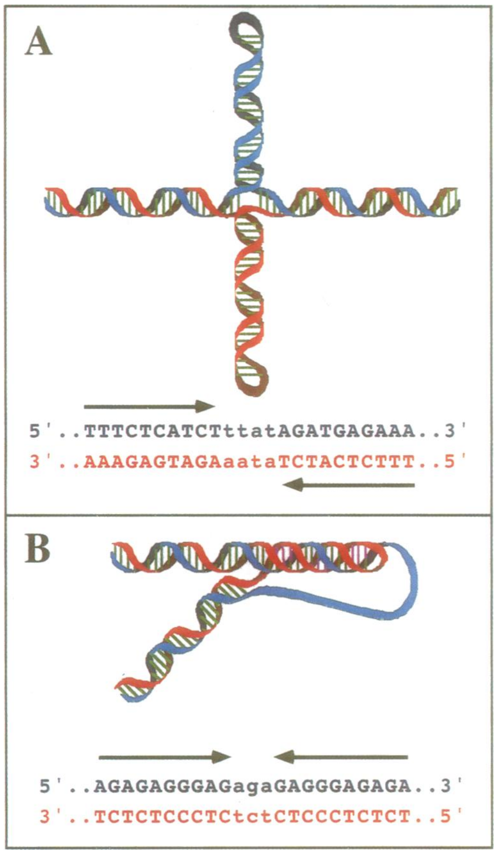
A
5'..TTTCTCATCTttatAGATGAGAAA..3'
3'..AAAGAGTAGAaataTCTACTCTTT..5'
B
5'..AGAGAGGGAGagaGAGGGAGAGA..3'
3'..TCTCTCCCTCtctCTCCCTCTCT..5'

Figure 1. Structure and DNA sequence requirements for cruciform and H-DNA conformations. Opposing strands of the standard duplex form of DNA are shown in red and blue. Standard Watson–Crick base pairing is represented in green. (**A**) A model of a cruciform structure. An IR from the sequence of the human serum albumin gene is also shown (GenBank accession no. M12523). The part of the sequence which is palindromic is capitalized, the spacer between the two palindromes is in lower case letters. The arrows highlight the 2-fold symmetry of the IR. (**B**) A model of an intramolecular DNA triplex, or H-DNA, structure. In this model the pyrimidine-rich red strand from one half of the mirror repeat folds back and forms Hoogsteen base pairs (represented in violet) with the other half of the repeat. A MR from the sequence of the human gastric ATPase gene is also shown (GenBank accession no. J05451). The sequence which is part of the MR is capitalized, the spacer between the repeats is in lower case letters. The arrows highlight the mirror symmetry of the sequence.

MATERIALS AND METHODS

Pattern searching and analysis programs

Both of the programs developed for these studies use similar search strategies which take advantage of the symmetry found in IRs and MRs (see Fig. 1). These programs scan each base within a given DNA sequence to determine whether it is a possible center of an IR or MR repeat. From the point of symmetry the program searches both forwards and backwards in the sequence for DNA bases that are either complementary (for IRs) or identical (for MRs). The program allows a certain number of bases (3–6 in this study) to act as a spacer that does not need to conform to the IR or MR symmetry. Since the spacer length between the two halves of the repeat can be an odd number of base pairs, the program also considers that the point of symmetry in the repeat could be located between base pairs. Thus these pattern matching programs search for repeats which must start at between 1.5 and 3.0 bp on both sides of the putative center of the repeat. If an IR or MR pattern is not observed within this range, the program continues to move along the DNA sequence to the next potential point of symmetry or until the entire sequence has been analyzed. The minimum length of the MR or IR sequences studied was 8 bp (for each half of the repeat), but there was no maximum size limit. Analysis of genomic DNA sequences was carried out on a Silicon Graphics Iris workstation.

Genomic sequences and datasets

The human, yeast and *E.coli* DNA sequences used in these studies were obtained from GenBank using the Intelligenetics software package. The human sequences consist of 157 complete individual genes totaling 1 086 110 bp. This set of human DNA contains overall 49.7% (dA–dT) bp. The genes selected were all genomic sequences (not cDNAs) and contained all of the exons, introns and a significant amount of the 5′ and 3′ flanking DNA. This human dataset contains many of the same sequences used in a previous study of Z-DNA in the human genome (22).

The yeast and *E.coli* DNA sequences used were each contiguous genomic sequences. The yeast DNA was the complete sequence of chromosome III from *Saccharomyces cerevisiae* (23), which contains 182 open reading frames and is 315 357 bp long (GenBank accession no. X59720). The yeast DNA has 61% (dA–dT) bp overall. The *E.coli* sequence consisted of the complete sequence from the 81.5–89.2 minute region of the *E.coli* chromosome (24–26), which contains 350 open reading frames and is a total of 324 146 bp in length (accession nos L10328, M87049 and L19201). The overall content of (dA–dT) bp of the *E.coli* DNA was 48%.

RESULTS

General rationale

The goal of this work was to use computer-assisted methods to study the occurrence of cruciform and H-DNA forming sequences in DNA from humans, yeast and *E.coli* (see Figure 1 for a description of the structures and examples of the sequence motifs discussed in this paper). First we studied the occurrence of all IRs and MRs in the three genomes, since these sequence patterns are a necessary, although not sufficient, prerequisite for forming cruciforms and triple-stranded H-DNA structures. Once the occurrence of all IRs and MRs had been determined, we used current biochemical data on cruciforms and H-DNA to aid in identifying the repeats with the highest potential for forming non-B-DNA structures. For example, the stability of both cruciforms and H-DNA is dependent upon the length of the repeat, with longer repeats having greater stability as non-B-DNA structures (27–30). The length of the spacer DNA between the two halves of the repeat is another important factor in stability, with a 3–6 bp loop region being most favored in both conformations (31,32). These concepts form the basic framework for these studies, in which we search genomic DNA for IRs and MRs of ≥8 bp in length which are separated by spacers of 3–6 bp.

Occurrence of IR sequence patterns in human, yeast and *E.coli* genomic DNA

Inverted repeats are palindromic DNA sequences with 2-fold symmetry, as shown in Figure 1. These repeats have the potential

Table 1. Occurrence of IR repeat patterns in human, yeast and *E.coli* genomic DNA[a]

	Occurrences per Million base pairs				Number of Occurrences		
IR Size (Each Half)	Expected Occurrence	Human	Yeast	E. coli	Human	Yeast	E. coli
≥ 8 bps	81	178	184	214	193	58	69
≥ 9 bps	20	52	82	93	57	26	30
≥ 10 bps	5	21	41	59	23	13	19
≥ 11 bps	1	9	13	40	10	4	13
≥ 12 bps	0.25	5	3	25	6	1	8
≥ 12 w/mismatch[b]	1	7	16	12	8	3	4

Numbers shown are for MRs found which have 3–6 bp spacers between the repeated sequences.

[a] This table gives the numbers of IRs found with given repeat sizes for each of the three data sets. On the right, the actual number of IRs found in the three data sets are given; on the left, the numbers have been normalized to 10^6 bp for comparison. Also shown is the calculated number of IRs of each size expected in 10^6 bp of random sequence DNA (25% each of A, C, G and T).

[b] The last row in the table gives the number of additional IRs found in each data set with an overall length of ≥12, if we allow the repeat sequences to have a single RY mismatch (i.e. either a T:G or C:A mismatched bp) every 10 bp.

to form both cruciforms in double-stranded DNA and/or hairpin structures in single-stranded DNA or RNA. We searched the human, yeast and *E.coli* genomic datasets for IRs which have a minimum length of 8 bases on each half of the repeat and a spacer of 3–6 bases. The results of these searches are listed in Table 1 as the total number of IRs found of each size for each dataset. The expected occurrence of IRs for each genome type are then normalized to 10^6 bp to allow for direct comparison of the three different data sets. For example, the right-hand side of Table 1 shows that a total of 26 IRs with a length ≥9 bp were found in the yeast dataset (total length 315 357 bp). This number is extrapolated on the left-hand side of the table to a value of 82, which is the number of IRs of this length that would be expected to be found in a total of 10^6 bp of yeast DNA. In Table 1 we also compare the number of IRs found in the three genomes with the calculated occurrence of IRs of these lengths in 10^6 bp of random sequence DNA (containing equal amounts of A, C, G and T bases).

The results show that IRs are more abundant in all three genomes than in random sequence DNA. The *E.coli* genome has the most IRs of the three and is especially enriched in longer IRs of 11 or 12 bp in length. In human DNA the occurrence of all IRs ≥8 bp ($178/10^6$ bp) is 2.2-fold higher than is found in a random sequence DNA ($81/10^6$ bp). The yeast and *E.coli* genomes are 2.3- and 2.6-fold enriched in IRs of ≥8 bp, when compared with random DNA. In terms of frequency of occurrence, IRs of ≥8 bp in length occur once every 5600 bp in humans, once every 5400 bp in yeast and once every 4700 bp in *E.coli*, whereas these sequences are found only about once every 12 300 bp in random sequence DNA.

Eukaryotic IRs are very (dA–dT)-rich, while prokaryotic IRs are relatively (dG–dC)-rich

Although the overall frequency of occurrence of IRs in the three genomes is similar (Table 1), we noticed distinct differences between the IRs from the two eukaryotes compared with prokaryotic *E.coli* DNA. As Figure 2 shows, IRs from yeast and human DNA are very (dA–dT)-rich, while the IRs from *E.coli* are relatively (dG–dC)-rich. In fact, there seems to be a demarcation at ~60% (dA–dT) between the eukaryotic and prokaryotic IRs (highlighted by the arrow in Fig. 2). As shown in Figure 2, 91% of the *E.coli* IRs are <60% (dA–dT)-rich, while 91% of all the eukaryotic IRs are >60% (dA–dT)-rich. The percent (dA–dT) richness does not correlate with the (dA–dT) composition of each genome, since the overall percent (dA–dT) of the human and *E.coli* genomes are similar and only the yeast genome is inherently (dA–dT)-rich.

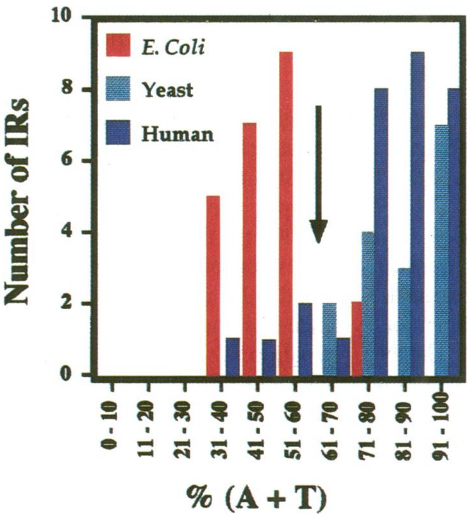

Figure 2. A graph comparing the (dA–dT) content of all IRs (10 bp from the human, yeast and *E.coli* genomes. The number of IRs are plotted against the percentage of (dA–dT) base pairs found in each repeat. The IRs from *E.coli* are shown in red, the yeast IRs are in light blue and the human IRs are in darker blue. The black arrow points to the 60% (dA–dT) point of the graph, which highlights the differences between prokaryotic and eukaryotic IRs.

The percent (dA–dT) richness of an IR is an important criterion in cruciform formation (4,18,27). Studies have shown that (dA–dT)-rich sequences form cruciforms more easily *in vitro* than (dC–dG)-rich sequences (18,33,34), although it is not clear how these observations apply *in vivo*. The first step in the formation of a cruciform involves the local melting of 8–10 bp at the center of the IR. Since IRs with relatively (dA–dT)-rich

Table 2. Occurrence of MR repeat patterns in human, yeast and *E.coli* genomic DNA[a]

	Occurrences per Million base pairs				Number of Occurrences		
MR Size (Each Half)	Expected Occurrence	Human	Yeast	*E. coli*	Human	Yeast	E. coli
≥ 8 bps	81	528	295	71	568	93	23
≥ 9 bps	20	306	158	15	329	50	5
≥ 10 bps	5	192	82	6	206	26	2
≥ 11 bps	1	124	38	0	133	12	0
≥ 12 bps	0.25	85	29	0	91	9	0

Numbers shown are for MRs found which have 3–6 bp spacers between the repeated sequences.

[a]This table gives the numbers of MRs found with given repeat sizes for each of the three data sets. On the right, the actual number of IRs found in each data set is given; on the left, the numbers have been normalized to 10^6 bp for comparison. Also shown is the calculated number of MRs of each size expected in 10^6 bp of random sequence DNA (25% each of A, C, G and T).

centers melt more readily, they have a lower kinetic barrier to extrusion as a cruciform (33,34). It should be noted that, as for the entire repeat, the central 10 bp of the eukaryotic IRs shown in Figure 2 are also very (dA–dT)-rich (data not shown).

Occurrence of MR sequence patterns in human, yeast and *E.coli* genomic DNA

Mirror repeats are DNA sequences which have mirror symmetry, as shown in Figure 1. A relatively small percentage of these type of sequences have the potential to form intramolecular triplexes, or H-DNA (5). For example, the abundant simple repeating sequence poly(dC–dA/dT–dG) fits the sequence requirements for a MR, but this sequence does not form H-DNA. In order to study H-DNA we must first study the occurrence of all sequences having MR symmetry, since this is a necessary, but not sufficient, criterion for H-DNA formation. Then we can limit the searches to MR sequences which are most likely to form H-DNA.

Table 2 gives the occurrence of all MRs in the human, yeast and *E.coli* genomes. This table is organized exactly like Table 1 was for IRs. The actual number of repeats found in the datasets are listed on the right-hand side of the table, while the normalized numbers of MRs expected in 10^6 bp of a particular genome are on the left-hand side of the table. The most striking observation shown in Table 2 is that MRs occur at a random frequency in the *E.coli* genome. The number of MRs expected to be found in 10^6 bp of *E.coli* DNA is nearly identical to the number expected in 10^6 bp of random DNA. This suggests that there has been very little evolutionary pressure either for or against the accumulation of MRs within the *E.coli* genome. In sharp contrast, both of the eukaryotes are highly enriched in the total number of MRs, especially for longer repeats. In human DNA, MRs of ≥8 bp in length are 6.5 times more abundant than in random DNA, while in yeast the occurrence of these MRs is 3.6 times that for random DNA. The frequency of occurrence above the expected levels for MRs of ≥12 bp are very high in both human and yeast DNA.

Occurrence of cruciform and H-DNA forming sequences in the human, yeast and *E.coli* genomes

The data given in Tables 1 and 2 represents the total number of IRs and MRs found in these genomes. However, not all IRs have the same ability to form cruciforms and certainly not all MRs have the potential of forming H-DNA. In order to determine which are potential cruciform or H-DNA forming sequences we needed to set criteria for defining repeats which are most likely to form non-B-DNA structures. As previously discussed, we have already limited our searches to sequences with 3–6 bp spacer regions. For both cruciform and H-DNA forming sequences it is known that longer repeats form more stable non-B-DNA structures (27–30). Therefore we must place a lower limit on the length of the repeat as a criterion for a cruciform or H-DNA forming sequence. For this discussion we have chosen a minimum length of 10 bp, since it has been shown that IRs and MRs of this length will form non-B-DNA structures *in vitro* (29,30,35). We would like to point out, however, that a 9 bp IR requires only slightly more supercoiling than a 10 bp IR of similar sequence to form a cruciform. The 10 bp limit is a conservative estimate chosen for this study in order to identify only the strongest cruciform and H-DNA forming sequences.

Another important consideration in assessing the potential of a repeat to form a non-B-DNA structure is its sequence, not just that it holds to an IR or MR pattern. H-DNA forms best in MRs which are homopurine/homopyrimidine-type sequences, but not in sequences that are very (dA–dT)-rich (5,36–38). The criteria used for considering a MR to be a strong H-DNA forming sequence were: (i) the minimum length of the MR was ≥10 bp in each half of the repeat; (ii) the sequence of the MR was 100% homopurine/homopyrimidine and was <80% (dA–dT)-rich; (iii) the repeat had no mismatches.

A similar criterion was used for screening all IRs and for choosing the strongest cruciform forming sequences. The criteria used for considering an IR to be a strong cruciform forming sequence were: (i) the minimum length of the IR was ≥10 bp in each half of the repeat; (ii) the sequence of the MR was ≥60% (dA–dT)-rich; (iii) the repeat had no mismatches, unless the IR was ≥12 bp in length on each half, in which case we allowed it to have one internal C–A or T–G mismatch every 10 bp. We expect that the extra length of the repeat should more than compensate for the slightly destabilizing pyrimidine/purine mismatches (39) and that the overall effect on the stability of the IR as a cruciform would be negligible.

We then used these criteria to screen the total number of IRs and MRs given in Tables 1 and 2 and produced a new list of only the strongest cruciform and H-DNA forming sequences in each genome. The observed frequencies, or probabilities, of finding cruciform and H-DNA forming sequences which fit our criteria are given in Table 3. These frequencies are simply the result of dividing the total length of the DNA in each dataset by the number of potential cruciform or H-DNA forming sequences that were found in each genome.

Table 3. Probability of finding strong cruciform and H-DNA forming sequences in human, yeast and *E.coli* genomic DNA[a]

Type of DNA	Probability of cruciforms	Probability of H-DNA
Human	1/41 800 bp	1/49 400 bp
Yeast	1/19 700 bp	1/ 315 217 bp
E.coli	1/162 073 bp	1/324 146 bp

The criteria for determining which repeats are strong cruciforms or H-DNA forming sequences is described in the text.

First consider the occurrence of these two structures in the *E.coli* genome. Using our criteria we found only two cruciform and one H-DNA forming sequence in the entire 324 146 bp dataset. Although total IRs are abundant in *E.coli* (Table 1), >90% of the IRs found in *E.coli* did not meet our 60% (A+T) threshold for being considered a strong potential cruciform. Nearly all of the IRs found in *E.coli* have already been recognized as transcription termination sequences (24–26). These termination sequences are known to be (dC–dG)-rich and function by folding RNA, not DNA, into stable hairpin structures (40). In terms of H-DNA forming sequences, we were quite surprised to find that one of the two MRs that was ≥10 bp long in the *E.coli* dataset (see Table 2) was also a perfect homopurine sequence which could form H-DNA. It is difficult to assess how well this would extrapolate to the probability of finding H-DNA in the entire genome and it seems likely that this single sequence represents an anomaly, rather than a rule. Nevertheless, we feel confident in predicting that potential H-DNA forming sequences are relatively rare in *E.coli* and probably occur at a frequency <1/324 146 bp over the entire genome.

We found that both human and yeast DNA had relatively high frequencies of cruciforms in their genomes, but that only human DNA seems to have high levels of H-DNA forming sequences (Table 3). The entire sequence of yeast chromosome III contains only one H-DNA forming sequence, while there were 16 potential cruciforms (1/19 700 bp). In the 1 086 110 bp human DNA dataset we found 26 cruciform and 22 H-DNA forming sequences. Therefore the frequency of finding cruciforms in the human genome is ~1/41 700 bp, while the frequency of finding H-DNA is ~1/49 400 bp.

What are the sequence properties of naturally occurring genomic cruciforms and H-DNA?

We found a significant difference in the (dA–dT) richness of IRs in eukaryotic as compared with prokaryotic DNA (Fig. 2). We also found other interesting sequence features in the cruciforms and H-DNA forming sequences. A partial listing of the potential cruciforms found in these studies is given in Table 4. This table focuses on the longest IRs found, many of which contain a C–A or T–G mismatch. By including this parameter in our searches we were able to identify a few very long, and most likely very stable, potential cruciforms in the human and yeast genomes. A particularly striking feature of cruciform sequences in yeast and humans is that many are also alternating pyrimidine/purine (APP) sequences. APP sequences are usually associated with Z-DNA formation (22). It is intriguing that for some of the sequences each half of the repeat could also be a strong Z-DNA forming sequence on its own (i.e. sequences 6 and 7 in the human DNA in Table 4). This raises the possibility that these IRs might be capable of multiple non-B-DNA conformations, depending upon the environmental conditions. Nevertheless, several studies have shown that alternating (dT–dG)- and (dC–dA)-containing IRs of this size prefer to form cruciforms and not Z-DNA (41,42).

Table 4. Examples of strong cruciform forming sequences

	Sequence of IR	Length of IR
Human DNA		
1)	atatg**ATATATATGTGTGTGTATATATGTATACATATATaT**gtgt- **AgATATATGTATACATATATACACACACATATATAT**gtaca	36
2)	tacat**ATATATATaTATATATAtGTATATAT**atacac- **ATATATACgTATATATAgATATATAT**gtata	26
3)	tacac**ACATATATgTATAtATATATGT**acac**ACATATATgTATAtATATATGT**acaca	22
4)	tacac**ACATATATgTAtATATATGT**acac**ACATATATgTAtATATATGT**acaca	20
5)	gtgta**TATATATATACATATATAT**atat**ATATATATGTATATATATA**cacat	19
6)	gtgtg**TGTGTGTATGCGTGTGTG**taga**CACACACGCATACACACA**tataa	18
7)	atata**TGTGTGTGTGTGTGTG**tata**CACACACACACACACA**catgt	16
8)	taagt**TATTTTATATATAT**aat**ATATATATAAAATA**tataa	14
9)	agggg**GTGTGTGcATGTG**tgca**CACATaCACACAC**acacac	13
10)	aaatt**CTCATACATAAA**catcac**TTTATGTATGAG**gcaaa	12
11)	aaaat**TATATATaTACA**catata**TGTAcATATATA**tgaat	12
12)	tatta**TTAAAAgATAAA**agtaaa**TTTATtTTTTAA**gatat	12
13)	ataat**TAtTTTTTAATT**gatg**AATTAAAAAgTA**tatat	12
Yeast DNA		
a)	tgtat**ATACATACaTTCTTATAC**aatacc**GTATAAGAAgGTATGTAT**gtatg	18
b)	ttcct**ATATATATgTATATA**tatc**TATATAgATATATAT**cccag	15
c)	tctag**GTAgTGAGATTGAT**gaa**ATCAATCTCAaTAC**taata	14
d)	acaaa**ATAtATATATATA**tata**TATATATATgTAT**gtcca	13
e)	agata**GTATATaTATAT**atat**ATATAcATATAC**atata	12
f)	attcc**TATATATATATA**tata**TATATATATATA**tcata	12
***E.coli* DNA**		
	atata**ACAAATCCCAATAA**ttaag**TTATTGGGATTTGT**ctggt	14

All of the H-DNA forming sequences found in the three genomes are listed in Table 5. Only five of the 22 human sequences are simple poly(dT–dC/dG–dA) repeats, even though these repeats are known to be very abundant in the human genome (13,17). Many of the sequences shown in Table 5, in fact, are not simple direct repeat sequence patterns. Another class of H-DNA forming sequences which has been intensely studied both *in vitro* and *in vivo* are the poly(dG) type of sequences (30). Although these sequences can form very stable triple-stranded structures, they were not found in any of the genomes studied. In fact, none of the potential H-DNA forming sequences found in these genomes were >80% (dG–dC)-rich, suggesting that poly(dG) sequences are rare in genomic DNA. For example, in the entire human data set we found almost 300 poly(dA) tracts of ≥10 bp in length, but only two poly(dG) tracts of this length (data not shown). This is reminiscent of the situation with Z-DNA forming sequences, in which alternating (dC–dG) sequences are best for Z-DNA formation, but long stretches of these sequences are rarely found in genomic DNA (16,22). It could be that poly(dC–dG) and poly(dG) sequences too easily convert to non-B-DNA conformations for them to serve a useful regulatory function inside the cell. Another possibility is that sequences which form very stable non-B-DNA conformations are so highly mutagenic that there is a strong evolutionary pressure against their accumulation in genomes (2).

DISCUSSION

The goal of this work was to study the occurrence of cruciform and H-DNA forming sequences in genomic DNA. First we studied the occurrence of all IRs and MRs in DNA datasets from the human, yeast and *E.coli* genomes (shown in Tables 1 and 2).

Table 5. Examples of strong H-DNA forming sequences

	Sequence of MR	Length of MR
HUMAN DNA		
1)	gtccc**TCCCCTCCCCTCCCCTCCC**ctc**CCCTCCCCTCCCCTCCCCT**tccct	19
2)	aagag**GAGAGAGAGAGAGAGAGAG**aga**GAGAGAGAGAGAGAGAGAG**aatga	19
3)	gaaag**GAGAGAGAGAGAGAGAG**aga**GAGAGAGAGAGAGAGAG**aacaa	17
4)	aggag**GAGGAGGAGGAGGAGG**atgt**GGAGGAGGAGGAGGAG**tgtca	16
5)	acaga**AAGAGAGAAAGAGAGA**gag**AGAGAGAAAGAGAGAA**tgcaa	16
6)	aagag**AGAGAGAGAGAGAGAG**att**GAGAGAGAGAGAGAGA**aggga	16
7)	ggaaa**GAAGGAAGGAAGGAAG**aaag**GAAGGAAGGAAGGAAG**gaaaa	16
8)	agaga**AAGAAAGAAAGAAAG**aaagg**GAAAGAAAGAAAGAA**tacgg	15
9)	tatat**AGAGAGAGAGAGAGA**gag**AGAGAGAGAGAGAGA**taccc	15
10)	ttccc**TCCCCTTCCCTTCC**ctcc**CCTTCCCTTCCCCT**agagg	14
11)	aaaaa**AAAAGAAAGAAAG**aaa**GAAAGAAAGAAAA**gaaat	13
12)	aaaga**AAAGAAAAGAGAA**gaa**AAGAGAAAAGAAA**gacag	13
13)	gatgg**AGAGAGGGAG**agagg**GAGAGGGAGAGA**aggga	12
14)	atttg**TTCTCTCTCTC**ccttt**CTCTCTCTCTT**ccccc	11
15)	gaaag**AAGAAAGAGAA**aaga**AAGAGAAAGAA**agaga	11
16)	aggga**GGAAGGAAGGA**aaga**AGGAAGGAAGG**gagag	11
17)	agccc**CTTTTTTTCTC**tct**CTCTTTTTTTC**ttttg	11
18)	tctcc**CTCTCCCTCTC**cca**CTCTCCCTCTC**tttct	11
19)	gatgg**AGAGAGGGAG**aga**GAGGGAGAGA**tgaag	10
20)	ttcct**TCCTTCTTCT**cctc**TCTTCTTCCT**gcatc	10
21)	tgatg**AGGAGGAGGA**agaag**AGGAGGAGGA**aggta	10
22)	tctcc**CTCTCTCTCT**ttctc**TCTCTCTCTC**tatat	10
Yeast DNA		
	agaaa**GAAGAAGAAGAAGA**aga**AGAAGAAGAAGAAG**ctgat	14
E.coli DNA		
	ccttt**GGGGAGAGGG**tta**GGGAGAGGGG**aaaac	10

This total population of repeats contained many sequences which, although they conform to either an IR or MR sequence pattern, cannot easily adopt a cruciform or H-DNA structure. We used published experimental data to help us select those IRs and MRs which are most likely to form cruciforms and H-DNA. Although the criteria we used to define the strongest non-B-DNA forming sequences were necessarily arbitrary, they were based very strictly upon published experimental evidence. Because our criteria were relatively conservative, the numbers shown in Table 3 are probably low estimates of the actual numbers of cruciform and H-DNA forming sequences in these genomes.

The point of this paper is not that H-DNA or cruciform structures exist every 40–50 kb in the human genome (as given in Table 3), but rather that relatively strong cruciform and H-DNA forming sequences can be found that frequently. Since both cruciforms and H-DNA are high energy conformations, none of the sequences identified in this paper will constitutively adopt these structures. The formation of these conformations requires high levels of negative supercoiling. We now understand that these levels of supercoiling are possible inside cells, most likely as a natural consequence of normal cellular processes such as transcription. Liu and Wang's 'twin-supercoiled domain' model shows how a transcribing polymerase is able to generate high levels of negative supercoiling in its wake (43). Subsequent work from many laboratories has shown that the negative supercoiling energy available from transcription can be transmitted into other biochemical reactions, including recombination events, chromatin reorganization and the formation of non-B-DNA structures (44,45) Therefore, at least transiently, it appears that the negative supercoiling required to form non-B-DNA structures does exist inside cells. The results presented in this paper clearly show that cruciform and H-DNA forming sequences are present in abundance in some genomes. The fact that there are high levels of dynamic supercoiling available inside cells and sequences which can easily form non-B-DNA structures in genomic DNA suggests that non-B-DNA structures probably do exist *in vivo*.

We found an interesting difference between *E.coli*, the only prokaryote in this study, and the two eukaryotic DNAs (yeast and human). Aside from the large number of IRs that were located at sites of transcription termination, the *E.coli* genome is relatively devoid of strong cruciform or H-DNA forming sequences. We suspect that the IRs found at termination sites are not good cruciform forming sequences, since in general they are very (dG–dC)-rich sequences. The (dG–dC)-rich IR sequences form stable hairpins in single-stranded RNA, which makes them good transcriptional terminators (40). These same (dG–dC)-rich properties, however, impose high kinetic barriers to cruciform formation in negatively supercoiled DNA (33–35). Finally, the suggestion that *E.coli* has few potential cruciforms or H-DNA forming sequences is entirely consistent with other studies noting that prokaryotes have little of the simple repeating DNA sequences [such as poly(dT–dA), poly(dC–dA/dT–dG) or poly(dT–dC/dA–dG) which can form non-B-DNA structures *in vitro*.

The lack of non-B-DNA forming sequences in *E.coli* is in sharp contrast to yeast, which had the highest frequency of potential cruciforms, or human DNA, which had relatively high levels of both cruciforms and H-DNA forming sequences (Table 3). Interestingly, these frequencies do not correlate strongly with gene density or with the amount of 'junk' DNA in the datasets. It is not surprising to find only a few non-B-DNA forming sequences in the *E.coli* genome, since the gene density is very high (~1 gene/900 bp), which means there is little room in the genome for non-coding DNA structural features. On the other extreme of this discussion is the human dataset, which has a gene density of ~1 gene/7000 bp of DNA sequence and which consists mostly of 'junk' DNA derived from introns and flanking sequences (these sequences constitute ~85% of the total human dataset used in these studies).

In this respect the results on yeast are quite intriguing, since the yeast genome is a relatively streamlined and compact genome, with 1 gene/1700 bp, and has very few introns (none were noted in the DNA sequence of chromosome III). In spite of the inherent limitations of an economical genome, *S.cerevisiae* seems to have evolved with a very high number of cruciform forming sequences. It could be that the accumulation of non-B-DNA forming sequences within a genome is a eukaryotic phenomenon, an idea which is consistent with the phylogenetic distributions of virtually all other types of simple repetitive DNA (13,14,17, 19–21).

In a previous study we showed that there was a non-random distribution of potential Z-DNA forming sequences in the human genome and that they tended to cluster near transcription start sites or promoters, as opposed to the 3′-end of the gene (22). In this study we did not observe a strong bias in the distribution of either IRs or MRs throughout these genomes (except, of course, for the noted location of IRs in *E.coli*). In human genes we found that ~76% of all the cruciform and H-DNA forming sequences were located either in introns or in regions encoding untranslated regions (UTRs) of the mRNA (introns and UTRs made up ~60% of the total human DNA in the dataset). Only 3% of the non-B-DNA forming sequences were found within protein coding regions (which made up ~15% of the total DNA). The remaining 20% were located in either the 5′ or 3′ flanking regions of the genes.

Now that we have identified sequences that can potentially form cruciforms and H-DNA in genomes and that conditions exist inside cells for these sequences to convert to non-B-DNA conformations (43–45), the challenge will be to determine what, if any, function these repeats have in cellular regulation. All non-B-DNA structures that form in response to high levels of negative supercoiling (including Z-DNA, H-DNA and cruciforms) may play similar roles in regulating the DNA topology in and around active genes. These structures could act as 'topological gauges' whose function is to respond to high levels of transcription-induced negative supercoiling. When transcriptional activity becomes too high, these unusual structures could form, which could slow down or pause transcription complexes. When the DNA becomes relaxed through the action of topoisomerases, the unusual structures would convert back to B-DNA and normal transcription would continue. This function could be useful for maintaining an ordered array of polymerases on genes or for affecting the chromatin structure of the region (similar models have been discussed in 1,3,22,46,47). Studying these functions will be an enormous challenge, since these structural transitions are likely to be dependent upon the dynamic and complex interplay between transcription, DNA supercoiling and non-B-DNA structure formation in sequences like those found in this paper.

ACKNOWLEDGEMENTS

This work has been supported by a grant from the American Cancer Society (NP-740C). G.P.S. is also supported by a Postdoctoral Research Fellowship from the American Cancer Society (Pf-3749).

REFERENCES

1 Sinden,R.R. (1994) *DNA Structure and Function*. Academic Press, San Diego, CA.
2 Sinden,R.R. and Wells,R.D. (1992) *Curr. Opin. Biotechnol.*, **3**, 612–622.
3 Rich,A. (1993) *Gene*, *135*, 99–109.
4 Murchie,A.I.H., Bowater,R., Aboul-ela,F. and Lilley,D.M.J. (1992) *Biochim. Biophys. Acta*, **1131**, 1–15.
5 Mirkin,S.M. and Frank-Kamenetskii,M.D. (1994) *Annu. Rev. Biophys. Biomol. Struct.*, **23**, 541–576.
6 Palecek,E. (1991) *Crit. Rev. Biochem. Mol. Biol.*, **26**, 151–226.
7 Lilley,D.M.J. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 6468–6472.
8 Panayotatos,N. and Wells,R.D. (1981) *Nature*, **289**, 466–470.
9 Mirkin,S.M., Lyamichev,V.I., Drushlyak,K.N., Dobrynin,V.N., Filippov,S.A. and Frank-Kamenetskii,M.D. (1987) *Nature*, **330**, 495–497.
10 Johnston,B.H. (1988) *Science*, **241**, 1800–1804.
11 Hanvey,J.C., Shimuzu,M. and Wells,R.D. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 6292–6296.
12 Htun,H. and Dahlberg,J.E. (1989) *Science*, **243**, 1571–1576.
13 Manor,H., Rao,B.S. and Martin,R.G. (1988) *J. Mol. Evol.*, **27**, 96–101.
14 Hamada,H., Petrino,M.G. and Kakunaga,T. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 6465–6469.
15 Haniford,D.B. and Pulleyblank,D.E. (1983) *Nature*, **302**, 714–716.
16 Trifonov,E.N, Konopka,A.K and Jovin,T.M (1985) *FEBS Lett.*, **185**, 197–202.
17 Gross,D.S. and Garrard,W.T. (1986) *Mol. Cell. Biol.*, **6**, 3010–3013.
18 Greaves,D.R., Patient,R.K. and Lilley,D.M.J. (1985) *J. Mol. Biol.*, **185**, 461–478.
19 Tripathi,J. and Brahmachari,S.K. (1991) *J. Biomol. Struct. Dynam.*, **9**, 387–397.
20 Wells,R.D., Collier,D.A., Hanvey,J.C., Shimuzu,M. and Wohlrah,F. (1988) *FASEB J.*, **2**, 2939–2949.
21 Lu,G. and Ferl,R.J. (1993) *Int. J. Biochem.*, **25**, 1529–1537.
22 Schroth,G.P., Chou,P.-J. and Ho,P.S. (1992) *J. Biol. Chem.*, **267**, 11846–11855.
23 Oliver,S.G. *et al.* (1992) *Nature*, **357**, 38–46.
24 Daniels,D.L., Plunkett,G., Burland,V. and Blattner,F.R. (1992) *Science*, **257**, 771–778.
25 Plunkett,G., Burland,V., Daniels,D.L. and Blattner,F.R. (1993) *Nucleic Acids Res.*, **21**, 3391–3398.
26 Burland,V., Plunkett,G., Daniels,D.L. and Blattner,F.R.(1993) *Genomics*, **16**, 551–561.
27 Murchie,A.I.H. and Lilley,D.M.J. (1992) *Methods Enzymol.*, **211**, 158–180.
28 Zheng,G., Kochel,T., Hoepfner,R.W., Timmons,S.E. and Sinden,R.R. (1991) *J. Mol. Biol.*, **221**, 107–129.
29 Lyamichev,V.I., Mirkin,S.M., Kumarev,V.P., Baranova,L.V., Vologodskii, A.V. and Frank-Kamenetskii,M.D. (1989) *Nucleic Acids Res.*, **17**, 9417–9423.
30 Kohwi-Shigematsu,T. and Kohwi,Y.(1991) *Nucleic Acids Res.*, **19**, 4267–4271.
31 Gough,G.W., Sullivan,K.M. and Lilley,D.M.J. (1986) *EMBO J.*, **5**, 191–196.
32 Shimuzu,M., Hanvey,J.C. and Wells,R.D. (1989) *J. Biol. Chem.*, **264**, 5944–5949
33 Courey,A.J. and Wang,J.C. (1988) *J. Mol. Biol.*, **202**, 35–43.
34 Zheng,G. and Sinden,R.R. (1988) *J. Biol. Chem.*, **263**, 5356–5361.
35 Singleton,C.K. (1983) *J. Biol. Chem.*, **258**, 7661–7668.
36 Hanvey,J.C., Klysik,J. and Wells,R.D. (1988) *J. Biol. Chem.*, **263**, 7386–7396.
37 Hanvey,J.C., Shimuzu,M. and Wells,R.D. (1989) *J. Biol. Chem.*, **264**, 5950–5956.
38 Fox,K.R. (1990) *Nucleic Acids Res.*, **18**, 5387–5391.
39 Spiro,C., Richards,J.P., Chandrasekaran,C., Brennan,R.G. and McMurray,C.T. (1993) *Proc. Natl. Acad. Sci. USA*, **90**,4606–4610.
40 Farnham,P.J. and Platt,T. (1980) *Cell*, **20**, 739–746.
41 McLean,M.J. and Wells,R.D. (1988) *J. Biol. Chem.*, **263**, 7370–7377.
42 Blaho,J.A., Larson,J.E., McLean,M.J. and Wells,R.D. (1988) *J. Biol. Chem.*, **263**, 14446–14455.
43 Liu,L.F. and Wang,J.C. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 7024–7027.
44 Freeman,L.A. and Garrard,W.T. (1992) *Crit. Rev. Eukaryot. Gene Expression*, **2**, 165–209.
45 Droge,P. (1994) *BioEssays*, **16**, 91–99.
46 van Holde,K. and Zlatanova,J. (1994) *BioEssays*, **16**, 59–68.
47 Wittig,B., Wolfl,S., Dorbic,T., Vahrson,W. and Rich,A. (1992) *EMBO J.*, **11**, 4653–4663.