A signal encoded in vertebrate DNA that influences nucleosome positioning and alignment

Arnold Stein* and Minou Bina1

Department of Biological Sciences and 1Department of Chemistry, Purdue University, West Lafayette, IN 47907-1392, USA

Received September 2, 1998; Revised and Accepted December 13, 1998

ABSTRACT

Evidence is provided that the nucleotide triplet consensus non-T(A/T)G (abbreviated to VWG) influences nucleosome positioning and nucleosome alignment into regular arrays. This triplet consensus has been recently found to exhibit a fairly strong 10 bp periodicity in human DNA, implicating it in anisotropic DNA bendability. It is demonstrated that the experimentally determined preferences for nucleosome positioning in native SV40 chromatin can, to a large extent, be predicted simply by counting the occurrences of the period-10 VWG consensus. Nucleosomes tend to form in regions of the SV40 genome that contain high counts of period-10 VWG and/or avoid regions with low counts. In contrast, periodic occurrences of the dinucleotides AA/TT, implicated in the rotational positioning of DNA in nucleosomes, did not correlate with the preferred nucleosome locations in SV40 chromatin. Periodic occurrences of AA did correlate with preferred nucleosome locations in a region of SV40 DNA where VWG occurrences are low. Regular oscillations in period-10 VWG counts with a dinucleosome period were found in vertebrate DNA regions that aligned nucleosomes into regular arrays in vitro in the presence of linker histone. Escherichia coli and plasmid DNA, which fail to align nucleosomes in vitro, lacked these regular VWG oscillations.

INTRODUCTON

The vertebrate genome is packaged into chromatin, consisting largely of nucleosomes. Each nucleosome contains 147 bp of tightly wrapped DNA (1), a variable length (generally ranging from 20 to 80 bp) of linker DNA, and linker histone H1 (2). Nucleosomes in chromatin aggregate into higher-order structures, the exact nature of which is currently unclear (3,4). In addition to its obvious role in organizing DNA, chromatin structure is thought to be an important factor in gene regulation (5–7). It is reasonable to suppose that the DNA base sequence influences the nucleosome arrangements in chromatin, and that this influence could be of functional significance. DNA sequences that position nucleosomes have been known for some time (8). However, the extent to which the DNA sequence contributes to the nucleosome arrangements in chromatin remains unknown.

This problem has been examined by a number of different approaches. There have been high resolution mapping studies performed on a few well-studied genes. From such studies, it seems clear, for example, that nucleosome positioning is involved in the regulation of the mouse mammary tumor virus long terminal repeat promoter (9–13) and in the Xenopus 5S rRNA genes (14). In another approach, sequence alignment studies of nucleosomal DNA have revealed a weak out-of-phase 10 bp periodicity of A·T and G·C bp (15–17) and a 10 bp periodicity in the occurrence of AA dinucleotides (18–20), consistent with in-phase DNA sequence-dependent anisotropic deformability playing a role in nucleosome positioning (21,22). Competitive reconstitution and sequence selection experiments using nucleosome-size DNA fragments have additionally demonstrated affinity differences among different genomic DNA sequences for core histones (23,24). Because >95% of bulk genomic DNA differed in free energy for nucleosome formation, under those conditions, by only ±0.2 kcal/mol relative to synthetic random DNA, it was concluded that the DNA sequence has no appreciable influence on nucleosome positioning for the vast majority of nucleosomes (25). However, this type of experiment does not include contributions to positioning preferences resulting from the low energy positional shifting along DNA, which can readily occur at physiological conditions, where histones do not dissociate from DNA to an appreciable extent (13,25). Pointing to the importance of such nucleosome sliding mechanisms is the demonstration that nucleosome reconstitution on a 4.4 kb DNA fragment containing the chicken adult β-globin gene resulted in extensive variation in the affinity of core histones for different DNA regions. The range of variation was about two orders of magnitude greater than that of the initial binding of the histone octamer (26). In still another type of experiment, in vitro chromatin assembly using Drosophila embryo extracts, it was concluded that DNA sequence plays, at most, a minor role in nucleosome positioning (27). However, it is not clear that the ATP-dependent nucleosome mobilization that occurs in Drosophila early embryos (28) is reflective of most animal cells. Finally, the presence of linker histone H1, not included in most studies, can dramatically influence the nucleosome arrangement on DNA in vitro (29,30).

Despite the complexities of the experimental data, the preponderance of evidence obtained over many years from both in vitro and in vivo experiments favors the view that the DNA sequence does influence nucleosome positioning. However, it has been found that nucleosomes generally do not occupy unique positions

*To whom correspondence should be addressed. Tel: +1 765 494 6546; Fax: +1 765 494 0876; Email: astein@bilbo.bio.purdue.edu
on DNA. Rather, there appear to be preferred nucleosome locations. An example is provided by SV40 chromatin. In one experiment (31) native SV40 chromatin was isolated, digested with micrococcal nuclease, and the mononucleosome DNA fragments were isolated and cloned. By repeatedly selecting random clones from this library, and by sequencing the SV40 inserts, the extent to which nucleosomes overlapped each SV40 map position was determined. The peaks and valleys obtained for the clone overlap function (Fig. 1A, solid curve) reflect the tendency of each region of the SV40 genome to form or avoid forming nucleosomes, respectively. The variations observed might reflect DNA sequence preferences for nucleosome formation (31), but it has not thus far been possible to predict the observed nucleosome forming preferences from analysis of the SV40 DNA sequence (19).

In vitro chromatin assembly studies using a purified system have also revealed that there are preferred nucleosome arrangements on different DNA sequences. These experiments showed that some 2–3 kb vertebrate DNA sequences assembled into highly ordered, physiologically spaced, nucleosome arrays in the presence of linker histone, whereas other 2–3 kb sequences did not, under identical conditions (32–36). Moreover, an excellent correlation has been found for the tendency of different DNA sequences to package into ordered nucleosome arrays in vitro and in animal cell nuclei (34–36). These studies suggest that the DNA base sequence can influence nucleosome array formation, and it is of interest to know what base sequence patterns could be responsible for this phenomenon.

Recently, a fairly strong consensus sequence: non-T(A/T)G, which will be referred to as VWG (with complement CWB), with a period of 10 nt was found throughout human DNA. This unexpected discovery was made when a powerful machine learning approach was applied to a large set of human genes in order to train the computer to better recognize splice junctions and to find possible distinctive characteristics of intron and exon sequences (37). Because the 10 nt period could be reflecting an influence that nucleosomes may have on the rotational positioning of DNA in the nucleosome. Figure 2 shows the results of computations performed using: AA/TT (16,17) or TT, AA (19,20). By visual inspection, it is clear that VWGs occurring in a window ± approximately 50 bp (5 periods) from each VWG position. Overlapping occurrences were relatively rare, and were counted separately. The University of Wisconsin GCG Sequence Analysis Software (running on a VAX system) was used to find the pattern VWG, or other patterns when desired. All other programs were written in VAX FORTRAN. The period was allowed to vary within a small specified range. For chicken genomic DNA this range was 10.00–10.33; for SV40 DNA the range was as stated. The period variation was implemented by counting as periodic only nucleotide numbers, which satisfy the relation

\[ P_L \leq \text{nucleotide no.}/\text{period no.} \leq P_H, \]

where the nucleotide no. is the integer number of nucleotides measured from any VWG/CWB position to another VWG/CWB position, the period no. is the integer 1, 2, 3, 4 or 5, \( P_L \) is the low end of the period variation and \( P_H \) is the high end. For example, for a period range of 10.00–10.33, \( P_L = 10.00, P_H = 10.33 \), and the periodic nucleotide numbers are: ±10; ±20; ±30, ±31; ±40; ±41; ±50; ±51. Thus, 41 is a periodic nucleotide number for this range because 41/4 = 10.25, whereas 42 is not periodic because 42/4 = 10.50.

The histogram data were then averaged in a sliding (5 bp increments) 100 bp window to generate a continuous curve of average period-10 VWG count versus GenBank nucleotide number. Closely similar results were obtained using window sizes over the range 50–150 bp. It should be pointed out that an average count of 10 would be obtained only for a perfectly periodic region of DNA (i.e., a 100 bp or larger region having a VWG every 10 bp and nowhere else, which is quite rare in vertebrate DNA). Moreover, because the VWG consensus includes six common nucleotide triplets, it occurs frequently (one every 5.6 bp in SV40 DNA on average), and this leads to an appreciable background level of the average period-10 VWG count throughout vertebrate DNA. The total number of VWG occurrences in a sliding (5 bp increments) 600 bp window was also computed. While deviations from the average value of VWG occurrences were usually <10%, occasionally VWG-rich or VWG-poor regions of DNA were encountered. Thus, the period-10 VWG counts were normalized by a factor equal to the average VWG count for 600 bp/the VWG count in the center of a 600 bp window surrounding the nucleotide of interest. The rationale for normalization is that DNA regions poor (or rich) in VWG will statistically have fewer (or greater) period-10 VWGs. The 600 bp window size is arbitrary, but large enough to accommodate several nucleosomes and to damp out fluctuations. It was found that this normalization had only a small effect on the computational results.

**RESULTS AND DISCUSSION**

Period-10 VWG variations correlate with nucleosome forming preferences in SV40 chromatin

When this analysis was performed with SV40 DNA using a periodicity range of 9.2–10.5 bp (Fig. 1A), the VWG variations (dotted curve) appeared to follow the peaks and valleys of the experimentally determined curve of nucleosome forming preferences (solid curve) fairly well for nucleotides 300–3600. Using a smaller periodicity range of 10.0–10.5 bp or a range of 10.0–10.7 bp produced qualitatively very similar curves, but the relative peak heights changed somewhat (not shown). The total occurrences of VWG in a sliding 600 bp window are shown in Figure 1B. The nucleotide region 3600–4100, where the greatest deviation between the theoretical and experimental curves in Figure 1A occurs, corresponds to a region of the SV40 genome that is relatively depleted in VWGs. Correspondence between the experimental and theoretical curves is consistent with the hypothesis that nucleosomes tend to form in DNA regions that are relatively depleted in VWGs. For chicken genomic DNA this range was 10.00–10.33; for SV40 DNA the range was as stated. The period variation was implemented by counting as periodic only nucleotide numbers, which satisfy the relation

\[ P_L \leq \text{nucleotide no.}/\text{period no.} \leq P_H, \]

where the nucleotide no. is the integer number of nucleotides measured from any VWG/CWB position to another VWG/CWB position, the period no. is the integer 1, 2, 3, 4 or 5, \( P_L \) is the low
Figure 1. Experimental nucleosome positioning preferences in native SV40 chromatin, periodic VWG/CWB counts along SV40 DNA, and VWG occurrences in SV40 DNA. (A) Comparison of the variation in the clone overlap (solid curve), an experimental measure of nucleosome positioning preferences in native SV40 chromatin, with the variation in the window-averaged count of periodic VWG/CWB (dotted curve) along the SV40 DNA sequence (strain 776). Clone overlap denotes the fragment overlap function defined precisely in (31). It is essentially the number of clones with SV40 nucleosomal DNA inserts found in the 412 clone set that overlapped each site ±5 bp along the SV40 DNA sequence. The solid curve reproduces the curve shown in figure 5 of ref. 31. Periodic VWG/CWB denotes the number of VWG nucleotide triplets on both strands spaced at intervals consistent with a periodicity range of 9.2–10.5 bp ±52 bp from any VWG. The VWG counts were averaged in a 100 nt sliding window, and normalized as described in the text. The analysis was performed using a linear map of SV40 DNA; the computation of periodic VWG/CWB values began and ended ∼300 bp in from each end. (B) VWG occurrences on both strands of SV40 DNA in a sliding 600 bp window.

The AA/TT peaks and valleys do not generally correspond to those of the experimental curve (Fig. 2A), in contrast with VWG/CWB (Fig. 1). It is also clear that there is a lack of correspondence between the period-10 TT variations and the experimental curve (Fig. 2B). The period-10 AA variations do not follow the experimental curve well for nucleotides 1–2600 (the SV40 late region), but do seem to follow the experimental curve fairly well for nucleotides 2600–5000 (the SV40 early region) (Fig. 2C). Interestingly, the AA periodicities seem to complement the VWG/CWB periodicities. The period-10 VWG/CWB variations follow the experimental curve more closely in the SV40 late region, whereas the period-10 AA variations follow the experimental curve more closely in the SV40 early region, which is somewhat deficient in VWG triplets (Fig. 1B). A previous study had noted that AA dinucleotides exhibited a dominant 10.2 bp periodicity in a collection of cloned nucleosomal DNA fragments prepared from SV40 chromatin, whereas TT dinucleotides did not (19).

In order to provide a more objective and a quantitative measure of how well the experimental and theoretical curves correspond in Figures 1 and 2, the y axis values of the experimental and theoretical curves were plotted as points for each common x axis value (nucleotide position). This produces a scatter plot from which a correlation coefficient can be computed. If the peaks and valleys always correspond, the points should fall on a straight line.
with a positive slope, and the degree of scatter should be small. On the other hand, if there is no correspondence between the peaks and valleys of the two curves, the degree of scatter should be large, and the slope of the least squares line through the points should be near zero. The magnitude of the correlation coefficient must be between 0 and 1, with 0 meaning no correlation and 1 meaning superimposable curves. Superimposable curves would not be expected because of the statistical nature of the experimental curve alone. For example, if random 200 clone subsets of the 412 clone set (31) are selected, the relative peak heights for the solid curve in Figure 1 change somewhat, but the general trends are unchanged.

Figure 3 shows scatter plots for some of the data presented in Figures 1 and 2. For AA/TT, over the nucleotide range from 300 to 3600 (Fig. 3A), the points are quite scattered and the slope of the least squares line is negative. In this case, \( r = -0.22, P = 0.2 \), a weak negative correlation. In contrast, for VWG/CWB over the nucleotide range from 300 to 3600 (Fig. 3B), the points are less scattered and the slope of the least squares line is positive. The correlation coefficient is 0.52, with a \( P \) value of <0.001 (assuming that points >50 nt apart are uncorrelated), indicating that the correlation is statistically significant. Over the whole 5000 bp, \( r = 0.36, P < 0.02 \), and over nucleotides 300–2300, \( r = 0.60, P < 0.001 \). With the smaller periodicity range of 10.0–10.5 bp, correlation coefficients decreased by ~15%. Weak negative correlations were also obtained for TT, with \( r = -0.21, P = 0.2 \) over the nucleotide range 300–3600, or \( r = -0.26, P = 0.1 \) over the entire 5000 bp (scatter plots not shown). For AA, over the nucleotide range 300–2600, the SV40 late region, \( r = -0.11, P > 0.2 \), no significant correlation. However, over nucleotides 2600–4900, the SV40 early region, \( r = 0.52, P < 0.001 \). For VWG/CWB over the SV40 early region, \( r = 0.20, P = 0.2 \) (scatter plots not shown). Thus, VWG/CWB correlates best with the experimental data over the entire SV40 genome or over the SV40 late region. The dinucleotide AA correlates best over the SV40 early region, where the total number of VWG occurrences is low. For example, in a 600 bp window centered on SV40 nucleotide 3600, the number of VWG occurrences (periodic plus non-periodic) is only 86, whereas the average count is 107 (Fig. 1B). It may be of significance that the AA occurrences in a sliding 600 bp window vary widely in the SV40 late region (not shown), where the correlation with the experimental data was poor.

**Periodic oscillations in period-10 VWG predict nucleosome alignment and the value of the dinucleosome spacing periodicity**

We next examined whether VWG oscillations can provide an explanation for why some, but not other, 2–3 kb regions of vertebrate DNA assemble into highly ordered, physiologically spaced, nucleosome arrays, as was shown for the chicken ovalbumin gene (32). Using *in vitro* chromatin assembly, it was shown that the 2–3 kb region within the gene assessed by probe 3 formed a highly regular nucleosome array with a 195 bp spacing periodicity when linker histone was present (32). Twelve multiples of this unit repeat were resolved for the micrococcal nuclease ladder using this probe. In contrast, the 2–3 kb region immediately downstream was considerably less ordered (32). The variations in the period-10 VWG count across the chicken ovalbumin gene are shown in Figure 4. In the region extending ~1.5 kb upstream and downstream from the midpoint of probe 3 of (32) (Fig. 4), the oscillations appear to be more regular than in other regions. This is the region of the gene where the strong 195 bp nucleosome ladder was detected. The extent to which the oscillations are regular and the period can be readily assessed by Fourier analysis. It can be seen in Figure 5A that the region of the gene (GenBank positions 4200–7500 bp), where strong nucleosome ordering with a 195 bp periodicity occurred, exhibits a strong predominant Fourier amplitude at a period very close to 2 × 195 bp = 390 bp (arrow), the value of the dinucleosome length in the ordered array. The adjacent disordered region does not have a strong predominant Fourier amplitude (Fig. 5B). This result suggests that periodic dinucleosome oscillations in period-10 VWG might be responsible for the nucleosome ordering that occurs in this region of chicken ovalbumin DNA.

We then analyzed the chicken β-globin gene, for which the nucleosome ladders have been examined both *in vitro* (33) and *in vivo* (38). In both cases, this DNA assembles into fairly well ordered chromatin with an ~180 bp nucleosome spacing periodicity. Interestingly, in erythrocytes the bulk nucleosome spacing periodicity is >200 bp, noticeably longer than on the β-globin sequences (38), consistent with the β-globin gene DNA base sequence being responsible for the shorter repeat. Fourier analysis revealed a predominant amplitude peak at a period very close to 2 × 180 bp = 360 bp throughout the β-globin locus, again consistent with a dinucleosome period oscillation in the period-10 VWGs. The analysis of a 3 kb region from the 5′ end of the locus is shown in Figure 6A. In contrast, periodic oscillations in period-10 VWGs do not occur in pBR322 DNA (Fig. 6B), or in 10 randomly selected 3 kb regions of *Escherichia coli* DNA (not shown). These results are consistent with the finding that plasmid pBR322 and *E. coli* DNA fail to assemble into ordered nucleosome arrays in the purified *in vitro* system (36).

Another example of a correlation between a regular dinucleosome period oscillation in the period-10 VWG count and DNA sequence-dependent nucleosome alignment is provided by SV40 DNA. *In vitro* and in transfected cells, the SV40 early region assemblies into
Figure 4. Variations in the average periodic VWG/CWB count throughout the chicken ovalbumin gene. Periodic VWG/CWB denotes the number of VWG triplets on both strands spaced at intervals consistent with a period range of 10.00–10.33 bp ±52 bp from any VWG. The VWG counts were averaged in a 100 nt sliding window, and normalized as described in the text. Position denotes GenBank locus CHKOV AL nucleotide numbers. Probe 3, from (32), was a 235 bp restriction fragment that spanned the intron E–exon 5 junction.

Figure 5. Fourier analysis of the oscillations in periodic (10.00–10.33 bp) VWG/CWB in two regions of the chicken ovalbumin gene (Fig. 4). (A) The approximate ordered region assessed by probe 3 (32); GenBank (locus CHKOV AL) nucleotide positions 4200–7500. The arrow marks the period corresponding to 2× the value of the experimentally determined nucleosome repeat detected by probe 3 (2 × 195 bp = 390 bp). (B) Adjacent disordered downstream region; GenBank (locus CHKOV AL) nucleotide positions 6790–8790.

Figure 6. Fourier analysis of the oscillations in periodic (10.00–10.33 bp) VWG/CWB. (A) DNA region just upstream of the chicken β-globin gene cluster; GenBank (locus CHKHBBRE) nucleotide positions 6000–9000. The arrow marks the period corresponding to 2× the approximate value of the experimentally determined nucleosome repeat in this DNA region in nuclei of erythroid cells (38) (2 × 180 bp = 360 bp). (B) pBR322; nucleotide positions 300–3300.

Figure 7. Fourier analysis of the oscillations in periodic (9.2–10.5) VWG/CWB for SV40 DNA. (A) SV40 early region (nucleotides 2600–4900). The arrow marks the period corresponding to 2× the value of the experimentally determined nucleosome repeat in this region (34) (2 × 200 bp = 400 bp). (B) SV40 late region (nucleotides 300–2600).

Conclusion

The findings reported here suggest that the period-10 VWG signal, identified serendipitously by a machine learning approach does in fact contribute to nucleosome positioning and alignment, a significantly more regular nucleosome array than the late region; the nucleosome spacing periodicity was 200 bp (34). Consistent with these findings, the SV40 early region (Fig. 7A) exhibits a strong predominant peak (arrow) at 2 × 200 bp = 400 bp, the dinucleosome length, whereas the late region (Fig. 7B) does not exhibit a single predominant peak. More experimental data are needed to further assess the generality of the finding that regular dinucleosome period VWG oscillations are present in highly ordered regions of chromatin.

Recent work by others has suggested that nucleosome arrangements might affect chromatin higher-order structure. In non-solenoid models (39,40), disordered nucleosome arrays should form more globular (superbead-like) chromatin fibers than highly ordered arrays, which should generally form extended (solenoid-like) structures (39). Variations in nucleosome arrangements along large contiguous regions of DNA could lead to chromatin structures of many different types and shapes. For example, some chromatin structures might dictate that distal enhancers be placed close to appropriate promoters in 3D-space, obviating the need for a solenoid to be disrupted to allow DNA looping to occur. It is as postulated (37), and also that the dinucleosome may have some structural significance, as has been claimed previously (2). These findings also suggest that it may be possible to deduce information about chromatin structure computationally.
plausible that the DNA base sequence, through its effect on nucleosome arrangements, contributes to functionally significant chromatin structure variations along the chromosome.

ACKNOWLEDGEMENTS

We wish to thank Rick Westerman for help using the GCG software and VAX system, and Raúl Alvarez for assistance. This work was supported by US Public Health Service grants GM48341 (to A.S.) and AI29121 (to M.B.).

REFERENCES