JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 13, Number 7, 2006
© Mary Ann Liebert, Inc.
Pp. 1289–1298

Wavelet Analysis of DNA Walks

ADRIAN D. HAIMOVICH,[1] BRUCE BYRNE,[1] RAMAKRISHNA RAMASWAMY,[2,3,4]
and WILLIAM J. WELSH[1,5]

ABSTRACT

A wavelet transform of the DNA "walk" constructed from a genomic sequence offers a direct visualization of short and long-range patterns in nucleotide sequences. We study sequences that encode diverse biological functions, taken from a variety of genomes. Pattern irregularities in the transform are frequently associated with sequences of biological interest. Exonic regions, for example, visualize differently under wavelet analysis than introns, and ribosomal RNA regions display distinct universal signatures. DNA walk wavelet analysis can provide a sensitive and rapid assessment of the putative biological significance of genomic DNA.

Key words: DNA walk, genomic analysis, sequence analysis, wavelet transform.

1. INTRODUCTION

The deluge of sequence information following the completion of the human and other genome projects has created the need for diverse and efficient analytic tools. Finding patterns of interest within DNA sequences remains an objective in the field of bioinformatics; new mathematical and computational tools are increasingly required to facilitate the rapid, accurate and computationally driven identification of regions of biological significance such as exons or repetitive DNA (Anastassiou, 2001; Claverie, 1997; Fickett, 1982; Tiwari et al., 1997).

The totality of DNA sequence comprises the genome. In complex organisms, segments of the genome that specify biological function (e.g., genes, regulatory regions) are interspersed in a background sequence of noncoding DNA. Although much of the nonfunctional DNA is apparently random in sequence, some has descended from functional DNA: genomes have evolved as a result of events such as genome duplication and gene transfer. Some parts of the functional DNA are involved in gene regulation. The majority of the functional DNA encodes RNA of four different types: messenger (m) RNA, which is translated into proteins, transfer (t) RNA, which is involved in amino acid transport, ribosomal (r) RNA, which makes up the ribosome, and small nuclear (sn) RNA, which is involved in regulating gene expression.

[1] The Informatics Institute of the University of Medicine and Dentistry of New Jersey, Newark, New Jersey.

[2] Center for Systems Biology, Institute for Advanced Study, Princeton, New Jersey.

[3] School of Physical Sciences, Jawaharlal Nehru University (JNU), New Delhi, India.

[4] Center for Computational Biology and Bioinformatics, School of Information Technology, JNU, New Delhi, India.

[5] Department of Pharmacology, University of Medicine and Dentistry of New Jersey, New Brunswick, New Jersey.

Since functional DNA is strongly constrained by evolution, the arrangement of its bases is known to be very specific. The pattern of such constraints includes how change might affect secondary structure of resulting RNA molecule, how specific short recognition sequences might be affected, or how a sequence change would otherwise impact the specific physiologic role of the molecule. In contrast, nonfunctional parts of DNA are under no selection pressure and have undergone random unchecked mutations in the course of evolution.

In the present paper, we apply wavelet analysis to DNA sequences. Our motivation is to uncover pattern irregularities in the DNA: these often result from constraints and are therefore frequently associated with function. A number of studies have been devoted to examining the short and long range correlations inherent in DNA sequences, the latter manifested, for example, as a power-law decay of the sequence correlation function (Azbel, 1995; Buldyrev et al., 1993, 1995). Indeed, a number of studies have suggested that there are long range correlations in DNA sequences that make it *scale-free* or fractal (Arneodo et al., 1996; Peng et al., 1992). This suggests that wavelet analysis could find fruitful application here, since it is well known that fractal time-series are characterized by regular wavelet patterns, against which background the contrasting short range correlations in DNA should be evident.

Analysis of both long and short-range structure in DNA often uses digital signal processing (DSP) tools that focus on intrinsic characteristics of DNA sequences. A given DNA sequence is converted into a digital signal (Tiwari et al., 1997) that can be subject to a variety of mathematical transforms. This has led to discoveries about the three-base periodicity in exons (Anastassiou, 2001; Tiwari et al., 1997). Alternately, correlation functions that compare each base of the DNA sequence against its adjoining bases have also been proposed and subsequently analyzed using Fourier transforms (Dodin et al., 2000) and other methods (Anastassiou, 2001). Wavelets, which have recently been applied to a variety of biomedical problems with great success (Aldroubi and Unser, 1996; Lio, 2003; Unser and Aldroubi, 1996), can be described as building blocks that can decorrelate data with great speed (Sweldens, 1996). Using the wavelet transform, a signal can be analyzed in space as well as in time, and is represented by a series expansion in terms of a translated and dilated version of a "mother" wavelet, usually denoted ψ (Vegte, 2002). The ability to obtain statistical indicators via wavelets has been used earlier, in combination with the DNA walk (Peng et al., 1992), to verify the existence of long range correlations in intronic DNA sequences (Arneodo et al., 1995). The wavelet transform has proven to be highly adept in characterizing the scaling properties of sequences (Berger et al., 2002) and it has also facilitated the extraction of fractal, scale-independent, information from genome sequences (Arneodo et al., 1996).

Here, we apply the wavelet transform to the analysis of DNA walks. The advantage of a localized signal in time and frequency provided by the wavelet transform is supplemented by its highly visual nature (Dodin et al., 2000). By examining a number of sequences and their DNA walk wavelet transforms, we show that irregularities in the wavelet pattern map over regions with varying complexities along the sequence, and are often associated with regions of particular biological interest, including those encoding exons in mRNA and other significant RNA classes.

2. METHODS

DNA walk

The usual *random walk* (Feller, 1968) is a stochastic process represented by a sequence of partial sums of random variables. A one-dimensional random walk is constituted of a sequence of independent displacements, either left or right at each time step. The *DNA walk* denotes a special case, where the partial sums are obtained by aggregating the numerical values associated with the components of a DNA sequence (Berger et al., 2004). The walk is created by defining an incremental variable that associates to the time step k the value $x(k) = \pm 1$, depending on whether the base at position k along the sequence is respectively, a pyrimidine (cytosine, thymine) or a purine (adenine, guanine). The DNA walk is then defined as

$$F(k) = \sum_{n=1}^{k} x(n).$$

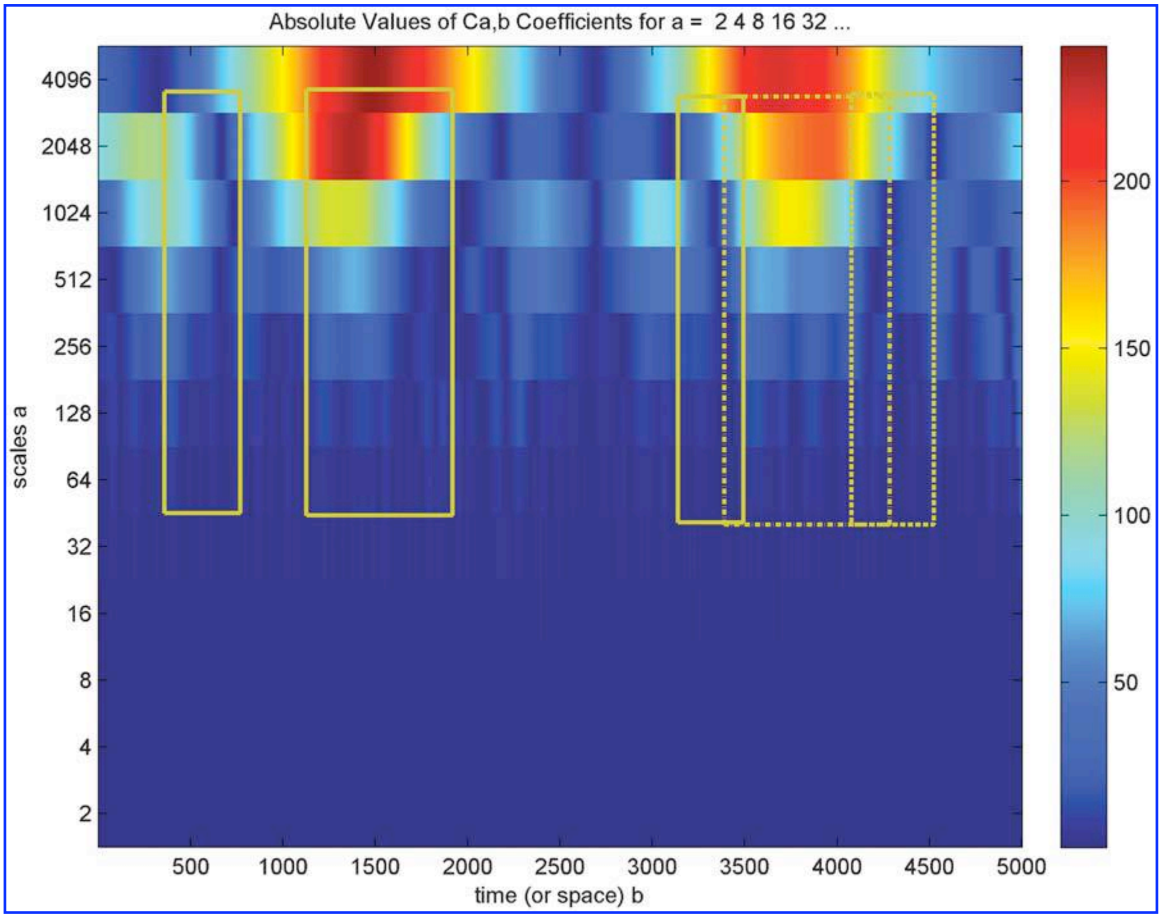

FIG. 1. *Shigella flexneri* large virulence plasmid pWR501 (accession no. AF348706 gi:13310487) Y-R DNA walk. Coding sequence (cds) data appears in Table 2. Solid boxes represent open reading frames (ORFs) of one DNA strand; dashed boxes represent antisense strand ORFs.

The function $F(k)$ expresses the relative content of pyrimidines (C, T) over purines (A, G), specifically, $F(k) = Y - R = (C + T) - (A + G)$. Similar walks can be constructed based, for example, on $(G + C) - (A + T)$ counts that weigh the GC bias in coding regions, or via a mapping to the complex plane, using the four cardinal points $\{+1, -1, +j, -j\}$ (Berger et al., 2004; Feder, 1988).

The quantity of interest is the exponent (H) governing the variance of increments of $F(t)$,

$$\mathrm{var}(F(t) - F(s)) = C|t - s|^{2H}, \tag{1}$$

where C is a positive constant, and $0 \leq H \leq 1$ is the so called Hurst exponent (Feder, 1988). Equation (1) represents a fractal process in that the variance of the increments is proportional to a power of the increment step.

The DNA walk has been applied in the study of long-range correlations in intron containing genes (Arneodo et al., 1995). This analysis was accomplished by fitting the variance of increments in the DNA walk to a power-law (Anastassiou, 2001). In other words, long-term correlations in DNA sequences were demonstrated by fitting the data to values of H different than 0.5. These correlations were not observed in coding DNA sequences or in genes without any introns (Peng et al., 1992). Correlations suggest the possibility of large scale patterns in introns over thousands of bases, a pattern that cannot necessarily be observed over short regions. It follows logically that regions with different correlation properties will result in a diversity of patterns. If long term correlations are restricted to introns, then pattern irregularities could indicate regions of biological interest.

Fractional Brownian motion

Once a signal of the DNA strand is constructed using the DNA walk technique, a study of fractional Brownian motion estimates can be initiated. Fractional Brownian motion is a random walk with a defined Hurst exponent. As mentioned previously, a random walk without memory has a Hurst exponent of 0.5. In the presence of some pattern, the Hurst exponent value will deviate from 0.5. Fractional Brownian motion estimates allow for the reverse process, finding the Hurst exponent of a defined signal as opposed to creating a signal from a defined exponent. The estimates presented in the results section were done using Matlab software. In addition, the range was calculated through a comparison of the greatest and least values for $F(k)$ in a given sequence.

Wavelets

A continuous one-dimensional Daubechies wavelet (Dodin et al., 2000) was applied to analyze the DNA walk to detect regions of irregular patterns. The continuous wavelet profile offers data decomposition at several scales (Kawagashira et al., 2002). The continuous wavelet transform (CWT) of the signal $F(t)$ with respect to the wavelet $\psi(t)$ is defined as

$$C_{a,b} \equiv \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} F(t)\psi\left(\frac{t-b}{a}\right) dt$$

where a (the scale parameter) > 0 and b (the translation parameter) is a real number (Lio, 2003).

A continuous wavelet transform maps a one-dimensional signal, like the DNA walk, to a two-dimensional time-scale representation. The transform itself is calculated by continuously shifting a continuously scalable function over the signal. The result is the correlation between the scale function and the signal (Lio, 2003). The size of the window sliding across the signal over time is the scale a defined as

$$a = \frac{\alpha}{frequency},$$

where α is a positive constant (Vegte, 2002).

The CWT allows for the investigation of the scale of a signal, while maintaining the time aspect. This feature gives wavelet transforms a distinct advantage over many other tests including the Fourier transform. It was this advantage that motivated this study towards the use of the CWT. Given the varying size of biological structures, a variable window size was thought to maximize chances of finding irregular patterns.

Guidelines for analysis

To begin, fractional Brownian motion estimates were utilized in the comparison of the Hurst exponents of entire sequence DNA walks to those of the sequences' notated exons. The DNA walk was then analyzed using the Daubechies wavelet (Daubechies, 1992; Matlab, 1984–2004) with the aid of the Matlab Wavelet Toolbox. No significant differences were seen when the analysis was undertaken with other wavelets, but for the sake of uniformity, all the figures reported here were generated using the Daubechies wavelet. Candidates for regions of biological interest were identified by high values of the wavelet coefficients $|C_{a,b}|^2$. These regions were then corroborated from NCBI Genbank data.

3. RESULTS

Fractional Brownian estimates

As discussed above, the study of fractional Brownian motion estimates enables investigation of uniformity in sequences. We started by comparing Hurst exponent estimates for regions of biological interest with larger portions of the original nucleotide sequence seeking to establish the possibility of distinct regions within the DNA walk. Table 1 shows the results of this exercise. In *R. norvegicus* beta-hO-r gene, *Ateles geoffroyi* haptoglobin, *Saccharomyces cerevisiae* chromosome IV, *Saccharomyces cerevisiae* chromosome

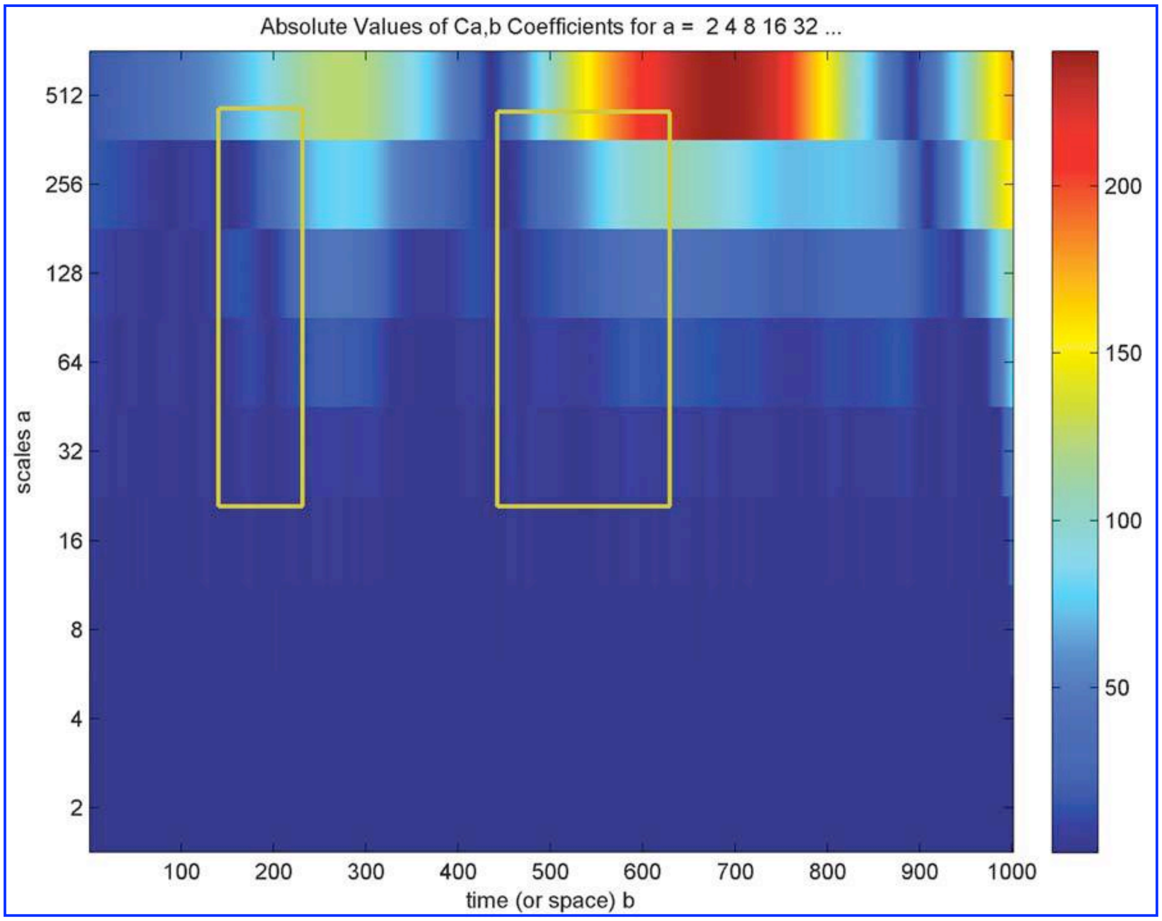

FIG. 2. *Homo sapiens* Xq pseudoautosomal region (accession no. AJ271736 gi:8979791) from the human Y chromosome containing an exon and a repeat region. The exon outlined in the left box is 69 bases long, whereas the repeat region framed by right box is 136 bases long.

VII, and *Shigella flexneri*, we estimated and compared the H exponents of exons and the surrounding sequence. The other comparisons were done in a DNA sequence containing transfer RNA (tRNA) and in a RNA sequence containing ribosomal RNA (rRNA). The results from the Hurst exponent estimates suggest that regions of biological interest have different pattern structures than their surrounding nucleotide segments. The pronounced differences led the investigation towards a method that would allow for the visualization of the fractal pattern along the length of the sequence, namely the wavelet transform.

Exons and coding regions

To verify that pattern irregularities elucidated by wavelet transforms indeed correspond to regions of biological interest, we have tested a number of different sequences of differing length from a variety of species. Figure 1 displays the wavelet DNA walk analysis of the first 5000 bases of the large virulence plasmid of *Shigella flexneri* (Venkatesan et al., 2001). The top portion of each of the figures generated by the Matlab Wavelet toolbox plots the $C_{a,b}$ wavelet coefficients versus the base index number, while the bottom portion is a "temperature" plot of $|C_{a,b}|^2$ versus the scale and base index number. High temperatures correlate to high intensities of $|C_{a,b}|^2$. Regions of high intensity in the wavelet transform are matched to the annotated coding regions in the sequence in all cases, including two open reading frames (ORFs) in the antisense strand (the dashed boxes; the boxed regions indicate the locations and length of exons).

With a focus on structure rather than on pattern, the DNA walk can thus be used to discover ORFs in any position and direction. Figure 2 shows the results of the DNA walk wavelet analysis on a region of the human Y sex chromosome. In this case, the two high intensity areas map to an exonic sequence

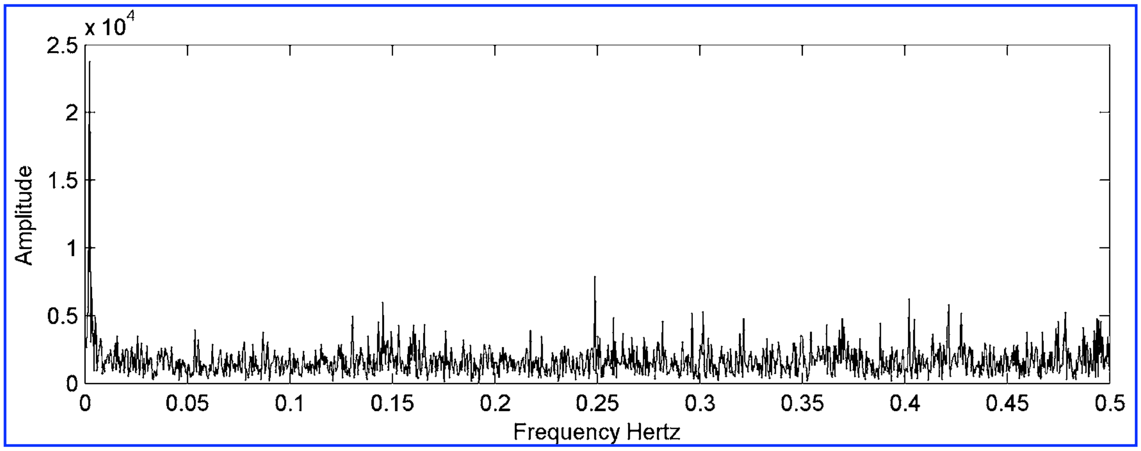

FIG. 3. Fourier power spectrum of tRNA-Leu from Mouse tRNA gene cluster (accession no. X00229 gi:54000). No single peak is apparent.

and a DNA repeat region. While not currently encoding information, many DNA repeat regions arise from ancient ORFs and therefore share many of the structural features of biologically significant genomic sequences.

Although coding regions typically have distinct local GC content, wavelet analysis of the GC or other mappings, such as the complex walk, did not reveal additional structure. This may be due to the fact that GC richness of exonic regions varies from species to species (Sankoff and Nadeau, 2000). To put our findings in perspective, it is well known that spectral transforms such as the Fourier transform can frequently locate generalized coding regions (Fickett and Tung, 1992), though the sensitivity can be quite poor. The simple Fourier transform is not designed to probe hierarchical structure, and thus cannot identify a number of different coding regions within a larger sequence. Size limitations strongly affect the Fourier transform, and here we find the wavelet transform superior in its ability to locate pattern irregularities regardless of their length and location on the sequence.

High intensity regions in the wavelet transform, in some cases, do not mark the entire length of the subsequence of interest: in Figure 1, for example, the first boxed exonic region is flanked by two high intensity areas. Conversely, the boxes map out most of the exon lengths in that region. One reason for the flanking intensity regions may be that the intron's pattern is being broken by the exon and then the newly set exonic pattern is being broken in turn by the introns on the other side. Most frequently, however, when the high intensity areas mapped to exons, they mapped to the entire length of exons.

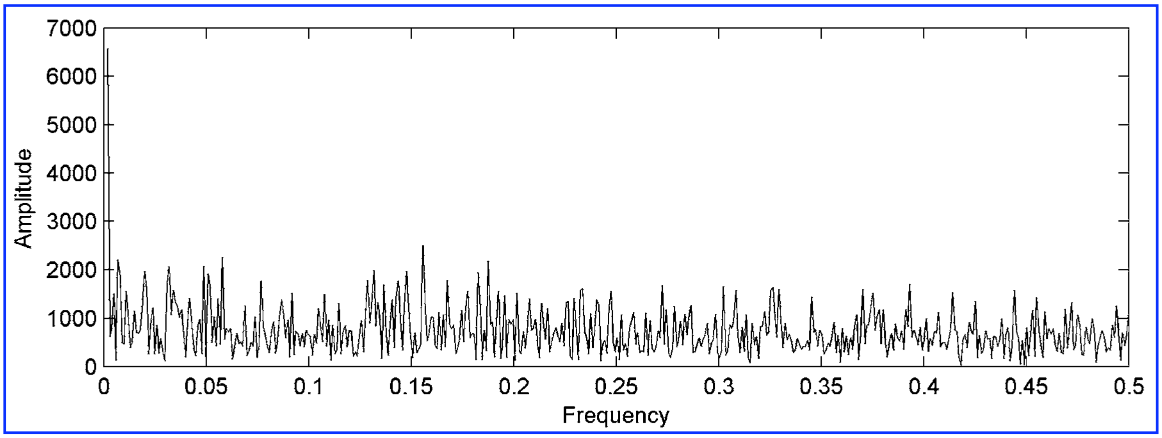

FIG. 4. Fourier power spectrum of *Arabidopsis* genomic sequence spanning two NCBI-predicted exons and a putative snRNA region. There is no indication that there may be mRNA or snRNA located within the sequence.

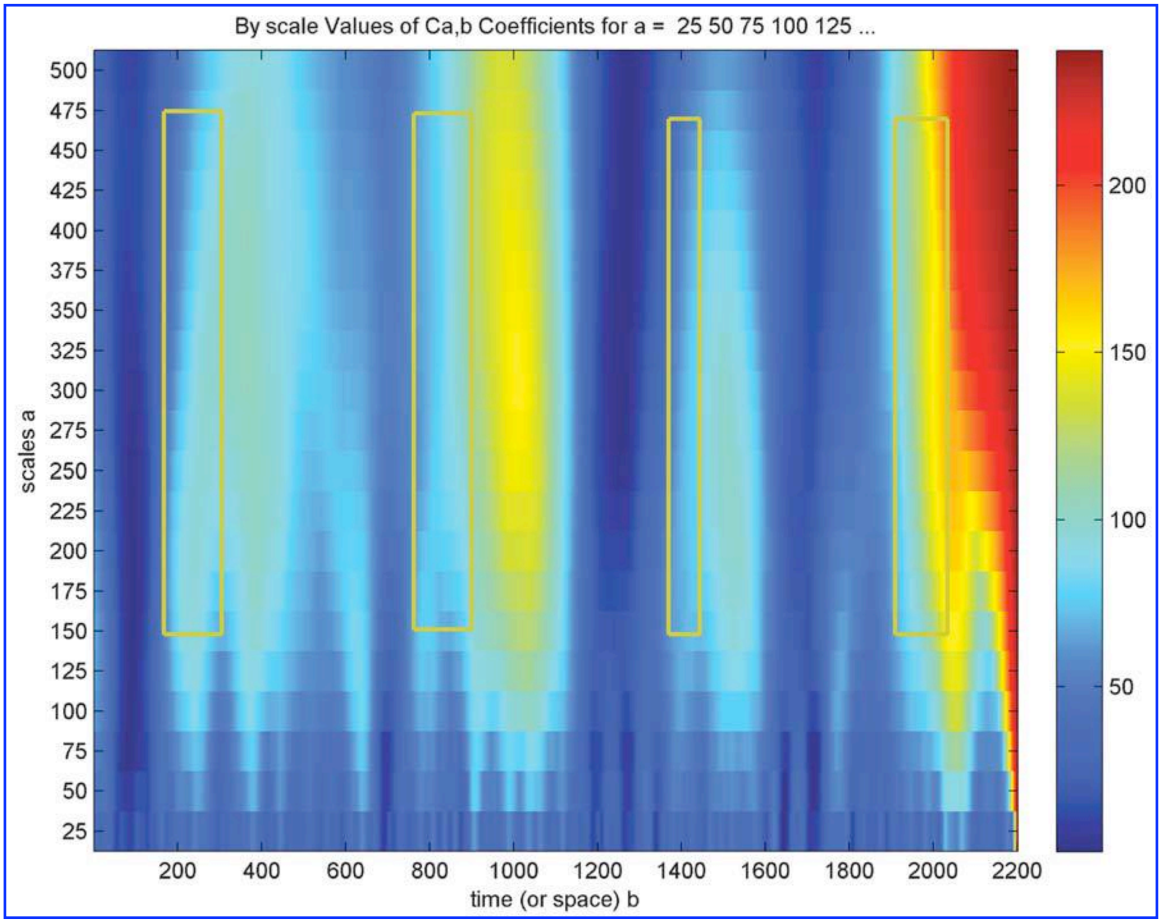

FIG. 5. Mouse tRNA gene cluster for tRNA-Leu, tRNA-Asp, tRNA-Gly and tRNA-Glu (accession no. X00229 gi:54000) tRNA: 137.. 219, 778.. 849, 1437.. 1508, 1945.. 2016.

RNA

We have also examined the utility of the wavelet transform in analyzing sequences encoding other functional types of RNA. For reference, Figure 3 shows the Fourier power spectral transform of a typical tRNA sequence. Note that no outstanding peaks are apparent. Figure 4 is a similar analysis of a 1000 base sequence that contains exons as well as a snRNA. This plot does not show a peak at 1/3 even though there are mRNA and snRNA present.

Once again, in RNA, regions without significant biological import (like introns in the previous section) are observed to have a regular pattern. In the earlier investigation into the irregular structure of exons, we saw that the intron patterns were broken by the exon ORFs. We subjected other sequences to analyses to determine if sequence pattern imposed by evolutionary constraints might be revealed by wavelet analysis. Indeed, DNA sequence encoding tRNA, rRNA, and snRNA generates irregular patterns in the DNA walk of their surrounding sequence; they display as high intensity regions in the wavelet transform. Figure 5 shows the wavelet analysis of a sequence of a mouse tRNA gene cluster. Within the cluster there are four tRNA regions which appear in the following order: leucine, aspartic acid, glycine, and glutamic acid. The four regions are clearly shown as having patterns that were not similar to those of the remainder of the sequence.

A genomic sequence containing two exonic segments and a small nuclear RNA segment was investigated using the same techniques as the previous examples. The sequence, a portion of the *Arabidopsis thaliana* (commonly known as thale cress) chromosome IV, was chosen for its unique combination of structures. The original Fourier transform (Fig. 4) does not give any indication that exons may be present over the span.

Figure 6 shows the wavelet transform of the *Arabidopsis* sequence. The wavelet transform's first high intensity area maps to the end of the first exon and the start of the second while the second and third intensity areas do not appear to map to mRNA. Instead, there is a snRNA sequence beginning at the end of the second region and continuing through to the start of the third.

Table 1. Results of Fractional Brownian Motion Estimates

Name	*Range*	*Entire sequence range*	*Average Hurst parameters*	*Entire sequence average Hurst parameter*
R. norvegicus beta-hO-r gene	35	108	0.5747	0.556
Ateles geoffroyi haptoglobin	109	306	0.4457	0.5591
Saccharomyces cerevisiae IV	86	426	0.5376	0.4695
Saccharomyces cerevisiae VII	40	302	0.4845	0.4438
TRK1 tRNA-Lys	7	97	0.6907	0.5269
TRQ1 tRNA-Gln	6	97	0.5381	0.5269
TRL1 tRNA-Leu	9	97	0.6402	0.5269
Candida sp. 153M 18S ribosomal RNA gene 18S ribosomal RNA	4	49	0.7925	0.4525
Candida sp. 153M 18S ribosomal RNA gene 5.8S ribosomal RNA	17	49	0.2382	0.4525
Candida sp. 153M 18S ribosomal RNA gene 26S ribosomal RNA	5	49	1.18	0.4525
Shigella flexneri	50	1920	0.4109	0.4485
Shigella flexneri	33	1920	0.2993	0.4485
Shigella flexneri	81	1920	0.4539	0.4485
Shigella flexneri	115	1920	0.5088	0.4485

The samples used represent a wide variety of genomic data types. The range values calculate the distance from the greatest value in the DNA walk signal to the least value as an indication of the sequence complexities. The Hurst parameters show that the correlations of the regions of biological interest differed from that of their surrounding regions.

Table 2. ORF for *Shigella flexneri* Large Virulence Plasmid p WR501 (accession no. AF348706 gi:13310487)

ORF	*Begin codons*	*End codons*	*First codon*	*G+C content (%)*	*No. of amino acids*
S0002	485	949	ATG	51.0	154
S0003	1176	2042	ATG	32.8	288
S0005	3272	3526	ATG	42.3	84
S0006	4441	3470	TTG	31.2	323
S0007	4742	4344	GTG	32.3	132

Data from Venkatesan et al. (2001).

4. DISCUSSION

The wavelet transform of DNA walks provides a simple analytic tool that can rapidly pinpoint regions of biological interest within genomic sequences. The multiscale nature of the technique is especially valuable in analyzing large sequences, where there may be structure on a variety of length scales. Although related techniques, such as the Fourier transform, have been used in different applications in sequence analysis, these have inherent limitations that can be overcome to some extent by wavelet transforms, which display pattern irregularities regardless of location and length. Since biologically significant structures can be

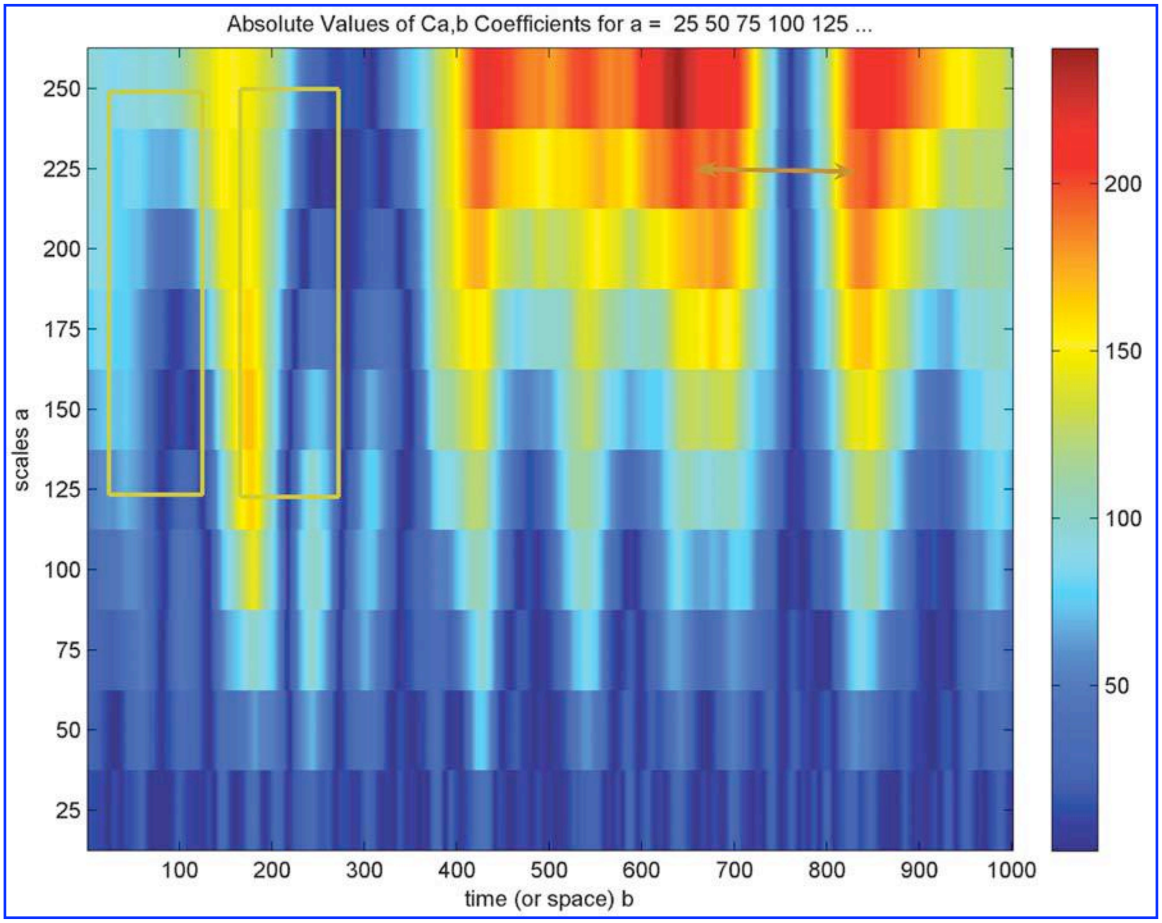

FIG. 6. *Arabidopsis thaliana* sequence containing two exons and an snRNA. The exons are notated using the boxes, whereas the solid line represents the snRNA region. Note that the flanks of the snRNA are shown. Here, the two exonic segments flank a region of high intensity.

shorter than ten bases and can extend to lengths in the thousands, multiscale techniques of increased sensitivity are valuable: here wavelet analysis is of great utility.

The present work is an initial application of wavelet transform of DNA walk signals. This appears to be a sensitive probe of DNA patterns that correlate with biological function. Further tests and experimentation is needed to uncover the variety of transform patterns that correspond to different features of biological interest (Collantes et al., 1997).

ACKNOWLEDGMENTS

Support for this research was provide to B.B. and W.J.W. by the NIH-NLM through Integrated Advanced Information Management Systems (grant 2G08LM06230-03A1).

REFERENCES

Aldroubi, A., and Unser, M. 1996. *Wavelets in Medicine and Biology*. CRC Press, Boca Raton, FL.

Anastassiou, D. 2001. Genomic signal processing. *IEEE Signal Proc*. 8–20.

Arneodo, A., Bacry, E., Graves, P.V., et al. 1995. Characterizing long-range correlations in DNA sequences from wavelet analysis. *Phys*. *Rev*. *Lett*. 74, 3293–3296.

Arneodo, A., d'Aubenton-Carafa, Y., Bacry, E., et al. 1996. Wavelet based fractal analysis of DNA sequences. *Phys. D* 96.

Azbel, M. 1995. *Phys. Rev. Lett.* 75, 168–171.

Berger, J.A., Mitra, S.K., Carli, M., et al. 2002. New approaches to genome sequence analysis based on digital signal processing. *IEEE Workshop on GENSIPS* 1–4.

Berger, J.A., Mitra, S.K., Carli, M., et al. 2004. Visualization and analysis of DNA sequences using DNA walks. *J. Franklin Inst.* 341, 37–53.

Buldyrev, S.V., Goldberger, A., Havlin, S., et al. 1993. On long-range power law correlations in DNA. *Phys. Rev. Lett.* 71, 1776.

Buldyrev, S.V., Goldberger, A.L., Havlin, S., et al. 1995. Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. *Phys. Rev. Lett.* 73, 5084–5091.

Claverie, J.-M. 1997. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* 6, 1735–1744.

Collantes, E.R., Duta, R., Welsh, W.J., et al. 1997. Preprocessing of HPLC trace impurity patterns by wavelet packets for pharmaceutical fingerprinting using artificial neural networks. *Anal. Chem.* 69, 1392–1397.

Daubechies, I. 1992. Ten lectures on wavelets. Presented at S.I.A.M., Philadelphia.

Dodin, G., Vandergheynst, P., Levoir, P., et al. 2000. Fourier and wavelet transform analysis, a tool for visualising regular patterns in DNA sequences. *J. Theor. Biol.* 206, 323–326.

Feder, J. 1988. *Fractals*. Plenum Press, New York.

Feller, W. 1968. *An Introduction to Probability Theory and Its Applications*. Wiley, New York.

Fickett, J.W. 1982. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.* 10, 5303–5318.

Fickett, J.W., and Tung, C.S. 1992. Assessment of protein coding measures. *Nucleic Acids Res.* 20, 6441–6450.

Kawagashira, N., Ohtomo, Y., Murakami, K., et al. 2002. Wavelet profiles: their application in *Oryza sativa* DNA sequence analysis. Presented at *IEEE Comp. Soc. Bioinformatics Conference (CSB'02)*. Stanford, CA.

Lio, P. 2003. Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinform. Rev.* 19, 2–9.

Matlab. 1984–2004. *Matlab*. MathWorks, Inc., Natick, Massachusetts.

Peng, C.K., Buldyrev, S.V., Goldberger, A.L., et al. 1992. Long-range correlations in nucleotide sequences. *Nature* 356, 168–170.

Sankoff, D., and Nadeau, J.H. 2000. *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Sweldens, W. 1996. Wavelets: What next? *IEEE Proc.* 84, 680–685.

Tiwari, S., Ramachandran, S., Bhattacharya, A., et al. 1997. Prediction of probable genes by Fourier analysis of genomic sequences. *CABIOS* 13, 263–270.

Unser, M., and Aldroubi, A. 1996. A review of wavelets in biomedical applications. *Proc. IEEE* 84, 626–638.

Vegte, J.V. 2002. *Fundamentals of Digital Signal Processing*. Prentice Hall, Upper Saddle River, NJ.

Venkatesan, M.M., Goldberg, M.B., Rose, D.J., et al. 2001. Complete DNA sequence and analysis of the large virulence plasmid of *Shigella flexneri*. *Infect. Immun.* 69, 3271–3285.

Address correspondence to:
Adrian D. Haimovich
Informatics Institute
University of Medicine and Dentistry of New Jersey
65 Bergen St., Suite 517
P.O. Box 1709
Newark, NJ 07101-1709

E-mail: adrian.haimovich@gmail.com