

# A SPECTRAL VIEW OF SUPERCOILING IN BACTERIAL GENOMES

Robert H. Baran<sup>1</sup> and Hanseok Ko<sup>2</sup>

<sup>1</sup>Genostat, 4508 Cheltenham Drive, Bethesda, MD 20814, USA

<sup>2</sup>Department of Electronics and Computer Engineering, Korea University, Seoul 136-701, Korea

## ABSTRACT

Spectral densities of genomic nucleotide sequences typically show peaks around the helical free period of DNA (10.5 base pairs) and deviations from this value are associated with over/under-winding. The bacterial chromosome is divided into topologically isolated loops called supercoiling domains that can relax or wind more tightly within boundaries. Prior investigations assumed the homogeneity of prokaryotic “DNA bending signals.” We exhibit spectrograms that illustrate their modulation with respect to amplitude and frequency, providing an indirect view of supercoiling that conforms to domain theory.

**Index Terms**—*Spectral analysis, Scientific visualization*

## 1. INTRODUCTION

Spectral analysis is a fundamental tool for discovery and measurement of periodic components in serial data. The spectral density of a stationary random process is the Fourier transform of its autocorrelation function (ACF). Periodicities that may be hard to see in the ACF, and virtually invisible in the sequence itself, are transformed to spectral lines or peaks and thus resolved more clearly. Such analyses of genomic DNA reveal (A) strong 3-periodicity in coding sequences of all species and (B) relatively weak signals with periods around 10.5 base pairs (bp) that pervade the genomes of bacteria, archaea, and lower eukaryotes. The latter (B-signals) are associated with supercoiling of the molecule [1]. The helical free period of DNA is 10.5 ( $\pm 0.1$ ) bp. Supercoiling can be modulated by the quasi-periodic placement of slightly non-parallel base pairs (e.g., ApA dinucleotides) that act as “wedges” to bend the DNA axis [2]. B-signals with shorter periods ( $\sim 10$  bp) have been attributed to excess supercoiling density as the molecule is overwound in the nucleosomal structure [3]. Longer periods ( $\sim 11$  bp) suggest that negatively supercoiled DNA is under-wound in most bacteria where nucleosomal structure is presumed absent [4].

The first step in creating a spectral representation is to reduce the nucleotide sequence to a series of

numbers, usually binary indicators of mono- or dinucleotide identity or type. Such a reduction is made by taking each single nucleotide to be strong (C and G = +1) or weak (A and T = -1) with respect to hydrogen bonding. Next, the positional ACF is computed. B-signals can be estimated by curve fitting to the ACF instead of computing the Fourier transform. Taking this approach, Schieg and Herzel (2004) found dominant periods ranging from 10.7 to 11.5 bp in 67 bacterial genomes, omitting others that produced a significant lack of fit [4].

An alternative approach was taken by Worning et al. (2000), who computed spectra from sequences of dinucleotide structural parameters, propeller twist and stacking energy, and obtained essentially concordant estimates of the dominant periods [5]. Spectral peaks were sometimes indistinct, especially for estimated periods near 10.5 bp, and double peaks were attributed to horizontal gene transfer from archaea. They found the main peak of the weak B-signal at 10.6 bp in *Borrelia burgdorferi* (omitted in [4]).

## 2. OBJECTIVE

A major theme in the extensive literature on DNA supercoiling is that the tendency of the molecule to bend in a preferred direction could be local and variable from one region to another. In this view, not yet directly supported by microscopic imaging, the bacterial chromosome is divided into topologically isolated loops, called supercoiling domains, of size about 100 kilobases (kb). The chromatin in one loop can relax or wind more tightly independent of events in neighboring loops as bound proteins hold the domain boundaries in place [6]. Curved DNA is found at loop tips near the domain centers where the most highly expressed genes are concentrated. If this view is correct, and if B-signal frequencies really reflect topological parameters, then the domain structure of the chromosome could be illuminated by plotting spectral density versus location of a sliding window that is somewhat narrower than 100 kb. Such a plot is called a spectrogram in diverse applications of spectral analysis, speech and sonar signal processing being prominent examples.

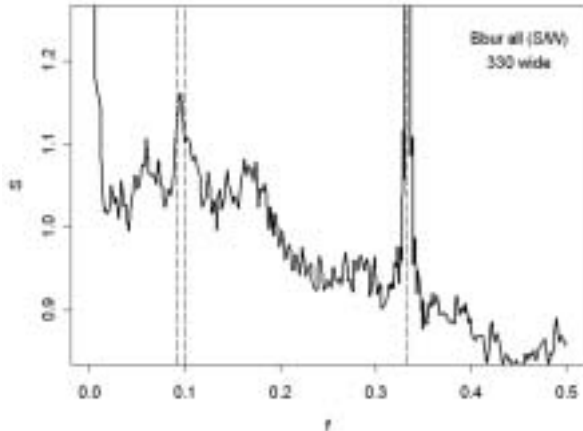


Figure 1. Spectral density ( $S$ ) at 330 frequency points ( $f$ ) for the nucleotide sequence of the linear chromosome of *B. burgdorferi* (NC\_001318) reduced to strong/weak indicators. Dashed vertical bars are drawn at  $f = 1/11, 1/10, \text{ and } 1/3 \text{ bp}^{-1}$ .

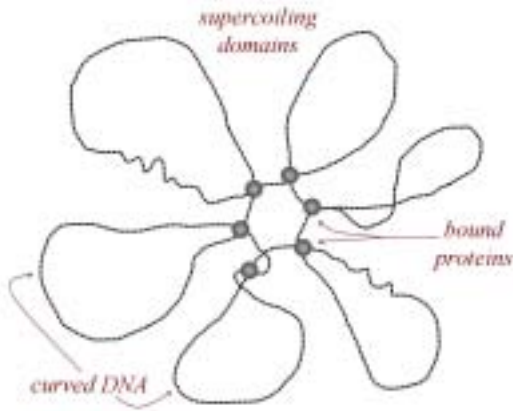


Figure 2. Cartoon illustration of supercoiling domains in a circular bacterial chromosome after Willenbrock and Ussery [6].

A focused examination of these issues has been neglected in prior studies on the grounds that B-signals are generally so weak as to require combining all the data to obtain the best estimate of a dominant period. Tomita et al. (1999), for example, found that 200 to 300 kb of data is required for reliable determination of ApA periodicity [3]. Using larger samples and complete genomes will improve the accuracy of the estimator provided that bending is fairly uniform and that the distribution of the supercoiling period is unimodal. But if the molecule is wound more or less tightly in different places then a single peak in the spectrum of the whole genome may be less distinct than sharper peaks in certain places. Moreover a flattened peak could appear close to the equilibrium

value in the combined spectrum if the bending period actually drifts above and below 10.5 bp along the sequence. These possibilities can be assessed by examining spectrograms of prokaryotic genome sequences, focusing on a narrow range of frequencies centered about  $f = (10.5 \text{ bp})^{-1} \approx 0.095 \text{ pb}$ .

### 3. METHOD

Our interest in the subject is tangential but our observations may be of general interest. We used genomic data to illustrate some basic principles of communications signal processing for an undergraduate course. The canonical procedure for computing the power spectral density of a real-valued series  $\{X_n: n = 1, \dots, N\}$  is to take the discrete cosine transform (DCT) the autocorrelation function,  $\text{ACF}(k) = E(X_n X_{n+k}), |k| = 0, 1, \dots, N-1$ , after multiplying it by a lag window,  $w(k)$ , defined for  $|k| = 0, 1, \dots, L < N$ . If  $L = N-1$  then the spectrum is computed at  $\frac{1}{2}N$  positive frequencies and smoothing in the frequency domain is generally needed to discern the structure of the spectrum. This can be accomplished by narrowing the lag window to  $L \ll N$  which has the same affect as convolution (smearing) in the frequency domain. But the number of floating point operations (flops) to compute an autocorrelation is still  $O(LN^2)$  as  $\text{ACF}(k)$  depends on all the sampled data. Therefore the usual procedure bypasses the ACF (and the Wiener-Khinchin Theorem) by using a fast algorithm (the FFT) to transform the data series in  $O(N \log N)$  flops and then smoothing the squared magnitude of the result.

As a student pointed out, however, the ACF of a series of signed binary digits can be obtained without multiplication, because

$$\text{ACF}(k) \equiv E(X_n X_{n+k}) = E(|X_n + X_{n+k}| - 1)$$

when every  $X = \pm 1$ . Hence the spectrum of a nucleotide sequence, reduced to strong/weak indicators, is rapidly obtained via the canonical route in almost any computing environment, including statistics packages that lack the FFT. We compute the normalized autocovariance,

$$c(k) = [\text{ACF}(k) - (EX)^2] / \text{var}(X),$$

multiply it by a triangular lag (Bartlett) window,

$$w(k) = 1 - |k|/L, |k| = 0, 1, \dots, L < N,$$

and then compute the spectral density as  $S(f) = \text{DCT}\{c(k)w(k)\}$  at the  $\frac{1}{2}L+1$  frequencies  $f = 0, 1/L, 2/L, \dots$ , up to  $\frac{1}{2} \text{ pb}$  (the Nyquist rate). Parameter  $L$ , called resolution, is the half-width of the lag window, which is adjusted to smooth the spectrum as illustrated by

Figure 3. Prior studies found the influence of B-signals to be strongest in the first 10 cycles of the ACF, corresponding to about  $10 \times 10.5$  bp = 105 lags. Widom (1996) used  $L = 1024$  with a lag window of unspecified shape [2]. In the following we generally choose  $L = 330 \approx (105 \times 1024)^{1/2}$ . This choice is divisible by 11, 10, and 3, so that spectral lines at any of these key periods will be appropriately centered.

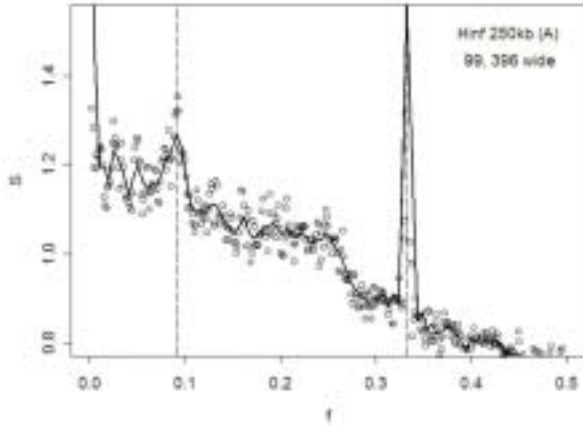


Figure 3. Spectral density  $S(f)$  of the first 250 kb of the *Haemophilus influenzae* genome sequence (NC\_000907), reduced to A-indicators, at 99 (solid curve) and at 396 (plotted points) frequencies. Dashed vertical bars are drawn at  $1/11$  and  $1/3$  pb.

Now let the total length  $N$  of the data series be replaced by the width  $\Omega$  of a sampling window that is displaced in increments of  $\Delta$  from the sequence start. The spectrogram is a series of spectral densities  $S_j(f)$  where  $j = 0, 1, 2, \dots, J - 1$  is the number of displacements, and  $J$  is the first integer such that  $J\Omega$  exceeds  $N$ . (Since the bacterial chromosome is circular in most cases, the genome sequence is allowed to wrap around, and the last sampling window revisits the first  $J\Omega - N$  nucleotides.) In the examples shown below, the sampling window is  $\Omega = 3\Delta$ , and  $\Omega < 100$ kb for reasons noted previously. To plot the spectrogram, select a band of frequencies (from  $f_{\min}$  and  $f_{\max}$ ), and let  $\sigma$  be the full range of  $S$  (over all  $j$ ) in the band. If  $\mu$  is the median of  $S$  in band then the monochrome plotting routine (in SPlus version 4.5) draws contours at base level  $\mu - \frac{1}{4}\sigma$ , mid-level  $\mu$ , and high level  $\mu + \frac{1}{4}\sigma$ .

#### 4. RESULTS AND DISCUSSION

Figure 4 shows spectrograms of the complete nucleotide sequences of four microbial chromosomes using the strong/weak reduction in every case. These spectrograms are focused on the band  $(f_{\min}, f_{\max}) = (1/12, 1/9.5)$  pb and a dashed centerline drawn at  $(10.5$

bp) $^{-1} \approx 0.095$  pb. With 330 resolution, there are 15 frequencies in the band. In two bacteria, (a) *H. influenzae* and (b) *Helicobacter pylori*, B-signal power above the base level is clearly concentrated below the centerline; but power above the mid-level is distributed unevenly with respect to location, which is measured in kb from sequence start. In (c) *Archaeoglobus fulgidus*, the same kind of picture is translated up above the centerline, consistent with the documented spectral contrast between archaeal and bacterial kingdoms [3,4,5]. But the linear chromosome of (d) *B. burgdorferi* shows islands of high spectral density on both sides of the centerline. Similar patterns of location-dependent frequency modulation are also evident in the genome sequences of other bacterial species, particularly in the phylum of actinobacteria (e.g., *Bifidobacterium longum* and genus *Streptomyces*, results not shown). Whether horizontal gene transfer from archaea can explain these observations is a question that could merit more detailed investigation.

Each contour level of the genomic spectrogram produces an archipelago of irregular islands. If islands roughly correspond to supercoiling domains then we can attempt to resolve the domain structure of the chromosome by subjecting the spectrogram to some post-processing. First partition the frequency band into sub-bands  $F(b) = (f_{\min}, 1/10.5]$  and  $F(a) = [1/10.5, f_{\max})$ , which lie respectively below and above the centerline, indicating under- and over-winding. Consider the series of differences

$$u_j = \sum_{f \in F(b)} S_j(f) - \sum_{f \in F(a)} S_j(f)$$

between the integrated spectral densities in the two sub-bands at the  $j$ -th location. A measure of excess under-winding, or *underwinding index*, is obtained by smoothing the de-meaned series  $\{u_j\}$  with a running median and then taking the positive part. Large values of this underwinding index are associated with domain centers and zeros are likely locations of bound proteins at domain boundaries.

Figure 5 shows a polar plot of the under-winding index in the *H. pylori* genome, with a positive offset of 0.1, drawn so that location advances clockwise from sequence start at  $90^\circ$  (12 o'clock), consistent with standard practice in genomics. We observe six major lobes and two more minor lobes that could represent domains of 205kb average extent in this sequence of total length 1,644kb. A cursory examination of such plots in other bacterial genomes suggests that the extent of supercoiling domains is highly variable and closer to about 290kb on average. While gene density in bacteria is generally close to one per kilobase, intra-genomic and inter-species variation in domain size

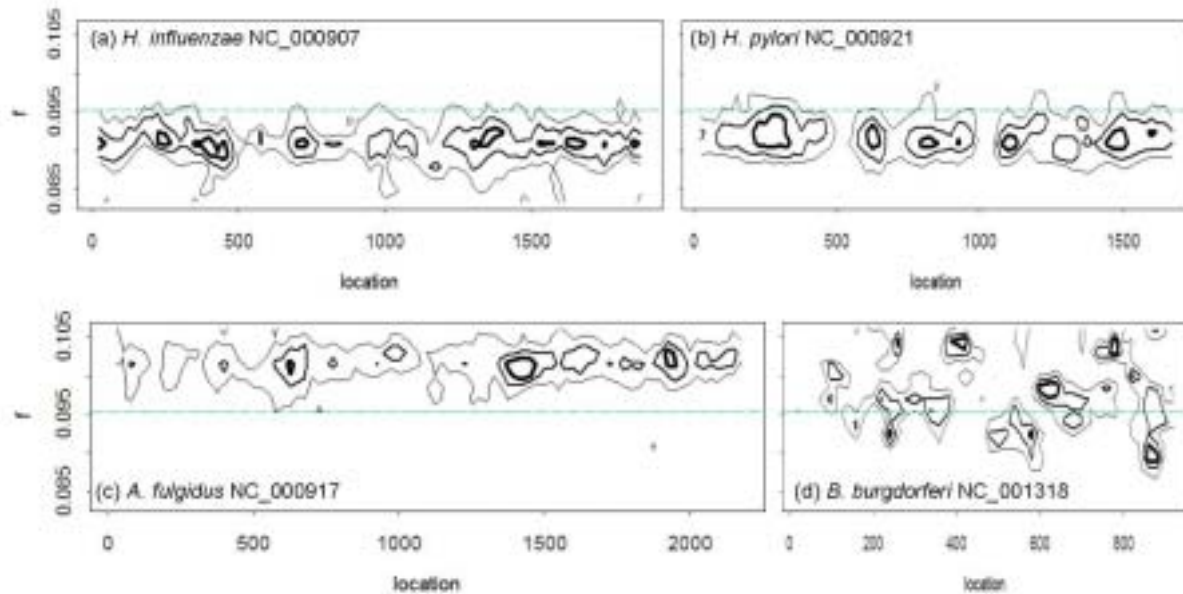


Figure 4. Spectrograms of four microbial genome sequences, labeled by name and GenBank access number, from  $f = 1/12\text{bp}$  to  $1/9.5\text{bp}$  in frequency with 330 resolution, for sampling windows of (a through c) 75kb and (d) 60kb.

must be expected if domain formation is driven by coherent transcription [6] as reflected in the statistics of co-oriented gene clusters [7].

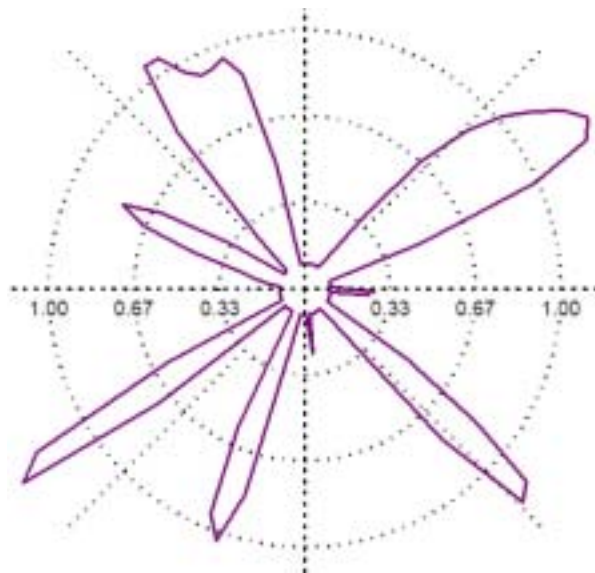


Figure 5. Polar plot of the underwinding index in the *H. pylori* genome sequence, from the spectrogram of Figure 4(b), proceeding clockwise from sequence start at 1200 hours.

## 5. REFERENCES

- [1] Trifonov E.N. and Sussman J.L., "The pitch of chromatin DNA is reflected in its nucleotide structure, *Proc. Natl. Acad. Sci. USA*, 77, pp.3816-3820, 1980.
- [2] Widom J., "Short-range order in two eukaryotic genomes: relation to chromosome structure," *J. Mol. Biol.*, 259, pp.579-588, 1996.
- [3] Tomita M., Wada M. and Kawashima Y., "ApA dinucleotide periodicity in prokaryote, eukaryote, and organelle genomes," *J. Mol. Evol.*, 49, pp.182-192, 1999.
- [4] Schieg, P. and Herzel, H., "Periodicities of 10-11 bp as indicators of the supercoiled state of genomic DNA, *J. Mol. Biol.*, 343, pp.891-901, 2004.
- [5] Worning P., Jensen L.J., Nelson K.E., Brunak S. and Ussery D.W., "Structural analysis of DNA sequence: evidence for lateral gene transfer in *T. maritima*," *Nucleic Acids Res.*, 28, pp.706-709, 2000.
- [6] Willenbrock H. and Ussery D.W., "Chromatin architecture and gene expression in *E. coli*," *Genome Biol.*, 5, pp.252-261, 2004.
- [7] Baran R.H. and Ko H., "An Ising model of transcription polarity in bacterial chromosomes," *Physica A*, 362, pp.403-422, 2006.