Species-specific patterns of DNA bending and sequence

Jeffrey D. VanWye, Edward C. Bronson and John N. Anderson*

Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA

Received July 12, 1991; Revised and Accepted September 12, 1991

ABSTRACT

Nucleotide sequences in the GenEMBL database were analyzed using strategies designed to reveal species-specific patterns of DNA bending and DNA sequence. The results uncovered striking species-dependent patterns of bending with more variations among individual organisms than between prokaryotes and eukaryotes. The frequency of bent sites in sequences from different bacteria was related to genomic A + T content and this relationship was confirmed by electrophoretic analysis of genomic DNA. However, base composition was not an accurate predictor for DNA bending in eukaryotes. Sequences from *C. elegans* exhibited the highest frequency of bent sites in the database and the RNA polymerase II locus from the nematode was the most bent gene in GenEMBL. Bent DNA extended throughout most introns and gene flanking segments from *C. elegans* while exon regions lacked A-tract bending characteristics. Independent evidence for the strong bending character of this genome was provided by electrophoretic studies which revealed that a large number of the fragments from *C. elegans* DNA exhibited anomalous gel mobilities when compared to genomic fragments from over 20 other organisms. The prevalence of bent sites in this genome enabled us to detect selectively *C. elegans* sequences in a computer search of the database using as probes *C. elegans* introns, bending elements, and a 20 nucleotide consensus sequence for bent DNA. This approach was also used to provide additional examples of species-specific sequence patterns in eukaryotes where it was shown that $(A)_{\geq 10}$ and $(A \cdot T)_{\geq 5}$ tracts are prevalent throughout the untranslated DNA of *D. discodium* and *P. falciparum*, respectively. These results provide new insight into the organization of eukaryotic DNA because they show that species-specific patterns of simple sequences are found in introns and in other untranslated regions of the genome.

INTRODUCTION

Sequence-directed bending of DNA causes local variations in the structure of genomes (1–7). The major determinant of bent DNA arises from oligo (dA) tracts which deflect the axis of the helix from a straight line. In bent sites, the tracts are spaced near the

helical repeat of DNA which causes their deflections to sum producing large unidirectional curvature. Bent molecules are most often identified by their anomalously slow electrophoretic mobility through polyacrylamide gels and this technique has been used to show that bent helices are characteristic of some promoters (3, 8–11), DNA attachment sites (12), and origins of replications (3, 13–16). The importance of the structure to function is also supported by deletion analysis (8, 9, 11, 13, 17) and by the observations that synthetic bent DNA can substitute functionally for natural bending elements in control regions for replication (17) and transcription (18).

Although the basic rules that govern intrinsic DNA bending are well established (1–7), our knowledge of the biological significance of this unusual structure is incomplete. For example, there is considerable variation in the extent of bending in promoters (10) and origins of replication (14, 16), and bent DNA has been found in genomic sites that are distal to known regulatory elements (4). In fact, due to the limited number of natural molecules that have been examined, it is not clear whether the structure is related to a common locus in the genome, to a common function, or even to common evolutionary origin. An approach that could be used to begin to address these questions consists of identifying the major bending elements in all genes and genomes.

We have previously described algorithms for predicting helical bending from nucleotide sequence and have shown that the calculated extent of bending is the same as that deduced from gel electrophoretic analysis of synthetic and natural molecules (14, 19, 20). The method has also yielded predictions (data not shown) which agree with recently published results of circularization experiments (15, 21, 22) and electron microscopy (23, 24). The uncertainties concerning the significance of bending elements in genomes prompted us to extend the investigation of DNA bending to include all nucleotide sequences in the GenEMBL database. Below we report our findings concentrating on bacteria and selected eukaryotes with A + T-rich introns. The results of this investigation uncovered striking species-specific patterns of DNA bending and simple repeated sequences in non-protein coding DNA.

METHODS

Computer Analysis

The algorithm for calculating DNA bending from nucleotide sequence has been previously described (14). Three dimensional

* To whom correspondence should be addressed

coordinates of the helical axis are calculated along the sequence using parameters of the wedge model for bent DNA (25). From these projections, a measure of curvature called ENDS ratio is computed. The ENDS ratio is defined as the ratio of the contour length of a nucleotide segment along the axis to the shortest distance between the ends of the segment. ENDS ratios were computed at a window width of 200 nucleotides and at a window step of 10 nucleotides. An example of this analysis for two sequences is shown in Figure 5. The analysis was also performed at a window width of 120 nucleotides and the results (data not shown) were consistent with those presented here.

ENDS ratios were computed for sequences from Release 67 of the GenBank and EMBL Sequence Libraries (denoted GenEMBL). The subdivisions of the library that were used are shown in Table 1. Sequences from the structural RNA, synthetic, and unannotated categories were omitted. Although this database consists of a total of 41,704 sequences, ENDS ratio analysis could only be performed on sequences with lengths greater than 200 bp. 31,497 sequences (76% of GenEMBL) with an average length of 1,626 nucleotides were analyzed. Preliminary work, conducted on Release 66 of GenEMBL, gave results which were consistent with those reported here. For random sequence DNA, 1,000 segments containing 1,500 nucleotides each were generated for each of the A+T contents given in Table 1 and Figure 3. For the data in Table 1 and Figure 2, the average of all ENDS ratios per sequence and the 5 highest ENDS ratio peaks per sequence were computed. As each peak was located during analysis, all ENDS ratio windows with centers that fall within the windows of any of the located peaks were eliminated from consideration during the search of the next highest peak. Sequence coordinates and ENDS ratio values were determined for the 5 peak centers. The positions of the peaks were then correlated with the positions of relevant landmarks in the sequences. Landmarks were identified from GenEMBL documentation and from the original literature. Known and suspected bacterial promoters were grouped together in Figure 2. Sequences were included in this category if they satisfied at least one of the criteria listed in the legends of Figures 1 and 2 of reference 26 or if they were similar in sequence to a known promoter. In order to evaluate relationships between bending and phylogeny, GenEMBL was divided into the major subdivisions listed in Table 1 and then sequences from individual organisms were retrieved by using the first 3 letters of the locus name (e.g. ECO = *E. coli*; CEL = *C. elegans*). The reliability of this retrieval process was verified in all cases. For each locus name abbreviation, the total number and average length of the sequences, the number of ENDS ratio peaks > 1.2 and > 1.5 per 100 Kb of DNA, and the average A+T content of the peaks were computed. No attempt was made to delete duplicate sequences from data presented in Table 1 and in Figures 1 and 3. Duplicate sequences were deleted from the data in Figures 2, 6, 8, and 9.

GenEMBL resides on a cluster of Digital Equipment Corp. (DEC) 3000 series VAXstations running the VAX/VMS operating system. DNA trajectory and ENDS ratio analyses in Table 1 and Figures 1–3 were performed on these computers using programs written in the C Programming Language along with FORTRAN subroutines from the Genetics Computer Group (GCG) Sequence Analysis Software Package (27). Additional processing was performed on a DEC VAX 11/ 780 computer and a Sun Microsystems, Inc. Sun-4/490 computer running the UNIX operating system using standard UNIX software tools and programs written in the C Programming Language augmented

with subroutines from the IMSL mathematics library (28). For the results presented in Figure 8, sequence searches were carried out by the method of Pearson and Lipman (29) using the FastA program included in the GCG Sequence Analysis Software Package. Positional autocorrelation analysis was performed as described previously (4).

Electrophoretic Analysis

Bacteria, *S. cerevisiae*, and *C. elegans* were grown in liquid culture prior to isolating genomic DNA by standard procedures. The *Z. mays* DNA was prepared from seedling root tips and DNAs from the vertebrates mentioned in the text were from erythrocytes or liver. In order to obtain an index of genomic bending, each DNA sample was digested separately with EcoRI, BamHI, and HaeIII and the fragments were analyzed by two-dimensional electrophoresis as described by Anderson (4). Fragments (30–40 μ g) produced by EcoRI and BamHI digestions were separated according to size by electrophoresis through 1.1% agarose gels at 22°C and then according to size and shape by electrophoresis at 4°C for 36 hours through 3.5% polyacrylamide gel slabs. HaeIII fragments were analyzed on gels containing 1.7% agarose and 7% polyacrylamide. The gels containing the HaeIII fragments and most of those with BamHI digests are not shown in the paper because they gave results which were consistent with those in Figures 4 and 7. To provide a reference diagonal, the genomic fragments were run in the presence of a 123-bp DNA ladder from Bethesda Research Laboratories. In some experiments, the first dimension agarose gels were incubated at 22° for 30 min with 20 μ g/ml of 4, 6-diamidino-2-phenylindole (DAPI) prior to electrophoresis on polyacrylamide. We have shown that DAPI is effective at straightening of the curvature of A-tract bending structures (30).

RESULTS

Strategies for Searching the GenEMBL Database

Table 1 shows an analysis of DNA bending for all nucleotide sequences in the major subdivisions of the GenEMBL database and for computer generated random sequences with various base compositions. The magnitude of bending is expressed as the ENDS ratio, a computer derived value defined as the ratio of the contour length of a segment of the helical axis to the shortest distance between its ends. The average ENDS ratios for the GenEMBL sequence subdivisions are similar to each other and are also similar to random sequences with 50–70% A+T. An ENDS ratio > 1.5 is used as a criteria for a strong bending element in this study. This value is, on the average, 7 standard deviations above the ENDS ratio means of the GenEMBL subdivisions. For comparison, the ENDS ratios that characterize the bending loci in kinetoplast DNA from *L. tarentolae* (31), the lambda origin of replication (15), the major control region in adenovirus (4, 19), and yeast ARS I (4, 13, 14) are 2.96, 1.37, 2.01, and 1.54, respectively. The frequencies of strong bending elements in the GenEMBL libraries also fall within the range of random sequence DNA.

The general strategy of grouping database sequences from several organisms into single categories as in Table 1 has been traditionally used for comparing sequences in regulatory elements, genes, and genomes. However, it is likely that species differences are masked by this approach since sequences from a single organism typically comprise a small fraction of the sequences within a category. In this paper, we present the alternative strategy

Table 1. DNA Bending in GenEMBL Sequences and Random Sequence DNA

	ENDS RATIO (Mean $\pm$ SD)	Frequency of Strong Bending Elements Per 100 Kb
GenEMBL Subdivision		
Phage	1.075 $\pm$ 0.059	1.97
Bacterial	1.071 $\pm$ 0.066	3.09
Invertebrates	1.081 $\pm$ 0.106	5.29
Vertebrates		
Nonmammalian	1.066 $\pm$ 0.080	1.39
Rodents	1.055 $\pm$ 0.046	0.54
Primates	1.058 $\pm$ 0.052	0.82
Other Mammals	1.055 $\pm$ 0.051	0.78
Plants	1.080 $\pm$ 0.068	2.62
Organelle	1.095 $\pm$ 0.084	5.64
Viruses	1.068 $\pm$ 0.056	1.20
Random Sequences		
0% A + T	1.000 $\pm$ 0.000	0.00
20% A + T	1.000 $\pm$ 0.000	0.00
40% A + T	1.035 $\pm$ 0.054	0.00
50% A + T	1.056 $\pm$ 0.039	0.06
60% A + T	1.075 $\pm$ 0.052	0.40
70% A + T	1.097 $\pm$ 0.068	3.13
80% A + T	1.114 $\pm$ 0.080	6.66
100% A + T	1.125 $\pm$ 0.083	9.53

of grouping sequences based on individual species. Although the analysis in Table 1 failed to reveal major differences in predicted bending among the GenEMBL subdivisions, striking differences were seen when sequences from individual species within the categories were compared as shown in Figure 1. The frequency of strong bending elements in the sequences from bacterial and eukaryotic species varied greatly from a low of zero to a high of 26 peaks per 100 Kb of DNA. In addition, the variation was nearly as great among different bacteria as it was among the eukaryotes. These species-specific patterns in DNA bending were investigated in detail.

DNA Bending in Bacteria

There were 130 strong bending elements in the sequences that comprise the bacterial database. The ENDS ratio values and positions of these bent sites relative to protein coding DNA are shown in Figure 2. For clarity, the map in the figure has been drawn to correspond to an averaged bacterial sequence as described in the legend. The distribution of bent sites is decidedly nonrandom with 17% of the sites found at the ends of genes and 70% of the sites in nongenic DNA. The nongenic sequences were divided into promoters, intergenic plus 3' gene flanking, and all other untranslated DNA. The promoter fraction contained 43% of the total sites and the 8 highest ENDS ratio peaks in the bacterial database. The preference for nongenic DNA extended to curvature scores that were as low as 1.2 for the sequences in the figure (see legend) as well as for 30 randomly chosen sequences from *E. coli* (data not shown).

About 1/4 of the bent sites in Figure 2 are from *E. coli* and *E. coli* sequences represent about 1/4 of the total sequences in the bacterial database. However, sequences from some bacteria are over represented in Figure 2 as compared to their frequency in the database while others are conspicuously absent (see Figure 1). Among species of bacteria, the average A+T content of genomic DNA varies over a wide range (~25%–80% A+T). Since we have previously reported a correlation between the A+T

Figure 1. Frequency of Strong Bending Elements in Nucleotide Sequences from Different Organisms. Each square gives the number of ENDS ratio peaks > 1.5 per 100 Kb of sequences from a single species. Open squares represent sequences from 25 different bacteria and squares with lines are from 40 eukaryotes. The first 3 letters of the locus sequence name was used to retrieve a species' sequences from the appropriate database subdivision listed in Table 1. All organisms in the GenEMBL categories whose sequences total > 50 Kb are shown except those in the phage, virus, and organelle subdivisions. Squares for selected organisms are identified by 3 letters: AFA = *A. faecalis*, CHR = chicken, MZE = *Z. Mays*, ECO = *E. coli*, YSC = *S. cerevisiae*, PFA = *P. falciparum*, BTH = *B. thuringiensis*, DDI = *D. discoideum*, and CEL = *C. elegans*.

content of a limited set of natural DNA molecules and their bending as measured by electrophoresis and computer analysis (4, 14), it was of interest to determine if this relationship also holds for bacterial genomes. In Figure 3A and B, the A+T content of genomes from 69 species is plotted against the average frequency of bent sites in their database sequences. For comparison, the open circles in the figures show the frequency of bent sites in computer generated random sequence DNA with varying A+T contents. A relationship between genomic A+T content and bending is evident for the bacterial DNA with sequences from G+C rich genomes exhibiting little or no bending to sequences from A+T rich genomes showing marked A-tract bending characteristics. In addition, the frequency of bent sites in DNA from nearly all of the organisms that exhibit some bending is substantially greater than the frequency of sites in random sequence DNA. This difference is especially evident with the strong bending elements plotted in Figure 3B. One factor that is responsible for this difference is that the A+T contents of the bending elements in the bacterial sequences averages 7% higher than the A+T contents of the genomic DNAs (data not shown). However, when this factor is taken into account by plotting the actual A+T contents of the bent segments in place of the genomic A+T values, (Figure 3C and D), the frequencies of bent sites in most of the bacterial groups are still higher than the frequency of sites in random sequences with the same A+T contents. Thus, although there is a direct relationship between genomic A+T content and bending in bacteria, the frequency of bent sites is higher in the natural DNA than in random sequence DNA of the same base composition.

The marked differences in predicted genomic bending shown in Figure 3 is supported by the electrophoretic analysis presented

Figure 2. Position of High ENDS Ratio Peaks Along Bacterial Sequences. Each symbol represents an ENDS ratio peak > 1.5 from a bacterial sequence and circled symbols are from *E. coli*. Plasmids were omitted from the analysis. The values for the peaks are given on the vertical axis. The sequence entries containing the peaks were used to derive the map at the bottom of the figure. These entries, on the average, contained 70% protein coding DNA with a mean gene length of 1,394 bp. Fifty randomly chosen sequences gave similar values. The horizontal arrow below the coding DNA was inserted so that all bent sites could be positioned on a single map. Bent sites to either side of the arrow were positioned in bp relative to the ends of genes. Those above the arrow were positioned by their fractional distances along individual coding sequences. 43% of the peaks were within 200 bp of known or suspected promoters and those peaks are indicated by the small vertical arrows. The remaining sites were within the 200 bp segments at the ends of genes (17%), in the central regions of coding DNA (13%), in intergenic plus 3' flanking sequences (18%), and in 'Other' untranslated DNA that was > 350 bp from known coding DNA (9%). The position of all intergenic sites are given in bp from the 3' ends of coding DNA. In these sequences, the preference of bent sites for nongenic DNA extended to ENDS ratio scores as low as 1.2: the relative frequencies of sites within nongenic DNA for ENDS ratio intervals of 1.2–1.4, 1.4–1.6, 1.6–1.8, 1.8–2.0, and > 2.0 were 45%, 56%, 66%, 82%, and 100%, respectively. A value of about 30% is expected if the sites were distributed at random (see above). Upon request, the magnitude and nucleotide position of the 5 highest ENDS ratio peaks in all GenEMBL bacterial sequences files can be provided in IBM-PC-readable form.

in Figure 4. In the experiment, genomic DNA fragments from *B. thuringiensis* (66% A+T), *E. coli* (49% A+T), and *M. luteus* (28% A+T) were analyzed by electrophoresis in two-dimensional gels in which the first dimension separates DNA on the basis of chain length and the second dimension separates molecules according to both length and shape (4). A 123-bp standard ladder was included in the *B. thuringiensis* and *M. luteus* samples so that migration of the genomic fragments could be compared. All fragments from *B. thuringiensis* migrated slower than the standard arc in the second dimension which is a characteristic of DNA molecules with some bent loci (4). There was an average of 8 ENDS ratio peaks > 1.2 per 10 Kb in the GenEMBL sequences from this organism. In contrast, none of the fragments from the *M. luteus* genome migrated slower than the standard arc and there were no ENDS ratio peaks > 1.2 for this species in the database. The position of these fragments relative to the standard is similar to that reported for poly dG·poly d·C and poly dG·C (4) which is consistent with the apparent lack of bending in the *M. luteus* genome. The *E. coli* sequences in the database exhibited intermediate values of bending (4.6 ENDS ratio peaks > 1.2 per 10 Kb) and approximately 75% of the fragments from *E.*

coli migrated slower than the DNA standard. Genomic DNA fragments from *A. faecalis* and *B. cereus* were also analyzed (data not shown) by the electrophoretic method and the relative migration of the bulk of these fragments, like those in Figure 4, could be predicted by the frequency of ENDS ratio peaks in their database sequences.

Bent DNA has been found in a number of bacterial promoters (3, 8–11) and the results in Figure 2 show that nearly half of the strongest bending elements in the bacterial database are in the vicinity of promoter-containing DNA. The centers of bending at these sites was, on the average, -50 bp relative to the site of transcription initiation (where known) and each promoter segment exhibited strong 10–11 base periodicities in A_3 and/or T_3 tracts (data not shown). Our results are also consistent with previous reports showing that not all promoters contain regions of strong intrinsic curvature (3, 10). For example, out of 22 *E. coli* promoters selected at random from a list compiled by Teng and Harvey (32), 11 had ENDS ratio peaks > 1.2 and only one had a value that was > 1.5 . For comparison, the average ENDS ratio for all *E. coli* sequences was 1.073 ± 0.059 . Previous computer studies that focused primarily on *E. coli* promoters

Figure 3. Relationship Between A+T Content and DNA Bending in Bacteria. Panels A and B—Genomic A+T content of 69 different bacteria is plotted against the frequency of bent sites in their sequences. 69 out of 71 organisms in the bacterial database whose sequences totaled > 15 Kb were included for calculating the > 1.2 ENDs ratio values in Panel A while a subset of these organisms whose sequences totaled > 50 Kb were used for calculating the > 1.5 ENDs ratio values in Panel B. Two species were omitted because their genomic base compositions have not been determined to our knowledge. Plasmids and insertion sequences were also omitted, when possible, because of the uncertain origins of these molecules. Genomic base compositions were obtained from references (33, 34). The open circles give the frequency of ENDs ratio peaks > 1.2 (Panel A) and > 1.5 (Panel B) in computer generated random sequences of the indicated base compositions. Panels C and D—The average A+T contents in the windows of the ENDs ratio peaks in the bacterial and random sequences from Panels A and B are plotted on the horizontal axes of Panels C and D, respectively.

revealed a rough correlation between the magnitude of promoter curvature and promoter strength (10). The results in Figures 3 and 4 seem to argue against the universality of this relationship because DNA fragments from G+C-rich genomes exhibit little or no bending. For example, in the 142 sequence entries from the 5 genomes with the lowest A+T contents in Figure 3A and B, there were no sites with ENDs ratios > 1.2. Since it is likely that promoters of all classes are represented in such a large ensemble of sequences (259 Kb), the results imply that not all strong promoters exhibit A-tract bending characteristics. To address this question more directly, we compared the promoter regions (± 300 bp) for ribosomal operons from *E. coli* to those from several bacteria with G+C-rich genomes (26–37% A+T). All P1 promoters from *E. coli* exhibited high bending scores ($\bar{X}$ ENDs ratios = 1.42 ± 0.09) in agreement with previous reports (10), while none of the promoter containing segments from the other organisms (GenEMBL accession nos. M26924, M35377, M55343, M58020, X00872, X53853, X53854 and X53855) had an ENDs ratio value > 1.12 ($\bar{X}$ = 1.08 ± 0.02). In addition, unlike the *E. coli* promoters, none of the others exhibited periodic patterns in A_3/T_3 . Since ribosomal RNA operons are presumably controlled by very active promoters in all bacteria, the results suggest that high curvature scores in promoters do not necessarily equate with high transcription rates.

Figure 4. Two-dimensional Separation of DNA Fragments from Bacteria. DNA fragments from *B. thuringiensis* (A), *E. coli* (B), and *M. luteus* (C) were separated on agarose gels (right to left) in the first dimension and polyacrylamide gels (top to bottom) in the second dimension. In Panels A and C, the 123-bp DNA ladder indicated by the arrows was incorporated into the samples to provide a reference diagonal for the genomic fragments. In a separate experiment (data not shown), the *E. coli* DNA was co-electrophoresed with the 123-bp ladder and the position of one fragment from the ladder is approximated by the arrow in B. Fragments were produced by digestion with EcoRI (A and B) or BamHI (C).

Species-Specific Bending in Eukaryotes

The large subunit of the RNA polymerase II gene from the nematode worm *C. elegans* (35) contained the 5 highest ENDs ratio peaks in the invertebrate subdivision of GenEMBL. Moreover, there were only two higher peaks in the entire GenEMBL database and these peaks were the archetypal bending elements from kinetoplast minicircles of the trypanosome *C. fasciculus* (36, 37). Figure 5A and B show projections of the helical axes of the segments with high ENDs ratios from the kinetoplast and from the polymerase gene. The predicted circular structures are strikingly similar with both showing near planar bending throughout the segments. This kinetoplast DNA has been visualized as a circle in electron micrographs with smooth unidirectional bending distributed over the entire region (23). The figure predicts that the *C. elegans* segment will share these properties. Figure 5C shows a plot of ENDs ratios and A+T content along the nematode gene. The gene contains 12 introns which range in length from 49 to 802 bp. The bent pattern extended throughout the length of each intron as shown by the ENDs ratio plot in the figure and by strong 10.2 base periodicities of A_4 tracts exhibited by each intron and by segments from each intron (Figure 5E). Strong bending was also observed throughout the 700 bp region upstream from the initiator codon and within the 3' untranslated DNA. Although the introns from this gene are A+T rich, the striking bending is not related to A+T richness *per se*, since the magnitude of curvature greatly exceeds the bending exhibited by A+T-rich random sequences as well as by all other A+T-rich introns in the database (Table 1, Figure 1, and see below). In addition, the bending is not directly related to exon sequences since the introns and gene flanking DNA from the RNA polymerase II gene of *D. melanogaster* (Figure 5D) and from all other organisms whose polymerase sequences have been entered into the GenEMBL database (data not shown), show little or no bending.

As a group, the sequences from *C. elegans* exhibited the highest frequency of strong bending elements in the GenEMBL database (Figure 1) with 43% of the sequences > 500 bp containing at least one site with an ENDs ratio > 1.5. In addition, sequences from this worm have a higher average ENDs ratio than sequences from all other organisms in GenEMBL (data not shown). In order to study the genomic distribution of the bent sites, all *C. elegans* sequences in the database with at least one intron > 200 bp were analyzed for ENDs ratios as in Figure 5C and the results are

Figure 5. Analysis of the RNA Polymerase II Gene. Panel A—Two dimensional projections of the helical axes of a 180 bp segment from a kinetoplast minicircle of *C. fasciculata* (AS. #M18394, reference 36). Each consecutive projection is rotated by 40 degrees. Panel B—Projections of 180 bp intron D segment of the RNA polymerase II gene from *C. elegans* starting at position 3810 in AS. #M29235 (reference 35). Panel C—ENDS ratios (top) and A+T content (bottom) of 200 bp windows are plotted as functions of position along the *C. elegans* RNA polymerase II gene. The site of the initiator codon (▲), poly A site (△), exons (■), introns (□), and exon flanking DNA (▨) are indicated. Panel D—ENDS ratio plot (top) and A+T content (bottom) along the RNA Polymerase II gene of *D. melanogaster* (AS #M27431, reference 38). Symbols are as in Panel C. Panel E—Positional autocorrelation analysis of A_4 tracts in segments of the *C. elegans* RNA polymerase II gene: 1—all introns, 2—all exons, 3—the 5' 50% of the intron sequences, 4—the 3' 50% of the intron sequences. The dashed vertical lines indicate 10 base intervals. Results similar to those shown in Panel E were obtained with T_4 tracts (data not shown).

summarized in Figure 6A. For comparison, the analysis was also performed on sequences from selected organisms with A+T-rich introns (Figure 6, B–D). Little bending occurred in translated DNA since the highest ENDS ratio peaks in 75–90% of the exons from each organism were less than 1.2. Likewise, little bending was observed in introns and gene flanking DNA from yeast and especially from corn and the cellular slime mold *D. discodium*. In contrast, over 80% of the introns and gene flanking segments from *C. elegans* had at least one ENDS ratio peak > 1.2 and over 40% were > 1.5. In most cases, the pattern of bending was similar to the RNA polymerase locus (Figure 5C) in that regions with elevated ENDS ratios extended throughout each intron and each gene flanking segment (data not shown). Frequently, different *C. elegans* genes exhibited multiple bending elements with similar ENDS ratio values (data not shown) as

Figure 6. DNA Bending in Intron-Containing Genes. All GenEMBL sequences from *C. elegans* (A), *Z. mays* (B), *D. discodium* (C), and *S. cerevisiae* (D) containing at least one intron > 200 bp were analyzed for ENDS ratios as in Figure 5. The sequences were divided into segments containing exons (■), introns (□), and flanking DNA (▨). The % of total segments examined is given on the vertical axis and the highest ENDS ratio peak per segment is plotted on the horizontal axis. The flanking segments contain both 5' and 3' untranslated regions of transcription units and nontranscribed gene flanking DNA. ENDS ratios and A+T content of both classes of untranslated DNA were similar to each other (data not shown) and similar to intron segments (see figure). The number, A+T content in %, and length in bp of the introns were: Panel A—59, 63, 672; Panel B—25, 60, 481; Panel C—5, 88, 595; and Panel D—18, 66, 359. The average A+T content of the exons in Panels A through D were 53, 47, 66, and 57%, respectively.

exemplified by the multiple regions of unusually high curvature in different introns and flanking segments along the RNA polymerase II locus shown in Figure 5C. However, the 43 *C. elegans* genes analyzed in Figure 6 are found on all six chromosomes of the worm and there is no clear relationship between chromosome map position and extent of bending (data not shown).

The two-dimensional electrophoretic procedure was used to provide independent evidence for the strong bending character of the *C. elegans* genome. As shown in Figure 7A and B, a substantial fraction of the genomic DNA from *C. elegans* exhibited marked bending characteristics since a large fraction of the fragments migrated substantially slower on polyacrylamide gels than the DNA standards and the bulk of the fragments from yeast, corn, and chicken (Figure 7D–F). The magnitude of the extreme electrophoretic anomaly exhibited by some of these fragments is at least as great as that reported for bending elements from kinetoplast DNA (31, 33, 34) and considerably greater than any of the bacterial fragments shown in Figure 4. Electrophoresis in the presence of DAPI, which is effective at straightening the curvature of bent DNA (30), largely eliminated the abnormal

Figure 7. Two-dimensional Electrophoretic Analysis of Eukaryotic DNAs. EcoRI fragments from *C. elegans* (A–C), *S. cerevisiae* (D), *Z. Mays* (E), and chicken (F) were separated by two-dimensional electrophoresis as in Figure 4. Note the multiple slow moving fragments from *C. elegans* in A and in the close up photograph of the top of the gel in B. In C, the first dimension agarose gel was incubated with DAPI prior to electrophoresis on polyacrylamide. In F, the 2 small spots near the top of the gel are bent repetitive DNAs (30) and the smudge at the bottom of the picture resulted from an imperfection in our camera lens. In separate experiments (data not shown), the DNAs were co-electrophoresed with the 123-bp ladder. The position of one fragment from each ladder is approximated by the arrows. The nature of the *C. elegans* DNA that is below the arrows in A–C is currently under investigation.

migration of the *C. elegans* fragments (Figure 7C). The general gel patterns exhibited by yeast, corn, and chicken DNA were also observed with genomic DNA from 18 other eukaryotes which represented the major vertebrates classes and two invertebrate phyla (data not shown). These results provide additional evidence for the unusual structure of the *C. elegans* genome.

Species-Specific Sequence Patterns

The sequence motif that gives rise to bent DNA is prevalent in the *C. elegans* genome (Figures 1, 5, 6). Our goal was to derive a simple sequence for this motif that would preferentially recognize *C. elegans* DNA in a computer-aided search of the database. The oligo A_3-6 tracts in the introns of the *C. elegans* RNA polymerase II gene are spaced at a 10.2 bp periodicity (Figure 5E) and most of the tracts are flanked on the 3' side by TTC/G and on the 5' side by C or G. Using this information, the 20 base sequence shown in Figure 8A was tested for its ability to match *C. elegans* sequences in the invertebrate subdivision of GenEMBL. The probe preferentially matched *C. elegans* sequences when compared to the bulk of the database sequences and when compared to sequences from *D. discoideum* and the malarial parasite *P. falciparum*. The specificity for *C. elegans* DNA was also seen with probes consisting of natural bent sites from *C. elegans* (see, for example, Figure 8B) and from other organisms (data not shown) as well as with 5 out of 6 complete introns that were randomly chosen from *C. elegans* sequences

Figure 8. Species-Specific Sequence Patterns. The four sequence probes shown at the bottom of Panels A–D were used to search the top strands of the sequence entries in the invertebrate database using the FastA program of Pearson and Lipman (29). The intron probe in Panel B was a 200 base bent segment from intron F (sequence positions, 5620–5820) of the *C. elegans* RNA polymerase II gene (see Figure 5C). The % of total database sequences for C, *C. elegans*; D, *D. discoideum*; P, *P. falciparum*; and T, total sequences in the best 30, 60, and 120 matches are plotted on the vertical axis. The 3 bars over each small letter give these values from left to right, respectively. The bars for total sequences are repeated 4 times for clarity. 25 of the 32 perfect matches to the (A·T)₁₅ probe were to *P. falciparum* sequences while 22 out of the 26 perfect matches to (T)₃₀ were for *D. discoideum*. Results similar to those shown in Panel C were obtained using (A)₃₀ in place of (T)₃₀ as a probe (data not shown). Over 98% of the matched sequences in the figure were in regions of untranslated DNA.

(data not shown). Visual inspection of these matches revealed sequence patterns characteristic of bent DNA.

The same strategy was used to provide additional examples of species-specific sequence patterns in invertebrates. Five randomly selected introns from *D. discoideum* preferentially matched sequences from this slime mold and 4 out of 5 introns from *P. falciparum* detected selectively sequences from this organism (data not shown). Visual inspection of the matched sequences revealed a preponderance of runs of (A)_{10–40} and (A·T)_{5–20} for *D. discoideum* and *P. falciparum*, respectively. Thus, the simple sequence probes (T)₃₀ and (A·T)₁₅ preferentially recognized sequences from these two organisms in a search of the invertebrate library (Figure 8C and D) as well as in a search of all GenEMBL sequences (data not shown). The specificity illustrated by the results in Figure 8 extended to closely related organisms in some instances. For example, 48 of the 120 matches to the (A·T)₁₅ probe were to *P. falciparum* and 13 were to 6 other species within this genus. However, this pattern was not observed for all Sporozoans since there were no A·T tracts longer than 5 nucleotides in the 10 sequence entries from *Eimeria* (data not shown). These results imply that the simple strategy outlined above can be used not only to identify species-related sequence patterns but also as a tool to study their phylogeny.

The program used for Figure 8 detects only the best match in a sequence and does not yield information on the prevalence of the patterns in individual sequences. Since a large fraction of

Figure 9. Frequency of Heteropolymeric and Homopolymeric A and T Tracts in Introns. The frequency of $(A \cdot T)_n$ tracts (Panel A) and $(A)_n + (T)_n$ tracts (Panel B) in the introns from C, *C. elegans*; D, *D. discoidium*, and P, *P. falciparum* are shown. The numbers on the horizontal axis give the length of the tracts in bp. The intron data shown here and the exon and gene flanking data discussed in the text (data not shown) were collected from all sequences from these organisms in the GenEMBL invertebrate database that contained at least one known intron.

the introns and gene flanking segments from *C. elegans* contain the sequence pattern that gives rise to bending (Figure 6), it was of interest to determine if such an extensive distribution of $(A)_n$ and $(A \cdot T)_n$ stretches occurred in the sequences from *D. discoidium* and *P. falciparum*. In the first approach taken to address this question, the number of tracts in sequences from the two organisms was determined for all sequences in the invertebrate library. There were 240 $(A \cdot T)_{\geq 5}$ tracts in the total of 248 Kb of *P. falciparum* DNA and 261 $(A)_{\geq 15} + (T)_{\geq 15}$ tracts in the 206 Kb of DNA from *D. discoidium*. These frequencies (about 1.1 tracts per Kb) are about 10-fold higher than the estimated frequency of Alu sequences in human genomic DNA (39). In the second approach, we determined the position, frequency, and length of tracts in all sequence entries containing at least one intron. The values for the intron regions are given in Figure 9. The species-specific nature of tracts is clearly evident from the analysis. There were 6.1 $(A \cdot T)_{\geq 5}$ tracts per Kb of *P. falciparum* intron DNA while these tracts were absent from the introns of *D. discoidium* and nearly absent from those of *C. elegans* (Figure 9A). In contrast, introns from *D. discoidium* were highly enriched in $(A)_{\geq 10}$ and $(T)_{\geq 10}$ (12.3 tracts per Kb of intron DNA) as compared to introns from *C. elegans* and *P. falciparum* (Figure 9B). Most of the introns from *D. discoidium* and *P. falciparum* (70% and 87%, respectively) contained at least one tract and nearly half contained two or more. In addition, values of about 1–2 tracts per 200 bp of intron DNA underestimate the true intron densities of the patterns because there were 5–6 fold more tracts containing 5–9 nucleotides (data not shown) than tracts with more than 9 nucleotides (Figure 9) and these shorter tracts also exhibited a degree of species specificity (data not shown). For these reasons, the matching sequences should be viewed as patterns rather than as discrete elements of fixed lengths. There were no tracts in exons (data not shown) and no more than 2% of the matches in Figure 8 occurred in exon DNA. Results similar to those shown in the

Figure 9 were seen with 3' and 5' untranslated regions of transcription units and nontranscribed gene flanking DNA (data not shown) which suggests that these intron patterns, like the pattern in *C. elegans*, are dispersed throughout the untranslated regions of the genomes.

DISCUSSION

Species-related patterns in intron length, number, and base composition have been noted previously (40–42). Simple sequence DNA has also been identified in a few introns (43, 44) and intron and gene flanking segments have been shown to exhibit a reduced sequence complexity as compared to exon DNA (45, 46). However, it has generally been assumed that most intron sequences are unique (see, for example reference 47). The results presented in this work provide fresh insight into the problem because they clearly show that species-specific sequence patterns are prevalent in introns from *C. elegans*, *D. discoidium*, and *P. falciparum* and further, that the patterns extend into other untranslated regions of the genomes. The absence of these patterns in exons from eukaryotes (Figures 5, 6, 8, 9) and the reduced frequency of bent sites in genes from bacteria (Figure 2) is expected since these simple sequences would impose severe coding constraints.

The sequences that form the patterns described here are reminiscent of highly repeated sequences described in other systems. For example, the satellite from crabs in the genus *Cancer* contains long $(A \cdot T)$ tracts (48, 49) and this simple sequence is prevalent in the genome of *P. falciparum*. A number of satellites are also composed of bent DNA (30, 50, 51) and bent sequences from white dove (30) and mouse (51) satellites preferentially matched *C. elegans* sequences when they were used as probes in an analysis like the one shown in Figure 8 (data not shown). However, satellites typically consist of tandem repeats which form sequence blocks in constitutive heterochromatin (39) while the simple sequences described here are apparently dispersed throughout the untranslated regions of the genomes.

Short stretches of moderately repetitive DNA are interspersed with single-copy DNA in most eukaryotes (52–54) and this genomic organization has been previously observed for *C. elegans* and *D. discoidium* (55, 56). The simple sequences characterized in this study appear to exhibit this short-period interspersion pattern with the repeated sequences in introns and gene flanking segments linked to unique sequences in the exons (Figure 5, 6, 9). Like moderately repeated DNA, each pattern forms a family of molecules whose sequences would be expected to appear related, but not identical, in a hybridization reaction using 200–600 bp fragments. Like many other highly and moderately repeated families, the sequences also differ in different organisms but are homogeneous within species which implies that either they arose late during evolution of these eukaryotic groups or arose early but were modified or amplified uniformly in a species-dependent manner. In either case, since these patterns comprise a substantial fraction of the untranslated sequences in these genomes (Figures 5–9 and text), it is clear that massive non-random changes must have occurred in the DNA since the divergence of these organisms from common stocks. During the preparation of this manuscript, we identified two additional species-specific sequence patterns by the general methods shown in Figure 8 (data not shown). An important goal is to continue these studies in order to determine whether the middle-repetitive fractions that have been found in nearly all eukaryotic genomes

by hybridization techniques can be defined in terms of simple sequence patterns by computer analysis.

The comparative approach used here has revealed limitations to current strategies used to study bent DNA. The unusual structure is most often defined by electrophoresis as an anomalously slow moving fragment or by sequence analysis as a series of oligo-A tracts spaced at 10–11 base intervals. Such DNA segments are indeed unusual in G+C rich genomes of bacteria but are prevalent in A+T rich bacterial genomes and especially in *C. elegans* DNA (Figure 1–7). The species-specific variation extends into presumptive regulatory regions at least in some cases such as in the ribosomal RNA promoters in bacteria (see text) and the upstream regions of the RNA polymerase II gene in eukaryotes (Figure 5C, D, and text). In addition, we suspect that the variation extends into the bulk of the regulatory loci at least in bacteria because essentially no bending was detected by computational or electrophoretic procedures in DNA fragments from G+C rich genomes (Figures 3 and 4). Thus genomic context effects should be considered when classifying bent molecules according to their unusual structures or possible functions.

More than a decade ago, Trifonov and Sussman used a computer-aided search to show that certain dinucleotides including AA and TT tend to be periodically distributed at 10–11 base intervals in selected eukaryotic but not prokaryotic sequences (1). These pioneering studies formed the foundation for much of the subsequent work on bent DNA. In addition, they have led to the now widely accepted view that bending serves to facilitate the packaging of eukaryotic DNA into nucleosomes. The results of the present study seem to question the original basis for this view since they clearly show that there is much more variation in bending among individual organisms than there is between prokaryotes and eukaryotes (compare Table 1 to Figures 1–7). Although the contribution of intrinsic bending to nucleosomal packaging remains unclear, the problem could be effectively addressed by studies centered around the genome of *C. elegans*.

ACKNOWLEDGEMENTS

We gratefully acknowledge critical reading of the manuscript by P.T.Gilham, S.F.Konieczny, and I.Tessman and helpful discussions with I.Tessman, H.E.Umbarger, B.L.Warner, H.L.Weith, R.Weissman, and especially P.T.Gilham. We also acknowledge C.Philbrook for typing the manuscript. Computer facilities were provided by the Purdue Engineering Computer Network and the Purdue AIDS Center Laboratory for Computational Biochemistry under Grant #NIH AI27713. E.C.B. was supported by Training Grant #1-T32-CA09634-01A1. This work was supported by Grant #R01 GM41708-03 from the National Institutes of Health.

REFERENCES

- Trifonov, E. N. and Sussman, J. L. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 3816–3820.
- Hagerman, P. J. (1985) *Biochemistry* **24**, 7033–7037.
- Trifonov, E. N. (1985) *CRC Crit. Rev. Biochem.* **19**, 89–106.
- Anderson, J. N. (1986) *Nucl. Acids. Res.* **14**, 8513–8533.
- Diekmann, S. (1986) *FEBS Letters* **195**, 53–56.
- Koo, H.-S., Wu, H.-M., and Crothers, D. M. (1986) *Nature* **320**, 501–506.
- Crothers, D. M., Haran, T. E., and Nadeau, J. G. (1990) *J. Biol. Chem.* **265**, 7093–7096.
- Bossi, L. and Smith, D. M. (1984) *Cell* **39**, 643–652.
- Gourse, R. L., deBoer, H. A., and Nomura, M. C. (1986) *Cell* **44**, 197–205.
- Plaskon, R. R. and Wartell, R. M. (1987) *Nucl. Acids. Res.* **15**, 785–796.
- McAllister, C. F. and Achberger, E. C. (1989) *J. Biol. Chem.* **264**, 10451–10456.
- Ross, W., Shulman, M., and Landy, A. (1982) *J. Mol. Biol.* **156**, 505–529.
- Snyder, M., Buchman, A. R., and Davis, R. W. (1986) *Nature (London)* **324**, 87–89.
- Eckdahl, T. T. and Anderson, J. N. (1987) *Nucl. Acids Res.* **15**, 8531–8545.
- Zahn, K. and Blattner, Z. K. (1987) *Science* **236**, 416–422.
- Eckdahl, T. T. and Anderson, J. N. (1990) *Nucl. Acids Res.* **18**, 1609–1612.
- Williams, J. S., Eckdahl, T. T., and Anderson, J. N. (1988) *Mol. Cell. Biol.* **8**, 2763–2769.
- Bracco, L., Kotlarz, D., Kolb, A., Diekmann, S., and Buc, H. (1989) *The EMBO J.* **8**, 4289–4296.
- Eckdahl, T. T. and Anderson, J. N. (1988) *Nucl. Acids Res.* **16**, 2346.
- Eckdahl, T. T., Bennetzen, J. L., and Anderson, J. N. (1989) *Plant Mol. Biol.* **12**, 507–516.
- Ulanovsky, L., Bodner, M., Trifonov, E. N., and Choder, M. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 862–866.
- Koo, H.-S., Drak, J., Rice, J. A., and Crothers, D. M. (1990) *Biochemistry* **29**, 4227–4234.
- Griffith, J., Bleyman, M., Rauch, C. A., Kitchin, P. A., and Englund, P. T. (1986) *Cell* **46**, 717–724.
- Muzard, G., Theveny, B., and Revet, B. (1990) *The EMBO J.* **9**, 1289–1298.
- Ulanovsky, L. and Trifonov, E. N. (1987) *Nature (London)* **326**, 720–722.
- Hawlye, D. K. and McClure, W. R. (1983) *Nucl. Acids Res.* **11**, 2237–2255.
- Devereux, Haeberli, and Smithies (1984) *Nucl. Acids Res.* **12**, 387–395.
- International Mathematical and Statistical Library MATH/LIBRARY, Fortran Subroutines for Statistical Analysis, Version 1.0, IMSL, Inc., Houston, Texas, 1987.
- Pearson, W. R. and Lipman, O. J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
- Williams, J. S., Dryden, G. L., and Anderson, J. N. (1991) submitted.
- Marini, J. C., Levene, S. D., Crothers, D. M., and Englund, P. T. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 7678.
- Teng, C.-S. and Harvey, S. C. (1987) *Nucl. Acids Res.* **15**, 4973–4980.
- Holt, J. G. (ed.), (1984) *Bergey's Manual of Systematic Bacteriology*. Williams and Wilkins, Baltimore/London, Vols. 1–4.
- DeMarsac, N. T. and Houmard, J. (1987) In Fay, P. and Van Baalen, C. (ed.), *The Cyanobacteria*. Elsevier, Amsterdam, 10, pp 251–302.
- Bird, D. M. and Riddle, D. L. (1989) *Mol. Cell Biol.* **9**, 4119–4130.
- Kitchin, P. A., Klein, V. A., Ryan, K. A., Gann, K. L., Rauch, C. A., Kang, D. S., Wells, R. D., and Englund, P. T. (1986) *J. Biol. Chem.* **261**, 11302–11309.
- Ray, D. S., Hines, J. C., Sugisaki, H., and Sheline, C. (1986) *Nucl. Acids Res.* **14**, 7953–7965.
- Jokerst, R. S., Weeks, J. R., Zehring, W. A., and Greenleaf, A. L. (1989) *Mol. Gen. Genet.* **215**, 266–275.
- Singer, M. F. (1982) *Int. Rev. of Cyto.* **76**, 67–111.
- Goodall, G. J. and Filipowicz, W. (1989) *Cell* **58**, 473–483.
- Csank, C., Taylor, F. M., and Martindale, D. W. (1990) *Nucl. Acids Res.* **18**, 5133–5141.
- Aissani, B., D'Onofrio, G., Mouchiroud, D., Gardiner, K., Gautier, C., and Bernardi, G. (1991) *J. Mol. Evol.* **32**, 493–503.
- Slightom, J. L., Blechl, A. E., and Smithies, O. (1980) *Cell* **21**, 627–638.
- Maroteux, L., Heilig, R., Dupret, D., and Mandel, J. L. (1983) *Nucl. Acids Res.* **11**, 1227–1242.
- Trifonov, E. N. (1990) In Sarma, R. H. and Sarma, M. H. (eds.), *Structure and Methods Vol 1: Human Genome Initiative and DNA Recombination*, Adenine Press, New York, Vol. 1, pp 69–78.
- Konopka, A. K. (1991) In Sarma, R. H. and Sarma, M. H. (eds.), *Structure and Methods Vol. 1: Human Genome Initiative and DNA Recombination*, Adenine Press, New York, Vol 1, pp 113–126.
- Lewin, B. (1990) *Genes IV* Cell Press, Cambridge, Mass, p 408.
- Sueoka, N. (1961) *J. Mol. Biol.* **3**, 31–40.
- Skinner, D. M. (1967) *Proc. Natl. Acad. Sci. USA* **58**, 103–110.
- Kodama, H., Saitoh, H., Tone, M., Kuhara, S., Sakaki, Y., and Mizuno, S. (1987) *Chromosoma* **96**, 18–26.
- Radic, M. Z., Lundgren, K., and Hamkalo, B. A. (1987) *Cell* **50**, 1101–1108.
- Davidson, E. H., Hough, B. R., Amenson, C. S., and Britten, R. J. (1973) *J. Mol. Biol.* **77**, 1–23.
- Davidson, E. H. and Britten, R. J. (1979) *Science* **204**, 1052–1059.
- Bouchard, R. A. (1982) *Int. Rev. Cyto.* **76**, 113–193.
- Firtel, R. A. and Kindle, K. (1975) *Cell* **5**, 401–411.
- Emmons, S. W., Rosenzweig, B., and Hirsh, D. (1980) *J. Mol. Biol.* **144**, 481–500.