On the Relationship Between DNA Periodicity and Local Chromatin Structure

Sheila M. Reynolds, Jeff A. Bilmes, and William Stafford Noble

University of Washington,
Seattle, Washington, USA
{sheila,bilmes}@ee.washington.edu
noble@gs.washington.edu

Abstract. DNA periodicity and its relationship to the formation of nucleosomes has been investigated extensively using autocorrelation and Fourier transform methods. We provide a precise treatment of the mathematical foundation for this type of analysis, and we apply the resulting method to quantify dinucleotide periodicity in several datasets. We begin by demonstrating, via simulation, the sensitivity of our method relative to previous methods. We then provide evidence of pervasive ∼10 bp periodicity in *S. cerevisiae*, with stronger periodicity in sequences associated with positioned nucleosomes. In human, although repeat-masked sequences do not exhibit significant periodicity on average, we find that experimentally determined nucleosome positions show a periodicity of the AA dinucleotide similar to that found in *S. cerevisiae*. Furthermore, transcription start sites in the human genome are marked by a sharp drop in the 10 bp periodicity of the AA dinucleotide, while occupied CTCF sites are surrounded by a local increase.

Key words: DNA periodicity, nucleosome, dinucleotide, chromatin

1 Introduction

A relationship between DNA sequence periodicity and local curvature of the DNA molecule and hence local chromatin structure has been hypothesized for over 30 years, going back to a "kinky helix" model proposed by Crick and Klug in 1975 [1]. The strongest periodic component in DNA sequences (excluding simple repeats) is generally observed in coding regions and is induced by the codon length and the bias in both amino acid usage and codon usage. This codon effect results in a spike in the spectrum at the normalized frequency $f = 1/3$, corresponding to a period of $T = 3$ bp. A second, weaker periodicity near 10 bp has also been observed in many different organisms, in coding, non-coding and repeat-masked sequences. This periodicity has been linked to the pitch of the DNA helix [2] as well as to the alternation of hydrophobic and hydrophilic amino acids in protein sequences [3]. The earliest evidence of this periodicity was based on just 36 kb of DNA sequence collected from several different eukaryotes as well as certain viruses [2]. Satchwell *et al.* isolated and sequenced 177 chicken nucleosomes, and observed a 10 bp periodicity of both AA/TT and GG/CC with the minor groove of AA/TT facing predominantly inward toward the histone and GG/CC facing outward [4]. Similar ∼10 bp periodicity signals derived from short patterns (up to tetra-nucleotides) have been observed in a wide variety of organisms and datasets, both natural and artificial [5], [6], [7], [8], [9]. These ∼10 bp periodicities are strongly correlated with nucleosome positioning, supporting the hypothesis that periodic sequence elements in phase with the DNA helix are related to large-scale bending of the DNA molecule. DNA bendability has been extensively modeled [10], [11], [12], and experimentally measured [13], with a general consensus that poly(dA:dT) tracts are extremely stiff [14], while some short sequences are very flexible, particularly the CA/TG dinucleotide and the CAG/CTG trinucleotide [15], and others such as TA are context-dependent [10].

Most previous approaches to quantifying periodicity in DNA sequences have been based on Fourier techniques which require that the symbolic DNA sequence be translated first into a numerical sequence. The spectrum of the numerical sequence can then be estimated directly using the Fourier transform [7], [16], [17], or by first computing the autocorrelation function [2], [18], [19], [20]. Sequence periodicity has also been studied using a machine-learning based approach which sought to learn a periodic nucleotide pattern in an unsupervised fashion using a cyclic hidden Markov model (HMM) [23], [24]. In the current work, we describe the drawbacks to these previously described approaches to characterizing DNA periodicity and we introduce a mathematically precise approach to evaluating the spectral content of DNA sequences.

With respect to the HMM analysis [24], we report that the scarcity of the CpG dinucleotide appears to be the main contributing factor that led the 10-state cyclic HMM to learn the apparently periodic (not-T)(A/T)(G) pattern. Using a dynamic Bayesian network (DBN) similar to the HMM described by Baldi *et al.* [24], we were able to reproduce this pattern by training on the repeat-masked ENCODE regions of the human genome. However, we found that the same pattern is learned from random DNA generated according to a second order Markov model which reproduces only the single, di- and trinucleotide statistics of human DNA. Additionally we found that the same pattern can be learned by smaller cyclic models constrained to allow period lengths as short as 4-6 bases.

We introduce ACSE (AutoCorrelation Spectral Estimation), a method for quantifying the periodicity of a specified fixed length pattern in DNA, which includes a null model and an estimation of the variance in the estimated spectral amplitude, and we demonstrate that our method is more sensitive to weak evidence of periodicity than previous spectral methods. We have applied our method to a variety of datasets from the human and yeast genomes, including the complete yeast genome, the human ENCODE regions, experimentally identified nucleosome sequences from human and yeast, and human DNA sequences proximal to transcription start sites (TSSs) and CTCF binding sites. We show that nucleosome sequences in both genomes exhibit increased ∼10 bp periodicity, especially of the AA/TT dinucleotide. This result contrasts with a recent study that found no evidence of 10-bp periodicity in human nucleosome sequences [25].

In addition, we report a sharp decrease of the 10-bp AA/TT periodicity in the vicinity of TSSs, and a local increase around binding sites of the CTCF insulator protein. These results indicate a strong relationship between DNA periodicity and local chromatin structure and are consistent with the classical statistical positioning theory of nucleosome organization [26], [27], which posits that nucleosomes are stochastically positioned along the genome and are distributed between boundary events that comprise nucleosome-free regions, such as those known to be found at active promoters, insulators or enhancers.

2 Methods

Spectral estimation seeks to estimate the frequency content of a time-dependent waveform, representing it as the weighted sum of a family of sinusoids. In this application, the dimension of "time" is genomic position, and in order to use classical spectral estimation techniques, the DNA sequence must first be translated into a numerical sequence. Our method uses a common approach to represent a DNA sequence as a binary sequence: given a DNA sequence s of length S and a k-mer m of length k, we create a binary sequence b of length $S-k+1$ such that

$$\begin{aligned} &b_i = 1 \text{ if } s_{i:i+k-1} = m \text{ and} \\ &b_i = 0 \text{ otherwise,} \end{aligned} \tag{1}$$

where $s_{i:i+k-1}$ is the length k substring in s, starting at position i. The number of 1's in the binary sequence is equal to the total number of occurrences of the k-mer, including overlapping copies. For example, the 10 base sequence GCAAAGCTAA becomes 001100001 for the dinucleotide AA.

The binary sequence can then be transformed from the "time" domain to the frequency domain in two different ways. Some methods [7], [16], [17] convert directly to the frequency domain via the Fourier transform, and then use the magnitude squared of the resulting complex spectrum. Another approach is to compute the autocorrelation function out to some maximum lag (typically in the range 50 to 500 bases) [2], [18], [19], [20], and then the power spectrum is obtained by a Fourier transform of the autocorrelation function. Under certain conditions, namely that the time-series is wide sense stationary[1] (WSS) [21], and that the two-sided, symmetric autocorrelation function is used, these two approaches are mathematically equivalent. The previously cited approaches have, however, routinely truncated the autocorrelation function which, as we will show, negatively impacts the sensitivity and accuracy of the resulting spectrum.

We prefer the autocorrelation-based approach over the direct Fourier transform approach for three reasons. The first is simply that the autocorrelation provides easily interpretable information about the self-similarity of a sequence,

[1] A process is said to be WSS if its first and second moments are time-invariant, resulting in an autocorrelation function that depends only on the time lag. This assumption does not generally hold for DNA segments, *e.g.* those spanning GC-rich and GC-poor regions, but it is not an unreasonable approximation.

irrespective of subsequent transforms. Second, this approach provides a natural way to combine any number of sequences of varying lengths such that each sequence will contribute in proportion to the number of times the k-mer is present in the sequence. Third, because the variance of the spectral estimates increases as the square of the Fourier transform length [22], the direct transform of several hundred kilobases of sequence requires some form of smoothing in the frequency domain. In our approach, the symmetric autocorrelation function is windowed[2] prior to the Fourier transform in order to reduce the variance in the resulting spectrum and to minimize the spectral leakage from strong periodic signals (such as the 3 bp codon periodicity). It is also useful to extend a time series with zeros prior to computing the Fourier transform in order to increase the resolution in the frequency domain. (This practice is known as *zero-padding*, and although it cannot increase the information content in a signal, it results in a smoothly interpolated spectrum.) In the results presented here, we have generally used a Fourier transform size of 720 which produces spectra with samples at several integer values of T, thereby reducing the amount of spectral leakage due to strong peaks that may exist at those positions, and with a resolution of 0.14 bp near T=10 bp, which allows us to distinguish fairly subtle variations in periodicity.

In the case of the binary sequence described here, the autocorrelation function has a probabilistic interpretation. Since $E[b_i]$ (defined in (1)) is the probability of observing the k-mer of interest at position i in the DNA sequence, the autocorrelation function $R(d)$ is equivalent to the joint probability of observing two 1s in the binary sequence, one at position i and the second at a relative lag d :

$$R(d) = R(-d) = E[b_i b_{i-d}] = Pr[b_i = 1 \text{ and } b_{i-d} = 1] \tag{2}$$

If the binary sequence represents occurrences of the AA dinucleotide, then:

$$\begin{aligned} R_{AA}(d) &= Pr[b_i = 1 \text{ and } b_{i-d} = 1] \\ &= Pr[s_i = A,\ s_{i+1} = A,\ s_{i-d} = A,\ s_{i-d+1} = A] \end{aligned} \tag{3}$$

In order to test whether a particular DNA sequence shows evidence of periodicity, a null model of the spectrum is required. We derive an analytic null model by modeling a random DNA sequence with no periodicity in which the nucleotides are independent and identically distributed. The autocorrelation function of the binary sequence representing the AA dinucleotide derived from such a random DNA sequence can be described by the following three equations:

$$R_{AA}(0) = Pr[A]^2 \tag{4}$$

$$R_{AA}(\pm 1) = Pr[A]^3 \tag{5}$$

$$R_{AA}(d) = Pr[A]^4 \quad \text{for all } |d| > 1. \tag{6}$$

A dinucleotide composed of two distinct bases (*e.g.* GC), results in an autocorrelation function with a value of 0 at ±1 :

$$R_{GC}(0) = Pr[G]Pr[C] \tag{7}$$

[2] A point-by-point multiplication of the symmetric autocorrelation function with a symmetric, tapered window such as the Hann window.

$$R_{GC}(\pm 1) = 0 \tag{8}$$

$$R_{GC}(d) = Pr[G]^2 Pr[C]^2 \quad \text{for all } \; |d| > 1. \tag{9}$$

In both of these cases, the autocorrelation function for dinucleotide m can be expressed as the sum of four terms:

$$R_m(d) \; \propto \; \delta(d) + U_m\delta(d+1) + U_m\delta(d-1) + V_m \tag{10}$$

where U_m is positive for a dinucleotide such as AA or CC and negative for a dinucleotide such as TA or GC. (U_m and V_m are constants that depend only on $R(0)$, $R(\pm 1)$ and $R(d)$, and $\delta(d)$ is the Kronecker delta.) The Fourier transform of (10) yields the power spectrum:

$$S_m(k) \; \propto \; 1 + 2U_m cos(2\pi k/K) + V_m\delta(k) \tag{11}$$

where K is the length of the symmetric autocorrelation function being transformed, and k is an integer in the range $[-K/2, +K/2]$. The normalized frequency f is defined as k/K and has a range $[-1/2, 1/2]$, and corresponds to periods of length $T = 1/f$ with a range of $[2, \infty]$. The cosine term in the spectrum for a dinucleotide such as AA results in a local maximum at $f = 0$ and a local minimum at $f = 1/2$, meaning more energy at lower frequencies (longer periods) and less energy at higher frequencies (shorter periods). For a dinucleotide such as TA, the reverse is true, with more spectral energy at higher frequencies and less at lower frequencies. An intuitive explanation of this effect is that consecutive AA dinucleotides result in the low frequency binary signal 1111111111 ($f = 0$ and $T = \infty$), while consecutive TA dinucleotides result in the high frequency signal 1010101010 ($f = 1/2$ and $T = 2$).

Real DNA sequences are of course not random as described above, and yet the autocorrelation function of the binary sequences generated from real DNA sequences is dominated by the same three components: (i) $R(0) \gg R(d) \; \forall \; d \neq 0$; (ii) either $R(\pm 1) = 0$, or $R(\pm 1) > R(d) \; \forall \; |d| > 1$; and (iii) a relatively flat or slowly-decaying background level with small-amplitude variations for all $|d| > 1$. The extent to which the dinucleotide spectrum estimated from real DNA deviates from the random model described above can be estimated by fitting a cosine to the spectrum (11) using a linear least-squares fit in cosine space. This fitted cosine will serve as our null model for the dinucleotide of interest.

Finally, in order to estimate the variance in the spectral estimate, we generated random DNA sequences of varying lengths, applied ACSE to each one and computed the variance in the estimate. The random DNA sequences varied in length from 150 bases to 500 kb. For each length, we generated 10,000 random sequences and computed the estimated ACSE spectrum. A linear regression ($R^2 > 0.989$) of the ratio of the standard deviation to the mean amplitude against five independent variables yielded the following error model:

$$\begin{aligned} log_{10}\left(\frac{\sigma}{\mu}\right) \;=\; & -0.513 log_{10}N \;+\; 0.000791 D_{max} \;-\; 0.124b \\ & -\; 0.572p \;+\; 0.226f \;+\; 1.042 \end{aligned} \tag{12}$$

where N is the total sequence length used in estimating the spectrum; D_{max} is the maximum lag in the autocorrelation function; $b=1$ if the k-mer is a repeat (*e.g.* AA, CC) and $b=0$ if it is not (*e.g.* AT, GC); p is the average probability of observing the k-mer $(0<p<1)$; and f is the normalized frequency ($f=1/T$, and $f<1/2$). As expected, increasing N by a factor of four reduces the standard deviation by approximately a factor of 2. The remaining terms are less significant assuming N is large, *e.g.* $\geq$ 100 kb.

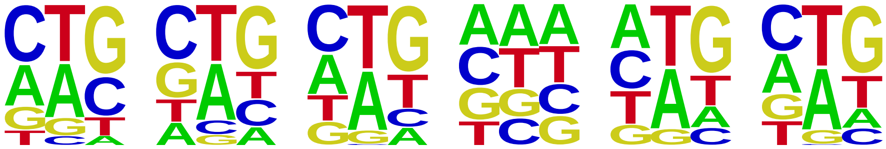
3 Results

3.1 A Cyclic HMM Identifies Spurious Periodicity

Baldi *et al.* [24] used a cyclic HMM to identify an apparently periodic signal in human genomic DNA. The signal is roughly characterized by a three-letter pattern of the form (not-T)(A/T)(G) and was reported to occur approximately every 10 bp. The HMM that identified this pattern consists of 10 states, each of which emits a single nucleotide. The state-transition matrix is defined such that each state $i \in \{0, \ldots, 9\}$ can transition to one of three states: i (a self-loop), $i+1$ (the next consecutive state), or $i+2$ (skipping over one state), with the addition being modulo-10 to create a cycle. Because of the potential for a state to self-loop, the number of nucleotides emitted prior to transitioning to a different state follows a geometric distribution, and the duration of one complete cycle can range between a minimum length of 5 nucleotides (if every other state is skipped) and an unbounded maximum length. The number of free parameters in this model is 50: 30 for the emission distributions and 20 for the transition probabilities.

Working within the framework of dynamic Bayesian networks (DBN) which include and extend HMMs, we reimplemented the HMM described above, and developed two additional models based on this idea in an attempt to both recreate and expand upon this previous result. In both of our models, we sought to constrain the complete cycle length by allowing only one or two "background" states to self-loop (*i.e.*, emit more than one nucleotide before transitioning to a distinct state), and by not allowing any states to be skipped. These background states are not implemented in the typical manner of a self-looping HMM state; rather, we use the DBN framework to specify a histogram of allowed lengths for these states (for details of a similar DBN with finite length models, see [28]).

Model-I is an eight-state model based on the hypothesis that there is a three nucleotide pattern that occurs roughly every 10 bases, and that there may be a second, complementary pattern that also occurs every 10 bases but out of phase with the first. States 1-3 represent the primary pattern and states 5-6 the complementary pattern; each of these states emits a single nucleotide. States 0 and 4 are "background" states and emit between one and four nucleotides, according to a shared length distribution. Each state i transitions directly to the next state, $i+1$ (modulo-8). The length of one complete cycle is therefore constrained between 8 and 14 nucleotides. The total number of free parameters



(a) Baldi exons (b) Baldi introns (c) Model-I ENCODE (d) Model-I 0th order (e) Model-I 1st order (f) Model-I 2nd order

Fig. 1. Patterns learned by (a) the Baldi HMM trained on exons, (b) the Baldi HMM trained on introns, (c) Model-I trained on the ENCODE repeat-masked data, and Model-I trained on (d) 0th order (e) 1st order and (f) 2nd order random DNA sequences.

in this model is 27: 24 for the emission distributions and an additional 3 for the length distribution.

Model-II is one of a class of N-state models, also similar to the cyclic HMM, but with additional constraints. In each model, state 0 is the background state and emits 1, 2 or 3 nucleotides, while states 1 through $N-1$ each emit a single nucleotide. Again each state i transitions only to state i+1 (modulo-N). The total cycle length is thus constrained between N and $N+2$ bases. The total number of free parameters in these models is $3N+2$ ($3N$ for the emission distributions, and an additional 2 for the length distribution for state 0).

When trained on the repeat-masked ENCODE regions, Model-I learned a three position pattern (Fig. 1c) very similar to the ones reported by Baldi *et al.* (Fig. 1a and b). By "pattern" we mean the learned emission probabilities from any three consecutive states in the model that best matches the (not-T)(A/T)(G) pattern. However, in trying to understand this result with the goal of expanding upon it, we discovered that the pattern could be replicated by training on simulated DNA generated according to a second order Markov model, *i.e.* a generative model (trained on the same repeat-masked ENCODE sequences) in which the probability of each nucleotide depends on the previous two nucleotides. DNA generated according to such a model reproduces the statistics of the training data for single, di- and trinucleotides. The pattern in Fig. 1(d) was learned from 0th order random DNA (independent and identically-distributed nucleotides). As expected, no distinctive pattern was found. However, Figs. 1(e) and (f) show the corresponding patterns learned from 1st and 2nd order random DNA. Both patterns closely resemble the one learned from the ENCODE repeat-masked data. Furthermore, using Model-II, we found that the same pattern is learned by cyclic models of different sizes (data not shown), with as few as 4 states (which allows periods of length 4-6 bp), showing again that the perceived periodicity is spurious. Based on these experiments, we conclude that it is the rarity (on average) of the CpG dinucleotide that causes this type of sequential-state, probabilistic model to learn this C(A/T)G pattern. The fact that the learned pattern is stronger (lower entropy) when trained on human DNA than when trained on yeast DNA also supports this, as CpG's are far more rare in human than in yeast.

Despite the fact that these results show that C(A/T)G does not exhibit ∼10 bp periodicity, there are other indications that this trinucleotide may be related to DNA bending and nucleosome formation. Numerous papers (most recently [29], for example) describing a relationship between the periodicity of the (not-T)(A/T)G (also referred to as VWG) motif and chromatin structure have cited the work by Baldi *et al.* as a starting point. Perhaps the relationship between C(A/T)G and chromatin structure is instead due to the extreme flexibility of this trinucleotide [14]. It should also be noted that CTG, one of the six codons that code for leucine, is the most commonly used codon in human, while CAG, one of just two codons that code for glutamine, is the third most common – together they account for over 7% of all codons. Overall, in the ENCODE repeat-masked data, CTG/CAG is the fourth most common trinucleotide (after AAA/TTT, AAT/ATT, and AGA/TCT), accounting for 4.4% of all trinucleotides. Intriguingly, repeats of the CTG/CAG trinucleotide are also responsible for several degenerative disorders [30].

3.2 Simulation Results Show ACSE Has Greater Sensitivity and Accuracy

Next, to demonstrate the power of ACSE to identify subtle periodic signals, we applied the method to simulated DNA with embedded periodic signals. Our simulation creates a random DNA sequence with equal representation of all four bases, and then adds to the DNA sequence two noisy periodic signals: one at 3 bp, and another with a period length that slowly increases from 10 bp to 11.5 bp across the length of the sequence. Figure 2 compares the output of ACSE with the two previously described methods: the direct Fourier transform approach, and the truncated autocorrelation approach for the simulated DNA described above, and for *S. cerevisiae* chromosome IV. The output of ACSE and the direct FFT have both had the null model spectrum subtracted, and all three have been normalized such that the regions between 4 and 9 bp and 14 and 24 bp combined have zero mean and unit standard deviation. For the simulated data, ACSE correctly shows a relatively smooth, broad peak between 10.1 and 11.4 bp, while the truncated autocorrelation method predicts two distinct peaks, one at 9.7 bp and the other at 10.9 bp. The truncated autocorrelation approach also errs on the exact position of the 3 bp peak, placing it instead at 2.9 bp. The direct FFT approach (after smoothing) yields a curve comparable to ACSE, although somewhat noisier. For yeast chromosome IV, the truncated autocorrelation method largely misses the evidence of the ∼10 bp periodicity of the AA dinucleotide.

3.3 Identifying Coding Regions in Yeast

As a proof of concept that dinucleotide periodicity can be used to identify biologically meaningful regions in DNA segments, we developed a *period 3 score*, defined to be the sum over all 10 dinucleotides of the maximum 3 bp periodicity observed on either the forward or reverse strand. We applied this scoring technique to the genome of *S. cerevisiae*, computing the score on short blocks

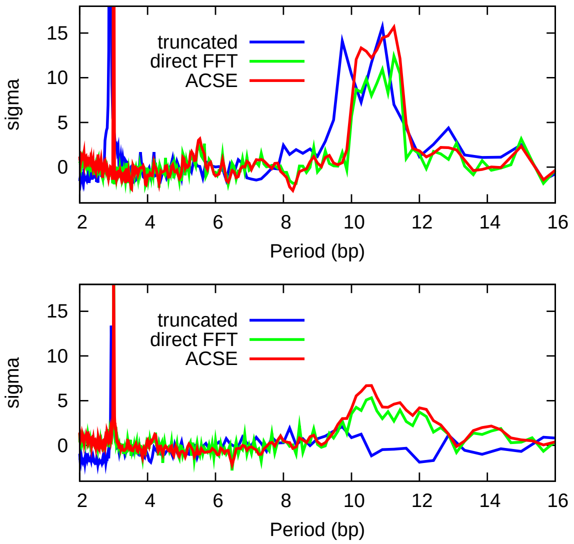


Fig. 2. Spectra obtained from the three different methods described: on synthesized DNA containing periodic signals at 3 bp and between 10 and 11 bp (top); and for the AA dinucleotide in *S. cerevisiae* chromosome IV (bottom).

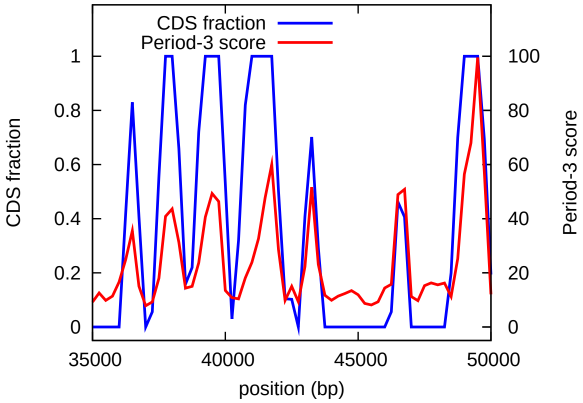


Fig. 3. Dinucleotide period 3 score (in red, y-axis on the right) computed on 500 bp blocks, compared to the fraction of the block that is coding (in blue, y-axis on the left). The plot shows a representative 15 kb region of the yeast mitochondrial chromosome.

of length 500 bp, and comparing the score to the fraction of bases within that block that lie in coding regions (excluding those coding regions that are annotated as “dubious”). The average Pearson correlation between the period 3 score and the coding fraction was 0.51, with the highest correlation observed for the mitochondrial chromosome (0.85), illustrated in Fig. 3.

3.4 Genome-Wide Evidence for ∼10 bp Periodicity of AA/TT Dinucleotide in *S. cerevisiae.*

Next, we computed the dinucleotide spectra from genome-wide autocorrelation functions for *S. cerevisiae*, as well as for each chromosome individually. As shown in Fig. 4, the strongest evidence for ∼10 bp periodicity is seen when analyzing the AA/TT dinucleotide. Each dinucleotide spectrum in Fig. 4 is accompanied by a null model curve, and deviations from that null model represent increases (above the null curve) or decreases (below) in periodic signal energy relative to what would be expected in a random signal. The other three dinucleotides show much less, if any, evidence of genome-wide 10 bp periodicity. The spectra computed separately for each chromosome show some differences from the genome-wide average (data not shown). Overall, for every chromosome in yeast, we observe a clear periodicity in the 10-11 bp range for the AA/TT dinucleotide.

3.5 Positioned Nucleosomes in Yeast Show Increased Periodicity

A comprehensive map of *S. cerevisiae* nucleosomes containing the histone variant H2A.Z in functionally important regions was described by Albert *et al.* [31], including a list of 41,103 nucleosome positions. More recently over one million nucleosomes obtained using antibodies against histones H3 and H4 were sequenced

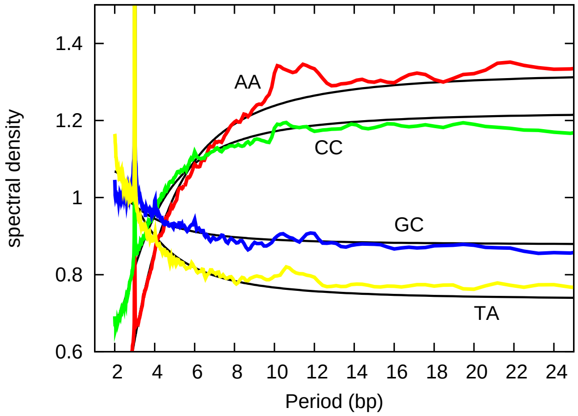


Fig. 4. Dinucleotide spectra averaged across the entire yeast genome (12.3 Mb) for AA, CC, GC and TA, with null model curves for each.

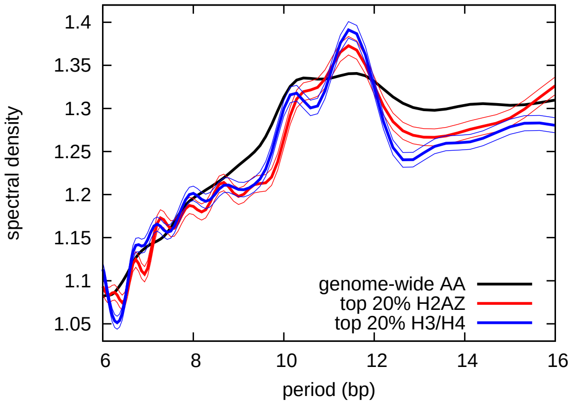


Fig. 5. AA dinucleotide spectra: genome-wide (black), highest scoring H2AZ (red) and H3/H4 nucleosome positions (blue). Each group of three curves corresponds to the mean and mean ± one (estimated) standard deviation.

[32], and 54,750 consensus nucleosome locations were identified and assigned confidence scores. After removing nucleosome positions that were within 157 bp of any higher-scoring positions, we were left with 38,356 H2A.Z nucleosome positions and 49,751 H3/H4 positions. (As all nucleosomes contain the H3 and H4 histones, although these two datasets are derived from different experiments, the H2A.Z set is essentially a subset of the H3/H4 set.) Several dinucleotide spectra for these nucleosome core sequences show greater evidence of periodicity between 10 and 12 bp than the genome-wide spectra. Using only the highest scoring nucleosome positions, which are thought to be the most stably positioned, yields spectra even more significantly different from the genome-wide background. Figure 5 shows the dinucleotide spectra computed using the highest scoring 20% from each dataset as compared to the genome-wide curve. Both dinucleotides show two distinct peaks, one at 10.4 bp and the other at 11.6 bp. These peaks may indicate different preferred periodicities in different portions of the core nucleosomal sequence, as the double helix underwinds at sites of major-groove bending and overwinds at sites of minor-groove bending [33], [34].

3.6 No Evidence of Genome-Wide Periodicity Human

Turning to the human genome, we began by analyzing the repeat-masked ENCODE regions. These regions comprise 1% of the human genome, selected to cover regions of varying gene content and varying concentrations of non-coding conserved elements [35]. Our analysis shows little to no evidence of dinucleotide periodicity when we average across the repeat-masked ENCODE regions. The AA dinucleotide spectrum deviates remarkably little from the null model curve, and none of the other dinucleotides exhibit any strong periodicity near 10 or

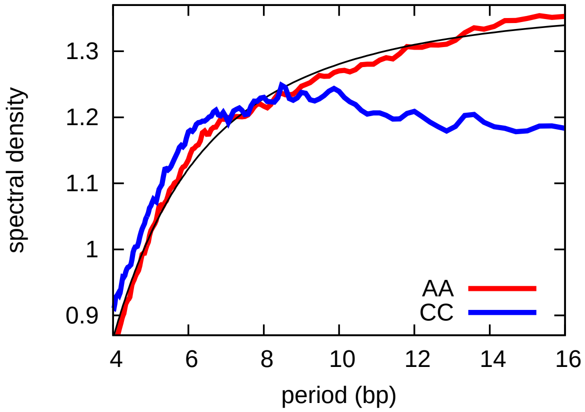


Fig. 6. AA and CC dinucleotide spectra based on the repeat-masked ENCODE regions, with the null model for AA in black.

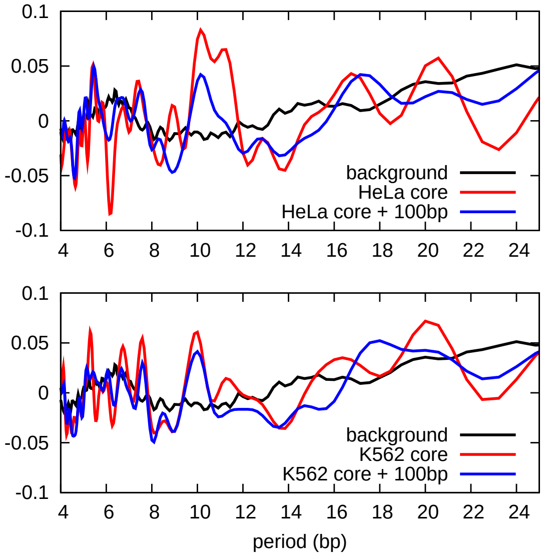


Fig. 7. AA dinucleotide spectra, after subtracting null model, for HeLa and K562 nucleosome sequences: genome-wide background, core 147 bp, and core+100 bp.

12 bp although some deviate significantly in overall shape from the null model. Figure 6 shows the spectra for AA and CC, with the null model for AA.

3.7 Human Nucleosome Sequences Show Evidence of Periodicity

In contrast, when we focus on nucleosome sequences, we observe strong evidence of periodicity. Using experimentally determined nucleosome positions in the HOX cluster [25], we extracted 147 bp nucleosome core sequences at 1086 positions based on HeLa data and 1169 positions for K562. The HeLa nucleosome cores show evidence of 10 bp periodicity of the AA dinucleotide, as shown in Fig. 7. The spectra in this pair of plots are shown after subtracting the null model curve, and are compared to the background curve for AA obtained from the repeat-masked ENCODE regions (Fig. 6). The positioned nucleosomes derived from the HeLa data show a significant increase over both the null model and the background between 10 and 12 bp, with local maxima at 10.14 and 11.25 bp. The shape of this peak is similar to the one observed for the AA dinucleotide in yeast (Fig. 4). More strikingly, if we expand the amount of sequence used by an additional 50 bases on either side of the original 147 bp core sequence, the 11.25 bp peak disappears completely while the 10.14 bp peak is somewhat reduced. The spectrum computed from the K562 nucleosome positions is not as striking although there is still a doublet peak with local maxima at 10.00 and 11.25 bp, and again the 11.25 bp peak disappears when the sequence window is widened. Restricting the analysis to the 589 nucleosome positions that are shared between the two subsets (data not shown), the periodicity signal is slightly stronger than the one shown for HeLa in Fig. 7, which may indicate that the remaining K562 positions represent less well positioned nucleosomes or nucleosomes that have been shifted from the most energetically favorable positions.

3.8 Decreased Periodicity near Transcription Start Sites

Using 20,334 transcription start sites from the RefSeq Genes track, we extracted short (150 bp) segments of DNA centered at positions relative to each TSS ranging from 10 kb upstream to 10 kb downstream, and staggered by 10 bases. For each set of sequences, we computed the AA dinucleotide spectrum and extracted the amplitudes at 10 and 11 bp. We then chose the regions between 5 and 10 kb away from the TSS (upstream and downstream) to serve as a background model, and normalized both traces relative to that background. As shown in Fig. 8, shortly before and after the TSS, the 11 bp amplitude is significantly elevated relative to the background while the 10 bp amplitude is relatively flat. Both drop sharply in the immediate vicinity of the TSS. We examined the average AA/TT counts within the same windows to see if this effect could be entirely attributed to local increases or decreases in the counts of these dinucleotides, but this was not the case. The sharp decrease in periodicity near the TSS will inhibit nucleosome formation, while the increase in the 11 bp periodicity upstream and downstream of the TSS will encourage the stable positioning of the canonical -1 and +1 nucleosomes.

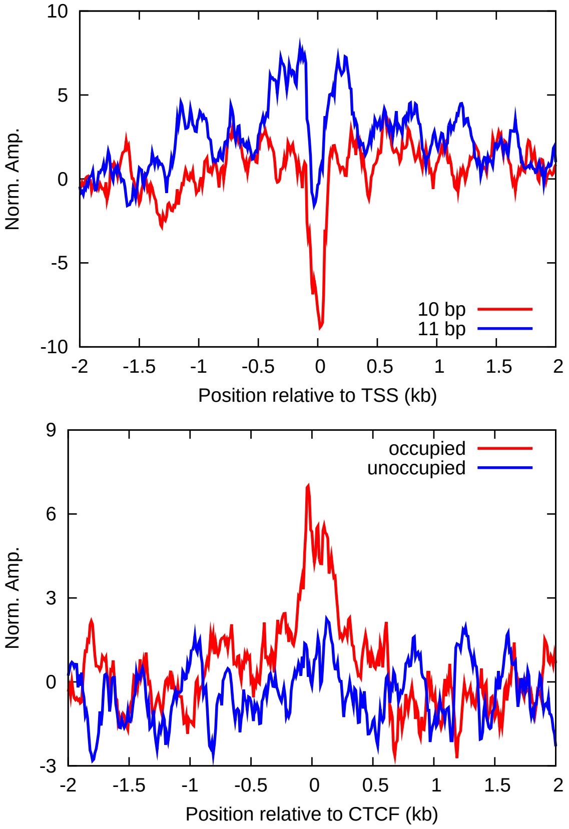


Fig. 8. (a) Average spectral amplitude at 10 and 11 bp for AA/TT dinucleotide relative to the TSS of 20,334 human genes. The amplitude has been normalized to have zero mean and unit standard deviation in the regions 5 to 10 kb from the TSS. (b) Average spectral amplitude for AA/TT dinucleotide at 11 bp for occupied and unoccupied CTCF sites, similarly normalized.

3.9 Dinucleotide Periodicity Increases Near Occupied CTCF Sites

CTCF is a DNA-binding protein which binds to insulator elements to restrict access to transcriptional promoters. CTCF is believed to play a wide-spread role in gene regulation [36], and a relationship between CTCF binding and strongly positioned nucleosomes has been shown experimentally [37]. We used a list of 6432 *occupied* and the same number of *unoccupied* CTCF binding sites [37] to examine the AA dinucleotide periodicity in the region surrounding CTCF sites using the same approach used above for transcription start sites. The spectral amplitudes at both 10 and 11 bp peak in the vicinity of the occupied CTCF sites, while there is little difference from the background for the unoccupied CTCF sites (Fig. 8b plots only the 11 bp spectral amplitude). The unoccupied

CTCF sites represent predicted sites and may include a large number of false positives, while the occupied CTCF sites have been experimentally proven to be true positives. Based on our results, a significant difference between true sites and false sites may be the local curvature of the DNA induced by AA periodicity, despite the local similarity in DNA sequence at the binding site. This increased local curvature may be necessary for CTCF to effectively bind the DNA.

4 Discussion

We have described a spectral estimation technique that is both mathematically precise and more sensitive than previously published approaches. Unlike previous methods, ACSE includes an analytic null model as well as a model of the variance in the spectral amplitude estimates. Provided sufficient data, ACSE can identify weak periodic signals in DNA sequences. Previous autocorrelation based approaches have truncated the autocorrelation function, using $R(d)$ only for $d \geq 1$ or $d \geq 2$. Truncating the autocorrelation function reduces the sensitivity of the spectral estimation to weak periodic signals and distorts the resulting spectrum. Based on truncated autocorrelation functions some studies have reported not finding any indication of periodicity near 10 bp either in complete human chromosomes [18] or nucleosome core sequences [25].

We have confirmed previously-reported genome-wide evidence of $\sim$10 bp periodicity in *S. cerevisiae.* The spectral pattern shows a doublet peak, with one local maximum near 10 bp and a second near 11 bp. Sequences associated with well positioned nucleosomes show stronger spectral peaks than the genome-wide average, in particular for the 11 bp peak. There is no similar widespread evidence of periodicity in the human genome; however, positioned nucleosomes within the HOX cluster show evidence of the same doublet peak for the nucleosome core sequence. Extending the analysis to include even just an additional 50 bases on either side of the core causes the 11 bp peak to disappear. In the neighborhood of transcription start sites, we observed a sharp decrease in the AA dinucleotide periodicity at both 10 and 11 bp, and it is interesting to note that the 11 bp periodicity is significantly increased both immediately before and after the TSS whereas the 10 bp periodicity is not. This apparent distinction between periodicities of 10 and 11 bases may indicate regions in the nucleosome core where the double helix is alternately under- or overwound [33], [34]. Finally, we find that CTCF sites show a local increase in the strength of the AA periodicity at both 10 and 11 bp, but more strongly at 11 bp. No crystal structure of the CTCF-DNA complex is available, but this evidence of periodicity at CTCF binding sites may indicate that some curvature of the DNA is required for effective binding. Alternatively, these sites may be occupied by nucleosomes when CTCF is not bound.

Overall, our analysis suggests a clear relationship between local periodic patterns in DNA and local chromatin architecture. The lack of a global periodic signal in human, and the correspondence between local periodic signals and functionally significant chromatin events, such as promoters and insulators, supports the classical statistical positioning theory of nucleosome organization.

References

1. Crick, F.H.C., Klug, A.: Kinky helix. Nature. 255, 530–533 (1975)
2. Trifonov, E.N., Sussman, J.L.: The pitch of chromatin DNA is reflected in its nucleotide sequence. Proceedings of the National Academy of Sciences of the United States of America. 77, 3816–3820 (1980)
3. Herzel, H., Trifonov, E.N., Weiss, O., Grosse, I.: Interpreting correlations in biosequences. Physica A. 249, 449–459 (1998)
4. Satchwell, S.C., Drew, H.R., Travers, A.A.: Sequence periodicities in chicken nucleosome core DNA. Journal of Molecular Biology. 191, 659–675 (1986)
5. Drew, H.R., Travers, A.A.: DNA bending and its relation to nucleosome positioning. Journal of Molecular Biology. 186, 773–790 (1985)
6. Shrader, T., Crothers, D.: Artificial nucleosome positioning sequences. Proceedings of the National Academy of Sciences of the United States of America. 86, 7418–7422 (1989)
7. Widom, J.: Short-range order in two eukaryotic genomes: relation to chromatin structure. Journal of Molecular Biology. 259, 579–588 (1996)
8. Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I.K., Wang, J.Z., Widom, J.: A genomic code for nucleosome positioning. Nature. 44, 772–778 (2006)
9. Fire, A., Alcazar, R., Tan., F.: Unusual DNA structures associated with germline genetic activity in *Caenorhabditis elegans*. Genetics. 173, 1259–1273 (2006)
10. Packer, M.J., Dauncey, M.P., Hunter, C.A.: Sequence-dependent DNA structure: dinucleotide conformational maps. Journal of Molecular Biology. 295, 71–83 (2000)
11. Packer, M.J., Dauncey, M.P., Hunter, C.A.: Sequence-dependent DNA structure: tetranucleotide conformational maps. Journal of Molecular Biology. 295, 85–103 (2000)
12. Beveridge, D.L., Dixit, S.B., Barreiro, G., Thayer, K.M.: Molecular dynamics simulations of DNA curvature and flexibility: helix phasing and premelting. Biopolymers. 73, 380–403 (2004)
13. Gabrielian, A., Simoncsits, A., Pongor, S.: Distribution of bending propensity in DNA sequences. FEBS Letters. 393, 125–140 (1996)
14. Bomble, Y.J., Case, D.A.: Multiscale modeling of nucleic acids: insights into DNA flexibility. Biopolymers. 89, 722–731 (2008)
15. Travers, A.A.: The structural basis of DNA flexibility. Philosophical Transactions of the Royal Society of London. Series A Mathematical, Physical and Engineering Sciences. 362, 1423–1438 (2004)
16. Thåström, A., Lowary, P.T., Widlund, H.R., Cao, H., Kubista, M., Widom, J.: Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. Journal of Molecular Biology. 288, 213–229 (1999)
17. Bailey, K.A., Pereira, S.L., Widom, J., Reeve, J.N.: Archaeal histone selection of nucleosome positioning sequences and the procaryotic origin of histone-dependent genome evolution. Journal of Molecular Biology. 303, 25–34 (2000)
18. Holste, D., Grosse, I., Beirer, S., Schieg, P., Herzel, H.: Repeats and correlations in human DNA sequences. Physical Review E. 67 (2003)

19. Hosid, S., Trifonov, E.N., Bolshoy, A.: Sequence periodicity of *Escherichia coli* is concentrated in intergenic regions. BMC Molecular Biology. 5:14 (2004)
20. Schieg, P., Herzel, H.: Periodicities of 10-11 bp as indicators of the supercoiled state of genomic DNA. Journal of Molecular Biology. 343, 891–901 (2004)
21. Grimmett, G.R., Stirzaker, D.R.: Probability and Random Processes. Oxford University Press, USA (2001)
22. Jenkins, G.M., Watts, D.: Spectral Analysis and Its Applications. Emerson-Adams Press, USA (1998)
23. Baldi, P., Brunak, S., Chauvin, Y., Englebrecht, J., Krogh, A.: Periodic sequence patterns in human exons. ISMB. 30–38 (1995)
24. Baldi, P., Brunak, S., Chauvin, Y., Krogh, A.: Naturally occurring nucleosome positioning signals in human exons and introns. Journal of Molecular Biology. 263, 503–510 (1996)
25. Kharchenko, P.V., Woo, C.J., Tolstorukov, M.Y., Kingston, R.E., Park, P.J.: Nucleosome positioning in human HOX gene clusters. Genome Research. 18, 1554–1561 (2008)
26. Kornberg, R.: The location of nucleosomes in chromatin: specific or statistical. Nature. 292, 579–580 (1981)
27. Kornberg, R.D., Lorch, Y.: Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromasome. Cell. 98, 285–294 (1999)
28. Reynolds, S.M., Käll, L., Bilmes, J.A., Noble, W.S.: Transmembrane topology and signal peptide prediction using dynamic Bayesian networks. PLoS Computational Biology. 4:11 (2008)
29. Takasuka, T.E., Cioffi, A., Stein, A.: Sequence information encoded in DNA that may influence long-range chromatin structure correlates with human chromosome functions. PLoS ONE. 3:7 (2008)
30. Wang, Y.H.: Chromatin structure of repeating CTG/CAG and CGG/CCG sequences in human disease. Front Bioscience. 122, 4731–4741 (2007)
31. Albert, I., Mavrich, T.N., Tomsho, L.P., Qi, J., Zanton, S.J., Schuster, S.C., Pugh, B.F.: Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. Nature. 446, 572–576 (2007)
32. Mavrich, T.N., Ioshikhes, I.P., Venters, B.J., Jiang, C., Tomsho, L.P., Qi, J., Schuster, S.C., Albert, I., Pugh, B.F.: A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Research. 18, 1073–1083 (2008)
33. Richmond, T.J., Davey, C.A.: The structure of DNA in the nucleosome core. Nature. 423, 145–150 (2007)
34. Tolstorukov, M.Y., Colasanti, A.V., McCandlish, D.M., Olson, W.K., Zhurkin, V.B.: A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning. Journal of Molecular Biology. 371, 725–738 (2007)
35. ENCODE Consortium: The ENCODE (ENcyclopedia Of DNA Elements) Project. Science. 306, 636–640 (2004)
36. Xie, X., Mikkelsen, T.S., Gnirke, A., Lindblad-Toh, K., Kellis, M., Lander, E.S.: Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. Proceedings of the National Academy of Sciences of the United States of America. 104, 7145–7150 (2007)
37. Fu, Y., Sinha, M., Peterson, C.L., Weng, Z.: The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. PLoS Genetics. 4:7 (2008)