Conservation and Periodicity of DNA Bend Sites in Eukaryotic Genomes

Yuko Wada-Kiyama and Ryoiti Kiyama
Department of Physiology, Nippon Medical School, Sendagi, Bunkyo-ku, Tokyo 113, Japan

Institute of Molecular and Cellular Biosciences, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113, Japan

(Received 11 January 1996)

Abstract

DNA bend sites appear every 680 bp on average in the human ε- and β-globin gene regions. Although most of their molecular nature has not been unraveled, a potential bend core sequence A2N8A2N8A2 (A/A/A) and its complementary T2N8T2N8T2 (T/T/T) appeared preferentially either in or very close to most of the bend sites, whereas other combinations of A2 and T2 dinucleotides, A/T/T + A/A/T, T/T/A + T/A/A and A/T/A + T/A/T, did not. The distances between any two of the core sequences in the entire β-globin locus showed a strong bias to a length of 701-800 bp and multiples thereof, suggesting that there is periodicity throughout the locus. This bias was not found for other combinations of A2 and T2. Again, this periodicity was identified in many eukaryotic genes, whereas the tendency was absent in mRNAs and prokaryotic as well as viral genomes.

Key words: periodicity; bent DNA; globin gene; nucleosome phasing

While mapping the DNA bend sites in the human ε-globin gene region, we found that they appeared at an average interval of 680 bp throughout the 7-kb region (ref. 1 and Fig. 1, upper). We also reported that the bend sites appeared in the immediate 5′ flanks of the human β-globin, ε-myc and erythropoietin receptor genes, the mouse βmn-globin gene, and two Alu family sequences in the human ε-globin gene region. Further analysis of the sites in the human β-globin gene region revealed that the distribution and spacing of the sites were very similar to those in the ε-globin gene region (ref. 2 and Fig. 1, lower). A total of 5 sites were mapped in the 5′ and 3′ flanks (βB-2, βB-1 and βB+3) and in the second intron (βB-1 and βB+2). Since most of the sites have three to over five repeats of short (dA)n (n ≥ 3) tracts at intervals of roughly ten or a multiple of ten nucleotides, we focused on the appearance of the short poly(dA) tracts with the consensus sequence A2N8A2N8A2 (A/A/A), and its complementary sequence T2N8T2N8T2 (T/T/T), which are potential bend core sequences. As shown in Fig. 1, the sequence A/A/A appeared roughly once in a few hundred bases and corresponded well with the bend sites in the human ε- and β-globin gene regions. Other sequences known for bending, A3N2T3N2 or G3N2C3N23, did not show such periodicity (data not shown). Similarly, other combinations of the A2 or T2 dinucleotides at ten base intervals, A/T/T (and A/A/T), T/T/A (and T/A/A) and A/T/A (and T/A/T), could not explain the bend sites. Further analysis of the sites in the Gγ-ε-δβ-globin gene region revealed that the periodicity was disturbed at the recombination junctions as well as at the 1st and 2nd exons, suggesting that the periodicity could be disturbed at regions of various biological functions (unpublished results). Therefore, analysis of the sites may reveal the potentially significant regions.

1. A/A/A Sequences in the β-globin Locus

To examine the periodicity of the A/A/A sequences in the human β-globin gene locus, we searched over 70 kb of the locus and scored the distances between any two of the sequences. Figure 2A shows the distribution of the distances in the range of 1 to 4000 bp. The distribution appears biased: major peaks were centered at 701-800 bp and its multiples (1301-1400 bp, 2101-2200 bp, 2801-2900 bp and 3501-3600 bp). Since the consensus A/A/A includes sequences that actually cannot bend because of being too A+T-rich, A2N8A2N8A2 (all Ns are As) for example, we subtracted A2N8A2N8A2 sequences where the total number of A and T in the total of 16 Ns exceeded eleven (referred to as A+T<11/16, Fig. 2B). Most of the peaks that appeared in Fig. 2A remained predominant. The relative frequency of the distances 701-800 bp and its multiples was calculated as follows: total frequency divided by the number of sequences.
DNA Bend Sites in Eukaryotic Genome

2. A/A/A Sequences in Eukaryotic, Prokaryotic and Viral Genes and mRNAs

We searched the nucleotide sequences from various sources for periodicity of the A/A/A sequences. Table 1 summarizes the survey of a total of 414,132 bp from human genomic sequences, 87,029 bp from other eukaryotic genomes, 96,005 bp from eukaryotic mRNAs, and 322,046 bp from *E. coli* and viral genomes. The normalized frequency of the periodicity of A/A/A (and its complementary T/T/T) sequences was 1.07 for human genes or 1.25 for other eukaryotic genes, suggesting that the bias for multiples of 701-800 bp is universal among eukaryotic genomes. On the other hand, the values for the eukaryotic mRNAs or *E. coli* and viral genes were 0.96 or 0.88, respectively, indicating no such bias. The relatively low correlation of periodicity among eukaryotic genomes could be explained by the fact that most of the sequences reported in the EMBL or GenBank databases focus upon the coding regions, which could interrupt the periodicity present in the intergenic region, as seen in the human ε-globin gene, where the distance between εB-1 and εB+1 was longer because of the presence of the 1st and 2nd exons (Fig. 1). This is characteristic of genes with many exons, such as the HPRT gene, which had relatively low scores. On the other hand, the human β-globin locus, which has a large intergenic region, scored relatively high (1.13). For A+T-rich sequences (β-globin, factor IX, serum albumin and factor XIIIb), the scores of the periodicity for less A+T-rich (A+T is equal to or less than 11, 10 or 11, respectively) A/A/A sequences were higher than those for the entire A/A/A sequences (Table 1). The presence of the bias in the human and other eukaryotic genes and human albumin (A+T≤11/16), human factor IX (A+T≤10/16), and human factor XIIIb (A+T≤11/16) genes, and the absence of bias in eukaryotic mRNAs and in *E. coli* and viral genes is shown in Fig. 3.

3. Biological Significance of the Bend Sites

We described here that sequences with a consensus A/A/A appeared periodically in the human β-globin locus and that the interval of the periodicity was strongly biased to 701-800 bp and multiples thereof. Since these sequences are located close to the most DNA bend sites, which also appear periodically with an average interval of about 700 bp, and some of them actually bend, it seems...
Figure 2. Distribution of the distances between any two of the A/A/A+T/T/T (A and B), A/T/T+A/A/T (C), T/T/A+T/A/A (D) and A/T/A+T/A/T (E) sequences in the range of 1 to 4000 bp for A/A/A (A and B) and 1 to 2000 bp for the others (C to E) in the human \(\beta\)-globin locus. A: A survey of all A/A/A+T/T/T sequences. B: A survey of the A/A/A+T/T/T sequences where \(A+T<11/16\). The positions of the peaks at 701-800 bp and its multiples are indicated by arrows.
Figure 3. Distribution of the distances between any two of the A/A/A sequences, or those where A+T is equal to or less than 11/16 (for human albumin and human factor XIIb) or 10/16 (for human factor IX). The positions of 701-800 bp and 1301-1400 bp are shown by arrows.
Table 1. Periodicity of A/A/A and T/T/T in DNA or RNA sequences.

| Sequence\(^a\) | Code | Length (bp) | GC\(\%\) | A/A/A + T/T/T (bp) | Score\(^c\) |
|---------------|------|-------------|-----------|---------------------|-------------|
| Human genes   |      |             |           |                     |             |
| factor IX     | HUMFIXG | 38059       | 39.0      | 214                 | 28         | 1.31 |
| [factor IX (A+T<10/16)] | | | | 86 | 12 | 1.49 |
| prosaposin    | HSPSAPA | 19985       | 49.1      | 46                  | 6          | 1.30 |
| pyruvate dehydrogenase | HSPDHtal | 17082 | 44.7 | 100 | 13 | 1.30 |
| \(\beta\)-globin | HSSHBB | 73326       | 39.5      | 1023                | 116        | 1.13 |
| [\(\beta\)-globin (A+T<11/16)] | | | | 533 | 64 | 1.20 |
| serum albumin | HSAHBGC | 19002       | 35.0      | 399                 | 45         | 1.13 |
| [serum albumin (A+T<11/16)] | | | | 122 | 17 | 1.39 |
| interferon \(\alpha/\beta\) RE | HSIFNAR | 32906 | 41.3 | 561 | 61 | 1.09 |
| HLA III       | HSHLA1467 | 10963 | 46.9 | 144 | 15 | 1.04 |
| TCR-C-\(\delta\) | HSTCRADCV | 97634 | 44.2 | 534 | 55 | 1.03 |
| factor XIIIb  | HSXBFIII | 33206 | 35.1 | 649 | 65 | 1.00 |
| [factor XIIIb (A+T<11/16)] | | | | 311 | 36 | 1.16 |
| von Willebrand factor | HSVWFAA | 21352 | 45.2 | 145 | 14 | 0.97 |
| HPRT          | HSHPRT8A | 56737 | 40.3 | 595 | 57 | 0.96 |
| tissue factor | HSTFPB | 13865 | 44.7 | 129 | 11 | 0.85 |
| Human genes (total) | | 414132 | 4539 | 486 | 486 | 1.07 |
| Other eukaryotic genes | | | | | |
| mouse \(\beta\)-maj-globin | MMBHBMJ | 6532 | 39.6 | 65 | 9 | 1.38 |
| chick emb. myosin HC | CHKMYHE | 31111 | 41.9 | 189 | 24 | 1.27 |
| rat GC-A       | RSGCA | 17517 | 50.5 | 48 | 6 | 1.25 |
| Dro. myosin HC | DROMHC | 22663 | 44.3 | 173 | 21 | 1.21 |
| chick ovalbumin | CHKOH | 9206 | 37.7 | 45 | 5 | 1.11 |
| Eukaryotic mRNAs (total) | | 87029 | 520 | 65 | 1.35 |
| E. coli and viral genes | | | | | |
| rpoBC         | ECPBOBC | 12357 | 51.3 | 20 | 3 | 1.50 |
| tRNA(thr) synthetase | ECTHRINF | 7784 | 50.6 | 16 | 2 | 1.25 |
| phage lambda  | LAMCG | 48502 | 49.9 | 142 | 15 | 1.06 |
| adenovirus 2   | AD2 | 35037 | 55.2 | 32 | 3 | 0.94 |
| SV40          | SV40XX | 5243 | 40.8 | 44 | 3 | 0.68 |
| \(rfB\)M, \(rfB\)K, \(gnd\) | ECGDPMPPP | 4278 | 41.8 | 13 | 0 | 0 |
| \(fda\), \(pg\), \(gapB\) | ECFDAFPK | 8029 | 50.3 | 12 | 0 | 0 |
| phage T7      | PODOT7 | 39936 | 48.4 | 11 | 0 | 0 |
| pyruvate dehydrogenase | ECACEX | 7749 | 52.6 | 5 | 0 | 0 |
| Herpes simplex virus | HEICG | 152260 | 68.3 | 1 | 0 | 0 |
| E. coli and viral genes (total) | | 322046 | 296 | 26 | 0.88 |

\(^a\) The sequences are shown where the total number of A/A/A + T/T/T is over 40 and the sequence length is over 5,000 bp for human and other eukaryotic genes, or 20 and 5,000 bp, respectively, for eukaryotic mRNAs. Further analysis of a total of 698,922 bp (human genes), 150,081 bp (other eukaryotic genes), 186,131 bp (eukaryotic mRNAs) and 322,046 bp (E. coli and viral genes), obtained by random sampling, revealed the same tendency (data not shown).  
\(^b\) Frequency of the distances of 701-800 bp and 1301-1400 bp/ that of the total.  
\(^c\) Relative frequency of the occurrence of distances of 701-800 bp and 1301-1400 bp normalized by random occurrence as 1.00.
that these A/A/A sequences are closely associated with the periodic appearance of the DNA bend sites. Furthermore, their periodicity seems to be universal among eukaryotic genomes. These bend sites might specify nucleosome phasing in the region rather than the actual folding angles, because the bending angles caused by each short poly(dA)-poly(dT) tract could be small at physiological temperatures.

A long-range correlation has been reported for nucleotide sequences of the genes with introns, and there was no correlation for genes without introns or mRNAs. The periodicity of nucleotide bases has also been shown by digesting genomic DNA with less base-specific nucleases or by computer analysis. Therefore, it seems natural to believe that there is an intrinsic signal, such as the bent DNA shown here, on genomic DNA that has been conserved through genome evolution. It should also be noted that the periodicity of DNA bend sites is disturbed at the introns, the junctions of genome rearrangements and the sites of open chromatin structures (unpublished results). Further analysis of the sites should reveal the functional regions that are important for replication, chromatin folding and other biological functions.

Acknowledgements: This work was supported by Grants-in-Aid from the Ministry of Education of Japan.

References

1. Wada-Kiyama, Y. and Kiyama, R. 1994. Periodicity of DNA bend sites in human z-globin gene region, J. Biol. Chem., 269, 22238–22244.
2. Wada-Kiyama, Y. and Kiyama, R. 1995. Conservation and periodicity of DNA bend sites in the human b-globin gene locus, J. Biol. Chem., 270, 12439–12445.
3. Shrader, T. E. and Crothers, D. M. 1989. Artificial nucleosome positioning sequences, Proc. Natl. Acad. Sci. USA., 86, 7418–7422.
4. Peng, C.-K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M., and Stanley, H. E. 1992. Long-range correlations in nucleotide sequences, Nature, 56, 168–170.
5. Hutchison, H. and Weintraub, H. 1985. Localization of DNAase I-sensitive sequences to specific regions of interphase nuclei, Cell, 43, 471–482.
6. Filipski, J., Leblanc, J., Youdale, T., Sikorska, M., and Walker, P. R. 1990. Periodicity of DNA folding in higher order chromatin structures, EMBO J., 9, 1319–1327.
7. Walker, P. R., Sikorska, M., and Witfield, J. F. 1986. Chromatin structure: Nuclear digestion profiles reflect intermediate stages in the folding of the 30-nm fiber rather than the existence of subunit beads, J. Biol. Chem., 261, 7044–7051.
8. Constanzo, G., Di Mauro, E., Salina, G., and Negri, R. 1990. Attraction, phasing and neighbour effects of histone octamers on curved DNA, J. Mol. Biol., 216, 363–374.
9. Bodnar, J. W. and Ward, D. C. 1987. Highly recurring sequence elements identified in eukaryotic DNAs by computer analysis are often homologous to regulatory sequences or protein binding sites, Nucl. Acids Res., 15, 1835–1851.