An overabundance of long oligopurine tracts occurs in the genome of simple and complex eukaryotes

Michael J. Behe

Department of Chemistry, Lehigh University, Iacocca Hall #111, Bethlehem, PA 18015, USA

Received September 30, 1994; Revised and Accepted November 29, 1994

ABSTRACT

A search of sequence information in the GenBank files shows that tracts of 15–30 contiguous purines are greatly overrepresented in all eukaryotic species examined, ranging from yeast to human. Such an overabundance does not occur in prokaryotic sequences. The large increase in the number of oligopurine tracts cannot be explained as a simple consequence of base composition, nearest-neighbor frequencies, or the occurrence of an overabundance of oligoadenosine tracts. Oligopurine sequences have previously been shown to be versatile structural elements in DNA, capable of occurring in several alternate conformations. Thus the bias toward long oligopurine tracts in eukaryotic DNA may reflect the usefulness of these structurally versatile sequences in cell function.

INTRODUCTION

The structure of DNA was long thought to have been finally solved by Watson and Crick (1), and the task of science in this area was expected to be the further refinement of their model. In the absence of known alternative structures of DNA the role of the polynucleotide in the cell was thought to be necessarily static: a long punch-tape of information that was read by the protein machinery which actually did the interesting work. That paradigm of DNA function was severely challenged in 1979 by the reported observation of Z-DNA in crystals of a synthetic double stranded hexanucleotide (2). Subsequent work showed that DNA was a much more flexible molecule than previously supposed, especially when under superhelical strain, and examples of alternative conformations increased quickly (3,4). The availability, at least in theory, of a number of conformations of DNA allows the polymer to play a more active role in the expression of the information it contains. However, for a number of reasons, conclusive evidence for the involvement of alternative DNA conformations in important cellular functions has been difficult to obtain.

All alternative DNA conformations have at least one feature in common: restriction to a small subset of nucleotide sequences. That is, although virtually any sequence can assume the canonical B conformation, Z-DNA requires an alternating purine–pyrimidine, GC-rich sequence, H-DNA requires a mirror-repeat oligopurine sequence, cruciforms require inverted-repeat sequences, etc. Therefore a persuasive, if indirect, method for showing that

such sequence-dependent conformations play a significant role *in vivo* is to demonstrate that the prerequisite sequences exist in the genome in numbers significantly greater than one would expect from statistical considerations. This has previously been done for alternating purine–pyrimidine sequences (5).

A search was also previously conducted for contiguous oligopurine sequences (6). Oligopurine tracts have been shown to be quite versatile in their conformational properties. Synthetic double stranded poly dA·poly dT and poly dG·poly dC were shown to differ from the canonical B form: the helical repeat of poly dA·poly dT was reduced to ~10.0 bp per helical turn from the normal value of 10.5 bp per turn in heterogeneous-sequence DNA (7), and crystallographic data from an oligonucleotide containing an oligo dG·oligo dC tract showed a wider helix diameter, reminiscent of the A form of DNA (8). Additionally, oligopurine tracts in longer DNA polymers were shown to be able to form triple helical structures under superhelical stress if the tract had mirror-repeat symmetry (9). More recently the formation of different triple helical structures was shown to be possible (10), especially in the presence of divalent cations, and four stranded structures have been observed in G-rich oligopurine tracts in the presence of monovalent cations (4).

The previous search for oligopurine tracts (6) (in both the previous search and in this paper both oligopurine tracts and oligopyrimidine tracts on the reported strand are counted together as one category and frequently called 'oligopurine tracts' for simplicity) was conducted 7 years ago on a limited number of sequences found in the GenBank database. Although the results showed that a strong sequence bias existed in favor of long oligopurine tracts in several human gene sequences, results from other species were merely suggestive. It appeared that, while higher eukaryotes such as mouse also shared the sequence bias, lower eukaryotes did not. It was noted at the time, though, that the conclusions concerning non-human species may have been biased by the small sequence base available for analysis.

Recently I analyzed large sequence files from the greatly expanded GenBank database. The results confirm the earlier conclusion that the human genome is strongly biased in favor of long tracts of contiguous oligopurines. However, the availability of large sequence files from lower eukaryotes shows, contrary to previous results, that the sequence bias is shared to approximately the same degree by all eukaryotic species examined. Conversely, prokaryotic DNA shows no such bias. The results strongly suggest that the alternative DNA structures available to oligopurine tracts play a role in the expression of the information in all eukaryotic genomes.

MATERIALS AND METHODS

DNA sequences were analyzed using the program TRACTS, written in Pascal for the Windows environment. The program keeps a running account of the number and length of all tracts consisting of one, two or three base combinations as it reads the input sequence from GenBank files, and is able to analyze 300 000 nucleotides/min. The source code for the program can be obtained from the author. Table 1 lists the GenBank sequence files that were analyzed, along with the number of nucleotides in the file. The analysis was restricted to files containing >20 000 bases in order to avoid biasing in favor of coding sequences. A total of >1.8 million nucleotides of human sequence data was analyzed, as well as over a million nucleotides from the nematode *Caenorhabditis elegans* and the yeast *Saccharomyces cerevisiae*. This represents ~0.05, 1 and 12% of the total genomes of the respective organisms. Additionally, over one million nucleotides of *Escherichia coli* sequences were analyzed, representing ~25% of its total genome.

An oligopurine tract of n nucleotides in length is defined as one in which the n purines are flanked on either end by a pyrimidine. Since purine tracts are the complement of pyrimidine tracts, I count both purine and pyrimidine tracts of the reported strand in a single category. Defining purine tracts in this manner allows the entire sequence file that is being analyzed to be categorized. Other categories of tracts, such as oligoadenosines and oligoguanosines, are treated analogously.

The following IUPAC abbreviations for deoxynucleotides are used throughout this paper: A, adenosine; C, cytosine; G, guanosine; T, thymidine; Y, either pyrimidine nucleotide; R, either purine nucleotide; H, any nucleotide except G; B, any nucleotide except A; D, any nucleotide except C; V, any nucleotide except T.

The number of actual occurrences of purine tracts of any particular length can be compared to the number that one would expect on a statistical basis. This expected value can be calculated in several ways. The first method is to make the pedestrian observation that in double-stranded DNA half of the nucleotides are purines and half are pyrimidines. Thus the expected number of occurrences of a purine tract, $-YR_nY-$, would simply be equal to the number of bases in the data set times 0.5^{n+2} . The second method is to take into account any strand asymmetry in the number of purines, but assume that the number of purine tracts that would appear in the absence of sequence bias depends simply on the base composition of the tracts. In this case the expected number of occurrences is equal to the number of bases in the data set $\times$ (fraction of pyrimidines)² $\times$ (fraction of purines) ^{n} . The first and second methods are essentially equivalent here since significant strand asymmetry does not occur in the sequence files under consideration.

The third method is to consider that the frequency of occurrence of dinucleotides in most organisms does not depend simply on base composition. For example the dinucleotide CG is greatly underrepresented in the DNA of higher eukaryotes. Such bias in dinucleotide frequency could affect the number of oligopurine tracts so that deviation from the 'expected' value based simply on base composition might actually be a secondary effect of the dinucleotide bias. Calculation of the expected number of oligopurine tracts of length n given biased dinucleotide frequency can be done as follows:

Figure 1. Counted number of occurrences of $-YR_nY-$ plus the complementary sequence $-RY_nR-$ in the reported strand of the *E.coli*, yeast, *C.elegans* and human sequence files listed in Table 1, plotted versus n , the tract length. The solid line in the figure shows the expected number of such tracts calculated from the dinucleotide compositions of the respective files. The dotted curves show the expected number ± 1 SD. The absence of a point at a particular value of n means that no tracts of that length were observed.

expected number of occurrences $[\Sigma[YR] \cdot \Sigma[RR]^{n-1} \cdot \Sigma[RY]] / \Sigma[R]^n$

where $\Sigma[YR]$ is the sum of the number of pyrimidine–purine nearest-neighbors, $\Sigma[RR]$ is the sum of the number of purine–purine nearest-neighbors, $\Sigma[RY]$ is the sum of the number of purine–pyrimidine nearest-neighbors, and $\Sigma[R]$ is the sum of the number of purine nucleotides. In all three cases the standard deviation is calculated simply as the square root of the expected number of occurrences.

In general, and especially for the higher eukaryotes, which have strong biases in dinucleotide frequencies, the expected values calculated from dinucleotide composition fit the counted data much more closely for tracts of length 1–10 than do the values calculated from base composition alone. So throughout this paper the expected number of occurrences of a tract is calculated from dinucleotide composition.

RESULTS

Oligopurine tracts

Figure 1 shows the sum of the number of oligopurine tracts as a function of length for the human sequence files listed in Table 1, along with similar data for *E.coli*, *S.cerevisiae* and *C.elegans*. In each of the panels the expected number of tracts is shown by a

solid line and the dashed lines represent one standard deviation from the expected value. Each of the eukaryotic species exhibits a strong increase in the number of oligopurine tracts of length 15–30 nucleotides over what is expected on statistical considerations. It should be kept in mind that the plot uses a log scale, so that the bias is stronger than it may appear to the eye. For example, the number of human oligopurine tracts of length 20 shown in Figure 1 is >25 standard deviations (SD) from the expected number. In contrast to the eukaryotic data, sequence analysis of *E.coli* files (Fig. 1) shows that the number of long oligopurine tracts is very close to that expected from statistical considerations. In addition to the species represented in Figure 1 the sequence files of several other eukaryotic and prokaryotic organisms were analyzed for the occurrence of oligopurine and oligoadenosine tracts. Those results are summarized in Table 2. It is seen that eukaryotic species have ratios of observed to expected number of long oligopurine tracts (15–30 contiguous purines) ranging from 3.6 to 5.8, whereas the several prokaryotic species all have ratios close to unity.

Table 3 shows an analysis of the base composition of the long oligopurine tracts of the eukaryotes. Tracts of 15–30 contiguous purines have AT contents (considering also the complementary strand) that range upwards from the AT content of the sequence file for the organism as a whole. The tracts of the nematode *C.elegans* are equal in %AT to the rest of the nematode sequences, but human and yeast tracts are substantially enriched in AT content. This appears to be the result of the overoccurrence in these organisms of oligoadenosine tracts, which are a subset of oligopurine tracts (see below). Within the tracts themselves the ratio of observed to expected number of nearest-neighbor dinucleotides, AA, AG, GA and GG, is very close to that expected from the base composition of the tracts (Table 3). The frequency of occurrence of adenosine at the 5'- and 3'-ends of long oligopurine tracts is also listed in Table 3. Although these values

Table 1. GenBank files used in the search

File	Size	File	Size	File	Size
human					
HSTCRBV	77743	GGVITIIG	20343	SCDPB3	32420
HSU07000	152141			SCHAPLAP	35757
HSIGLAMB	33737			SCXI207K	20723
HSIFVAR	32906	<i>D. melanogaster</i>			
HSBAT2	38324	DROABDB	80423	SCXI286K	28600
HSCSF1PO	35100	DROMHC	22663	SCRACH	69748
HSCAMHCA	31462	DROZFP	28627	SCPEKGA	38477
HSMHCAPG	66109	DROBXD	43548	SC22KBF	22469
HSG6PDGEN	52173	DRORYRH	25680	YSCLTESPO	29410
HUMHBB	73326	DROABDA1A	77444	YSCSYGP2	36772
HUMGHCSA	66495			YSCCHROMI	41987
HUMFIXG	38059	<i>C. elegans</i>			
HUMADAG	36741	CELUNC22	47081	YSCH9986	41663
HUMTPA	36594	CELB0303	41071	YSCH9177	52664
HUMHPTB	56737	CELC06E1	40216	YSCH9998	35600
HUMBEXIII	33206	CELC08C3	44025	YSCH8179	44113
HUMDYSTROP	38770	CELC14B9	43492	YSCH932	36176
HUMHPT2HPR	38542	CELC30C11	30865	YSCH8025	27816
HUMNEUROF	100849	CELC50C3	44733		
HUMRETBLAS	180388	CELF09G8	41645	<i>E. coli</i>	
HUMRTMD	32000	CELF10E9	32733	ECOUW82	136254
HUMTCRADCV	97634	CELF22B7	40222	ECOUW85U	91408
HUMVITDBP	55136	CELC12H4	38582	ECO110K	111401
HUMCC1S149	34379	CELR05D3	38810	ECOUW87	96484
HUMHDABCD	58864	CELC21	36087	ECOUW89	176195
HUMHDAC	40289	CELCZK112	38269	ECOUW76	225419
HUMHDAD	40103	CELCZK370	37675	ECOHU47	75888
HUMMDA	37314	CELCZK652	36052	ECOHU49	39149
HUMMMDB	68505	CELCZK688	36977	ECO82K	82727
HUMCOL2A1Z	31001	CEZK637	40699		
HUMFMR1S	61613	CEZK643	39534	<i>B. subtilis</i>	
HUMDXS455A	38409	CEF59B2	43782	BAC180K	180136
HUMRCCI	34641	CEB0464	40909	BACSRFAB	20535
		CEC38C10	34193	BSSRFAP	32802
		CEF54G8	31613	BACDIA	28206
		CEF58A4	38000	BSGENR	97015
		CEK01B6	34002		
mouse					
MMBGCDX	55856	CEK11H3	33000	<i>M. leprae</i>	
MMSYNDE1A	23600	CER107	40970	U00010	41171
MMU04056	21718	CEZC84	38955	U00011	40429
MUSPACOLL	30701	CEZK1098	37310	U00012	33312
MUSSLPSEXB	26307	CEZK512	36997	U00013	35881
MUSTCRA	94647	CEZK632	36000	U00014	36470
MUSTCRVAD	34476			U00015	42325
				U00016	42931
				U00017	42157
				U00018	42991
				U00019	36033
				U00020	36947
				U00021	39193
				U00022	36411
				U00023	36181
chicken					
CHKMYHE	31111	SCCHRII	315344		
CHKCRYD	25342	SCCEN11D	24743		
CHKHBBRE	23652	SCCHRIIFG	30981		
GDCOL6A1A	21180	SCIL260	32560		
GDCOL6A2G	27443	SCDCHRI1	36272		
GDLIPLIP	22257	SCDCHRL11	20592		

frequently differ noticeably from the AT composition of the tracts, there is no apparent regularity to the deviations.

Table 2. Analysis of oligopurine and oligoadenosine tracts

	Nucleotides analyzed	Observed no. of long tracts ^a	Expected no. of long tracts	Observed/expected	No. of very long tracts ^b	Very long tracts per 100 kb	No. of long A tracts	A tracts per 100 kb
human	1849259	1371	283	4.8	112	6.1	348	18.8
mouse	287304	199	50	4.0	28	9.7	16	5.6
chicken	171317	103	20	5.1	2	1.2	6	3.5
fruit fly	278385	82	14	5.8	0	0	8	2.9
<i>C.elegans</i>	1194499	554	156	3.6	8	0.7	3	0.3
yeast	1054887	468	75	5.8	13	1.2	43	4.1
<i>E.coli</i>	1034925	44	32	1.4	0	0	0	0
<i>B.subtilis</i>	358689	66	75	0.9	0	0	0	0
<i>M.leprae</i>	542432	3	7	0.5	1	0.2	0	0

^aA 'long' oligopurine tract is defined as one that contains 15–30 contiguous purines.

^bA 'very long' oligopurine tract is defined as one that contains >30 contiguous purines.

Table 3. Base composition of tracts and tract termini

	%AT of sequence file	%AT of long purine tracts	Observed/expected nearest-neighbors in				%A at 5' end of purine tracts	%A at 3' end of purine tracts
			AA	AG	GA	GG		
human	55	66	1.1	0.9	0.8	1.3	70	57
mouse	54	56	1.1	0.9	0.9	1.1	60	68
chicken	53	56	1.1	0.9	0.9	1.0	51	62
fruit fly	57	60	1.1	0.7	1.1	1.1	58	54
<i>C.elegans</i>	64	64	0.9	1.1	1.1	1.0	55	54
yeast	61	70	1.0	1.0	1.0	1.0	50	62

Figure 2 shows the position of occurrence of oligopurine tracts along the lengths of three large sequence files, one each from human, yeast and *E.coli*. The position of coding regions in the files is also shown. The prokaryotic sequence file is dense with coding sequences but has only a few, widely-scattered oligopurine tracts of length 15–30 (all four tracts are 17 purines in length). For both eukaryotic files it is seen that the tracts are scattered along the length of the DNA, and their positioning bears no apparent relationship to the positioning of the coding regions. It is also easily seen by this graphical representation that, while the density of coding regions differs greatly between the simple and complex eukaryote sequences, the density of long oligopurine tracts is relatively constant. No long tracts appear in coding regions of the HUMRETB LAS sequence file, while one-third of long oligopurine tracts occur within coding regions in SCCHR III. However, coding regions account for only ~2% of the human sequence file but account for about two-thirds of the yeast sequence. Within coding regions of SCCHR III the 36 long oligopurine tracts arise from diverse sequences. The longer tracts, >20 contiguous purines, frequently code for charged amino acid tracts containing lysine, arginine and glutamic acid, all of which have codons consisting of three contiguous purines. Several long tracts code for regions rich in serine and proline, which have codons consisting of three contiguous pyrimidines.

Although the number of tracts of oligopurines of length 15–30 is similar for the lower and higher eukaryotes examined, the number of very long tracts, >30 nucleotides in length, is substantially greater for mammalian DNA than for other species. Table 2 shows that the number of such tracts per 10 000 nucleotides of sequence DNA is >5-fold greater in man and mouse than in chicken and lower eukaryotes. (The number of very long tracts is normalized to number of tracts per 100 kb, instead of to the ratio of counted/expected as for long purine tracts, because no very long tracts are statistically expected to occur.)

Oligohomopurine tracts

Besides base and dinucleotide composition, the expected number of long tracts of a particular sequence can also be affected by biases in the occurrence of subsets of the sequence. For example, if there occurred in a file 100 tracts of G₂₅T, then it may either be significant in itself or else explainable as an expected result given an over-occurrence of long G tracts. If we looked and saw that the number of G₂₅A, G₂₅T and G₂₆ tracts also was ~100, then no special significance could be attributed to the terminal nucleotide. In the case presently under consideration it is possible that any overrepresentation of oligopurine tracts could be the result simply of an overabundance of oligoadenosine and oligoguanosine tracts.

Figure 2. Position of long oligopurine tracts and coding sequences in three large sequence files. For each file the X axis shows the distance in nucleotides from the beginning of the sequence. For each sequence the line of points right above the X-axis shows the positions of coding regions. The second line of points above the axis shows the positions of long oligopurine tracts.

Figure 3. Counted number of occurrences of $-HG_nH-$ plus the complementary sequence $-DC_nD-$ in the reported strand of the *E.coli*, yeast, *C.elegans* and human sequence files listed in Table 1, plotted versus n , the tract length. The solid line in the figure shows the expected number of such tracts calculated from the dinucleotide compositions of the respective files. The dotted curves show the expected number ± 1 SD. The absence of a point at a particular value of n means that no tracts of that length were observed.

Figure 3 shows the sum of the number of oligoguanosine tracts found in *E.coli*, *S.cerevisiae*, *C.elegans*, and human sequence files as a function of tract length, along with the number of tracts expected based on dinucleotide composition. Figure 4 shows similar data for oligoadenosine tracts. In all cases the number of oligo dG tracts (Fig. 3) is very close to the number expected from statistical considerations, with a few stray tracts observed in the 10–20 nucleotide range of human and *C.elegans*. The number of oligo dA tracts (Fig. 4) is also very similar to the expected value for *E.coli* and *C.elegans* files; however, although the counted number of adenosine tracts of length 1–8 is very similar to the number expected from calculations based on dinucleotide composition, the counted number of long oligoadenosine tracts ($n = 15$ –30) is significantly greater than expected for *S.cerevisiae*, and vastly greater than expected for human DNA. Table 2 lists the total number of occurrences of long oligoadenosine tracts for each species listed in Table 1, and also the number of such tracts per 100 kb of nucleotide sequence. (No such tracts are expected based on the dinucleotide composition.) It is seen that human DNA contains several-fold more of such sequences per 100 kb than most other eukaryotes. Surprisingly, *C.elegans* sequences are sparsely populated with long oligoadenosine tracts, even though the nematode sequence files have a base composition of 64% AT. None of the prokaryotic sequence files examined here contained a long oligoadenosine sequence.

Figure 4. Counted number of occurrences of $-BA_nB-$ plus the complementary sequence $-VT_nV-$ in the reported strand of the *E.coli*, yeast, *C.elegans* and human sequence files listed in Table 1, plotted versus n , the tract length. The solid line in the figure shows the expected number of such tracts calculated from the dinucleotide compositions of the respective files. The dotted curves show the expected number ± 1 SD. The absence of a point at a particular value of n means that no tracts of that length were observed.

For human DNA, then, it is possible that the overabundance of oligopurine tracts is simply a secondary effect of the overabundance of oligoadenosine tracts. To examine this possibility the number of long oligopurine tracts that also contain tracts of 10 or more contiguous adenosines was counted. It was seen that such tracts account for only about one-third of the total number of long human oligopurine tracts (data not shown), and therefore cannot explain the large majority of oligopurine tracts. It should also be noted that the excess of oligoadenosine tracts cannot be accounted for by the excessive number of oligopurine tracts. From a consideration of base composition alone, only ~1 in 8000 tracts of 15 contiguous purines is expected to be an uninterrupted adenosine tract. As seen by comparison of the human data plotted in Figures 3 and 4 the actual number of long oligoadenosine tracts is substantially larger than that.

It should also be noted that tracts of the simple alternating dinucleotide A–G do not contribute significantly to the total number of oligopurine sequences. In the human sequence files examined there are only 28 tracts of alternating A–G ≥ 15 bases in length out of 1371 total oligopurine tracts. In mouse there are 16 AG tracts, in chicken one AG tract, in *Drosophila* four AG tracts, in *C.elegans* 12 AG tracts and in yeast one AG tract. In all cases such tracts are a very small percentage of total long oligopurine tracts. Thus it can be emphasized that the overabun-

dance of oligopurine tracts observed in eukaryotes can not be attributed to simple repeating sequences.

DISCUSSION

Oligopurine tracts in eukaryotes

From the data presented in this paper it is clear that an overabundance of long oligopurine tracts is a common feature of eukaryotic DNA, ranging from simple, unicellular organisms with relatively small genomes, through organisms of intermediate complexity, up to mammals with relatively large genomes (Fig. 1; Table 2). This phenomenon can not be explained as a secondary result of biases in base composition, dinucleotide frequency, or the occurrence of oligoadenosine tracts. The ratio of observed/expected long oligopurine tracts is remarkably similar for all eukaryotic species examined (Table 2). It should be emphasized that the wide occurrence of this phenomenon, along with the quantitative similarity across vast phylogenetic distances (Fig. 2), is itself a strong argument that such sequences serve an important role in the expression of eukaryotic cell function.

In contrast, the number of oligopurine tracts is not greater than would be expected on the basis of dinucleotide frequency in the prokaryotic organisms examined here, even though for one organism, *E. coli*, ~25% of the total genome has been analyzed (Fig. 1; Table 2). The fact that yeast DNA possesses such a bias even though the yeast genome is only about twice as large as the *E. coli* genome indicates that the overabundance of oligopurine tracts is not dependent on genome size or on the density of structural genes in the genome. It has previously been shown that no increase in oligopurine tracts is found in the prokaryotic bacteriophages T7 and λ (6). Thus, in no prokaryotic organism examined so far has a strong bias in favor of oligopurine tracts been observed. Although the number of prokaryotic organisms (including phages) examined is relatively small one can tentatively identify the overabundance of long oligopurine tracts as one way in which the genomes of eukaryotes differ from those of prokaryotes.

In this paper I have divided oligopurine tracts into two categories: 'long' (15–30 nucleotides) and 'very long' (>30 nucleotides). This is justified by the difference in the frequency of occurrence of these categories in the eukaryotes. Whereas, as mentioned above, an overabundance of long tracts occurs in all eukaryotic species examined, very long oligopurine tracts occur much more frequently in the two mammalian species than in the other eukaryotes (Table 2). Even the chicken sequences examined here are much poorer in very long oligopurine tracts, even though the genome size of mammals is only two to three times larger than that of birds (11). Thus, again, the occurrence of very long tracts seems to be uncorrelated with the size of the genome. Because of the much greater occurrence of very long sequences in the two mammalian species, generalization about the importance of the origin and role of such sequences is more problematic. Such sequences may have arisen serendipitously due to differences in mammalian polymerases or recombination events vis-a-vis non-mammalian organisms, and they may have no role to play in gene expression. Or, on the other hand, very long sequences may be utilized for a peculiarly mammalian cell function. The database is currently too small to make even a tentative judgment, although in individual cases there has been speculation about the importance of particular very long oligopurine tracts (12–16).

Structural implications

Oligopurine tracts are versatile structural elements. The synthetic, double-stranded homopurine–homopyrimidines poly dA–poly dT and poly dG–poly dC have structural parameters that are different from canonical B-form DNA (7,8). Shorter tracts of homopurines embedded in longer DNA polymers show varying properties: although it requires only four contiguous guanosines to achieve a structure similar to that of poly dG–poly dC, oligoadenosine tracts approach the structure of poly dA–poly dT only very slowly (17). Although at one time the synthetic polymers were thought to be unable to form nucleosomes, it has been shown by our laboratory that they can interact with histones in a normal manner (18–20). It has also been shown in recent years that oligopurine tracts with mirror-repeat or near-mirror repeat symmetry can form H-DNA, in which one repeat melts and the freed pyrimidine strand folds back to form a triple stranded complex with the second repeat (9). Further work has shown that other triple strand structures can be formed consisting of two purine strands and one pyrimidine strand or a mixture of purines and pyrimidines (10). Additionally, it has been shown that guanosine-rich oligopurine strands, similar to those found in the telomeric regions of eukaryotic genomes, can form four-stranded structures (4). Apparently several of the aforementioned structures can exist simultaneously in the same molecule of DNA (21).

H-form DNA and other multiple-strand structures proposed in the past few years have a significant limitation: the element which folds back must immediately precede or follow the element to which the melted strand will hydrogen bond. This avoids topological difficulties that would result if a constrained single strand had to wrap around a constrained double strand. Because the two elements are contiguous, the oligopurine region has to be fairly lengthy. Typically oligopurine tracts of 40–60 bp are used in experimental systems investigating such structures, and the demonstration of triple-stranded forms in naturally-occurring sequences has so far been limited to very long oligopurine tracts. However, as shown in this manuscript, very long oligopurine tracts are limited mostly to mammalian species, with few occurring in the sequenced DNA of chicken and lower eukaryotes. Long tracts of 15–30 nucleotides are found to be overabundant in all eukaryotic species examined, but they are too short to be candidates for H-form DNA or the other multiple-strand structures that have been elucidated in recent years. Furthermore, very few of the long oligopurine tracts examined in this study have mirror-repeat symmetry (data not shown).

Since the frequency of their occurrence (about one long oligopurine tract per 2 kb) is high, it is very likely that the tracts are structural elements. But because of their comparatively short length they seem unable to form the multi-strand structures reported in recent years. And although double stranded synthetic homopurine polymers have unusual structures, shorter mixed-sequence oligopurines do not (22). Therefore it would seem that long oligopurine tracts are useful to the cell because of a hitherto unknown structural property. One possibility is that these structural elements do form triple-stranded structures, but that, while close in space in folded chromatin, the third strand is remote along the DNA sequence from the double-stranded sequence to which it binds (Fig. 5). As mentioned earlier, this type of structure would present some topological difficulties since the third strand would have to unwind one turn for every turn it wraps around the double strand. However, if a significant length of the the third

Figure 5. Possible structural effect of long oligopurine tracts. In this speculative model one strand of a melted oligopurine tract forms a triplex with another tract which is close in space, but either distant along the sequence or else on another molecule entirely. The topology of the constrained DNA would then require the strand to 'disentangle' one turn for each turn of triple helix formed.

strand were melted, the entropic cost of such a structure might not be too high. If this were the case long oligopurine tracts might be able to act as 'hydrogen bond sinks' to stabilize melted structures that result from transcription, replication, and other cellular processes.

The demonstration of alternative DNA structures occurring naturally *in vivo* has proven over the years to be quite difficult. It would likely be the same for the structural scheme discussed above. However, progress has been made by using model systems *in vitro* and *in vivo* for other alternative structures, and our laboratory is currently constructing a model system to examine the energetics of remote triple helix formation. Of course, long oligopurine tracts might not exert their effect through the structural scheme outlined above. However, because of their overabundance and universal distribution throughout the eukaryotic kingdom it is nearly certain that they do have a role to play in the cell, which remains to be articulated.

Oligoadenosine tracts

Poly dA-poly dT has a structure significantly different from B form DNA (7). Circular dichroism studies of synthetic polydeoxynucleotides show that a tract containing as few as three contiguous oligoadenosines is in a conformation recognisably different from B-form DNA, but that the structure continues to change as the tract increases in length and does not stop changing until the tract is >50 adenosines long (17). It is seen in Figure 4 that the human genome is greatly biased to contain many more long oligoadenosine tracts of 15–30 nucleotides than would be expected to occur on the basis of dinucleotide composition alone. The overabundance of long oligoadenosine tracts is not a secondary result of the overabundance of oligopurine tracts: the two are independent phenomena. It seems likely that long oligoadenosine tracts play a positive role or roles in human cells, and that the structural properties of the tracts are exploited for this purpose.

Curiously, however, long oligoadenosine tracts are not over-represented in all eukaryotes examined, as are long oligopurine

tracts, nor do they occur to the same extent in species which overproduce them, as do oligopurine tracts. Figure 4 shows that *C.elegans* is similar to the prokaryote *E.coli* in that both have about the same number of long oligoadenosine tracts as predicted from dinucleotide composition. This occurs even though the nematode DNA has a base composition of 64 versus 49% for the prokaryote DNA. Yeast DNA contains significantly more long oligoadenosine tracts than does *C.elegans* even though both are eukaryotes of approximately the same base composition. The number of such tracts per 100 000 nucleotides varies significantly for the eukaryotic species examined (Table 2).

One conclusion that can be drawn from the differential occurrence of long oligoadenosine tracts is that some species find them useful and some species do not. Given the known structural properties of oligoadenosine tracts and their sequence homogeneity versus the unknown structural properties of oligopurine tracts and their sequence heterogeneity, the sporadic occurrence of the former serves to highlight the universal occurrence of the latter.

ACKNOWLEDGEMENT

This work was supported by grant GM36343 from the National Institutes of Health.

REFERENCES

- 1 Watson, J.D. and Crick, F.H.C. (1953) *Nature* **171**, 737–738.
- 2 Wang, A.H.-J., Quigley, G.J., Kolpak, F.J., Crawford, J.L., van Boom, J.H., van der Marel, G. and Rich, A. (1979) *Nature* **282**, 680–685.
- 3 Wells, R.D., Collier, D.A., Hanvey, J.C., Shimizu, M. and Wohlrab, F. (1988) *FASEB J.* **2**, 2939–2949.
- 4 Palecek, E. (1991) *Crit. Rev. Biochem. Mol. Biol.* **26**, 151–226.
- 5 Schroth, G.P., Chou, P.-J. and Ho, P.S. (1992) *J. Biol. Chem.* **267**, 11846–11855.
- 6 Behe, M.J. (1987) *Biochemistry* **26**, 7870–7875.
- 7 Peck, L.J. and Wang, J.C. (1981) *Nature* **292**, 375–378.
- 8 McCall, M., Brown, T. and Kennard, O. (1985) *J. Mol. Biol.* **183**, 385–396.
- 9 Mirkin, S.M., Lyamichev, V.I., Drashlyak, K.N., Dobrynin, V.N., Filippov, S.A. and Frank-Kamenetskii, M.D. (1987) *Nature* **330**, 495–497.
- 10 Cheng, Y.-K. and Pettitt, B.M. (1992) *Prog. Biophys. Mol. Biol.* **58**, 225–257.
- 11 Lewin, B. (1990) *Genes IV*, p 467, Oxford University Press, Oxford.
- 12 Poncz, M., Schwartz, E., Ballantine, M. and Surrey, S. (1983) *J. Biol. Chem.* **258**, 11599–11609.
- 13 Liu, Q.-R. and Chan, P.K. (1990) *J. Mol. Biol.* **212**, 453–459.
- 14 O'Neill, D., Borschlegel, K., Flamm, M., Castle, M. and Bank, A. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 8953–8957.
- 15 Bacolla, A. and Wu, F. (1991) *Nucleic Acids Res.* **19**, 1639–1647.
- 16 Lu, Q., Wallrath, L.L., Allan, B.D., Glaser, R.L., Lis, J.T. and Elgin, S.C.R. (1992) *J. Mol. Biol.* **225**, 985–998.
- 17 Gudibande, S.R., Jayasena, S.D. and Behe, M.J. (1988) *Biopolymers* **27**, 1905–1915.
- 18 Jayasena, S.D. and Behe, M.J. (1989a) *Biochemistry* **28**, 975–980.
- 19 Jayasena, S.D. and Behe, M.J. (1989b) *J. Mol. Biol.* **208**, 297–306.
- 20 Puhl, H.L., Gudibande, S.R. and Behe, M.J. (1991) *J. Mol. Biol.* **222**, 1149–1160.
- 21 Panyutin, I.G. and Wells, R.D. (1992) *J. Biol. Chem.* **267**, 5495–5501.
- 22 Jayasena, V.K. and Behe, M.J. (1991) *Biopolymers* **31**, 511–518.