

COMMUNICATION

Short-range Order in Two Eukaryotic Genomes: Relation to Chromosome Structure

J. Widom

Department of Biochemistry
Molecular Biology, Cell
Biology, and Department of
Chemistry, Northwestern
University, Evanston
IL 60208-3500, USA

Fourier transform techniques have been used to analyze the distributions of all ten independent DNA dinucleotide steps in two eukaryotic genomes and one prokaryotic genome, for periodicities of ≈ 2 to 500 bp. The results reveal systematic deviations from random expectation for certain dinucleotide steps over this entire range of periodicities, together with striking peaks at certain spatial periodicities for particular dinucleotide steps. Several dinucleotides yield peaks at a periodicity of ≈ 10.2 bp that are unique to the eukaryotic genomes. Certain members of this set of dinucleotide signals were previously identified as involved in nucleosome positioning, while others were previously unrecognized. In real-space, these dinucleotides are uncorrelated or even anticorrelated (relative to random expectation) at distances of 10 and 11 bp, despite having greater than random spectral power at the corresponding periodicity. Real-space correlations of these dinucleotides at distances of 10 and 11 bp are suppressed by another spectral component, a 3 bp periodicity attributed to codons, which has a local minimum probability at ≈ 10.5 bp. When the two eukaryotic genomes are encoded for the signal "AA or TT", the peak at ≈ 10.2 bp periodicity is strengthened, whereas for the prokaryotic genome such a peak remains absent. For the *Caenorhabditis elegans* genome, this peak becomes the dominant feature in the transform, surpassing a peak owing to the existence of codons in both height and integrated intensity. These results suggest that the requirements of chromosome structure place significant constraints on eukaryotic genome organization; they reveal additional signals that may be related to nucleosome positioning; and they reveal a wealth of additional new non-random aspects of genome sequence organization.

© 1996 Academic Press Limited

Keywords: chromatin; dinucleotides; Fourier transform; genomic DNA sequence; nucleosome

Genomic DNA sequences may contain many signals that have not yet been detected, including signals that represent important constraints on evolution. Recent advances made by genomic sequencing projects encourage a fresh analysis. One interesting possibility is that genomes could be under selective pressure because of requirements for particular DNA physical properties (Widom, 1985; Travers & Klug, 1987), as well as for particular sequences. Dinucleotide steps, which include one base-pair stacking interaction, are fundamental units of DNA structure and dynamics (Calladine & Drew, 1984; Yanagi *et al.*, 1991). Analysis of the distributions of dinucleotides in genomic sequences is complicated by the prevalence of codons. These impose a strong modulation with a 3 bp period that

effectively obscures other signals. Fourier transformation separates out this 3 bp modulation into a single peak, allowing other spectral regions to be examined free of its effects. With this approach, previously unrecognized periodic signal components are readily detected.

Fourier transform analysis has been used in other studies of genome organization (Li & Kaneko, 1992; Voss, 1992; Buldyrev *et al.*, 1995); however, those studies differ in focus from the present one. In particular, they are concerned chiefly with analysis of long-range correlations. They do not systematically investigate the different dinucleotide steps, nor do they compare the high-frequency regions of the spectra with random expectation.

I analyzed genomic sequences from two eukary-

otes and one prokaryote: ≈ 3.9 Mbp from eight completely sequenced *Saccharomyces cerevisiae* chromosomes (Bussey *et al.*, 1995; Feldmann *et al.*, 1994; Oliver *et al.*, 1992; Johnston *et al.*, 1994; Dujon *et al.*, 1994; and see Acknowledgements), omitting 1000 bp from each end, which include telomeric sequences; ≈ 18 Mbp from ≈ 460 genomic cosmid clones of *Caenorhabditis elegans* (Sulston *et al.*, 1992; and see Acknowledgements); and ≈ 1.8 Mbp from the complete genomic sequence of the prokaryote *Haemophilus influenzae* (Fleischmann *et al.*, 1995; and see Acknowledgements), in three separate sets of computations for each of the ten independent dinucleotide steps.

For a particular dinucleotide step, genomic sequences were encoded 1 at positions corresponding to the first nucleotide of that dinucleotide step, and 0 elsewhere. The power spectra of the encoded dinucleotide signals were evaluated over segments and summed over all segments for that genome. The results are shown in Figures 1 to 3, along with the means from ten randomized runs for each dinucleotide. The standard deviations of the means were evaluated but are too small to be visible on these plots (data not shown).

All of the power spectra reveal peaks that are statistically significant whether measured by standard deviations from the mean random spectra or

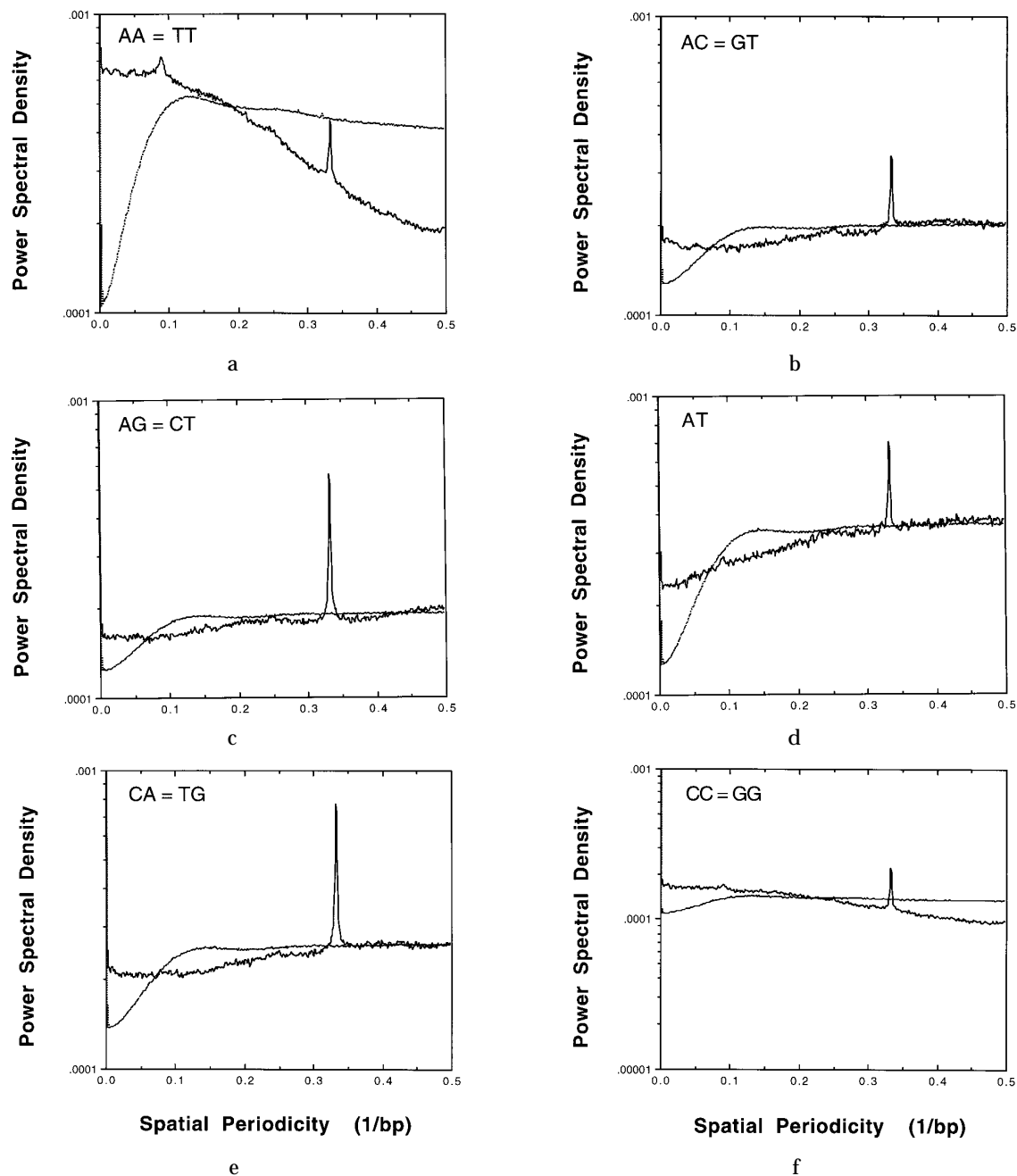


Figure 1a-f (legend opposite).

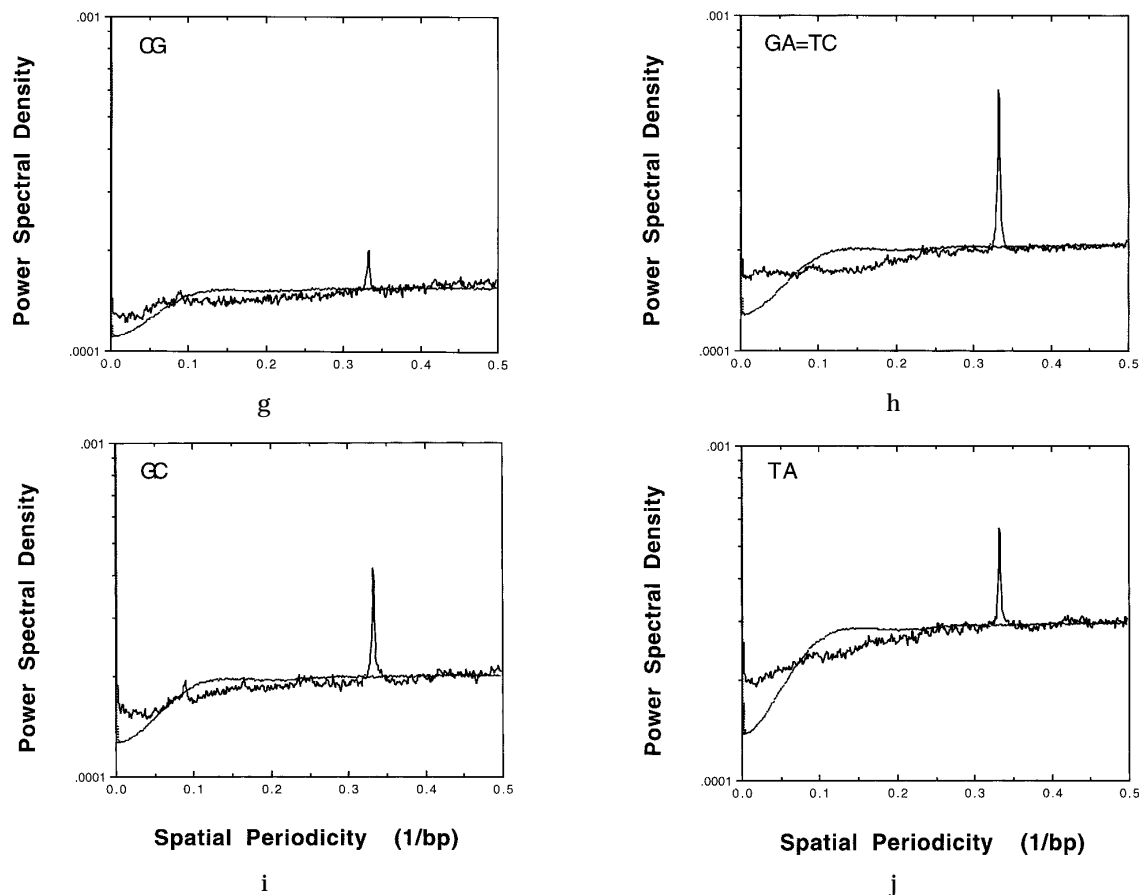


Figure 1g-j.

Figure 1. Power spectra (continuous lines) for all ten independent dinucleotide steps in the genome of the prokaryote *H. influenzae* (Fleischmann *et al.*, 1995; and see Acknowledgements) compared with the means (thinner, dotted lines) from calculations on ten randomized sequences. The standard deviations of the means were also evaluated but are too small to be visible on these plots. The apparent noise fluctuations in the experimental signals (in this and the following Figures) are much greater than the standard deviations between random trials (which are too small to represent in these Figures). Therefore, these fluctuations reflect other non-random aspects of genome organization that vary significantly between closely spaced spatial periodicities. The breadths of the many peaks found in this and the following Figures at periodicities other than 0.333 bp^{-1} exceed that of the 0.333 bp^{-1} peak, suggesting that they may represent imperfectly periodic signal components or a superposition of several components at closely spaced periodicities. The DNA sequence was obtained from Internet site <http://www.tigr.org/>. The sequence was encoded as described in the text. Power spectra were calculated using the program SPCTRM (Press *et al.*, 1986) in consecutive half-overlapping 1024 bp segments, and were summed for all segments from the encoded sequence and its reverse complement. Hence, results for any particular dinucleotide are identical for the reverse complement of that dinucleotide (parentheses). For creation of the randomized sequences, the number of occurrences of a particular dinucleotide within the real sequence was counted, and then an equal number were placed at random locations chosen by the random number generator RAN1 (Press *et al.*, 1986), with proper handling of permitted and forbidden adjacencies. As for the real sequences, both the “forward” and reverse-complement of each randomized sequence were included in the calculations of the power spectra. The power spectra were evaluated separately for each of ten sets of randomized sequences, and the mean and standard deviation in the power spectral density were evaluated at each reciprocal lattice point. Depending on the nature of disorder in the sequences, power spectra may vary somewhat with the segment length. A segment length of 1024 bp provided a good balance between signal to noise ratios (which improve as the number of segments is increased; i.e. as the segment length is decreased) and spatial scope and resolution (which improve as the segment length is increased). Control calculations used much longer and much shorter segment lengths; the results obtained are quantitatively similar and qualitatively equivalent to those presented here. As another control, entirely independent programs based on “maximum entropy” methods for estimating the power spectra were used to repeat certain of the calculations presented in subsequent Figures.

compared with neighboring fluctuations within the signal itself. The nature of the transforms suggests that the peaks represent first-order maxima from differing periodic components of the signal, rather

than subsidiary maxima from a single larger structural unit.

A sharp peak at a periodicity of 0.333 bp^{-1} is observed in all of the power spectra; it corresponds

to 3.00 bp in real-space, and presumably reflects the existence of strings of codons. An important benefit of analyzing the signal in reciprocal space *via* the power spectra, instead of in real-space, is that the 3 bp modulation is condensed into a single peak, allowing other regions of the spectra to be examined free of its effects.

For many but not all of the dinucleotide steps, there are systematic discrepancies between the experimental power spectra and the mean random spectra over the full range of periodicities. The close agreement between the experimental signal and the mean random signal found for many cases (e.g. Figure 2b, d, e, g and i, and Figure 3b, e and h)

confirms that the randomization algorithm is appropriate. We conclude that the many examples of systematic deviations from the mean random signals represent real properties of genomes. These systematic deviations represent non-random distributions of those dinucleotides over the full range of lengthscales investigated here, 2 to ≈ 500 bp.

Novel non-random features at specific spatial periodicities

Additional striking new findings are the peaks in the power spectra at periodicities other than 0.333 bp^{-1} . For the prokaryotic genome (Figure 1),

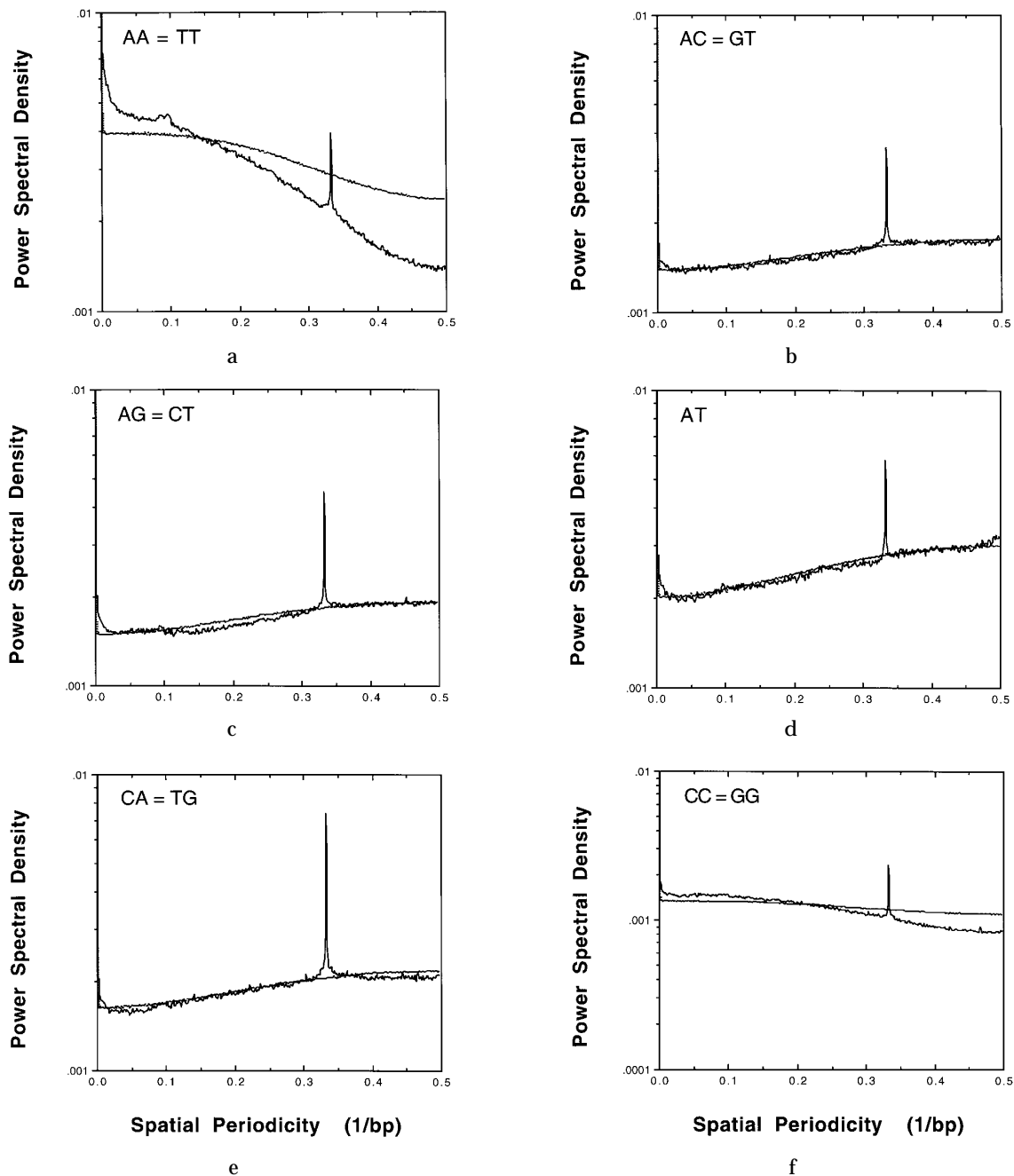


Figure 2a-f (legend opposite).

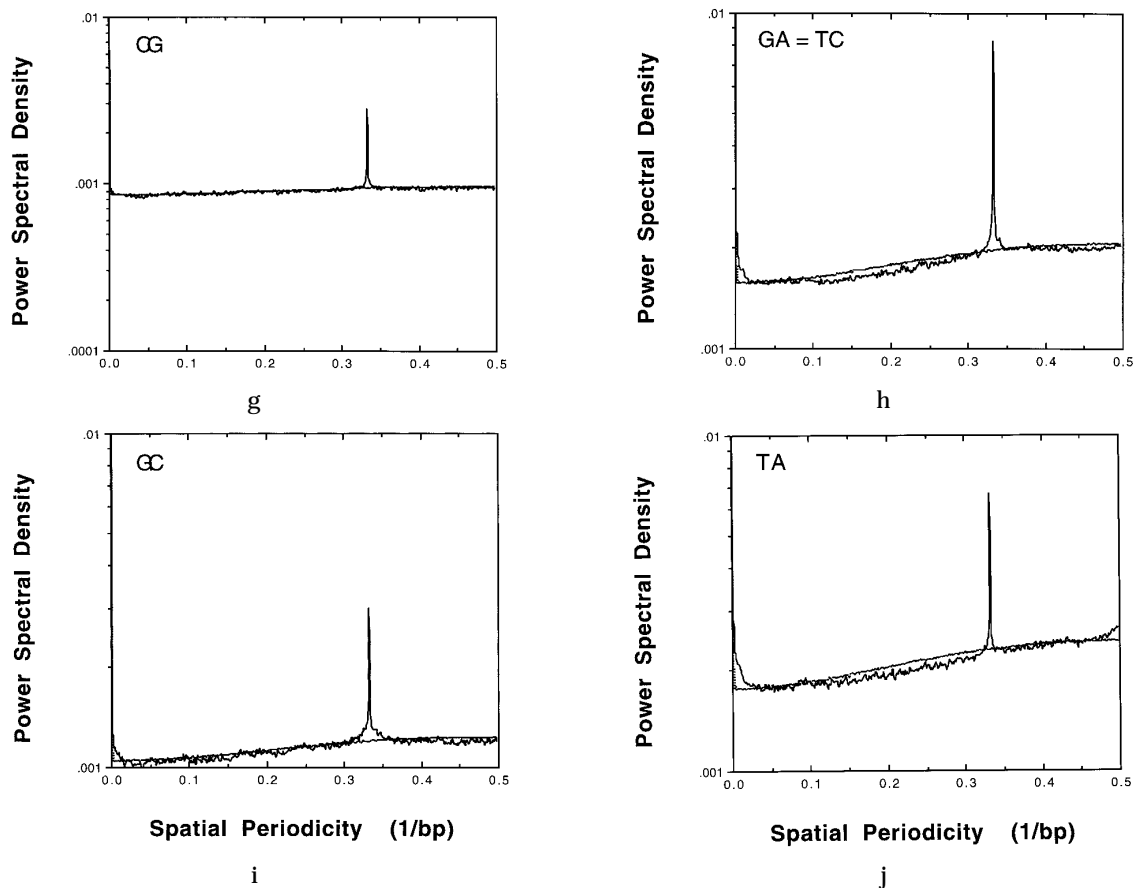


Figure 2g-j.

Figure 2. Power spectra (continuous lines) for all ten independent dinucleotide steps in the genome of the eukaryote *S. cerevisiae*, chromosomes I, II, III, V, VI, VIII, IX and XI (Bussey *et al.*, 1995; Feldmann *et al.*, 1994; Oliver *et al.*, 1992; Johnston *et al.*, 1994; Dujon *et al.*, 1994; and see Acknowledgements) compared with the means (thinner, dotted lines) from calculations on ten randomized sequences. The standard deviations of the means were evaluated but are too small to be visible on these plots. DNA sequences were obtained from Internet site <http://genome-gopher.stanford.edu/>: 1000 bp (which include the telomeres) were eliminated from each end. Power spectra were summed over all overlapping segments from all eight chromosomes. Other methods are the same as Figure 1. In certain cases, substantially longer and shorter segment lengths were investigated. Certain additional calculations were carried out using the programs MEMCOF and EVLMEM for estimation of power spectra using the maximum entropy method (Press *et al.*, 1986). Other calculations investigated individual chromosomal sequences or subsets of them, and sequences from which the centromeres were also deleted. Variants of the computer programs were written to calculate the transforms from non-overlapping segments, allowing identification of the genomic regions contributing the highest intensity to the peak for "AA or TT" integrated over the range 0.0967 to 0.0986 bp^{-1} .

notable additional peaks are found with the dinucleotide AA (=TT) at a periodicity of $\approx 0.089 \text{ bp}^{-1}$; with CC (=GG) at a periodicity of $\approx 0.09 \text{ bp}^{-1}$; perhaps with CG at a periodicity of $\approx 0.089 \text{ bp}^{-1}$; and with GC at periodicities of $\approx 0.089 \text{ bp}^{-1}$, $\approx 0.164 \text{ bp}^{-1}$ and ≈ 0.24 to 0.25 bp^{-1} . For the eukaryotic *S. cerevisiae* genome (Figure 2), notable additional peaks are found for the dinucleotide AA (=TT) at two periodicities, $\approx 0.088 \text{ bp}^{-1}$ and $\approx 0.097 \text{ bp}^{-1}$. For the eukaryotic *C. elegans* genome (Figure 3), notable additional peaks are found with the dinucleotide AA (=TT), which reveals a doublet at periodicities of $\approx 0.093 \text{ bp}^{-1}$ and ≈ 0.098 to 0.099 bp^{-1} ; with CC (=GG) at periodicities of $\approx 0.090 \text{ bp}^{-1}$ and $\approx 0.101 \text{ bp}^{-1}$; weakly, with GA (=TC) at a periodicity

of $\approx 0.100 \text{ bp}^{-1}$; and with GC at periodicities of $\approx 0.098 \text{ bp}^{-1}$ and $\approx 0.166 \text{ bp}^{-1}$.

The doublet observed for AA (=TT) in the *C. elegans* genome at periodicities of $\approx 0.093 \text{ bp}^{-1}$ and ≈ 0.098 to 0.099 bp^{-1} is a real doublet. The two peaks are formally well resolved, being separated by five reciprocal lattice points. Moreover, this doublet is not simply an artifactual result of noise superimposed on a single, broader, peak. Repeating the calculation using only 40% of the available sequence information leads to an identical appearance to this region of the power spectrum (results not shown).

The majority of these peaks appear to fall in three groups: a peak near ≈ 0.089 to 0.093 bp^{-1} , which appears for several dinucleotides in both

prokaryotic and eukaryotic genomes; a peak near ≈ 0.164 to 0.166 bp^{-1} , which is found with both prokaryotic and eukaryotic genomes but only for the dinucleotide GC; and a peak near ≈ 0.097 to 0.101 bp^{-1} , which appears for several dinucleotides but only for the eukaryotic genomes.

It is conceivable that the peaks at ≈ 0.089 to 0.093 bp^{-1} and ≈ 0.164 to 0.166 bp^{-1} , found for both prokaryotes and eukaryotes, could arise from periodic aspects of protein structure. In real-space, these peaks are centered at periodicities of $\approx 11 \text{ bp}$ and $\approx 6 \text{ bp}$, which corresponds to ≈ 3.7 and 2 codons, relatively close to the repeat length of the α -helix, and quite close to that of the β -sheet, respectively. Alternatively, a reviewer notes that the

peak centered at $\approx 11 \text{ bp}$ is close to the relative helical repeat for superhelical DNA in *Escherichia coli*.

Origin of the $\approx 10.2 \text{ bp}$ periodicity

The peak at ≈ 0.097 to 0.101 bp^{-1} appears most prominently in both eukaryotic genomes for the dinucleotide AA (=TT), where it is centered at $\approx 0.098 \text{ bp}^{-1}$, although in the larger *C. elegans* genome it is detected also for CC (=GG), GA (=TC) and GC. I carried out several additional investigations on the origin of this repeat for AA (=TT) in the yeast genome. This peak does not arise from telomeric sequences, since these are omitted from

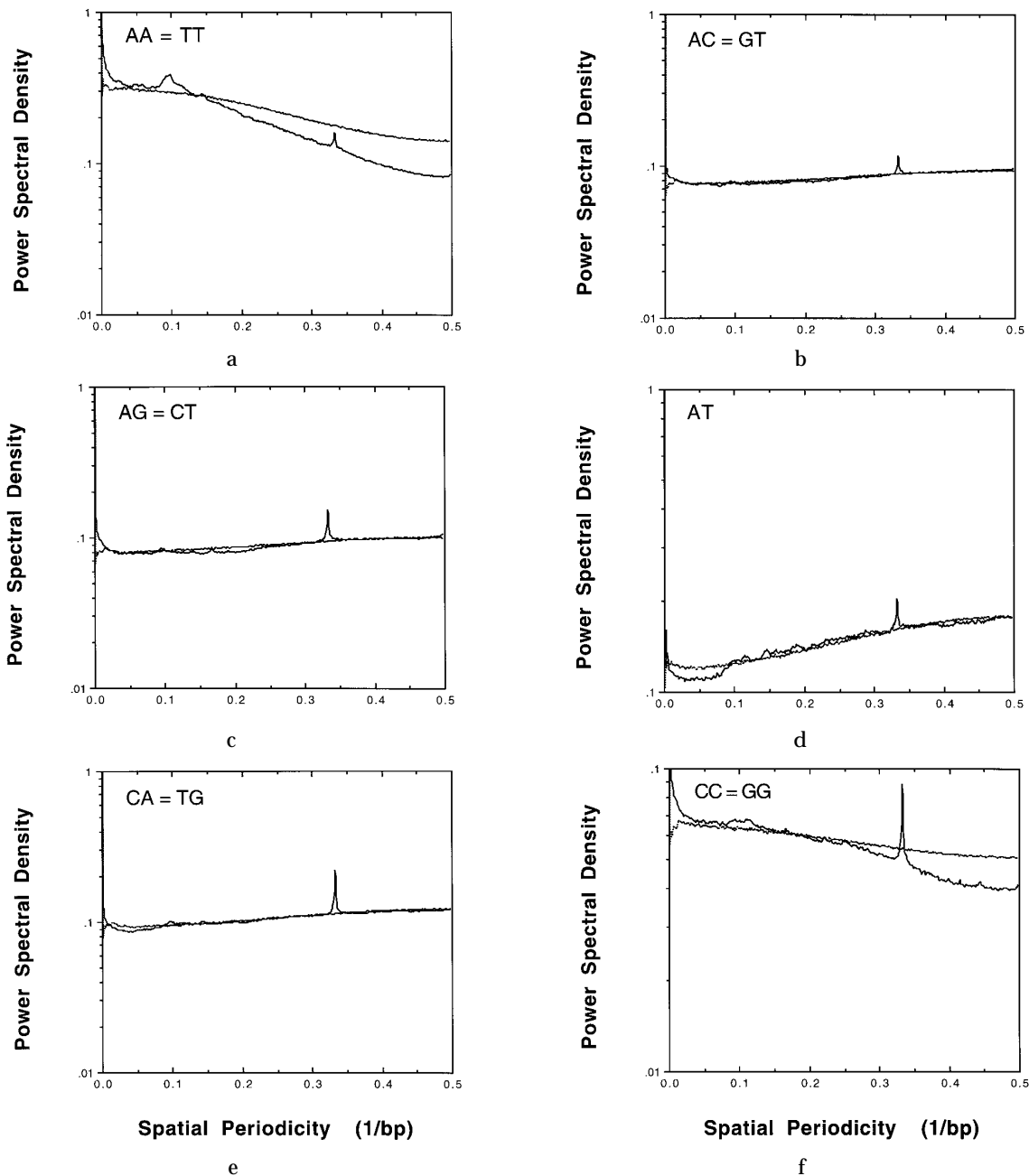


Figure 3a-f (legend opposite).

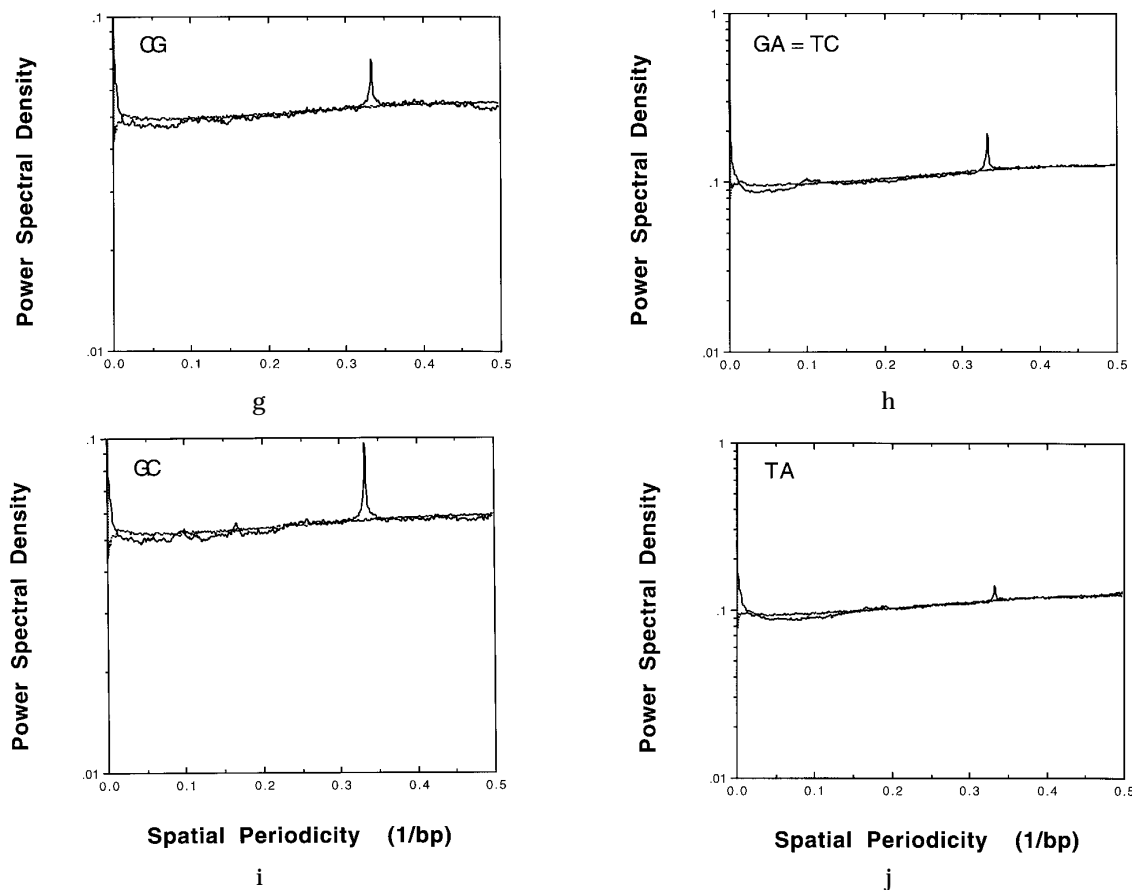


Figure 3g-j.

Figure 3. Power spectra (continuous lines) for all ten independent dinucleotide steps in the genome of the eukaryote *C. elegans* (Sulston *et al.*, 1992; and see Acknowledgements) compared with the means (thinner, dotted lines) from calculations on ten randomized sequences. The standard deviations of the means were evaluated but are too small to be visible on these plots. Genomic DNA sequences from individual cosmids were obtained from Internet sites http://www.sanger.ac.uk/~sjj/C.elegans_Home.html and <ftp://genome.wustl.edu/pub/gsc1/sequence/st.louis/elegans/genbank/>. The first 200 bp in these sequences are potentially duplicated in another cosmid (for contig construction) and were eliminated. Other methods are the same as for Figure 1.

the calculations. It originates in multiple chromosomes, because it can be detected in distinct individual chromosome sequences and in subgroups of chromosome sequences (results not shown). An additional calculation on six yeast chromosomal sequences from which the centromeres were deleted revealed this peak still to be present with comparable intensity (results not shown). Finally, as will be seen below, equivalent peaks that are obtained after encoding "AA or TT" cannot be attributed to known repetitive elements in the genome. Taken together, these results lead to the conclusion that the peak at ≈ 0.097 to 0.101 bp^{-1} obtained for the two eukaryotic genomes and with a variety of dinucleotides reflects properties of the bulk of eukaryotic DNA.

The periodicity of this peak's center corresponds to a real-space repeat of $\approx 10.2 \text{ bp}$. It is noteworthy that this periodicity is close to the average helical repeat reported for DNA in a nucleosome (Lutter, 1978; Hayes *et al.*, 1990), as well as to the repeat length by which linker DNA lengths are preferen-

tially quantized in chromatin (Widom, 1992; Yao *et al.*, 1993). Moreover, repeats of AA, CC (=GG) and GC having similar periodicities have been detected in eukaryotic DNA sequences that have been physically selected for formation of stable nucleosome core particles (Satchwell *et al.*, 1986) or for nucleosome positioning (Ioshikhes *et al.*, 1992; Bina, 1994; Staffelbach *et al.*, 1994; Bolshoy, 1995). Selected DNA molecules are necessarily non-random in some manner; importantly, the present analysis reveals that these particular non-random aspects of those selected DNA molecules are also predominant periodic signals present in unbiased samples of entire eukaryotic genomes. This analysis also reveals a weak signal at this periodicity for the additional dinucleotides GA = TC, which were not detected in the selected sequences.

In the selected sequences, the periodic signal can be enhanced by encoding "AA or TT" rather than AA or TT alone (Satchwell *et al.*, 1986). The results of Fourier transform calculations using such encodings for the *H. influenzae*, *S. cerevisiae* and

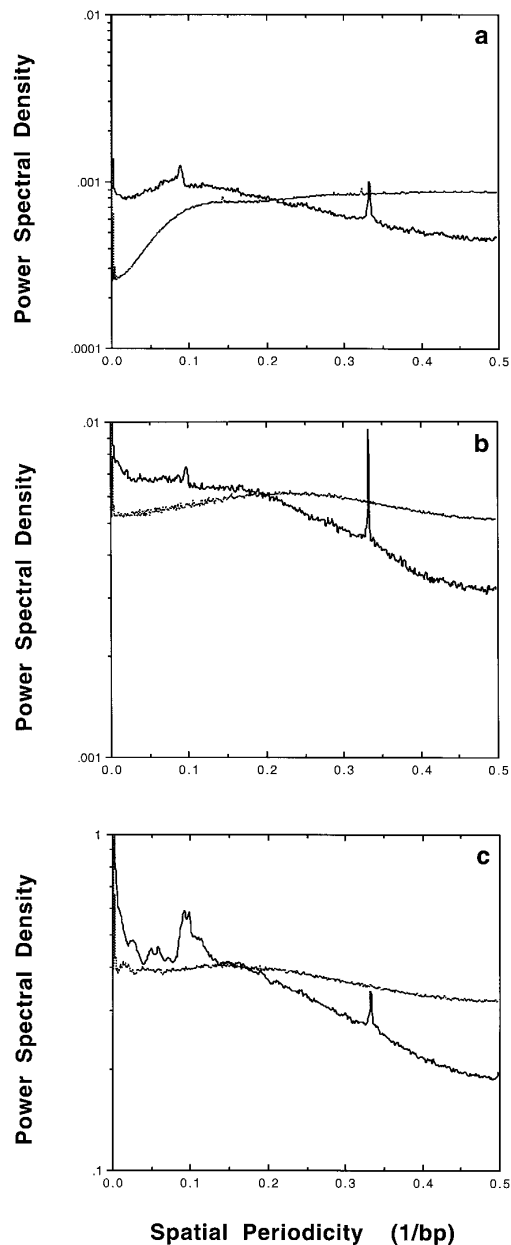


Figure 4. Power spectra (continuous lines) for genomes encoded "AA or TT" compared with the mean (thinner, dotted lines) from calculations on ten randomized sequences for the same genomes. a to c, Results for *H. influenzae*, *S. cerevisiae* and *C. elegans* genomes, respectively. Methods were as for Figures 1 to 3, except that the genomic sequences were encoded 1 at positions corresponding to the first nucleotide of all AA or TT dinucleotide steps, and 0 elsewhere.

C. elegans genomes are shown in Figure 4a to c, respectively.

For the prokaryotic genome (Figure 4a), there is little change compared with the results obtained encoding AA (=TT) alone (compare with Figure 1a). The peak at a periodicity of ≈ 0.089 bp⁻¹ remains at the same position and with comparable intensity relative to the background on which it is superimposed. The position of this peak corresponds to a real-space periodicity of ≈ 11.3 bp,

distinct from the peak at ≈ 10.2 bp (the outermost peak of the ≈ 0.09 to 0.1 bp⁻¹ doublet) obtained with the two eukaryotic genomes.

In striking contrast to these results with the prokaryotic genome, there are substantial changes to the ≈ 0.09 to 0.1 bp⁻¹ (≈ 10 - 11 bp in real-space) region of the power spectra when the two eukaryotic genomes are encoded "AA or TT". For the *S. cerevisiae* genome (Figure 4b; cf. Figure 2a), the innermost (≈ 0.088 bp⁻¹, or ≈ 1.4 bp in real-space) peak of the doublet near ≈ 0.1 bp⁻¹ periodicity is slightly suppressed, while the outermost peak of the doublet (≈ 0.098 bp⁻¹, or ≈ 10.2 bp in real-space) is substantially strengthened. For the *C. elegans* genome (Figure 4c; cf. Figure 3a), the results are even more striking. The doublet character of the peak remains and is emphasized. The two peaks of the doublet occur at real-space periodicities of ≈ 10.9 bp (innermost peak) and ≈ 10.1 bp (outermost peak). Remarkably, this doublet has become the dominant feature in the transforms, surpassing the peak attributed to codons in both height and integrated intensity.

In summary, the two eukaryotic genomes, which reveal a peak at ≈ 10.2 (or 10.1) bp for the AA (=TT) encoding, do so again with the "AA or TT" encoding, and the intensity of this peak is strengthened in the "AA or TT" encoding. In contrast, for the prokaryotic genome, no such peak is detected for AA (=TT) encoding, and the "AA or TT" encoding leaves this region of the power spectra essentially unchanged.

Chromosomal origins of the ≈ 10.2 bp periodicity

A variant of the computer program monitored the integrated intensity of this peak (with "AA or TT" encoded) for every consecutive 1024 bp segment in the eight yeast chromosome sequences. As could be anticipated, every segment contributed some intensity to this region of the power spectrum. The top-scoring group of 20 1024 bp segments were further analyzed. This group does not contain known repetitive DNA elements. Interestingly, it includes segments that are completely contained within protein-coding sequences, in addition to segments that are completely contained between two adjacent coding sequences, and segments that partially overlap with coding sequences. These findings suggest that if this ≈ 10.2 bp periodicity reflects genomic sequences that are selected for nucleosome positioning, then nucleosome positioning has roles at locations throughout genes, not just at their promoters.

Real-space analysis

The detection of the peak near ≈ 0.097 to 0.101 bp⁻¹ (≈ 10.2 bp periodicity in real-space) provides an illustration of the power of the present analysis. In real-space, this signal is buried by the 3 bp modulation. This property is examined in

Table 1. Real-space analysis of dinucleotide correlations in the yeast genome

Signal	λ (bp)	Occurrence	Random	σ_{random}	$ \delta $
AA:AA (=TT:TT)	9	101,040	98,324	148.2	11.7
	10	96,680	98,573	238.2	8.2
	11	96,820	98,478	181.1	7.1
	12	99,339	98,332	332.5	4.3
AA:TT	9	75,212	98,469	256.9	100.4
	10	75,920	98,588	217.8	97.8
	11	76,772	98,460	215.0	93.6
	12	78,092	98,545	313.4	88.3
CC:CC (=GG:GG)	9	13,996	11,734	72.5	25.6
	10	11,860	11,768	85.6	1.0
	11	11,825	11,726	106.7	1.1
	12	14,200	11,746	81.7	27.7
GC:GC	9	13,876	10,328	66.9	43.9
	10	10,146	10,287	82.5	1.7
	11	10,394	10,317	79.6	1.0
	12	14,736	10,279	104.7	55.1

Dinucleotide (wx) correlated with dinucleotide (yz) at distances of λ bp. λ is defined as distance in bp of first nucleotide of the second dinucleotide (y) from first nucleotide of the first dinucleotide (w). Occurrence, actual number of occurrences in seven yeast chromosomes and their reverse complements (minus 1000 bp from each end;—see the text); only seven chromosome sequences (instead of eight) used for this analysis. Random, mean number of occurrences in ten randomized runs; randomizations were carried out as for the Fourier transform analysis. σ_{random} , Standard deviation of number of occurrences over the ten random runs. $|\delta|$, Absolute value of the difference between the actual number of occurrences and random expectation, in multiples of the standard deviation. (Note: there is much noise in σ with only ten random runs; hence δ is computed using the mean value of σ , which is calculated using actual σ values obtained from $\lambda = 1$ to $\lambda = 200$.)

Table 1, for the dinucleotides that were previously implicated in nucleosome positioning (Satchwell *et al.*, 1986; Ioshikhes *et al.*, 1992; Bina, 1994; Staffebach *et al.*, 1994; Bolshoy, 1995). Direct analysis of real-space correlations in the *S. cerevisiae* genome (Table 1) reveals that AA dinucleotides recur at distances of 10 bp or 11 bp less often than random expectation, by ≈ 8 and ≈ 7 standard deviations, respectively. In other words, AA (and TT) dinucleotides are anticorrelated at distances of 10 and 11 bp in the yeast genome. Nevertheless, the transforms reveal enhanced power at the corresponding spatial periodicity. How can both of these statements be true at once? The real-space correlations at distances of 10 to 11 bp are suppressed by the 3 bp oscillation, which has a local minimum at ≈ 10.5 bp (note the positive real-space correlations, which have maxima at λ values equal to every integral multiple of 3 bp (Table 1 and data not shown). The actual signal represents a superposition of this 3 bp modulation on top of the signal having a ≈ 10.2 bp periodicity; their amplitudes and phases are such that the net actual signal happens to have a minimum at $\lambda = 10$ to 11 bp, despite the enhanced spectral power at that and other periodicities.

Table 1 reveals other examples of related signals that cannot be detected in real-space but are revealed by the transforms. AA dinucleotides are strongly anticorrelated with TT (by ≈ 90 to 100 standard deviations) at distances of 10 and 11 bp in

real-space. Thus, AA is anticorrelated with both AA and TT at these distances; nevertheless, as seen above, the transforms detect a strong peak at the corresponding periodicity for the signal encoded "AA or TT". CC dinucleotides are positively correlated with other CC dinucleotides (equivalently, GG with GG) at distances of both 10 and 11 bp; but the difference from random expectation is only ≈ 1 standard deviation, so this correlation has only weak statistical significance. GC is anticorrelated with itself at a distance of 10 bp (by ≈ 2 standard deviations, modest statistical significance) and positively correlated with itself at a distance of 11 bp (by ≈ 1 standard deviation, weak statistical significance). Nevertheless, all of these dinucleotides, and others, reveal significant peaks in the corresponding regions of the power spectral density functions.

Conclusions

The present study plainly illustrates the power of the reciprocal space analysis and reveals a wealth of previously unrecognized non-random aspects of genome organization over lengthscales of ≈ 2 to 500 bp in both prokaryotic and eukaryotic genomes. Most importantly, the results suggest that the requirements of nucleosomal organization may contribute significant constraints on the sequences of eukaryotic genomes, and they reveal new dinucleotide steps that may contribute toward this organization. If these signals do indeed reflect DNA sequence-directed nucleosome positioning (or biasing of positioning), then the present results imply that such positioning has significance at locations throughout genes, not just at promoters. It will be of great interest to extend the analysis to new and larger genome databases as these become available, as well as to longer lengthscales and to new encoding schemes. Finally, there remains the task of experimentally testing the interpretation of each new signal.

Acknowledgements

The author is grateful to Drs S. Scherer and M. Johnston for comments on the manuscript, and acknowledges his indebtedness to the many investigators involved in genome sequencing projects that made this study possible: for *S. cerevisiae*, the *Saccharomyces* Genomic Information Resource at Stanford University (WWW site <http://genome-gopher.stanford.edu/>) and the investigators cited therein for the sequences of chromosomes I, II, III, V, VI, VIII, IX and XI; for *C. elegans*, the investigators of the Sanger Center (WWW site http://www.sanger.ac.uk/~sjj/C.elegans_Home.html) and the Washington University Saint Louis *C. elegans* sequencing project (WWW site <ftp://genome.wustl.edu/pub/gsc1/sequence/st.louis/elegans/genbank/>); and for *H. influenzae*, The Institute for Genomic Research (WWW site <http://www.tigr.org/>). Research in the author's laboratory is supported by the NIH and by the

generous donation of equipment from the Hewlett Packard Foundation.

References

- Bina, M. (1994). Periodicity of dinucleotides in nucleosomes derived from simian virus 40 chromatin. *J. Mol. Biol.* **235**, 198–208.
- Bolshoy, A. (1995). CC dinucleotides contribute to the bending of DNA in chromatin. *Nature Struct. Biol.* **2**, 446–448.
- Buldyrev, S. V., Goldberger, A. L., Havlin, S., Mantegna, R. N., Matsu, M. E., Peng, C.-K., Simons, M. & Stanley, H. E. (1995). "Long-range correlation properties of coding and noncoding DNA sequences: Gen Bank analysis. *Phys. Rev. ser. E.* **51**, 5084–5091.
- Bussey, H., Kaback, D. B., Zhong, W., Vo, D. T., Clark, M. W., Fortin, N., Hall, J., Ouellette, B. F., Keng, T., Barton, A. B. *et al.* (1995). The nucleotide sequence of chromosome I from *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA*, **92**, 3809–3813.
- Calladine, C. R. & Drew, H. R. (1984). A base-centred explanation of the B-to-A transition in DNA. *J. Mol. Biol.* **178**, 773–782.
- Dujon, B., Alexandraki, D., Andre, B., Ansoerge, W., Baladron, V., Ballesta, J. P., Banrevi, A., Bolle, P. A., Bolotin-Fukuhara, M., Bossier, P. *et al.* (1994). Complete DNA sequence of yeast chromosome XI. *Nature*, **369**, 371–378.
- Feldmann, H., Aigle, M., Aljinovic, G., Andre, B., Baclet, M. C., Barthe, C., Baur, A., Becam, A. M., Biteau, N., Boles, E. *et al.* (1994). Complete DNA sequence of yeast chromosome II. *EMBO J.* **13**, 5795–5809.
- Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
- Hayes, J. J., Tullius, T. D. & Wolffe, A. P. (1990). The structure of DNA in a nucleosome. *Proc. Natl Acad. Sci. USA*, **87**, 7405–7409.
- Ioshikhes, I., Bolshoy, A. & Trifonov, E. N. (1992). Preferred positions of AA and TT dinucleotides in aligned nucleosomal DNA sequences. *J. Biomol. Struct. Dynam.* **9**, 1111–1117.
- Johnston, M. Andrews, S., Brinkman, R., Cooper, J., Ding, H., Dover, J., Du, Z., Favello, A., Fulton, L., Gattung, S. *et al.* (1994). Complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome VIII. *Science*, **265**, 2077–2082.
- Li, W. & Kaneko, K. (1992). Long-range correlation and partial $1/f^2$ spectrum in a noncoding DNA sequence. *Europhys. Letters*, **17**, 655–660.
- Lutter, L. C. (1978). Kinetic analysis of deoxyribonuclease I cleavage in the nucleosome core: evidence for a DNA superhelix. *J. Mol. Biol.* **124**, 391–420.
- Oliver, S. G., van der Aart, Q. J., Agostoni-Carbone, M. L., Aigle, M., Alberghina, L., Alexandraki, D., Antoine, G., Anwar, R., Ballesta, J. P., Benit, P. *et al.* (1992). The complete DNA sequence of yeast chromosome III. *Nature*, **357**, 38–46.
- Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1986). *Numerical recipes*, Cambridge University Press, Cambridge, UK.
- Satchwell, S. C., Drew, H. R. & Travers, A. A. (1986). Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191**, 659–675.
- Staffelbach, H., Koller, T. & Burks, C. (1994). DNA structural patterns and nucleosome positioning. *J. Biomol. Struct. Dynam.* **12**, 301–325.
- Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L. *et al.* (1992). The *C. elegans* genome sequencing project: a beginning. *Nature*, **356**, 37–41.
- Travers, A. A. & Klug, A. (1987). The bending of DNA in nucleosomes and its wider implications. *Phil. Trans. Roy. Soc. ser. B*, **317**, 537–561.
- Voss, R. F. (1992). Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences. *Phys. Rev. Letters*, **68**, 3805–3808.
- Widom, J. (1985). Bent DNA for gene regulation and DNA packaging. *BioEssays*, **2**, 11–14.
- Widom, J. (1992). A relationship between the helical twist of DNA and the ordered positioning of nucleosomes in all eukaryotic cells. *Proc. Natl Acad. Sci. USA*, **89**, 1095–1099.
- Yanagi, K., Prive, G. C. & Dickerson, R. E. (1991). Analysis of local helix geometry in three B-DNA decamers and eight dodecamers. *J. Mol. Biol.* **217**, 201–214.
- Yao, J., Lowary, P. T. & Widom, J. (1993). Twist constraints on linker DNA in the 30 nm chromatin fiber: implications for nucleosome phasing. *Proc. Natl Acad. Sci. USA*, **90**, 9364–9368.

Edited by P. E. Wright

(Received 31 January 1996; received in revised form 18 March 1996; accepted 2 April 1996)