J. Mol. Biol. (1986) 191, 659–675

Sequence Periodicities in Chicken Nucleosome Core DNA

Sandra C. Satchwell, Horace R. Drew and Andrew A. Travers

Medical Research Council
Laboratory of Molecular Biology
Hills Road, Cambridge CB2 2QH, England

(Received 14 April 1986)

The rotational positioning of DNA about the histone octamer appears to be determined by certain sequence-dependent modulations of DNA structure. To establish the detailed nature of these interactions, we have analysed the sequences of 177 different DNA molecules from chicken erythrocyte core particles. All variations in the sequence content of these molecules, which may be attributed to sequence-dependent preferences for DNA bending, correlate well with the detailed path of the DNA as it wraps around the histone octamer in the crystal structure of the nucleosome core.

The sequence-dependent preferences that correlate most closely with the rotational orientation of the DNA, relative to the surface of the protein, are of two kinds: ApApA/TpTpT and ApApT/ApTpT, the minor grooves of which face predominantly in towards the protein; and also GpGpC/GpCpC and ApGpC/GpCpT, whose minor grooves face outward. Fourier analysis has been used to obtain fractional variations in occurrence for all ten dinucleotide and all 32 trinucleotide arrangements. These sequence preferences should apply generally to many other cases of protein–DNA recognition, where the DNA wraps around a protein.

In addition, it is observed that long runs of homopolymer (dA)·(dT) prefer to occupy the ends of core DNA, five to six turns away from the dyad. These same sequences are apparently excluded from the near-centre of core DNA, two to three turns from the dyad. Hence, the translational positioning of any single histone octamer along a DNA molecule of defined sequence may be strongly influenced by the placement of (dA)·(dT) sequences. It may also be influenced by any aversion of the protein for sequences in the "linker" region, the sequence content of which remains to be determined.

1. Introduction

The principal unit of DNA packaging in eukaryotic chromosomes is the core nucleosome. In this structure, the DNA wraps twice around an octamer of histone proteins as a left-handed superhelix (Finch *et al.*, 1977; Richmond *et al.*, 1984). The midpoint of the bound DNA is termed the "dyad", at which point each copy of histones H2A, H2B, H3 and H4 can be related to its partner by an axis of twofold symmetry. The path of the DNA between histone octamers is not known, but it appears that variable lengths of "linker" DNA separate individual nucleosomes (Prunell & Kornberg, 1982; Widom & Klug, 1985). The location of the histone octamers on the DNA sequence can be determined with varying degrees of precision on different DNA sequences (Fitzgerald & Simpson, 1985; Thoma & Simpson, 1985; Ramsay, 1986; Thoma, 1986).

The structural and mechanical properties of DNA change according to its base sequence, and therefore the ability of a DNA molecule to bend around the histone octamer is thought to be a major determinant of nucleosome positioning. This view has been confirmed by the demonstration (Drew & Travers, 1985*a*) that one particular DNA sequence, which includes the *tyrT* promoter from *Escherichia coli*, adopts essentially the same bending geometry both when closed into a small circle and when bound to a histone octamer.

To describe the position of the DNA in such a protein–DNA complex, one must consider two variables: a translation, marking where along the DNA the histone octamer sits; and a rotation, which defines the local orientation of the DNA relative to the protein surface. Several particular DNA molecules, when reconstituted to form a nucleosome core, have exhibited a well-defined rotational setting (Simpson & Stafford, 1983; Ramsay *et al.*, 1984; Drew & Travers, 1985*a*; Rhodes, 1985). To determine the general nature of sequences that influence rotational position, a

representative population of core DNA molecules was isolated from chicken erythrocyte core particles and analysed by a technique known as statistical sequencing (Drew & Travers, 1985*a*). In this method the predominant locations of particular sequences, in a mixed population of DNA molecules, are detected by binding antibiotic drugs of known sequence-specificity to the DNA and then treating the drug–DNA complex with DNAase I. At sites where the drug binds, the rate of DNAase I cleavage is substantially reduced. By this method, the preferred distribution of such sites along the length of the DNA can be determined. An analysis of this sort showed that short runs of (A, T) are preferentially positioned with their minor grooves facing in towards the histone octamer, whereas short runs of (G, C) are positioned with their minor grooves facing out. The modulation of sequence content in these molecules was found to observe a period of 10·17(±0·05) bp†. In a uniform superhelix this period of 10·17 bp would correspond to the twist of the DNA helix in a local frame of reference.

The resolution of such an analysis allows a description of the most dominant sequence features of nucleosome core DNA, but it cannot describe the occurrences of all possible arrangements. Nor can it assess the detailed nature of helix curvature in regions such as those close to the protein dyad, where the path of the DNA deviates markedly from a uniform superhelix (Richmond *et al.*, 1984). To investigate these aspects of nucleosome core structure, we have cloned and sequenced 177 individual DNA molecules from the same sample that was analysed previously by statistical sequencing (Drew & Travers, 1985*a*). By use of these 177 examples, we have been able to identify regularities in the occurrence of certain dinucleotide and trinucleotide arrangements that may be equated with sequence-dependent preferences for bending a DNA double helix. By looking for irregularities in the occurrence of these same short sequences, we have been able to examine variations in the curvature of the DNA when it was bound originally to the protein. All of these variations agree qualitatively with the path of the DNA as seen in the crystal of the core nucleosome (Richmond *et al.*, 1984).

2. Materials and Methods

(a) *Cloning of core DNA molecules*

The same DNA preparation used by Drew & Travers (1985*a*) for analysis of sequence content by the method of statistical sequencing was cloned into M13 DNA for the purpose of sequencing many individual molecules. Nucleosome core particles were isolated in a pure form from chicken red blood cells by the method of Lutter (1978). We estimate that at least 80% of the available core particles were recovered from digestion of H5-stripped polynucleosomes (Drew & Travers, 1985*a*). The DNA in these particles was then removed from the protein by extensive digestion with proteinase K (0·5 mg/ml), 1% (w/v) sodium dodecyl sulphate at 37°C for 60 min, followed by extraction 3 times with phenol/chloroform and twice with ether. For the purposes of a previous experiment, this DNA was incubated with phage T4 polynucleotide kinase and [γ-^{32}P]ATP, and then applied to a 6% non-denaturing polyacrylamide gel. The radioactive core DNA migrated as a single band of about 146 bp. At least 90% of the radioactivity in this band was eluted from the gel, treated with alkaline phosphatase to remove ^{32}P from its 5′ ends, and then incubated with T4 polynucleotide kinase in the presence of 1 mM-ATP to place phosphate groups at its 5′ termini to facilitate the subsequent reaction with DNA ligase.

Phage M13mp8 in its double-stranded form was cut with *Sma*I (CCCGGG)‡ to expose a blunt-ended cloning site, treated with alkaline phosphatase to prevent self-ligation of the site, then incubated overnight at 15°C with 50 ng of purified core DNA and DNA ligase. This ligation mixture was used to transform *E. coli* cells of the TG1 strain that had been prepared by the following method (modified from Hanahan, 1983).

An overnight culture of TG1 cells in 2×TY medium (TY medium contains per litre, 10 g Bacto-tryptone, 10 g yeast extract and 5 g NaCl) was diluted to 1/100 concentration with 2×TY and incubated until the optical density at 650 nm reached 0·3 to 0·5. The cells were then centrifuged at 2000 revs/min for 10 min at 4°C, resuspended in 1/12 of their original volume with 10 mM-potassium 2-(*N*-morpholino)ethane sulphonic acid, 100 mM-KCl, 45 mM-$MnCl_2$, 10 mM-$CaCl_2$, 3 mM-cobaltic hexamine trichloride (pH 6·2). Dimethylsulphoxide was added to a concentration of 35 μl/ml and the mixture left on ice for 5 min. Next, a solution of 2·25 M-dithiothreitol, 40 mM-potassium acetate (pH 6·0) was added to a concentration of 35 μl/ml, and the mixture left on ice for 10 min. Dimethylsulphoxide was added for a second time (35 μl/ml) and the mixture left on ice for 5 min. Finally, the cells having been made competent, the ligation mixture of M13mp8 and core DNA was added to achieve transformation. Plating and subsequent preparation of clones were carried out as described by Bankier & Barrell (1983).

(b) *Length and alignment of the sequences*

A total of 239 clones was sequenced by the method of Sanger *et al.* (1977, 1980), which involves termination of a DNA polymerase extension reaction by dideoxynucleotides. The nucleotide [α-^{35}S]dATP was included as a radioactive label and the sequencing reaction analysed on a gradient gel, as described by Biggin *et al.* (1983). Of these 239 sequences, those that contained a double-length insert of about 290 bp, or those that contained a single-length insert of less than 142 bp, were not used. The mean length of the remaining 177 clones was found to be 145(±1·5) bp.

These 177 sequences range in length from 142 to 149 bp. The slight variation in length arises because of the method in which nucleosome cores are isolated from H5-stripped long chromatin. In this part of the isolation procedure, an enzyme (micrococcal nuclease) is used to trim both ends of the core DNA until it reaches the protein (Lutter, 1978). Since the activity of micrococcal nuclease is bidirectional, it would be incorrect to align the trimmed core DNA sequences about either their left-hand or right-hand borders: any alignment must be about their

† Abbreviation used: bp, base-pair(s).

‡ Sequence hyphens are omitted throughout for clarity.

centres. Thus, those clones of length 142, 143, 144, 147, 148 and 149 bp were shifted by one or two bases to achieve alignment about a central reference point of base 73·25. Those of length 145 or 146 bp were not shifted in any way.

All sequences were oriented as read from the sequencing gel, so that their 5′ ends correspond to the CCC side of the *Sma*I site, while their 3′ ends correspond to the GGG side. This point will prove important in the discussion below, concerning the nature of bases at either end of the cloned DNA. A listing of the 177 sequences is available on request to the principal author (S.C.S.) upon receipt of a tape suitable for writing in 1600 bpi VAX format.

(c) *Analysis by counting*

The frequencies of occurrence of specified short base sequences were determined by simple counting on a computer. For each of the 145 base steps in core DNA, the number of occurrences of each type of sequence arrangement were tabulated. In order to construct the plots shown below in Figs 1, 2 and 3, each of these arrays of count *versus* position was converted to a three-bond running average by averaging the counts at position x with those at positions $x-1$ and $x+1$, over positions $x = 2$ to 144. By definition, the *position* of any particular dinucleotide refers to the base-pair step, so that position 1·0 lies between bases 1 and 2, position 2·0 falls between bases 2 and 3, etc; whereas the position of any particular trinucleotide refers to the central base.

The 3-bond averages so obtained were then averaged about the centre of the sequence in order to improve the signal-to-noise ratio, and also to display the data concisely. This centre was taken to be 72·5 for dinucleotides or base 73·0 for trinucleotides. The assumption of such an analysis is that the centre of each sequence was originally very close to the axis of 2-fold symmetry through the histone octamer. When the DNA was removed from the protein, any information about its orientation relative to any asymmetry in the protein (e.g. amino acid modification) was lost.

In the course of our analysis, we have examined curves of occurrence for all 16 possible dinucleotides by treating separately such steps as ApA and TpT, or TpG and CpA. Thus, ApA steps in positions 1 to 145 can be studied separately from TpT steps in positions 1 to 145. Alternatively ApA steps in positions 1 to 72 can be considered in combination with TpT steps in positions 73 to 145; or else, TpT steps in positions 1 to 72 can be studied in combination with TpT steps in positions 73 to 145. We have studied such curves carefully, by both visual inspection and Fourier analysis, and find that the occurrences of 2-fold-related steps such as ApA and TpT are not significantly different from one another, except at the very ends of the molecule. There, in positions 1 to 3 and 143 to 145, we see a distinct preference for pyrimidines T and C at 5′ ends but purines G and A at 3′ ends. We attribute such an asymmetric distribution to use of the ligation site *Sma*I (CCCGGG) during the cloning step. It was noted in a previous experiment (Drew & Travers, 1985*b*) that pyrimidine bases ligate preferentially to the CCC end of a *Sma*I site, whereas purine bases ligate preferentially to its GGG end. Presumably the DNA molecules orient themselves in this way so as to maximize purine-to-purine overlap at each junction.

Thus, to the best of our knowledge, it appears that only 10 of the possible 16 curves of dinucleotide occurrence are unique within the present limits of the experimental data. These 10 unique steps are CpG, TpG = CpA, TpA, GpG = CpC, ApG = CpT, GpA = TpC, ApA = TpT, ApT, GpT = ApC and GpC. We have not examined all 64 possible trinucleotide arrangements, only the 32 unique ones, due to the infrequent occurrences of many of the trinucleotides even in a data set of 25,000 bases.

We have also considered the possibility of changing the present experimental alignment of our 177 core DNA molecules by various sorts of statistical algorithm. For this purpose, we constructed a 10×130 matrix of the likelihoods of occurrence for all possible dinucleotides in each of the central positions 8 to 137. When applied to the individual sequences that contributed to the matrix, this approach indicates that many sequences are misaligned by 1 to 2 bases in either direction (as would be expected). However, when such a likelihood matrix is applied to sequences on which the positions of histone octamers are known from experiments in solution (Simpson & Stafford, 1983; Ramsay *et al.*, 1984; Rhodes, 1985; Drew & Travers, 1985*a*), the calculation yields erroneous predictions of nucleosome location. So, until it can be shown that this alignment is in accord with known nucleosome locations, it would not appear trustworthy to alter the experimental alignment of our sequences by any statistical method.

(d) *Analysis by Fourier methods*

In order to provide a completely objective measure of the data, the raw occurrences of all 10 dinucleotide and 32 trinucleotide steps were analysed by Fourier transformation. In this procedure, one isolates the periodic waveforms that describe the occurrence of each and every step, in order to determine their amplitudes and phases. For the purposes of a Fourier calculation, the occurrences per position were not modified in any way from those obtained by a simple counting procedure: they were not converted into running 3-bond averages, nor were they averaged about the protein dyad.

All Fourier periodicities in the range 8 to 13 bp were sampled at intervals of 0·05 bp by the equation:

$$F = \sum_{x=3}^{x=143} C_x \exp(2\pi i hx/c).$$

Here the sum is taken over positions $x = 3$ to 143 of the aligned molecules in order to avoid end effects; the C_x are total occurrences for each particular dinucleotide step from which the mean value has been subtracted; $c = 2000$, and h is incremented from 150 to 250 to provide periodicities of 8·0 to 13·0 bases at intervals of 0·05. In a second set of calculations the Fourier periodicities were computed from positions $x = 3$ to 61, 84 to 143, thereby omitting the central 2 double-helical turns 62 to 83.

The results of such an analysis yield an amplitude F and a phase angle ϕ for each possible periodicity in the range 8·0 to 13·0. One may examine the variation in amplitude *versus* periodicity to see which waveform is the strongest, and describes the data most accurately. For some steps such as ApG or GpA, the amplitudes are of such a small value throughout the range that no waveform is favoured over another; but for five of the possible 10 steps (ApA, TpA, GpG, TpG and GpC), the strongest amplitudes are found in the range 10·15 to 10·31 bp, with a mean value of 10·2 bp.

How may one relate the amplitude and phase of such waves to the preferred positions of sequences in

nucleosome core DNA? The amplitude F must first be divided by the mean number of occurrences per step in the range $x = 3$ to 143, in order to place all dinucleotides on an equal basis, irrespective of their relative populations in chicken DNA. The normalized amplitude can then be converted to an amplitude per period by dividing by the number of periods = 141·0 (h/c). Finally, it can be converted to a fractional variation in occurrence, on the assumption that such variation follows a cosine wave, by dividing by the summation of $\cos^2(\phi)$ over one period of 10 steps, i.e. 5·0. The fractional variation in occurrence is thereby defined to be the maximal variation in occurrence of a particular sequence at any point along the curve, relative to a mean occurrence of 1·0.

The phase angle ϕ locates a series of maxima in the occurrence of a step at positions $x = 1{\cdot}0 + \text{period}\ (\phi/360°)$. Thus, for $\phi = 0°$ and a period of 10·2 bases, Fourier maxima are located at positions x = 1·0, 11·2, 21·4, 31·6, 41·8, 52·0, 62·2, 72·4 etc.

In the Fourier analysis of trinucleotide occurrences, the sum was taken over positions $x = 3{\cdot}5$ to 142·5 of the aligned molecules. The phase origin used for this analysis was position (or base-pair step) 0·5, which is equivalent to base 1·0. So that the phase angles for trinucleotides should be directly comparable with those listed for dinucleotides, all of the phase angles for trinucleotide listed in Tables 3 and 4 have been adjusted to a phase origin of position 1·0 by the subtraction of 18°.

3. Results

Since the result of previous experiments indicated that the distribution of bases in nucleosome core DNA would be non-random (Drew & Travers, 1985*a*), we have now cloned and sequenced 177 different core DNA molecules from chicken blood. These range in length from 142 to 149 bp about a mean of 145, to yield about 25,000 bases in all. Where different from 145 or 146 bp in length, these sequences have been aligned about their centres. Detailed analysis has shown that no single base or series of bases occupies the same position in all molecules. However, there are distinct preferences for certain dinucleotide and trinucleotide arrangements to lie in related positions separated by 10·2 bp, or one turn of the double helix. It appears, therefore, that the specificity of the histone octamer for different sequences in DNA is the sum of many different local interactions. One must consider the 145 bp of DNA as a whole, and not try to find a "consensus sequence" within some small part of the molecule.

We present the nature of sequence periodicities in nucleosome core DNA in two ways. First, we show simple plots of the occurrence of various dinucleotides and trinucleotides *versus* their position along the length of the aligned molecules. In order to bring out long-range periodicities in the data, every point has been averaged with its two nearest neighbours to reduce noise. In addition, the curves have been made symmetrical about the centre of the sequence, position 72·5, which was an axis of twofold symmetry when the DNA was bound to the histone octamer in solution. Having constructed such curves, we then look for periodicities of sequence at intervals of 10·2 bases, which is the helix repeat. We may also look for deviations from any regular periodicity as an indication of variations in the curvature of the DNA when it was bound to the protein (Richmond *et al.*, 1984).

Next, we provide a quantitative measure of the strength of such sequence preferences by taking the Fourier transform of the raw, unaveraged data points. This Fourier analysis yields the period, amplitude and phase of the waveform that best describes the variation in occurrence for each separate step.

It should be emphasized that the total occurrence of dinucleotides in our sample is closely equivalent to that of chicken blood DNA in bulk (Table 1). Both samples contain 57·5% A+T, and the occurrences of steps such as ApA/TpT, GpC and TpG/CpA are practically identical. Only one of the ten steps listed in Table 1 appears to be underrepresented: the step TpA, which exhibits a total occurrence in core DNA of 0·87 relative to chicken blood DNA in bulk. A possible explanation for this will be offered later.

(a) *Occurrences of the ten dinucleotides*

There are 16 possible dinucleotides, but some of these are related by the twofold axis that passes between the two strands of the double helix. Thus, for every ApA there is a TpT, for every TpG a CpA, and so on. It would not be unreasonable if such related steps were to occupy similar positions in core DNA in order to facilitate the tight bending of DNA around the protein. We have examined the separate occurrences of steps such as ApA and TpT (as described in Materials and Methods) and find that the differences between these occurrences do not appear to be significant at the present level of experimental resolution. In the analysis to follow,

Table 1

Dinucleotide frequencies in chicken erythrocyte DNA

Base step	Total occurrences	Frequency		Ratio core/total
		Core	Total	
GC	1257	0·049	0·052	0·94
CG	263	0·010	0·011	0·90
GG/CC	2470	0·096	0·102	0·94
TG/CA	4017	0·157	0·155	1·01
AG/CT	3804	0·149	0·145	1·03
GA/TC	3168	0·124	0·113	1·10
GT/AC	2602	0·102	0·104	0·98
AT	1724	0·067	0·072	0·93
TA	1392	0·054	0·062	0·87
AA/TT	4735	0·185	0·184	1·01
(A+T)		0·577	0·573	1·01

Experimentally determined dinucleotide frequencies for total erythrocyte DNA are by Swartz *et al.* (1962).

therefore, such related steps are taken to be equivalent, leaving us with ten unique examples.

The occurrences of all ten unique dinucleotides are plotted in Figure 1 as a function of their position in the core DNA molecule. Looking at these curves, the most striking result is a well-defined periodicity of the curve for ApA/TpT (Fig. 1(g)) between steps 1 and 56. In this region of the plot, the average spacing between successive peaks is approximately 10·2 bases. Maxima lie at positions 5, 16, 27, 36, 47 and 56, while minima are found at positions 10, 21, 32, 41 and 52.

From position 56 to the dyad at 72·5, the curve of occurrence for ApA/TpT does not follow the same periodicity. At positions 62 and 72, one would expect to find minima in a continuation of the previous pattern; instead, one finds maxima. At position 67 one expects a maximum; instead, there is a minimum. So, there has been an approximate reversal in phase of the ApA/TpT periodicity over a small region of DNA near the dyad.

There are further minor fluctuations in the shape of the curve at positions 26, 47 and 62, where the maxima look slightly "camel-humped". These minor fluctuations become more pronounced when a restricted set of sequences comprising only those clones in the size range 144 to 146 bp is examined (lower curve, Fig. 1(g)). A fourth notable point is the higher than average occurrence of ApA/TpT sequences near either end of the core DNA molecule, positions 1 to 20 (and the twofold-related 126 to 145).

The ApA/TpT maxima at steps 27, 36, 47 and 56 correspond well with the preferred placement of short runs of (A, T) as identified in a previous experiment: at positions 28·5, 38·5, 48·5 and 58·5 (Drew & Travers, 1985*a*). These correspond to places where the minor groove of the DNA originally faced approximately inward when the DNA was bound to the histone octamer. Similarly, the ApA/TpT minima at steps 10, 21, 32, 41 and 52 correspond to places where the minor groove faced approximately out and away from the protein. We can be sure of the true location of such steps from the X-ray structure analysis of the nucleosome core particle, where it is seen that the minor groove points out at position 72·5 (the dyad) (Richmond *et al.*, 1984). The true orientation of this DNA had also been deduced, prior to the crystal structure, from nuclease digestion measurements in solution (Lutter, 1978).

Of the nine other dinucleotides, the clearest periodicities are seen for the sequences GpC, GpG/CpC and TpG/CpA. All of these steps are modulated with a phase opposite to that of ApA/TpT. For example, the sequence GpC exhibits maxima at positions 12, 21, 31, 40 and 53. Such variations are less apparent, though still detectable, for sequences GpG/CpC and TpG/CpA. Step ApG/CpT is modulated weakly in a fashion similar to that of GpC, and step TpA similarly to that of ApA/TpT. Plots of occurrence *versus* position for sequences GpA/TpC, GpT/ApC and ApT appear almost flat. Lastly, the number of steps per position for the dinucleotide CpG is so low as to preclude interpretation.

To obtain a quantitative measure of amplitude and phase for all of the variations shown here, the raw counts of occurrence *versus* position (prior to any averaging) were analysed by Fourier transformation, separately for each of the ten dinucleotide steps. This analysis (Table 2) confirms and extends the conclusions drawn from simple visual inspection. The dinucleotides that produce periodic waveforms of the largest amplitude are GpC and ApA/TpT. For these dinucleotides the periods at which the maximum amplitudes are observed are 10·15 and 10·26 bp, respectively (Fig. 2), with corresponding fractional variations in occurrence of $\pm 20\%$ and $\pm 16\%$. The phase angle of $+13^\circ$ for GpC means that it is found where the minor groove of the DNA originally faced outward on the nucleosome core, whereas the phase of 180° for ApA/TpT means that this step is found where the minor groove faced inward. (By convention, a phase of 0° is "out" whereas a phase of 180° is "in" relative to the minor groove.) Other significant periodicities are apparent for the dinucleotides GpG/CpC and TpG/CpA, which exhibit the same phase ($+25^\circ$, -14°) as that of GpC but lesser variations of $\pm 8\%$ and $\pm 7\%$. The periodicity of TpA has the same phase ($+176^\circ$) as that of ApA/TpT, but a lesser variation of $\pm 12\%$. Finally, the step GpT/ApC exhibits a weak variation of $\pm 8\%$ that points both up and out (-68°).

It was noted above that sequence periodicities in the region between positions 62 and 83 of core DNA show an apparent reversal in phase, for sequences of the type ApA/TpT. Indeed, one may detect such an anomaly in other curves of occurrence *versus* position (e.g. the curve for GpC in Fig. 1(j)). Since the occurrences of sequence in this region can only detract from periodicities elsewhere in the core, it seemed advisable to do a second Fourier calculation omitting positions 62 to 83. The results are shown in the right-hand part of Table 2. By removing the irregularity near the dyad, most of the strong periodicities calculated for total core DNA (left-hand part) become even stronger than before. In addition, significant signals now are seen for steps CpG, ApG/CpT and ApT, the magnitudes of which are $\pm 15\%$, $\pm 5\%$ and $\pm 6\%$, respectively.

(b) *Occurrences of the 32 trinucleotides*

There are 64 possible trinucleotide arrangements, but for present purposes we consider complementary trinucleotides to be equivalent, thereby reducing the number to 32. The 177 core DNA molecules were analysed by Fourier transformation to obtain the best waveforms that describe the variations in occurrence of these 32 trimers. The results are shown in Table 3 and Figure 3. Following the discussion above, it would seem that the most reliable calculation of Fourier amplitudes is that omitting the dyad region, positions 62 to 83;

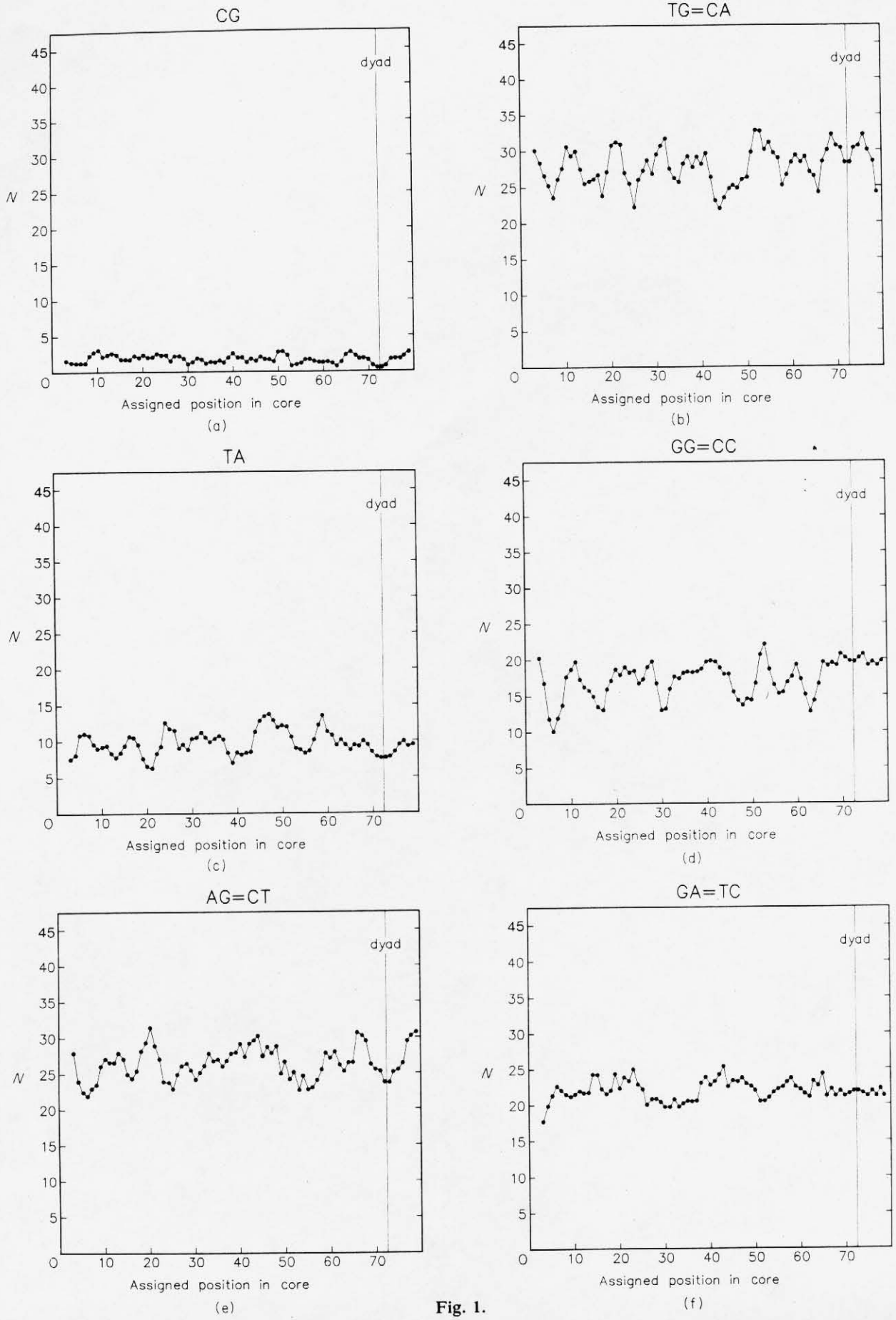
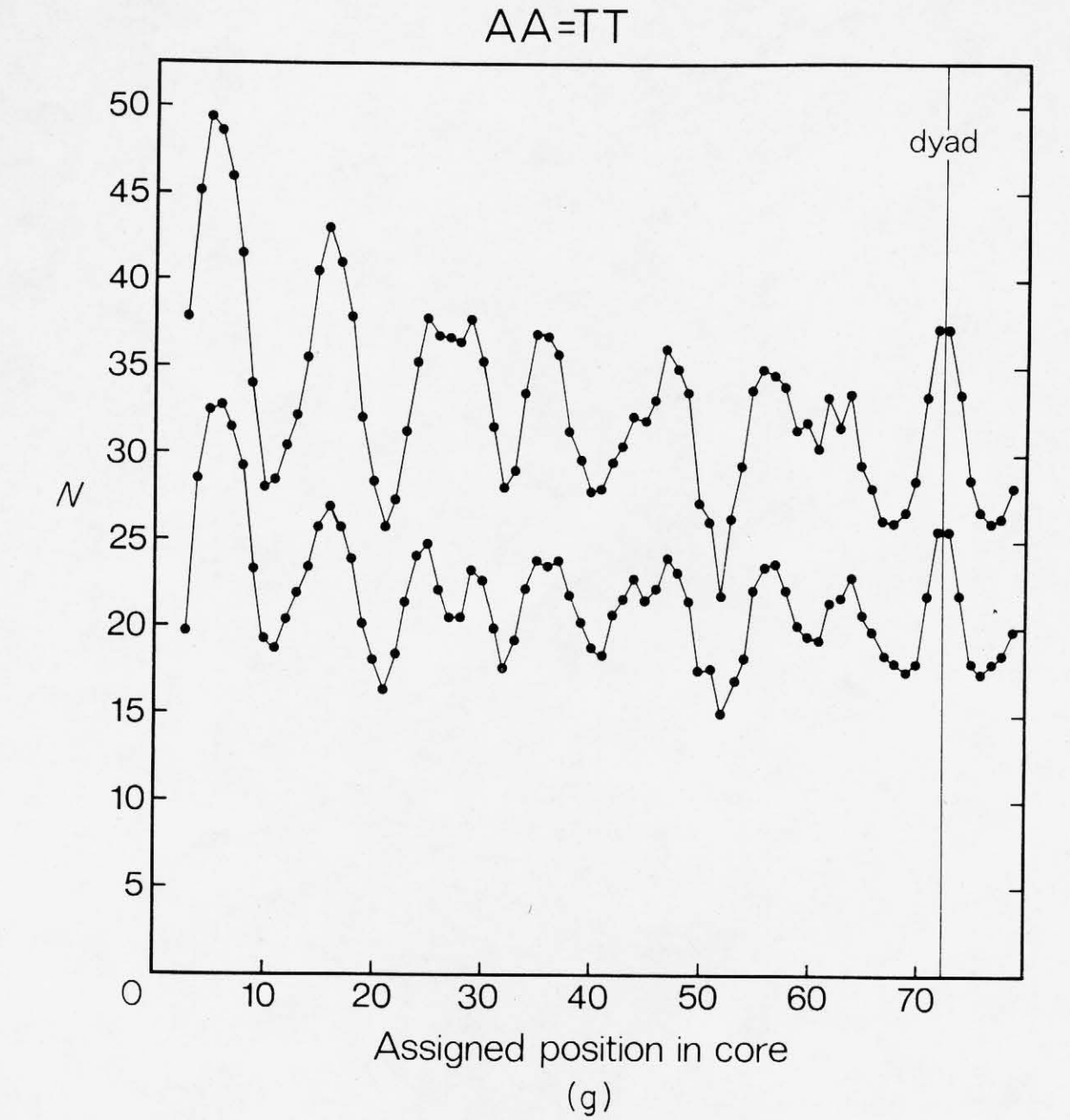
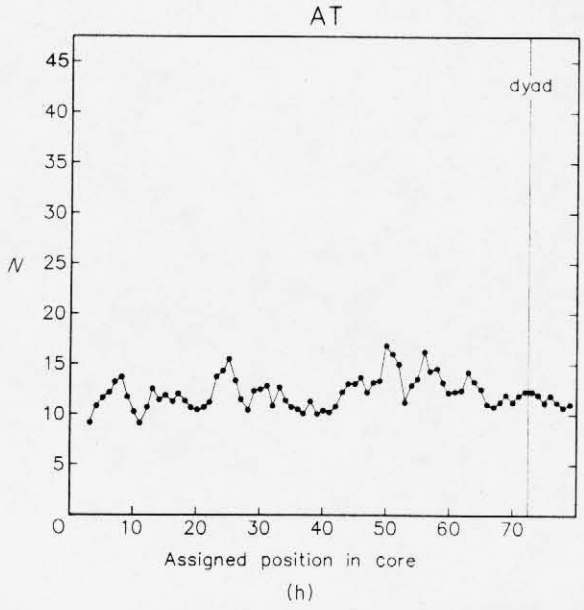

CG
dyad
N
Assigned position in core
(a)
TG=CA
dyad
N
Assigned position in core
(b)
TA
dyad
N
Assigned position in core
(c)
GG=CC
dyad
N
Assigned position in core
(d)
AG=CT
dyad
N
Assigned position in core
(e)
GA=TC
dyad
N
Assigned position in core
(f)

Fig. 1.


AA=TT
dyad
N
Assigned position in core
(g)


AT
dyad
N
Assigned position in core
(h)

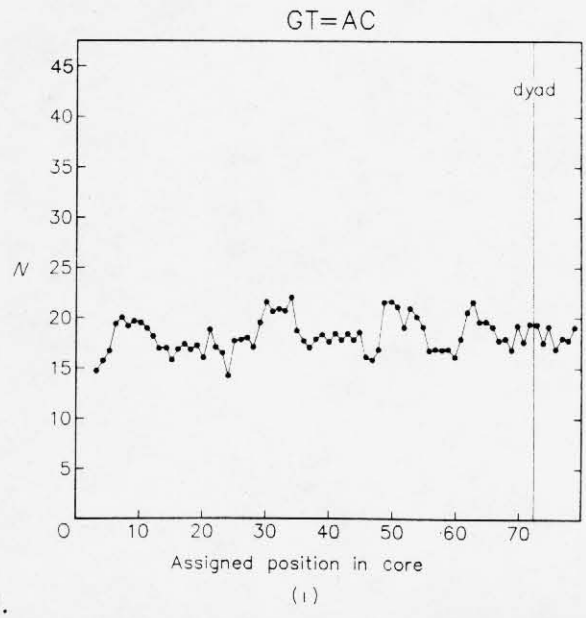

GT=AC
dyad
N
Assigned position in core
(i)

Fig. 1.

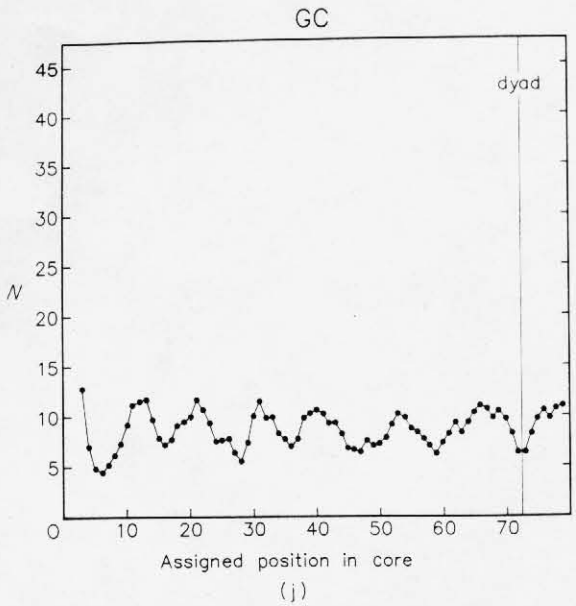


Figure 1. Variations in the occurrence of dinucleotides (N) *versus* position in the sequence. Data are presented as a running 3-bond average of occurrence averaged about base step 72·5: (a) CpG, (b) TpG/CpA, (c) TpA, (d) GpG/CpC, (e) ApG/CpT, (f) GpA/TpC, (g) ApA/TpT, (h) ApT, (i) GpT/ApC, and (j) GpC. In all cases, the total stepwise occurrence of the dinucleotide in 177 sequences is shown. For ApA/TpT, the total stepwise occurrence in a selected set of 117 sequences of length 144 to 146 bp is also shown (lower curve).

and it is these data that will be now discussed (although the results for both kinds of calculation have been listed).

Looking at the right-hand part of Table 3, it may be seen that the trinucleotides that produce waveforms of largest amplitude are GpGpC/GpCpC (±45%), ApGpC/GpCpT (±25%), ApApA/TpTpT (±37%) and ApApT/ApTpT (±30%). Both trinucleotides of the kind PupGpC exhibit a period of 10·15 bases and a phase angle of 26° to 29°, while the trinucleotides $ApAp\overset{A}{\underset{T}{}}$ exhibit a period of 10·26 to 10·31 bases and a phase of 154° to 181°.

Several large periodicities for trinucleotides containing the step CpG may be noted (Table 3), but these would not appear to be unduly significant in view of the low total occurrence of such sequences (column 2). Other significant periodicities (±18%) are found for the trinucleotides CpCpC/GpGpG and CpApT/ApTpG, both of which exhibit a period of 10·26 to 10·31 bases, with a phase of −11° to −21°. In addition, the sequences TpApA/TpTpA and TpApG/CpTpA exhibit variations of ±20% at a period of 10·31 to 10·15 bases with a phase of 179° to −170°. Another strong periodicity is observed for GpApC/GpTpC (±23%), but at a phase of 99° and an unusual period of 9·66 bases.

Clearly, if the periodicity of sequence content represents the twist of the double helix, as would be the case if the superhelix were uniform (Drew & Travers, 1985*a*), then one would like to examine the amplitude and phase for each of these 32 trinucleotides at a common period of about 10·2 bases. This would filter out much of the statistical noise from having a limited data set of 25,000 bases, and would select for the true periodic signal. Such a compilation will be listed below.

(c) *The occurrences of long sequences*

The occurrences of the tetranucleotide ApApApA/TpTpTpT and the pentanucleotide ApApApApA/TpTpTpTpT are distributed more selectively throughout core DNA than any other sequences of comparable length. In general, long runs of $(dA)_n \cdot (dT)_n$ from $n = 5$ to $n = 18$ are concentrated at either end of the molecule in positions 1 to 20 and 126 to 145. This accounts for

Table 2

Fourier analysis of variations in the occurrence of dinucleotides

		Steps 3 to 143			Steps 3 to 61, 84 to 143		
Dinucleotide	Mean steps/position	Fractional variation in occurrence	Phase angle (°)	Period (bp)	Fractional variation in occurrence	Phase angle (°)	Period (bp)
CG	1·76	No significant signal			0·15	−51	10·42
TG/CA	27·79	0·07	−14	10·31	0·08	−6	10·26
TA	9·75	0·12	176	10·26	0·13	174	10·26
GG/CC	17·15	0·08	25	10·15	0·12	25	10·15
AG/CT	26·40	No significant signal			0·05	−57	10·36
GA/TC	21·94	No significant signal			No significant signal		
AA/TT	33·17	0·16	±180	10·26	0·20	172	10·26
AT	12·15	No significant signal			0·06	147	10·26
GT/AC	18·24	0·08	−68	10·53	0·07	−69	10·47
GC	8·65	0·20	13	10·15	0·27	25	10·15

The fractional variation in occurrence is defined as the normalized amplitude per period divided by 5·0, which is the sum of $\cos^2\theta$ over 1 period of 10 steps where $\theta = 0°, 36°, 72°$ etc. The phase origin is set at step 1·0. Thus, a phase angle of 0° at this position, for a period of 10·214 bp, yields a phase angle of 0° also at position 72·5, which is coincident with an axis of 2-fold symmetry in the protein–DNA complex (Richmond *et al.*, 1984). The periods are given for Fourier maxima ≥0·04 within the range of 9·55 to 11·15 bases. The value for mean steps/position is calculated from the total number of occurrences between positions 23 to 123, inclusive, to avoid end effects.

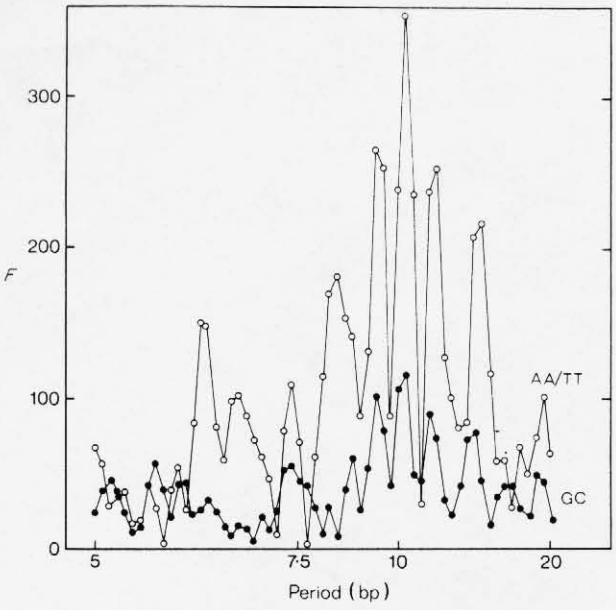


Figure 2. Variation of amplitude (F) against period, following Fourier transformation of the occurrences of the dinucleotides GpC and ApA/TpT, including those in the dyad region. F is as defined in Materials and Methods. Note that if the dyad region is omitted, the 10-fold periodicity is even more pronounced.

the higher than average occurrence of ApA/TpT dinucleotides in these regions, as was noted earlier.

Plots of occurrence *versus* position for several different lengths of $(dA)_n \cdot (dT)_n$, where $n = 2$, 3, 4, 5, are shown in Figure 4(a) and (b). In each case, steps have been selected that do not have any A or T neighbours on either side, which might lengthen the run of identical bases (e.g. for $n = 3$ choose CAAAG or GAAAT, but not CAAAA or AAAAT). It may be seen that the critical length for preferences of this sort to appear is about $n = 3$. The isolated dinucleotide ($n = 2$) shows little preference for where it is located on the core. Short runs of $n = 4$ are found both at the ends of core DNA and at many locations in the central positions 21 to 124, where their minor grooves originally faced in toward the protein. When $n = 5$ or greater, runs of $(dA)_n \cdot (dT)_n$ prefer to avoid the central portion of core DNA. This avoidance is most pronounced between positions 40 and 50, two to three turns from the dyad, where very few such homopolymer sequences can be found (Fig. 4(b) and (c)).

One might expect that long runs of $(dA)_n \cdot (dT)_n$, where $n = 10$ or greater, would prefer to be excluded entirely from core DNA: yet the total occurrence of such runs in our sample is not significantly less than that expected on a random basis, assuming a base composition of 57·3% A+T (Table 1). There is therefore no present evidence for the exclusion of $(dA) \cdot (dT)$ runs from the nucleosome core, although such sequences clearly occupy preferred positions.

Lastly, we have looked for preferences in the placement of runs of $(dG)_n \cdot (dC)_n$ such as GpGpGpG/CpCpCpC, but the number of these sequences in chicken blood DNA is too small to draw any firm conclusions. However, the total occurrence of such runs (up to 9 bp in extent) does not differ significantly from that expected on a random basis. One of the 177 clones examined was quite remarkable for another reason: it consisted almost entirely of alternating A and G residues, as in ApGpApGpApG. It seems likely, therefore, that the polymer poly(d-Ad-G)·poly(d-Cd-T) might be reconstituted successfully with the histone octamer, even though this has not proved to be possible with any other poly(purine)·poly(pyrimidine) DNA.

4. Discussion

(a) *Isolation and cloning of chicken core DNA*

The variations in sequence content that have been described here can only be related to nucleosome positioning, and likewise DNA bending, if it can be shown that the results have not been influenced by the techniques used to obtain the sequences. There are three separate issues with which we need to concern ourselves: (1) the relevance of isolated core DNA to the locations of nucleosomes *in vivo*; (2) the use of micrococcal nuclease to trim long chromatin down to nucleosome cores: and (3) any selection that may have occurred during the cloning procedure.

First, let us consider the relationship between the set of sequences found in isolated core DNA and the locations of nucleosomes *in vivo*. Do the histone proteins move or "slide" along the DNA as the protein–DNA complexes are isolated and purified? On the basis of our experience, this seems probable. A necessary step in the isolation and purification of nucleosome cores is the removal of histones H5 and H1 from long chromatin: this step involves exposure to 0·65 M-NaCl at 4°C, for three to six hours as the material passes through a column. It has been reported that, under these conditions of salt and temperature, no measurable modifications of the original repeating structure of chromatin can be detected (Spadafora *et al.*, 1979). Nevertheless, since the conditions under which H5 and H1 are removed also allow the exchange of histone octamers between DNA molecules (Drew & Travers, 1985*a*; Rhodes, 1985), it seems likely that some migration of the histone proteins may have occurred during this step. Therefore the population of core DNA molecules analysed here is not necessarily equivalent to the population of DNA sequences associated with histone octamers *in vivo*.

A second concern involves the use of micrococcal nuclease to trim long chromatin down to nucleosome core particles. Since this enzyme prefers to cut at the mononucleotides pA and pT (e.g. see Cockell *et al.*, 1983; Drew, 1984), one might expect that any excess digestion of H5-stripped long chromatin would produce a sample of nucleosome

Table 3

Fourier analysis of variations in the occurrence of trinucleotides

Trinucleotide	Mean steps/ positions	Steps 3·5 to 142·5			Steps 3·5 to 60·5, 83·5 to 142·5		
		Fractional variation in occurrence	Phase angle (°)	Period (bp)	Fractional variation in occurrence	Phase angle (°)	Period (bp)
CGA/TCG	0·86	0·15	37	10·20	0·31	22	10·26
CGT/ACG	1·08	0·35	−56	9·66	0·34	−58	9·66
CGG/CCG	0·83	0·25	25	9·52	No significant signal		
CGC/GCG	0·75	No significant signal			0·25	40	10·20
TGA/TCA	7·34	0·11	56	10·05	0·09	70	10·00
TGT/ACA	7·85	0·13	−70	10·70	0·12	−83	10·70
TGG/CCA	6·01	No significant signal			0·08	30	10·10
TGC/GCA	6·56	0·10	−14	10·26	0·14	−13	10·26
TAA/TTA	6·59	0·19	−175	10·31	0·21	179	10·31
TAT/ATA	5·47	0·10	74	10·36	0·15	81	10·36
TAG/CTA	3·83	0·19	137	10·15	0·19	−170	10·15
TAC/GTA	3·71	No significant signal			No significant signal		
GGA/TCC	5·71	0·07	−48	9·76	No significant signal		
GGT/ACC	3·75	0·12	89	9·95	0·11	93	10·87
GGG/CCC	3·97	0·12	57	9·71	0·14	56	9·85
GGC/GCC	3·59	0·36	30	10·15	0·45	26	10·15
AGA/TCT	8·14	0·12	−160	10·15	0·09	−127	10·10
AGT/ACT	5·54	0·14	−81	10·53	0·15	−77	10·53
AGG/CCT	6·34	No significant signal			No significant signal		
AGC/GCT	6·24	0·19	50	10·10	0·25	29	10·15
GAA/TTC	7·58	0·11	−175	11·05	0·13	−134	10·10
GAT/ATC	4·65	No significant signal			No significant signal		
GAG/CTC	6·39	No significant signal			No significant signal		
GAC/GTC	3·31	0·21	127	9·62	0·23	99	9·66
AAA/TTT	12·19	0·30	160	10·31	0·37	154	10·31
AAT/ATT	7·31	0·21	−169	10·26	0·30	−179	10·26
AAG/CTT	7·78	0·07	131	9·85	0·08	−156	9·71
AAC/GTT	5·86	No significant signal			No significant signal		
CAA/TTG	6·94	No significant signal			0·09	141	10·10
CAT/ATG	6·96	0·15	−20	10·31	0·18	−21	10·31
CAG/CTG	8·49	0·08	−29	9·71	0·05	−24	9·76
CAC/GTG	5·37	0·14	−4	10·26	0·18	−11	10·26
XAAAX/ZTTTZ	5·06	0·30	162	10·31			
XAAAAX/ZTTTTZ	1·67	0·34	159	10·31			
XAAAAAX/ZTTTTTZ	0·62	0·39	−103	10·00			
XGGGX/ZCCCZ	2·06	0·21	52	10·00			

The phase origin is set at position 1·0.

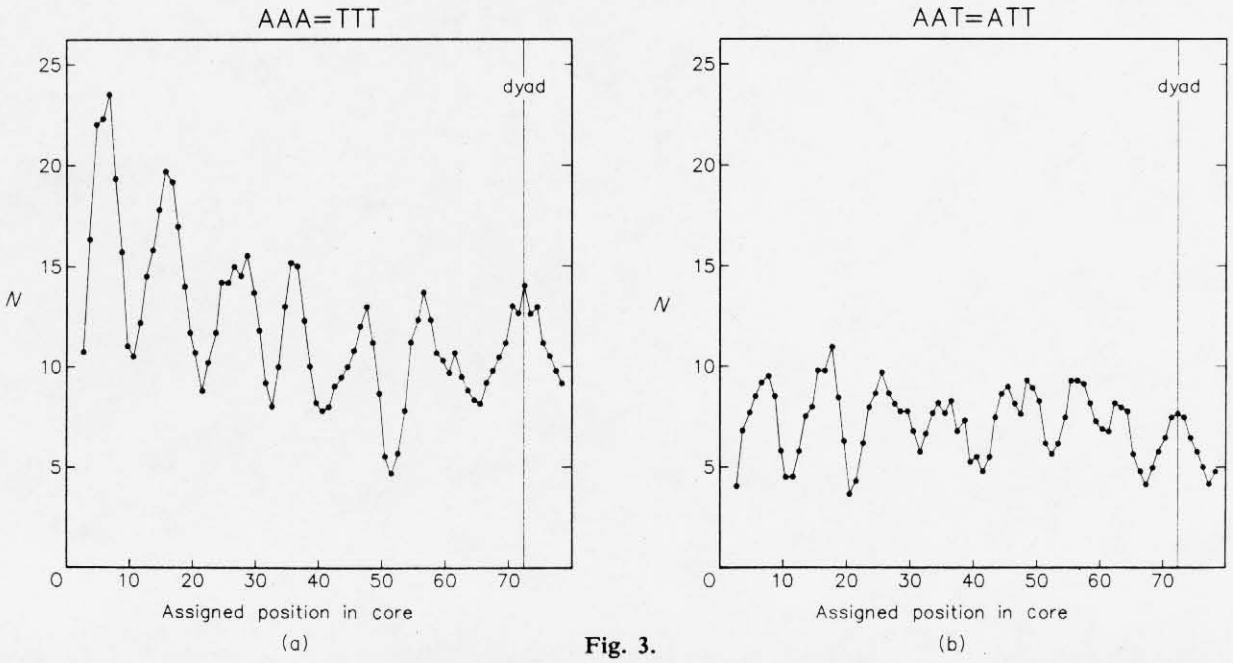


Fig. 3.

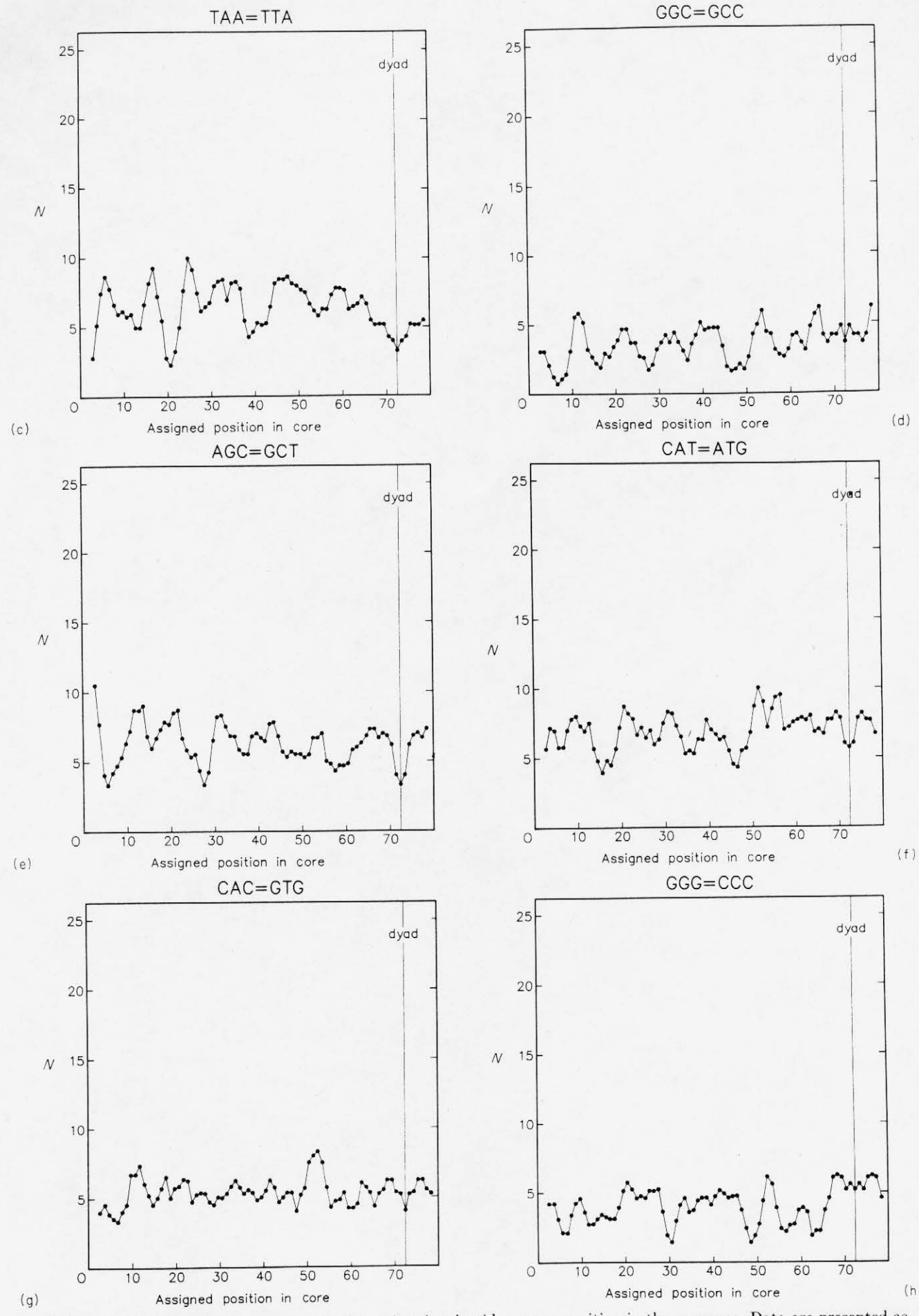


Figure 3. Variations in the occurrence of certain trinucleotides *versus* position in the sequence. Data are presented as a running 3-bond average of occurrence averaged about base 73 (base step 72·5). (a) ApApA/TpTpT.

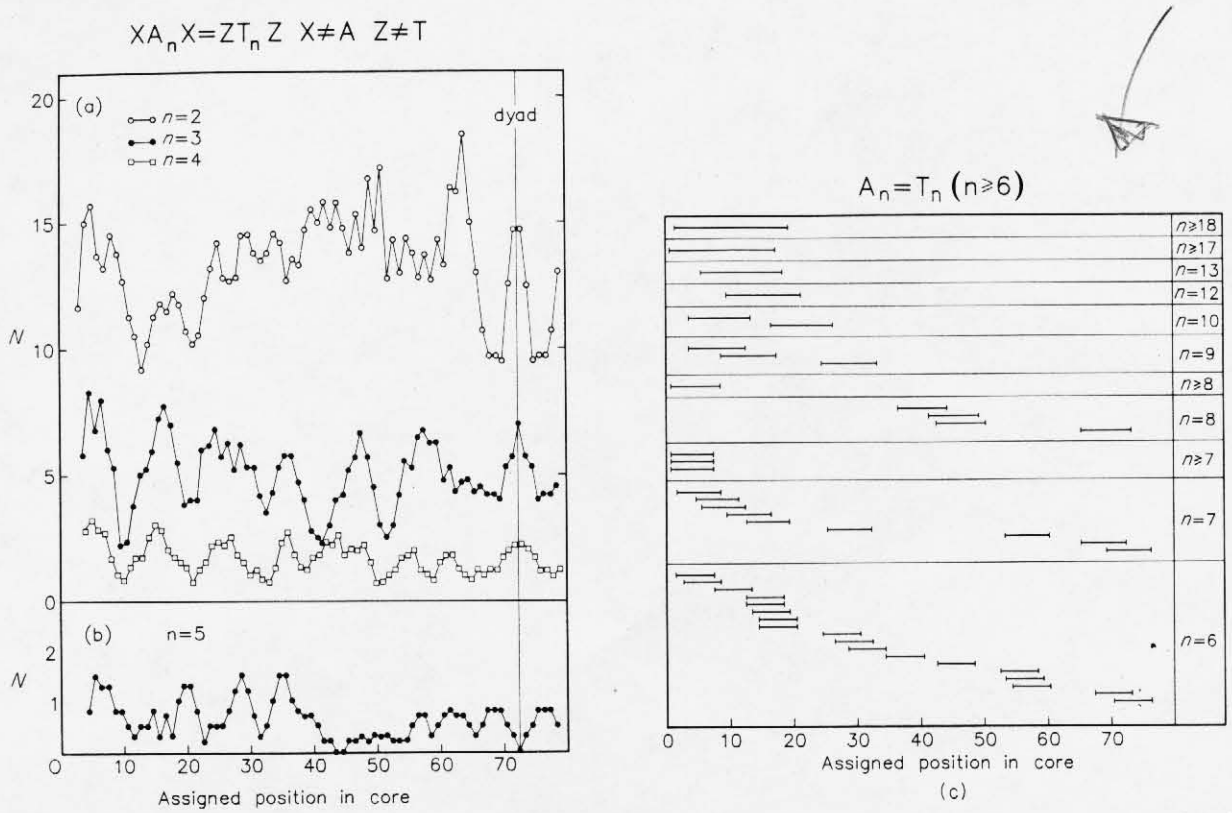


Figure 4. Variations in occurrence of XA_nX/ZT_nZ, where $X \neq A$ and $Z \neq T$. (a) $n = 2, 3, 4$; (b) $n = 5$; (c) $n = 6$ to 18. Data are presented as a running 3-bond average of occurrence averaged about base step 72·5. In those cases where (dA)·(dT) runs terminate a cloned DNA fragment the full length of the run in genomic DNA is unknown: such runs are indicated by the symbol ⩾ in (c).

cores which would be depleted in A and T residues. Such is not the case for our sample. As shown in Table 1, the overall A plus T content of 177 cloned DNA molecules is 57·7%, a value which does not differ significantly from the 57·3% of total chicken erythrocyte DNA.

Comparison of dinucleotide frequencies in cloned core DNA with those determined for chicken blood DNA in bulk (Swartz *et al.*, 1962) shows that, with one exception, the two sets of numbers do not differ by more than 10%. For example, in the case of ApA/TpT (a step which exhibits a position-dependent modulation of ±20% within each turn of double helix), the two values in Table 1 do not differ by more than 0·6%. The sole exception to this rule is the step TpA, the total occurrence of which decreases by 13% in going from chicken DNA to our core DNA sample. A similar deficiency is seen for the trinucleotide TpApG/CpTpA, but not for the other trinucleotides containing the step TpA. Since TpA is a strongly preferred substrate for micrococcal nuclease, it seems possible that these small discrepancies are a consequence of the use of micrococcal nuclease during the isolation procedure. Because of this possibility, we do not attribute any significance to the distributions of those sequences containing TpA that occur at significantly lower than expected frequencies.

McGhee & Felsenfeld (1983) have reported that certain sequence periodicities in chicken nucleosome core DNA could be induced artificially by a very extensive digestion with micrococcal nuclease, followed by a two-step fractionation of the remaining material. These results would not appear to be relevant to ours for three reasons: first, we obtain chicken cores in at least 80% yield from H5-stripped long chromatin, in contrast to a tenfold lower yield after prolonged micrococcal nuclease digestion (McGhee & Felsenfeld, 1983); second, the periodicities induced by extensive digestion with micrococcal nuclease are different from those reported here, since this enzyme cuts primarily at the two ends of the DNA along the upper and lower surfaces of the supercoil (Cockell *et al.*, 1983) and not uniformly along the outer surface; and third, the sequence preferences reported here account adequately for the rotational positioning of seven different reconstituted nucleosome cores in solution (Simpson & Stafford, 1983; Drew & Travers, 1985*a*; Rhodes, 1985; H. R. Drew & C. R. Calladine, unpublished results).

A third concern relates to any possible selection of sequences during the cloning step. We used the *Sma*I site of M13 DNA to select for recombinant core DNA molecules. The enzyme *Sma*I leaves two blunt ends of sequence CCC and GGG, so the choice of this cloning site will almost certainly select for core DNA molecules with flush ends. Indeed, the mean double-stranded length of our clones (145 bp) appears to be slightly less than the mean single-stranded length of the same sample before cloning (146 to 147 bases) (Drew & Travers, 1985*a*). More seriously, in our cloned DNA molecules we see a strong tendency for pyrimidine bases to be located

at the 5′ end, joined to CCC, and purine bases at the 3′ end, joined to GGG. In a previous experiment we found the same phenomenon: DNA ligase tends to join pyrimidines to pyrimidines and purines to purines (Drew & Travers, 1985*b*). So, it would not be safe to attribute any significance to the asymmetry of pyrimidine and purines at either end of a cloned molecule.

In addition, the cloned DNA molecules terminate in short runs of (dA)·(dT) more often than would be expected on a random basis. It seems less likely that this kind of sequence distribution could be influenced by the ligation procedure. There are two reasons for this opinion: first, because short runs of (dG)·(dC) are not distributed in the same way; and second, because the high occurrence of runs of (dA)·(dT) extends inward from position 1 to position 20 (and likewise from position 145 to 126).

(b) *Relation of sequence periodicities to DNA bending*

As mentioned above, the most notable feature in our collection of 177 DNA sequences is the strong periodicity of ±20% found for the stepwise occurrence of the sequence ApA/TpT. This may be seen again in Table 4A, where we have listed fractional variations in occurrence for all ten dinucleotides at a common period of 10·20 bases. When analysed by Fourier methods, it may be seen that there is also another very strong periodicity in the dinucleotides: that of ±27% for the sequence GpC. The periodic occurrence of GpC has a phase angle of +11°, which means that it prefers to lie about 180° or 5·1 bases away from ApA/TpT (−174°).

It was suggested previously that these periodic variations in sequence correlate with certain sequence-dependent properties of a DNA molecule, which facilitate its bending around the histone octamer in chromatin (Drew & Travers, 1985*a*). In particular, it was proposed that runs of (A, T) prefer to occupy positions where the minor groove of a double helix lies on the inside of a curve, so that the groove would be relatively narrow. Such a preferred location would accord with the characteristic structure of runs of (A, T), as evidenced both in crystal and in solution (Dickerson & Drew, 1981; Fratini *et al.*, 1982; Drew & Travers, 1984). Conversely, runs of (G, C) were said to prefer positions on the outside of a bent DNA molecule, where the minor groove can be relatively wide (McCall *et al.*, 1985).

These previous interpretations are strongly supported by the present data. Using the method of Fourier analysis, we have shown that all of the predominant variations in sequence of core DNA exhibit a period in the range 10·15 to 10·26 bases. Further, we have shown that the phases of these periodic waveforms fall into two classes: those including ApA/TpT, TpA and ApT with a phase angle close to 180°; and those including GpC, GpG/CpC and TpG/CpA with a phase close to 0°. The structural interpretation of such a result is

Table 4

Fractional variations of occurrence and phase angles of dinucleotides and trinucleotides, at a best period of 10·20 bp

A. *Dinucleotide* (at period = 10·20 bp)

	Fractional variation in occurrence	Phase angle (°)	
GC	0·27	11	"Out"
(CG)	0·14	28	
GG/CC	0·12	12	
TG/CA	0·08	8	
GT/AC	0·06	−27	
AG/CT	0·04	1	
			No rotational preference
GA/TC	0·01	−154	
AT	0·06	161	
TA	0·13	−173	
AA/TT	0·20	−174	"In"

B. *Trinucleotide* (at period = 10·20 bp)

	Fractional variation in occurrence	Phase angle (°)	
GGC/GCC	0·45	12	"Out"
(TCG/CGA)	0·31	35	
AGC/GCT	0·25	10	
(CGC/GCG)	0·25	37	
CAT/ATG	0·18	0	
CAC/GTG	0·17	0	
GGG/CCC	0·13	5	
TGC/GCA	0·13	−1	
AGT/ACT	0·11	−19	
GAG/CTC	0·08	8	
GGT/ACC	0·08	−10	
TGG/CCA	0·08	11	
(CGT/ACG)	0·08	7	
AGG/CCT	0·08	29	
GAC/GTC	0·08	−81	
TGA/TCA	0·08	13	
GAT/ATC	0·07	54	
TGT/ACA	0·06	−18	
AAG/CTT	0·06	1	
(CGG/CCG)	0·02	−17	
			No rotational preference
CAG/CTG	0·02	−106	
GGA/TCC	0·05	−173	
TAC/GTA	0·06	−98	
AAC/GTT	0·06	−150	
AGA/TCT	0·09	−159	
CAA/TTG	0·09	115	
GAA/TTC	0·12	126	
TAT/ATA	0·13	113	
TAG/CTA	0·18	177	
TAA/TTA	0·20	−155	
AAT/ATT	0·30	−169	
AAA/TTT	0·36	±180	"In"

The phase origin is set at position 1·0. Parentheses indicate sequences whose variations in occurrence may be influenced by noise, due to the infrequent occurrence of the dinucleotide CpG. In addition, the total occurrences of the dinucleotide TpA, and of the trinucleotide TpApG appear to be slightly underrepresented in our sample, relative to total chicken blood DNA. Dinucleotides and trinucleotides are listed in order of decreasing fractional variation of occurrence for phase angles in the range −90° to 0° to +90° and then in increasing fractional variation of occurrence for phase angles in the range +90° to ±180° to −90°. Finally, it should be noted that most of the CpG steps in chicken blood DNA are methylated at the 5 position of cytosine.

that, within each turn of double helix, the average rotational setting of sequences containing A+T is opposite to that of sequences containing G+C, and also opposite to that of TpG/CpA.

The weighted-average period of the predominant waveforms is 10·21 bp. For this period and a phase angle of 0°, maxima in the Fourier wave are located at positions 1·0, 11·21, 21·42, 31·63, 41·84, 52·05, 62·26, 72·47 etc. Position 72·5 corresponds to the location of the dyad axis in the nucleosome core particle, where the minor groove of the DNA faces directly out and away from the protein (Richmond *et al.*, 1984). This result defines the rotational setting of the phase angles on an absolute scale. It tells us that ApA/TpT steps, with a phase angle of 180°, prefer to occupy positions where the minor groove faces directly in towards the protein; whereas GpC steps, with a phase angle of 0°, prefer positions where the minor groove faces out. This conclusion is also consistent with the observed rotation of AT-rich and GC-rich sequences in several different DNA molecules that have been reconstituted with the histone octamer and studied in solution (Drew & Travers, 1985*a*; Rhodes, 1985).

(c) *Path of the DNA around the histone octamer*

On the strength of the data just presented, one might wish to turn the argument around: to ask whether a 10 bp periodicity in base sequence can be taken as evidence for the existence of curvature in a case where the three-dimensional structure is not known. The nucleosome core provides a simple test of this hypothesis. It is known from the crystal structure analysis at 7 Å resolution (Richmond *et al.*, 1984) that the path of the DNA around the histone octamer deviates substantially from that of a uniform superhelix. The most prominent deviation is seen in the neighbourhood of the particle dyad, where two turns of rather flat, left-handed superhelix (positions 1-61, 84-145) are joined by a region where the DNA angles upward at a steep pitch (positions 62-83). The overall shape of the DNA at low resolution thus resembles a "key ring".

If these periodicities in base sequence truly reflect the stresses of curvature sensed by the DNA when it was originally bound to the protein, then one would expect to see a rather regular, 10·2 bp oscillation in sequence outside of the dyad region and some sort of discontinuity in amplitude or phase as one proceeds toward the dyad. Indeed, this is precisely what is observed.

Using the ApA/TpT modulation as an example (Fig. 1(g)), one sees a periodic fluctuation once every 10·2 bases between positions 1 to 61 and 84 to 145. Thus, maxima occur at positions 5, 16, 27, 36, 47 and 56 and minima at 10, 21, 32, 41 and 52. If this periodicity were to continue into the dyad region, one would expect a maximum at position 67 along with minima at positions 62 and 72. Yet, in reality, one sees maxima at positions 63·5 and 72·5, along with a minimum at position 67·5. Therefore, in the region of the dyad, the periodic fluctuation in ApA/TpT content is shifted in phase by 120° to 180° from elsewhere in the core. Detailed inspection of the many different curves shown in Figure 1 suggests a similar reversal in phase for many of them, especially in cases such as GpC or TpG/CpA, where the amplitude of the normal periodicity is quite strong.

We hesitate to predict exactly what kind of curvature is present at the dyad on the basis of sequence information alone. There are two principal reasons for this: (1) the twist of the DNA may vary locally from 10·2 bp/turn to some other value, thereby altering the phase of sequence occurrence; (2) structural parameters of DNA other than curvature may influence sequence selection. The simplest model one could construct would be to place an "S" curve on the dyad, aligning the centre of the S with position 72·5 and joining its ends to positions 61 and 84. If it is assumed that the observed periodicity of sequence selection reflects only curvature, then this model predicts that the amplitude of periodicities near the dyad should be roughly one-half of that elsewhere, and also that the angle of phase should shift by 90° as the DNA first bends up and then down. The predictions of such a model do not agree especially well with the present experimental data.

At the two ends of core DNA, positions 1 to 20 and 125 to 144, the fluctuation in sequence content remains periodic for most of the curves shown in Figure 1, including that of ApA/TpT. In addition, the mean value of the ApA/TpT curve increases in going from position 20 to position 1 (Fig. 1(g)). It was mentioned earlier that this increase may be attributed to the preferred placement of long runs of (dA)·(dT) near the ends of the DNA. One could imagine that a mixed population of nucleosome cores in solution would be composed of two kinds of particle: those without (dA)·(dT) at the ends, in which the DNA continues to bend smoothly as it leaves the particle; and those with (dA)·(dT) at the ends, in which the DNA leaves the particle in a straight path.

(d) *Rotational positioning*

We have shown that those sequences whose frequency of occurrence is highly periodic are directly related to the rotational setting of the DNA on the nucleosome. In general, the observed periodic occurrence of a particular dinucleotide reflects that of only a subset of all possible trinucleotides containing that dinucleotide. So, the structural constraints that determine the placement of a DNA sequence about the histone proteins may depend not only on the presence of a certain dinucleotide step but also on the sequence context in which it is located.

The dinucleotide ApA/TpT is located preferentially where the minor groove faces inwards. This distribution is determined principally by the

occurrence of the sequences ApApA/TpTpT, ApApT/ApTpT and TpApA/TpTpA (Table 4B). The dinucleotides GpC and TpG/CpA are preferentially located where the minor groove faces out. Again these dinucleotides occur in a restricted context. Thus the dominant periodically repeating trinucleotides containing GpC are GpGpC/GpCpC and ApGpC/GpCpT, while the most important trinucleotides containing TpG are ApTpG/CpApT and GpTpG/CpApC.

The amplitudes of the periodic functions for the dinucleotides and trinucleotides can be used in principle to construct a predictive algorithm for nucleosome positioning that is independent of the assumption of any particular structural model. In such an algorithm, some of the amplitudes for dinucleotides add in a straightforward fashion to yield the amplitude for a trinucleotide. For example, AA(0·20) plus AA(0·20) yield 0·40, nearly the same as AAA(0·36). Similarly, AA(0·20) and AT(0·06) yield 0·26, close to AAT(0·30); but other steps do not behave in the same way. Thus, GG(0·12) and GC(0·27) add up to 0·39, which is approximately GGC(0·45), but TG(0·08) and GC(0·27) add up to 0·35, which is far from TGC(0·13).

It seems probable, that to describe the rotational preferences of DNA bending around the histone core, one would have to use somewhere between ten and 32 parameters to arrive at a correct summation of Fourier terms. Preliminary experiments indicate that 12 terms may be sufficient (H.R.D., unpublished results).

(e) *Translational positioning*

The curvature of a DNA molecule wrapped around the histone octamer varies in direction near the dyad. To accommodate this change in direction, the sequences that correlate with rotational positioning seem to occupy a different helical phase relative to those sequences elsewhere on the core DNA. Thus, the precise arrangement of rotational preferences may be one determinant of translational positioning. Similarly the preferred location of runs of (dA)·(dT), 5 to 18 bp in extent, near either end of the DNA, and their relative exclusion between positions 40 and 52, may also contribute to translational placement.

However, it seems possible that sequences in the "linker" region between histone octamers could be equally or more important than the core DNA itself in determining translational position. Such sequences could be conformationally constrained in any of several ways, and therefore be unable to fit into the shape required by a histone octamer. One kind of example would include very long tracts of homopolymer (dA)·(dT) and (dG)·(dC), neither of which has yet been reconstituted to form a nucleosome core (Rhodes, 1979; Simpson & Kunzler, 1979; Kunkel & Martinson, 1981; Prunell, 1982). A second kind of example would include the DNA that forms the binding site for transcription factor IIIA, and lies at the border of a nucleosome core (Rhodes, 1985). This DNA exhibits a strong periodicity of sequence once every 5·6 bp, rather than the presently reported 10·2 bp (Rhodes & Klug, 1986). It has been shown to have a structure resembling that of RNA, due to the preferred stacking of bases according to their nucleotide sequence (McCall *et al.*, 1986).

(f) *Relation of sequence periodicities to models of DNA bending*

By themselves, the sequence periodicities observed here do not allow a unique structural description of DNA bending. This requires reference to some particular structural model. The data do, however, place constraints on which models are possible.

Our data cannot easily be reconciled with the theoretical predictions of Trifonov (1980) or Mengeritsky & Trifonov (1983). These authors predict that purine–purine steps in core DNA should prefer to occupy positions 5 bp removed from pyrimidine–pyrimidine steps, in no specified orientation (Trifonov, 1985). Yet we have observed here that purine–purine steps such as ApA and GpG occupy dissimilar locations in core DNA, while steps ApA and TpT, and likewise GpG and CpC, occupy similar locations. Let there be no confusion on this point: if ApA and TpT were to occupy positions 5 bp apart, then the curve of ApA and TpT combined (Fig. 1(g)) should show a periodic fluctuation of 5 bp. In fact, no significant periodic signal is seen at 5 bp, but we can see a clear periodicity at 10·2 bp that is one of the strongest for any dinucleotide (Table 4A).

The occurrences of all such strongly periodic steps in our sample have been found to observe a phase of close to either 0° or 180°, which means that the base-pairs lie in an orientation that allows the major and minor grooves of the helix to open and close smoothly as the DNA winds around the protein. This motion is equivalent to the "roll" deformation of Dickerson & Drew (1981), Fratini *et al.* (1982), Dickerson *et al.* (1983) and Calladine & Drew (1984). It is not in agreement with the isotropic "wedge" model of Trifonov (1985) and Ulanovsky *et al.* (1986), where "roll" and "tilt" are given an equal footing.

There are insufficient data at present to relate the observed sequence periodicities to the many crystal structures of DNA that have been solved in the past few years. However, certain correlations are evident. For example, the sequence GpGpC in the crystal structure of d(GGGGCCCC) has a large, total roll of +20° that opens the minor groove (McCall *et al.*, 1985; see also Wang *et al.*, 1982). The sequence ApApTpT in the crystal structure of d(CGCGAATTCGCG) has a slightly negative roll and large "propellor twist" that closes the minor groove (Dickerson & Drew, 1981). The reversible bending of this molecule has been studied in the crystal at 2 Å resolution, and all of the deformation

was found in roll rather than in tilt or wedge (Fratini *et al.*, 1982; Dickerson *et al.*, 1983).

It has been reported that any short run of A or T nucleotides longer than 3 bp (such as AAAA or TTTTT), when periodically repeated at intervals of 10 to 11 bp, can confer a detectable amount of curvature on an isolated DNA molecule (Wu & Crothers, 1984; Hagerman, 1985, 1986; Koo *et al.*, 1986; Ulanovsky *et al.*, 1986). Such sequences make no contribution to the difference between linking number and twist unless they are deformed by supercoiling (Diekmann & Wang, 1985). Several models for DNA bending, including that of a wedge at ApA, have been ruled out by these experiments (Hagerman, 1986; Koo *et al.*, 1986). However, it remains unclear from these measurements by what amount the DNA bends, and in what direction. We have found, on the nucleosome core, there is a strong tendency for the centres of short runs of (dA)·(dT) such as AAA or AAAA to lie with their minor grooves along the inside of the DNA supercoil; runs of intermediate length such as AAAAA and AAAAAA tend to lie on the upper and lower surfaces of the supercoil; while long runs of (dA)·(dT) tend to avoid the bent region entirely and are found most often at the ends of core DNA (Figs 1(g), 3(a) and 4(a), (b) and (c)).

(g) *DNA bending in other protein–DNA complexes*

The nucleosome core is an example of a protein–DNA complex in which the DNA is bent over a long region of 145 bp. Are other such complexes between protein and DNA influenced by sequence-dependent preferences for DNA bending? The path of the DNA is not known in very many cases (Better *et al.*, 1982; Kirchhausen *et al.*, 1985), but it is possible to deduce the rotation of the DNA with respect to the protein by looking at patterns of DNAase I digestion. Detailed inspection of such data for six different proteins required for DNA replication, recombination or transcription suggests that, within such protein–DNA complexes, the DNA adopts a rotation similar to that found in the complex of core DNA with the histone octamer. (The particular examples are: DNA gyrase (Morrison & Cozzarelli, 1981; Kirkegaard & Wang, 1981). DnaA protein (Fuller *et al.*, 1984). Tn*3* resolvase (Grindley *et al.*, 1982; Sherratt *et al.*, 1984), λ integrase (Ross *et al.*, 1979), λ O protein (Zahn & Blattner, 1985), and RNA polymerase holoenzyme (Spassky *et al.*, 1985).)

We would also expect the same sequence-dependent preferences for DNA bending to occur when the double helix is constrained in a tight loop by the co-operative binding of proteins to separated sites. Such a structure has been proposed in the regulatory region of the *E. coli ara*BAD operon, whose repression depends on two binding sites for the AraC protein separated by ~225 bp (Dunn *et al.*, 1984). In a second case the λcI repressor binds co-operatively to two operator sites separated by six double helical turns (Hochshild & Ptashne, 1986). It should be possible to design experiments, using these proteins, to evaluate the energetic preferences involved in DNA bending.

In summary, it would appear that the sequence-dependent preferences for bending a DNA helix are of general occurrence. These sequence-dependent features constitute a further kind of information present in a DNA molecule, in addition to its well-known propensity to encode the amino acid sequence of proteins. A simple theory of DNA bending is to be presented in a forthcoming article (Calladine & Drew, 1986).

We thank Drs B. G. Barrell, C. R. Calladine, J. T. Finch, B. F. Luisi, M. J. McCall, J. R. Miller, D. Rhodes and T. J. Richmond for supplies and help. We are particularly indebted to Dr H.-C. Thøgerson, whose initial experiments on cloning DNA fragments from nucleosome crystals provided an impetus for this work, and especially to Dr A. Klug for his advice and encouragement. H.R.D. was supported by PHS grant CA06971-03 of the National Cancer Institute, DHHS.

References

Bankier, A. T. & Barrell, B. G. (1983). In *Techniques in Life Sciences*, vol. B508, pp. 1–34. Elsevier, Amsterdam.

Better, M., Lu, C., Williams, R. C. & Echols, H. (1982). *Proc. Nat. Acad. Sci., U.S.A.* **79**, 5837–5841.

Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983). *Proc. Nat. Acad. Sci., U.S.A.* **80**, 3963–3965.

Calladine, C. R. & Drew, H. R. (1984). *J. Mol. Biol.* **178**, 773–782.

Calladine, C. R. & Drew, H. R. (1986). *J. Mol. Biol.* In the press.

Cockell, M., Rhodes, D. & Klug, A. (1983). *J. Mol. Biol.* **170**, 423–446.

Dickerson, R. E. & Drew, H. R. (1981). *J. Mol. Biol.* **149**, 761–786.

Dickerson, R. E., Kopka, M. L. & Pjura, P. (1983). *Proc. Nat. Acad. Sci., U.S.A.* **80**, 7099–7103.

Diekmann, S. & Wang, J. C. (1985). *J. Mol. Biol.* **186**, 1–11.

Drew, H. R. (1984). *J. Mol. Biol.* **176**, 535–557.

Drew, H. R. & Travers, A. A. (1984). *Cell*, **37**, 491–502.

Drew, H. R. & Travers, A. A. (1985*a*). *J. Mol. Biol.* **186**, 773–790.

Drew, H. R. & Travers, A. A. (1985*b*). *Nucl. Acids Res.* **13**, 4445–4467.

Dunn, T. M., Hahn, S., Ogden, S. & Schleif, R. F. (1984). *Proc. Nat. Acad. Sci., U.S.A.* **81**, 5017–5020.

Finch, J. T., Lutter, L. C., Rhodes, D., Brown, R. S., Rushton, B., Levitt, M. & Klug, A. (1977). *Nature (London)*, **269**, 29–36.

Fitzgerald, P. C. & Simpson, R. T. (1985). *J. Biol. Chem.* **260**, 15318–15324.

Fratini, A. V., Kopka, M. L., Drew, H. R. & Dickerson, R. E. (1982). *J. Biol. Chem.* **257**, 14686–14707.

Fuller, R. S., Funnell, B. E. & Kornberg, A. (1984). *Cell*, **38**, 889–900.

Grindley, N. D. F., Lassick, M. R., Wells, R. G., Wityk, R. J., Salvo, J. J. & Reed, R. R. (1982). *Cell*, **30**, 19–27.

Hagerman, P. J. (1985). *Biochemistry*, **24**, 7033–7037.

Hagerman, P. J. (1986). *Nature (London)*, **321**, 449–450.

Hanahan, D. (1983). *J. Mol. Biol.* **166**, 557–580.

Hochshild, A. & Ptashne, M. (1986). *Cell*, **41**, 681–687.
Kirchhausen, T., Wang, J. C. & Harrison, S. C. (1985). *Cell*, **41**, 933–943.
Kirkegaard, K. & Wang, J. C. (1981). *Cell*, **23**, 721–729.
Koo, H. S., Wu, H.-M. & Crothers, D. M. (1986). *Nature (London)*, **320**, 501–506.
Kunkel, G. R. & Martinson, H. G. (1981). *Nucl. Acids Res.* **9**, 6869–6888.
Lutter, L. C. (1978). *J. Mol. Biol.* **124**, 391–420.
McCall, M. J., Brown, T. & Kennard, O. (1985). *J. Mol. Biol.* **183**, 385–396.
McCall, M. J., Brown, T., Hunter, W. & Kennard, O. (1986). *Nature (London)*, **322**, 661–664.
McGhee, J. D. & Felsenfeld, G. (1983). *Cell*, **32**, 1205–1211.
Mengeritsky, G. & Trifonov, E. N. (1983). *Nucl. Acids Res.* **11**, 3833–3851.
Morrison, A. & Cozzarelli, N. R. (1981). *Proc. Nat. Acad. Sci., U.S.A.* **78**, 1416–1420.
Prunell, A. (1982). *EMBO J.* **2**, 173–179.
Prunell, A. & Kornberg, R. D. (1982). *J. Mol. Biol.* **154**, 515–523.
Ramsay, N. (1986). *J. Mol. Biol.* **189**, 179–188.
Ramsay, N., Felsenfeld, G., Rushton, B. & McGhee, J. D. (1984). *EMBO J.* **3**, 2605–2611.
Rhodes, D. (1979). *Nucl. Acids Res.* **6**, 1805–1816.
Rhodes, D. (1985). *EMBO J.* **4**, 3473–3482.
Rhodes, D. & Klug, A. (1986). *Cell*, **46**, 123–132.
Richmond, T. J., Finch, J. T., Rushton, B., Rhodes, D. & Klug, A. (1984). *Nature (London)*, **311**, 532–537.
Ross, W., Landy, A., Kukuchi, Y. & Nash, A. (1979). *Cell*, **18**, 297–307.
Sanger, F., Nicklen, S. & Coulson, A. R. (1977). *Proc. Nat. Acad. Sci., U.S.A.* **74**, 5463–5467.
Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. J. H. & Roe, B. A. (1980). *J. Mol. Biol.* **143**, 161–178.
Sherratt, D., Dyson, P., Boocock, M., Brown, L., Summers, D., Stewart, G. & Chan, P. (1984). *Cold Spring Harbor Symp. Quant. Biol.* **59**, 227–233.
Simpson, R. T. & Kunzler, P. (1979). *Nucl. Acids Res.* **6**, 1387–1415.
Simpson, R. T. & Stafford, D. W. (1983). *Proc. Nat. Acad. Sci., U.S.A.* **80**, 51–55.
Spadafora, C., Oudet, P. & Chambon, P. (1979). *Eur. J. Biochem.* **100**, 225–239.
Spassky, A., Kirkegaard, K. & Buc, H. (1985). *Biochemistry*, **24**, 2723–2731.
Swartz, M. N., Trautner, T. A. & Kornberg, A. (1962). *J. Biol. Chem.* **237**, 1961–1967.
Thoma, F. (1986). *J. Mol. Biol.* **190**, 177–190.
Thoma, F. & Simpson, R. T. (1985). *Nature (London)*, **315**, 250–252.
Trifonov, E. N. (1980). *Nucl. Acids Res.* **8**, 4041–4053.
Trifonov, E. N. (1985). *CRC Crit. Rev. Biochem.* **19**, 89–106.
Ulanovsky, L., Bodner, M., Trifonov, E. N. & Choder, M. (1986). *Proc. Nat. Acad. Sci., U.S.A.* **83**, 862–866.
Wang, A. H.-J., Fujii, S., van Boom, J. H. & Rich, A. (1982). *Proc. Nat. Acad. Sci., U.S.A.* **79**, 3968–3972.
Widom, J. & Klug, A. (1985). *Cell*, **43**, 207–213.
Wu, H. & Crothers, D. M. (1984). *Nature (London)*, **308**, 509–513.
Zahn, K. & Blattner, F. R. (1985). *EMBO J.* **4**, 3605–3616.

Edited by R. Laskey