EURASIP Journal on Applied Signal Processing 2004:1, 29–42

Spectrogram Analysis of Genomes

David Sussillo
Department of Electrical Engineering, Columbia University, NY 10027, USA
Email: sussillo@ee.columbia.edu

Anshul Kundaje
Department of Electrical Engineering, Columbia University, NY 10027, USA
Email: abk2001@cs.columbia.edu

Dimitris Anastassiou
Department of Electrical Engineering, Center for Computational Biology and Bioinformatics (C2B2) and Columbia Genome Center, Columbia University, NY 10027, USA
Email: anastas@ee.columbia.edu

Received 28 February 2003; Revised 22 July 2003

We perform frequency-domain analysis in the genomes of various organisms using tricolor spectrograms, identifying several types of distinct visual patterns characterizing specific DNA regions. We relate patterns and their frequency characteristics to the sequence characteristics of the DNA. At times, the spectrogram patterns can be related to the structure of the corresponding protein region by using various public databases such as GenBank. Some patterns are explained from the biological nature of the corresponding regions, which relate to chromosome structure and protein coding, and some patterns have yet unknown biological significance. We found biologically meaningful patterns, on the scale of millions of base pairs, to a few hundred base pairs. Chromosome-wide patterns include periodicities ranging from 2 to 300. The color of the spectrogram depends on the nucleotide content at specific frequencies, and therefore can be used as a local indicator of CG content and other measures of relative base content. Several smaller-scale patterns are found to represent different types of domains made up of various tandem repeats.

Keywords and phrases: DNA spectrograms, frequency-domain analysis, genome analysis.

1. INTRODUCTION

Color spectrograms of biomolecular sequences were introduced in [1, 2] as visualization tools providing information about the local nature of DNA stretches. These spectrograms give a simultaneous view of the local frequency throughout the nucleotide sequence, as well as the local nucleotide content indicated by the color of the spectrogram. They are helpful not only for the identification of genes and other regions of known biological significance, but also for the discovery of yet unknown regions of potential significance, characterized by distinct visual patterns in the spectrogram that are not easily detectable by character string analysis. Further, they have been found to give global information about whole chromosomes as well.

In this paper, we discuss the features and patterns that such spectrograms reveal. We applied a slightly modified version (described below) of the spectrogram development tool introduced in [1, 2] that provides a more direct manifestation of the local relative nucleotide content in the color of the spectrogram, and explored the patterns characteristic in the genomes of various organisms. We created color spectrograms of various frequency bandwidths and sequence lengths. Although the genomes of these organisms vary greatly in size, chromosome number, and complexity, we found many interesting features, some of which are common to all organisms and some are unique to a particular organism. Some of the uncovered patterns relate to the overall chromosome structure or to protein coding. On some occasions, the specific function of a protein could be understood by visual comparison to other proteins.

We analyzed some parts of the genomes from *E. coli*, *M. tuberculosis*, *S. cerevisiae*, *P. falciparum*, *C. elegans*, *D. melanogaster*, and *H. sapiens*, viewing chromosomes and chromosome subsequences using the tricolor spectrogram with as much or as little frequency and sequence resolution as necessary. We allowed zooming in and out in both the frequency and sequence dimensions, thus facilitating easy navigation of DNA that is normally intimidating in its complexity. A set of colors was initially chosen for the four different bases to maximize the discriminatory power of the spectrogram. Depending on the pattern, we adjusted the frequency

and sequence resolutions so that the prominent frequencies were accurately highlighted and thus we were able to view different features of the chromosome with great precision. When possible, we referenced the subsequence from which the pattern was created with various public databases to further ascertain the function of the region. We then annotated the patterns with the type of pattern, prominent periodicities, position in the chromosomal DNA sequence, and corresponding position in the protein sequence if the DNA was coding. Thus, we related pattern shape and color to significant structural and functional elements in the genome. Most of our searches were exhaustive, and the patterns shown in this paper are exemplary of myriad patterns in the various genomes.

The spectrograms were developed using the short-time Fourier transform, that is, by applying the N-point discrete Fourier transform (DFT) over a sliding window of size N. The difficulty in creating DNA spectrograms results from the fact that DNA sequences are defined by character strings rather than numerical sequences. This problem can be solved by considering the *binary indicator sequences* $u_A[n]$, $u_T[n]$, $u_C[n]$, and $u_G[n]$, taking the value of either one or zero depending on whether or not the corresponding character exists at location n. These four sequences form a redundant set because they add to 1 for all n. Therefore, any three of these sequences are sufficient to determine the character string. In [1, 2], color spectrograms are defined by creating RGB superposition, using the colors red, green, and blue, of the spectrograms for the numerical sequences

$$\begin{aligned} x_r[n] &= a_r u_A[n] + t_r u_T[n] + c_r u_C[n] + g_r u_G[n], \\ x_g[n] &= a_g u_A[n] + t_g u_T[n] + c_g u_C[n] + g_g u_G[n], \\ x_b[n] &= a_b u_A[n] + t_b u_T[n] + c_b u_C[n] + g_b u_G[n], \end{aligned} \quad (1)$$

in which, to enhance the discriminating power of the visualization, the coefficients in the above equations are chosen by assigning each of the four letters to a vertex of a regular tetrahedron in the three-dimensional space. In the present implementation, we further improve the discriminating power by ensuring that all points in the tetrahedron have different absolute values with respect to any axis using the following choice of coefficients:

$$\begin{array}{lll} a_r = 0, & a_g = 0, & a_b = 1, \\ t_r = 0.911, & t_g = -0.244, & t_b = -0.333, \\ c_r = 0.244, & c_g = 0.911, & c_b = -0.333, \\ g_r = -0.817, & g_g = -0.471, & g_b = -0.471. \end{array} \quad (2)$$

To illustrate, we first consider three examples that demonstrate both the use of color and periodicity in the spectrogram. The horizontal axis indicates the location in the DNA sequence measured in base pairs (bp) from the origin and the vertical axis indicates the discrete frequency of the DFT measured in cycles per STFT window size. The corresponding period is equal to N/k, where k is the discrete frequency and N is the STFT window size.

Unlike the traditional spectrograms that employ pseudocolor to achieve greater contrast, the spectrograms that are used to visualize DNA sequences contain useful information

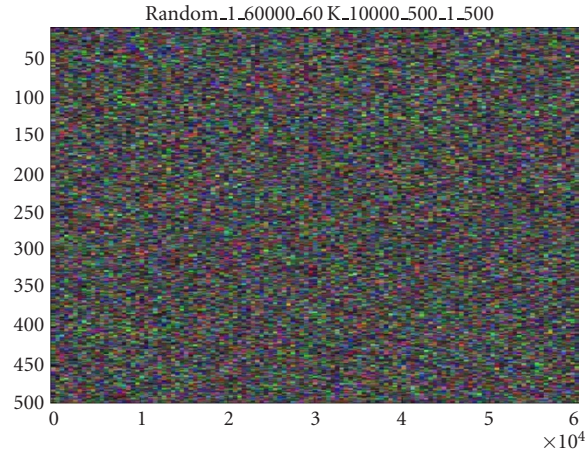

Figure 1: Spectrogram of a random DNA sequence of length 60 kbp. No obvious patterns are discernable. Spectrogram titles are annotated with a helpful name or accession tag, sequence-start index, sequence-end index, approximate sequence length, DFT window size, window overlap, lowest frequency shown in image, and highest frequency shown in image.

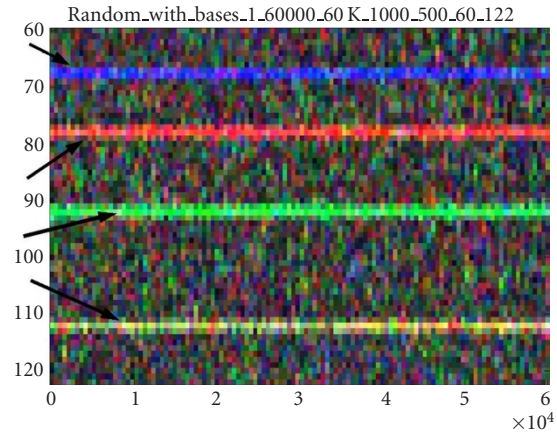

Figure 2: Spectrogram of random DNA of length 60 kbp with bases A, T, C, and G with periods 15, 13, 11, and 9, respectively. The nucleotide A is represented by the color blue, T by red, C by green, and G by yellow. Arrows mark the different periodicities.

encoded in color. The colors for the nucleotides A, T, C, and G are blue, red, green, and yellow, respectively. These colors were chosen to optimize the discrimination between different nucleotides. As a rule of thumb, the interaction between the various nucleotides is visualized as the superposition of colors representing those nucleotides. Thus, a sequence composed of ATATAT ... would have a purple bar at the frequency corresponding to period 2. The first spectrogram (Figure 1) shows a spectrogram created from a sequence of 60000 "totally random" nucleotides. The sequence was created from an independent identically and uniformly distributed random sequence model so that every position has equal chance of being an A, T, C, or G. No obvious patterns are noticeable. The second spectrogram (Figure 2) shows the same sequence as the first but with a modification

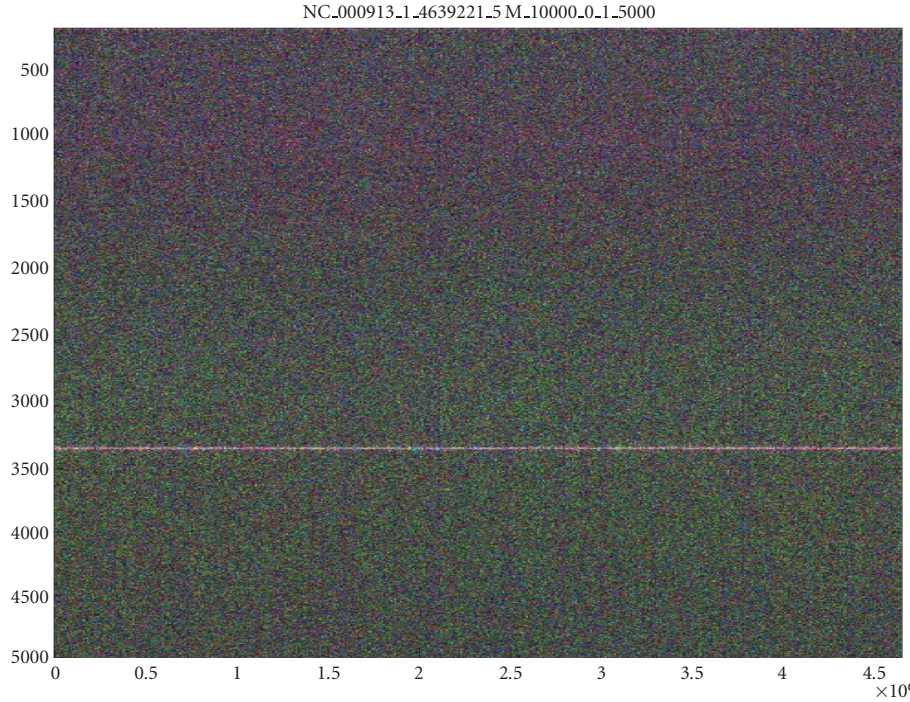

Figure 3: Spectrogram of the entire *E. coli* K12 chromosome (about 4.6 Mbp). The line marking the 3-base periodicity of protein-coding regions extends without a visible break across the entire chromosome. There is a change in color going from higher frequencies (greenish) to lower frequencies (purplish).

so that every 15 nucleotides, there is an *A*; every 13 nucleotides, there is a *T*; every 11 nucleotides, there is a *C*; and every 9 nucleotides, there is a *G*. This figure demonstrates that even in complicated sequences, *A* is mapped by the color blue, *T* by red, *C* by green, and *G* by yellow.

2. CHROMOSOME-WIDE PATTERNS

Distinguishing patterns by their size makes a simple categorization. Those patterns composed of millions of bp are considered large; those that are composed of up to several hundred thousand nucleotides are medium; and those patterns consisting of up to several thousand bp are small. Typically, larger patterns represent structural elements and smaller patterns are useful in visualizing something about a protein-coding region. Here, we focus first on large patterns. In doing so, we focus on the general characteristics of the chromosome-wide spectrogram.

2.1. *E. coli*

Figure 3 shows the spectrogram of the entire chromosome for the bacteria *E. coli* using STFT window size $N = 10\,000$. The count among all nucleotides in *E. coli* is roughly equal (A=1142136, T=1140877, C=1179433, G=1176775) and the total number of nucleotides is over 4.6 Mbp. The most salient feature is the strong intensity with periodicity 3 (frequency 3333) that corresponds to protein-coding regions. The fact that protein-coding regions in DNA typically have a peak at the frequency of 3 periodicity in their Fourier spectra is well known [3, 4, 5, 6]. The whiteness of this line shows that most of the bases are being used in protein coding, and this is clearly reflected by the continuity and intensity of the line with periodicity 3. Second, at regular intervals along the DNA sequence, there appear thin veins of purple, implying AT rich areas intermittently placed along chromosome. Finally, there is a general shift in hue as the frequency decreases. The larger frequencies are more greenish in hue and the lower frequencies are more purplish. The purplish hue extends over from about the 6.5-base periodicity and upwards and shows that even while apparently coding for genes almost everywhere on the chromosome, the chromosome is also preserving higher periodicities involving the nucleotides *A* and *T*. This is particularly interesting considering that the total number of each of the four bases in the genome is nearly equal. The purplish hue in the lower frequencies may be related to the twisting of the DNA molecule that leads to helical repeats.

2.2. *C. elegans chromosome III*

We now turn our attention to the multicellular organism *C. elegans*. Figure 4 shows the DNA spectrogram of chromosome III. The general hue of the spectrogram is darker than that of *E. coli*. This relates directly to the relative number of bases in chromosome III (A=4444502, T=4423430, C=2449072, G=2466240). The horizontal line of intensity marking the 3-base periodicity is much less pronounced than *E. coli* in that there are more gaps along the sequence. This is consistent with the general rule that eucaryotic DNA

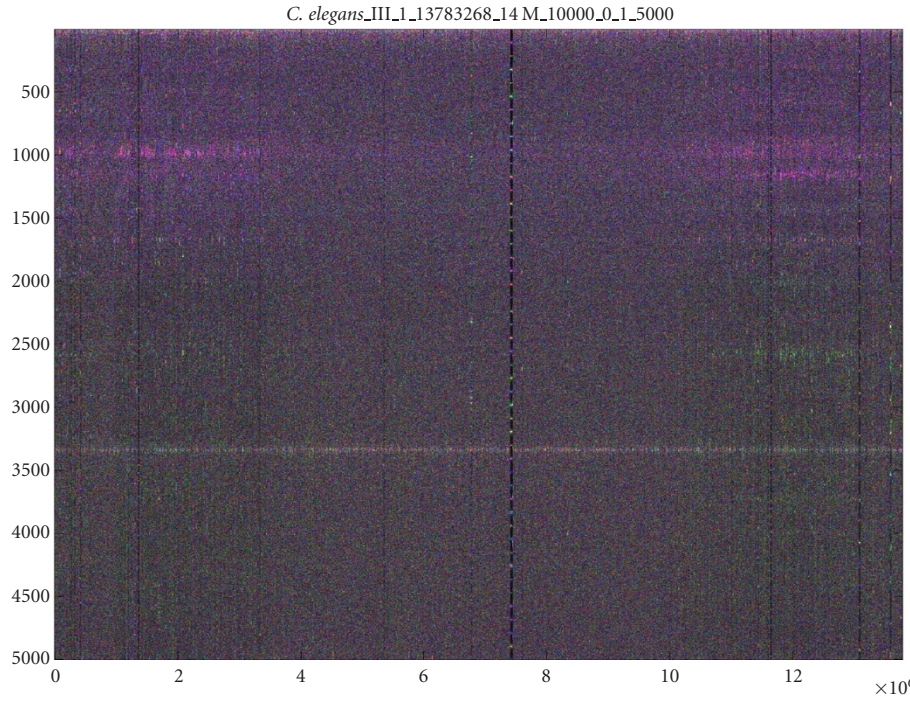

Figure 4: Spectrogram of the chromosome III of *C. elegans* (13.8 Mbp). The 3-base periodicity relating to protein coding is noted. A minisatellite is noticeable at 7.4 Mbp (see Figure 16). Various periodicities are noticeable, in particular, the purple 10+-base periodicity in both chromosome arms and coincident 8, 9-base and green 3.8-base periodicities in the right chromosome arms.

contains more noncoding DNA such as intergenic DNA and introns. In the middle of the spectrogram, there is a vertical bar that identifies a "minisatellite," roughly 50 kbp in length. The details of minisatellites are explained in Section 3.1. On some regions, there are strong horizontal bands of intensity between the frequencies representing the 8-base periodicity and 9-base periodicity (at 8.7) and also just above 10 (at 10.2, which we call the "10+ periodicity") throughout the entire chromosome. In the right part of the spectrogram, (close to 12 Mbp) there are strong periodicities involving the color green and thus the bases GC at 3.9.

The 10+ periodicity appears to be of special importance. Figure 5 shows the magnitude plot of the DFT for the four nucleotides in the subsequence 1456174 – 1596391. Each separate base is plotted with a different color. The frequency range shown corresponds to periods 8 through 12. The periodicities at 10+ are the strongest in the bases *A* & *T* (area indicated by arrow). This periodicity may relate to DNA helical structure, which has a periodicity of 10.4 bp on average [7, 8, 9, 10]. The 10+ periodicity may also be related to folding around nucleosomes, as the nucleotides *A* and *T* are preferred in the minor grove when binding to the nucleosome core. The DNA double helix kinks when wrapped around the nucleosome core, thus reducing its helical periodicity to 10.39±0.02 bp [9]. We found that the maximal intensity of this band has a 10.2-base periodicity.

We further searched chromosome III of *C. elegans* at much lower frequencies and found a 1.5 Mbp long (0.8 Mbp–2.6 Mbp subsequence) bubble centered on period 300. This was accomplished using a DFT window size of 40000. Figure 6 shows this spectrogram with the two bubbles centered at period 300 marked by arrows. This was the only example of a periodicity found around 300 and it is unclear what biological significance the bubble may have. Figure 7

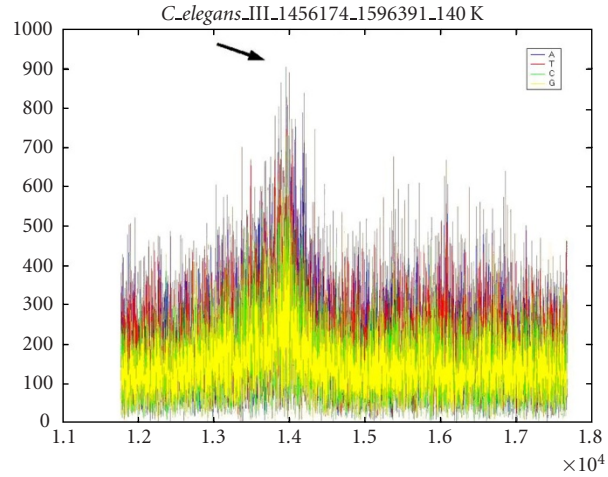

Figure 5: DFT magnitude plot of 140 Kbp section of *C. elegans* chromosome III showing higher values at period 10+ in all bases, but particularly *A* and *T*. An arrow marks the peak in the periodicity range of 9.9–10.5.

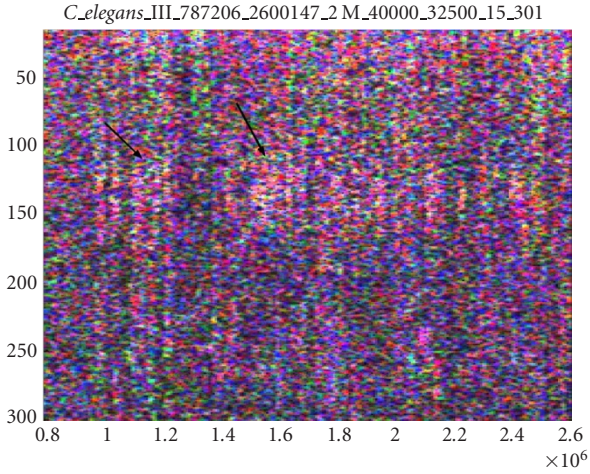

Figure 6: Spectrogram showing an intensity increase around a periodicity of 300 in *C. elegans* chromosome III. The sequence is roughly 2 Mbp in length. Arrows mark two such areas.

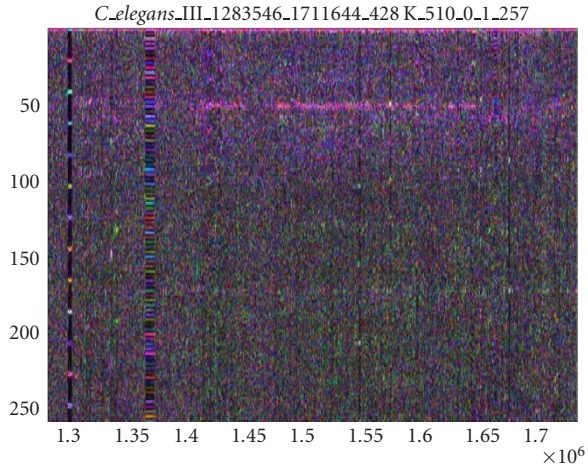

Figure 7: Spectrogram showing a strong coincident 10+-base periodicity in the same DNA sequence shown in Figure 6 (coincident with 300-base periodicity). This spectrogram corresponds to the rightmost arrow in Figure 6 and is 428 Kbp in length.

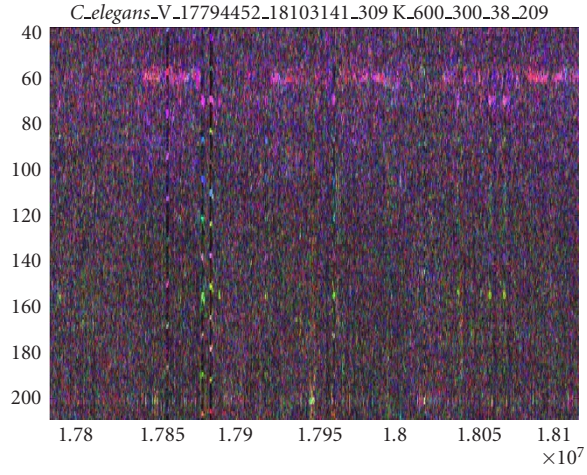

Figure 8: Spectrogram showing antagonism between 10+-base and 3-base periodicities in *C. elegans* chromosome III (300 Kbp). The 10+-base periodicity is at the top of the figure while the 3-base periodicity is shown at the very bottom.

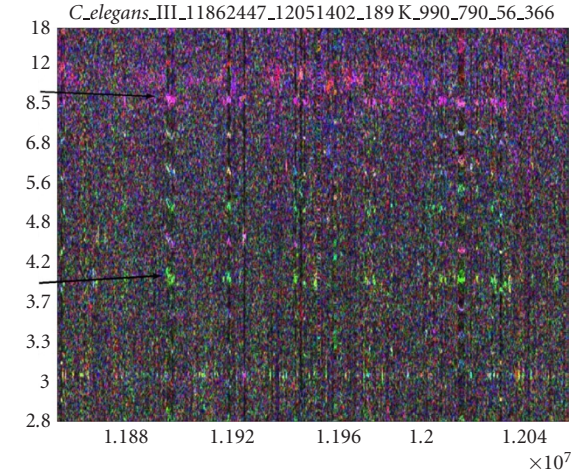

Figure 9: Spectrogram of 189 Kbp section of the right arm of *C. elegans* chromosome III. Note that the periodicity is shown on the vertical scale. The arrows point to sections of the spectrogram, showing a single instance of the highly dispersed repeat family. Variations of the pattern can be seen throughout the spectrogram. A purple 8.75-base periodicity, as well as a green 3.9-base periodicity, identifies this family of strings. The harmonics between 3.9 and 8.75 (the beads of color between 3.9 and 8.75) change color from one repeat to another, indicating that they are different but related strings. These tandem repeats are non-protein-coding regions. The 10+-base periodicity is antagonistic with the repeat family. This pattern is found over 3 Mbp of the right arm of the chromosome.

shows the same area of the chromosome (1.4 Mbp–1.6 Mbp) at higher frequency resolution, thus showing smaller periodicities. There appears to be coincident intensity at 10+ period in exactly the same area of intensity in the 300-period bubble.

In general, it appears that there are both "antagonism" and "cooperation" between various periodicities in all the chromosomes that we analyzed. For example, the arms of *C. elegans* chromosome III show obvious cooperation among many periodicities appearing simultaneously (Figure 7). Some cooperative periodicities are harmonics of a fundamental periodicity, indicating a repeat region (see Section 3.1). On the other hand, Figure 8, a subsection of chromosome V of *C. elegans*, shows an example of antagonism between the 3-base periodicity and the 10+-base periodicity. The brightest spots on the 3-base periodicity are the dimmest spots on 10+-base periodicity and vice versa. An explanation may be that in non-protein-coding regions, the periodicities due to structural constraints are more pronounced.

We identified a unique family of repeats in chromosome III via cooperation among periodicities. In the right arm of

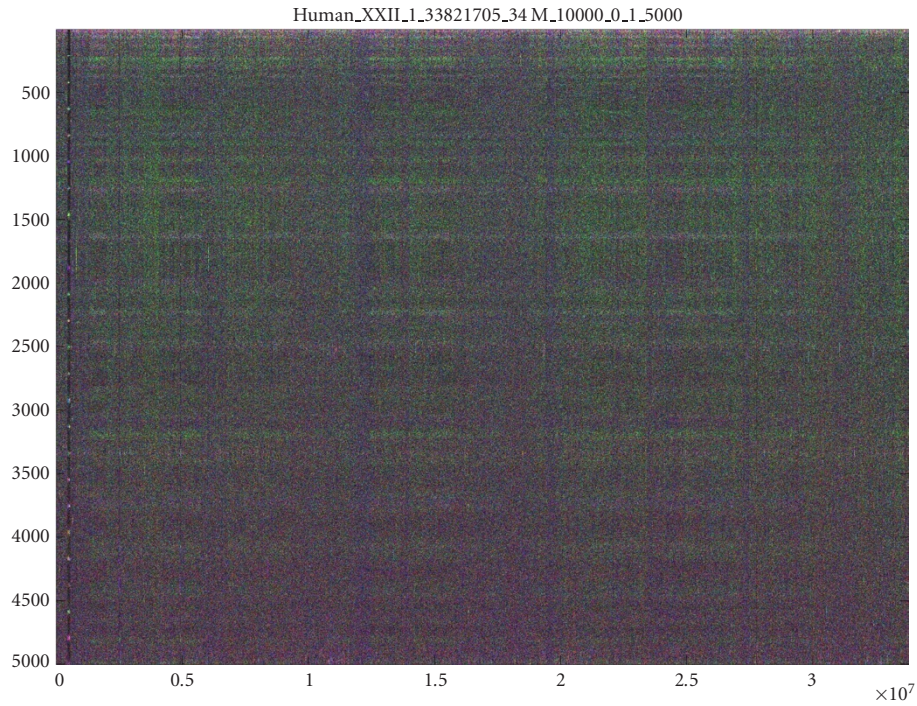

Figure 10: Spectrogram of human chromosome 22. Noticeably absent is the line representing the 3-base periodicity relating to protein coding. The 800 or so genes located on chromosome 22 simply do not cover enough of the chromosome to make a visible line at the resolution of 34 Mbp. Many periodicities are visible across the entire length of the chromosome.

chromosome III (10–13 Mbp), it appears that the AT rich 8.75-base periodicity almost always coincides with the GC-rich 3.9-base periodicity (Figure 4). In fact, the pattern found in the right arm of chromosome III, which shows cooperative periodicities at the chromosome level, is composed of a family of strings that are repeated in a very haphazard fashion. These strings are both heavily mutated and heavily dispersed throughout the chromosome. Yet throughout the many variations within the family, the 8.75-base and 3.9-base periodicities are always conserved. One instance of a repeat unit is "tttccggcaaattggcaagctgtcggaatttaaaa." Figure 9 shows how the family of strings manifests within the DNA. An instance of the family repeats for a hundred to a couple thousand bp, and these regions are interspersed among other DNA every 10 Kbp or so. Repeats of this family of mutated strings, unbelievably, are responsible for the macroscopic character of the right arm (3 Mbp region) of chromosome III (Figure 4). It is unclear whether or not the conserved periodicities imply a conserved biological function for the string, or whether it is simply a mathematical or biological property of this family of strings that certain of its periodicities are more easily preserved against mutation.

2.3. Human chromosome 22

The last full chromosome we analyzed was human chromosome 22. The actual sequence used was the correct reordering of contigs found in hs_chr22.fa from NCBI. This ordering is: NT_011516.5, NT_028395.1, NT_011519.9, NT_011520.8, NT_011521.1, NT_011522.3, NT_011523.8, NT_030872.1, NT_011525.4, NT_019197.3, and NT_011526.4. Figure 10 shows the 33 million-plus nucleotides of human chromosome 22. A strong bar of intensity representing the 3-base periodicity is strikingly absent. Closer inspection shows that there are many genes along chromosome 22 but they are far enough apart so that there is no noticeable band. There are around 30 easily noticeable, different periodicities that span the entire length of the chromosome. The biological function of these periodicities is unclear. Some periodicities may reflect higher periodicities in the form of harmonics.

The higher structures in DNA folding are unknown, so we were interested in determining whether or not spectrogram analysis would yield any insights into the DNA folding and superstructure. It is known that DNA has many orders of structure [11]. The simplest of such a superstructure is that of the nucleosome. Nucleosomes are an essential structural element in DNA: 146 bp wrap twice around a single nucleosome core particle, and between two nucleosomes, there is "linker" DNA that ranges in size but on the whole, nucleosomes repeat at intervals of about 200 bp. Nucleosome core particles will bind randomly along a sequence of DNA. However, AT rich sequences in the minor groove of DNA bind preferentially to the nucleosome core particle. Since euchromatic DNA is arranged in nucleosomes that require structural bending of the DNA, it is plausible that there might be some evidence of this structure in the form of a strong band with intensity of 200-base periodicity. We

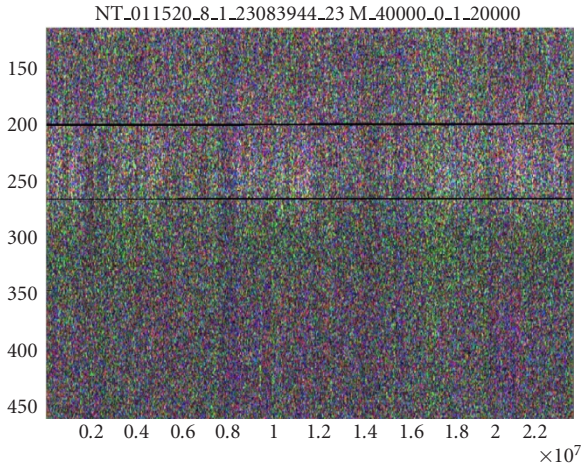

Figure 11: NT_011520.8 (23 Mbp in length) of human chromosome 22. The two artificial black lines mark the 150-base and 200-base periodicities. This band of intensity may relate to the folding of DNA into nucleosomes.

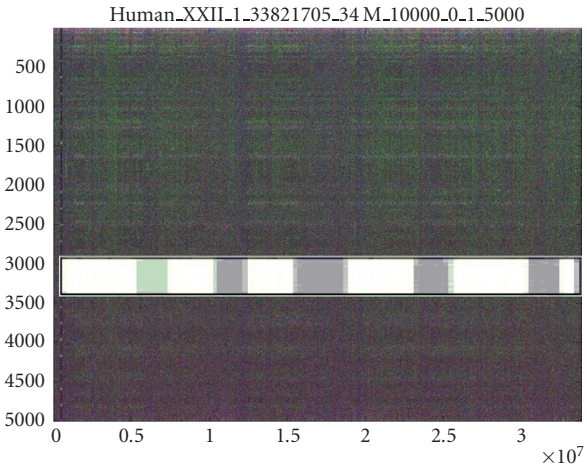

Figure 12: Spectrogram of human chromosome 22 matched up with a part of the Giemsa-stained schematic of the same chromosome. There is a visual agreement between AT-rich regions and dark bands of Giemsa staining.

viewed contig NT_011520.8 (23 Mbp in length) of chromosome 22 with a very large DFT window in order to get high-frequency resolution. Figure 11 shows contig NT_011520.8 in the frequency range to show the 200-base periodicity. Two dark lines mark the 150-base periodicity and the 200-base periodicity, indicating a band of increased intensity between these markers. This intensity band may represent periodicities involved in nucleosome-chromatin superstructure. This 150 − 200-base periodicity band was the only one found in our exploration of various chromosomes. The 150 − 200-base periodicity was the largest periodicity found in the human chromosome 22.

We found an interesting feature of human chromosome 22 in the variation of the CG versus AT rich regions. As mentioned earlier, the color of the DNA spectrogram reflects the ratio of different nucleotides in the sequence (Figures 1 and 2). Different genomes vary greatly in the percentages of nucleotides that compose the sequence. As shown in Figure 10, a single chromosome can have long expanses of a single distribution of bases. Figure 10 shows clear boundaries between areas of high CG content and areas with lower GC content. The laboratory technique of Giemsa staining is correlated to the relative content of CG nucleotides. The GC-rich regions of DNA are responsible for the light bands in Giemsa staining while GC-poor regions create the dark bands [12]. We matched up a schematic of human chromosome 22 marked by Giemsa staining with our DNA spectrogram and found a reasonable alignment between the dark bands of the Giemsa stained chromosome schematic and the darker, purplish AT regions of the spectrogram (Figure 12). The match was made by aligning the rightmost part of the spectrogram with the "bottom" of the chromosome, that is, contig NT_011526.4. Because the spectrogram encodes different colors for each different base, it is easy to get a feeling for the relative number of bases in a sequence.

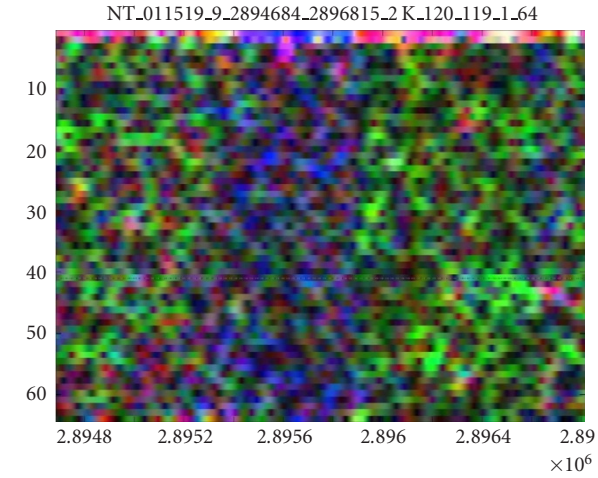

Figure 13: Spectrogram showing two CpG islands separated by a sequence very rich in the nucleotide *A*. Both islands yielded blast results showing T-box genes.

CpG islands [13] are DNA stretches in which a particular methylation process that normally reduces the occurrence of CG dinucleotides is suppressed, and therefore CG nucleotides appear more frequently than elsewhere. Such stretches are also readily identified using the DNA spectrogram. For example, we found two CpG islands simply by searching for the greenest subsequence we could locate in the genome. This simple color criterion yielded two CpG islands, shown in Figure 13. Figure 14 shows the results from the Emboss CpGplot program on the sequence that generated the spectrogram. The CpGplot figure shows that the CpG islands are located in exactly those sequences that are most green in the spectrogram. The subsequences from which the spectrogram was created were blasted on the NCBI website and both "green" sequences coded for T-box genes. The T-box genes

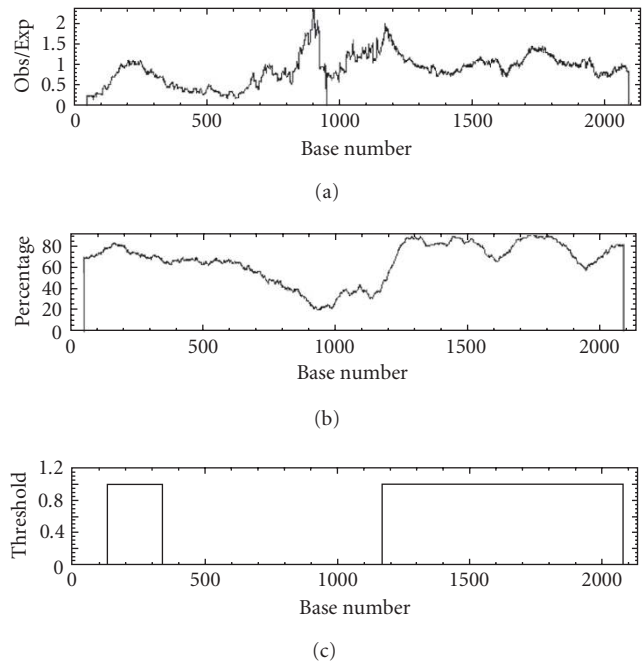

Figure 14: Graphs showing the results from the emboss CpGplot routine. (c) shows the predicted CpG islands (putative islands).

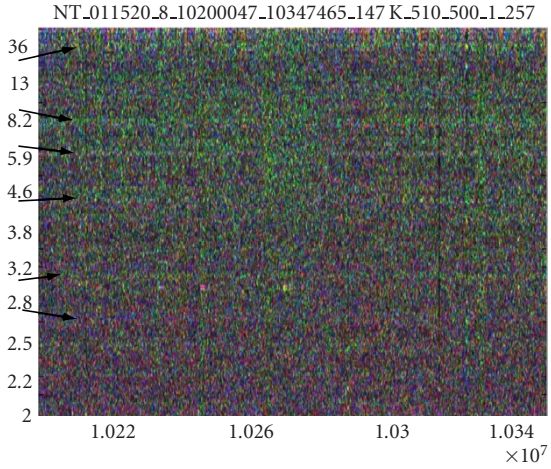

Figure 15: Spectrogram of a 147 Kbp section of human chromosome 22. Periodicity is shown on vertical scale. Contrasted with Figure 9, this spectrogram shows that the chromosome-wide periodicities found in human chromosome 22 are qualitatively different from those found in the right arm of C. elegans chromosome III. The periodicities here are much more finely embedded in the DNA and do not represent any obvious family of strings discretely interspersed throughout the region. Arrows point out some of the chromosome-wide periodicities found in Figure 10.

share a common binding domain, called the T-box. Finding this gene is in keeping with the idea that CpG islands encode for housekeeping genes.

Finally, we wondered whether or not the chromosome-wide periodicities found in human chromosome 22 are caused by a highly dispersed repeat family similar to that found in the right arm of *C. elegans* chromosome III. This appears not to be the case. The macroscopic appearance of periodicities in *C. elegans* is caused by widely placed repeats with such strong characteristics as shown at the macroscopic level. In the case of human chromosome 22, it appears as if the very fabric of intergenic DNA is woven with a string patterns that employs characteristic periodicities seen at the chromosome level (Figure 15). In other words, it appears as if the majority of intergenic DNA carries the periodicities found at the macroscopic level. Initial investigations show that these embedded periodicities are not found in chromosome 17 of the mouse.

3. SMALL PATTERNS

We now turn our attention to smaller subsequences of interest in various genomes. Color spectrograms can clearly identify, by their special signatures, several patterns including repetitive areas of biological significance such as particular triplet repeats [14], GATA repeats [15], or other characteristic repeating motifs in protein structures [16].

The sequences that we analyzed were typically several thousand bp in length, no more than a hundred thousand bp. The majority of smaller sequences we analyzed relates to protein-coding regions or repetitive sequences in non-protein-coding regions. The public databases were often helpful in matching spectrogram patterns to proteins. We annotated the spectrograms with the type of pattern, prominent periodicities, position in the chromosome, and corresponding position in the protein sequence if the DNA was coding.

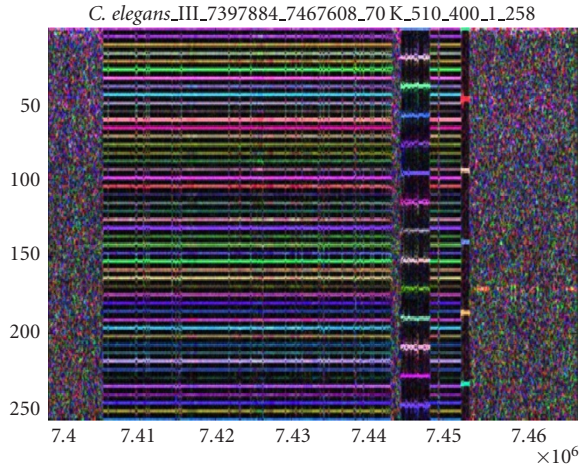

Figure 16: Spectrogram showing a minisatellite with repeat unit of length 95 bp in chromosome 3 of *C. elegans*. Slight variations in the basic repeat pattern can be seen as vertical lines that appear blurry. The minisatellite is interrupted by a small amount of nonrepeat DNA as well as an even simpler repeat unit of length 5 kbp.

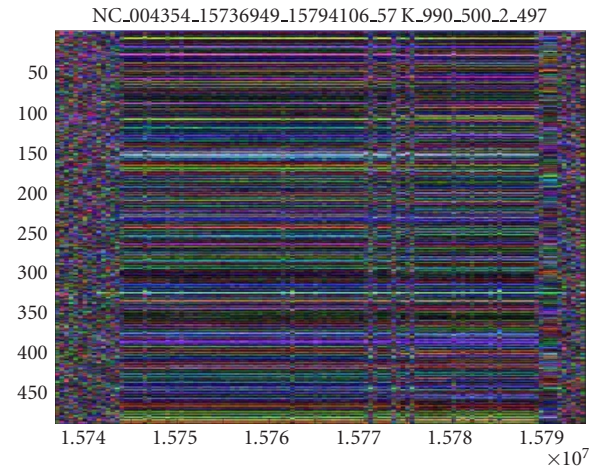

Figure 17: Spectrogram showing 40 kbp minisatellite in chromosome X of *D. melanogaster*. The repeat length is 298 bp. Three strong interruptions can be seen as vertical lines just right of the center.

We used a number of public databases during our analysis of DNA color spectrograms. The determination of whether or not a sequence was protein coding was accomplished using the SGD and GenBank databases. We also noted structural and functional details of the corresponding protein. Domains and motifs corresponding to the protein region were discovered using PFAM, CYGD, and SWISS-PROT databases for yeast, WormPD for *C. elegans*, and GenBank annotations for humans. Structural predictions were obtained using Pedant (CYGD) and GCG PepStruct (SGD). To test specifically the beta-helix supersecondary structure, the Betawrap program (Betawrap) was used.

At smaller length scales, the parameters of the STFT are very important in visualization; we initially experimented these parameters with different DFT window sizes for the spectrogram. It was found that using roughly 6 K nucleotides per spectrogram image with a DFT window size of 120 and an overlap of 119 gives the most optimal visualization of protein-coding regions. The choices of DFT window size and overlap were found to be particularly important in determining the pattern shape.

3.1. Minisatellites

The genome has repetitive regions varying in range from 500 bp to 100 kbp in length. These regions are composed of a smaller repeat unit that varies in length. If the length of the repeat unit is below 100, then the overall repeat region is called a minisatellite or variable number of tandem repeats (VNTR). Minisatellites have been found to vary in the number of tandem repeats in different germ cells and thus, make useful genetic markers [17]. A minisatellite composed of roughly 30 kbp was found in *C. elegans* chromosome III (Figure 16). It is also visible in the middle of Figure 4. The tandem repeat is composed of the 95 bp-long unit sequence "ttttgataattactgcctccagaaattgatgattttcccattgatttgtctacatagggca tcgaaaagcacccaatatttagagaacagaaga" and slight variants. According to "WormBase," this subsequence of chromosome III is completely unannotated. Another 40 kbp minisatellite was found in chromosome X of *D. melanogaster* (see Figure 17). The tandem repeat sequence is composed of the 298 bp-long unit sequence "ttcatttcaagaatccagtgcagaagaaaatcaaatgacagaa gtgccatggacactatcaacatcactttcccaatcaagttcaaaaacaaagaatatatttt tcgagtcaaagtgtaaatgaagacaacatttctcaagaagatacaaggacaccatcaa tatctgtcccacaatcaagtacaacagcaaatagattacttacaggttcgggtgcagaa gagccaacagctcaagaggagacatcggaactttcaaaatccttacctcaattaacaa cagaagagagcagttcattt." The GenBank file indicates that the location of the predicted gene CG32580 is in the region 15740143-15792683. Both minisatellites are large enough to be identifiable when viewed from a spectrogram of the entire chromosome.

Spectrogram visualization of DNA repetitive areas, including minisatellites, microsatellites, and the other smaller tandem repeats that we will discuss, gives an immediate indication of the repeat length T. If the DFT window size N is sufficiently large to capture the fundamental frequency $k = N/T$, then all the harmonics will appear as equally spaced horizontal lines at the integer multiples of N/T up to (and including if present) the "maximum" frequency $N/2$. Therefore, the number L of horizontal lines that appear in the spectrogram (without counting the omnipresent DC frequency) will be the integer part of half the repeat length T. Conversely, the repeat length can be deduced by inspection of the spectrogram as $2L$ if L is even, or $2L + 1$ if L is odd. The color of each harmonic shows the contribution from the different bases.

Intergenic tandem repeats are interesting because of their mutagenic properties. It is known that there are large numbers of intergenic tandem repeats in the form of microsatellites and minisatellites in higher organisms. In *C. elegans*, there are around 38 defined dispersed repeat families, many of which correspond to transposon-like elements. Many transposons have already been defined in *C. elegans*

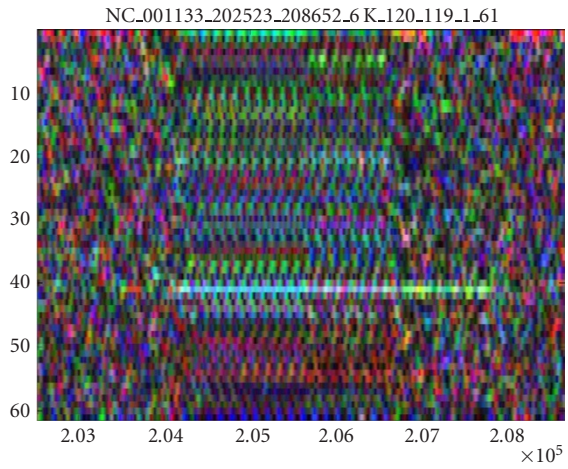

Figure 18: Spectrogram showing the quilt in protein FLO1 corresponding to the flocculin domain.

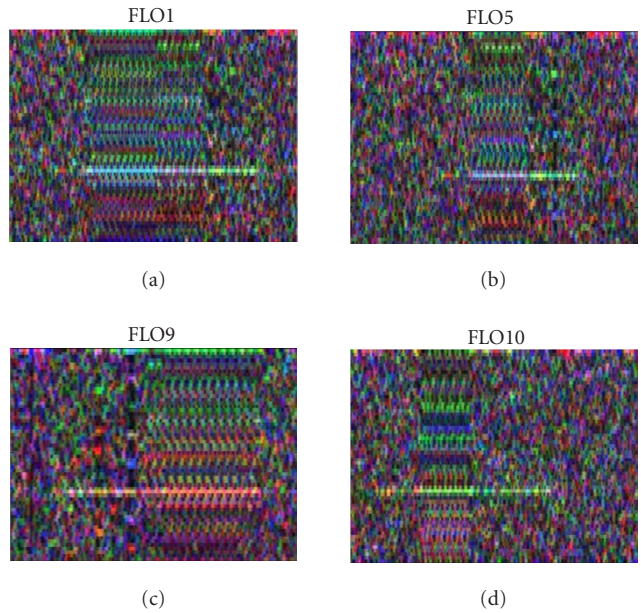

Figure 19: Four spectrograms of FLO genes 1, 5, 9, and 10. Quilts can be seen in all four genes. Close inspection of (a) and (b) shows that (b) is a subsection of (a). FLO9 (c) shows the same coloration as the other three upon reverse complementation.

as mutagenic elements. Many of the dispersed repeat families have been found to be relics of transposon families no longer active. Autosome arms tend to have high recombination rates as compared to the central regions. We found that spectrogram analysis confirms that there are relatively large numbers of repeat patterns in the autosome arms. Some of these repeat clusters were also found in closely related genes. This suggests that these regions may be sites of random mutations and may be rapidly evolving to give rise to new genes and gene families.

3.2. Smaller tandem repeats—quilts, shafts, and bars

After detailed analysis of all the 16 nuclear chromosomes of *S. cerevisiae* (GenBank accession numbers NC_001133-NC_001148) as well as sections of the *C. elegans*, *D. melanogaster*, and human genomes, we identified three basic types of patterns, to which we refer as "quilts," "shafts," and "bars," based on their appearance. All three patterns represent tandem repeats, but the repeat-unit length differs between them. These were not found to be exhaustive but merely illustrative of patterns in the various genomes. Many genes were found to be composites of these patterns. We discovered that quilts, shafts, and bars could be used to predict the homology, structure, and function of proteins. In yeast, most of these patterns were part of the protein-coding regions. However, in the higher organisms, the patterns were also found in the intergenic and intronic regions.

Quilts (Figure 18) are relatively rare patterns in the yeast genome. They appear as beating, repetitive patterns at almost all frequencies over relatively long stretches of DNA. If present in the coding regions of genes, quilts represent protein domains consisting of large tandem repeats. We found quilts representing repeats of up to 45 amino acids (135 bp).

Bars (Figures 20 and 21) and shafts (Figure 22) show strong periodicities uniformly over a stretch of coding DNA. Shafts differ from bars in that they are thin and have few dominant periodicities, causing black areas along most of the other frequencies in the spectrograms. In other words, the basic repeat sequence is smaller in shafts than bars. Bars and quilts with similar appearances, having similar frequency patterns and colors, were found to be homologous as confirmed by BLAST alignment scores, database annotations, and literature.

It should be noted that a quilt appears as a quilt and not as a bar because the DFT window size (typically 120 for viewing proteins) used to create these spectrograms is smaller than the base repeat unit length (135 bp in this case). Although the distinction between quilts and bars is artificial, we found the distinction to be useful since we could differentiate high complexity repeats from lower complexity repeats while still maintaining an appropriate sequence resolution for viewing protein-coding regions.

3.2.1. Quilts—yeast flocculation genes

The quilt observed in Figure 18 is an example of a yeast "flocculation" gene [18]. Yeast flocculation is an asexual, calcium-dependent, and reversible aggregation of cells into flocs. This phenomenon is thought to involve cell surface components. Yeast flocculation is under genetic control, and two dominant flocculation genes have been defined by classical genetics, FLO1 and FLO5. The other relevant FLO genes include FLO9 and FLO10. The functional active domain in these cell surface proteins is made of large tandem repeats up to 45 amino acids known as flocculin repeats. The flocculin region corresponds to the quilted region of the spectrogram. The quilted region was observed in all the FLO genes (Figure 19). The flocculin domain is serine-threonine rich and highly O-glycosylated, adopting a stiff and extended conformation. The efficiency of interaction of the FLO proteins is directly

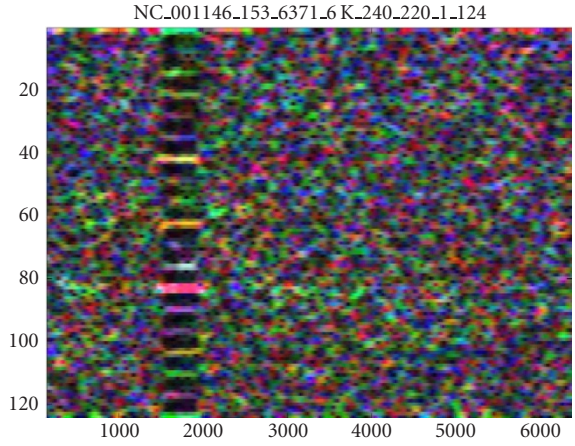

Figure 20: Spectrogram of the YRF1-6 gene. The bar region corresponds to a highly conserved domain in Y′-helicase subtelomeric open reading frames.

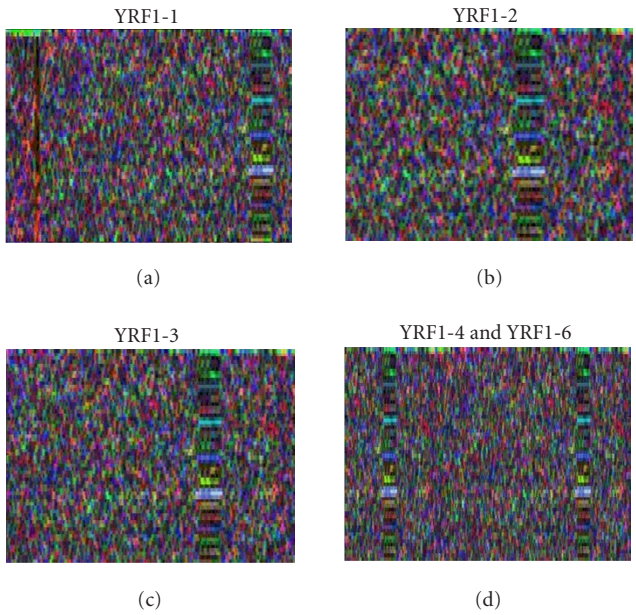

Figure 21: Four spectrograms showing similarity between YRF1 genes 1, 2, 3, 4, and 5. The genes have very similar spectrograms.

dependent on the length of the repeated sequences which are thought to act as spacers to expose a reacting domain at the cell surface. The flocculin repeats that endow the protein with a crucial part of its function are directly visible in the color spectrogram.

Other cell wall proteins whose DNA sequences show quilts are FIT1 (cell wall iron transport) and DAN4 (cell wall mannoprotein). The human MUC2 protein encoded in chromosome 11 of the human genome also shows a large quilt spanning several thousands of nucleotides. This protein is found to have a high BLAST alignment score with FLO1. It is a secreted surface protein that coats the epithelia of intestines, airways, and other mucus membrane containing organs. A common feature is that these proteins have their localization in and around the cellular membrane. Thus, it is possible that the domains represented by quilts cause their proteins to have particular conformations and/or binding sites that function along the cell surface or lead to cell surface localization.

3.2.2. Bars—the Y′-helicases

A large number of bars were found in all genomes, including the yeast genome. We found bars corresponding to protein domains of low complexity tandem repeat units. These repeat units are much simpler, compared to quilts or minisatellites.

The yeast Y′-element is a highly polymorphic repetitive sequence present in the subtelomeric regions of many yeast telomeres [19]. It has been reported that survivors arising from yeast mutants deficient in telomerase compensate for telomere loss by the amplification of Y′-elements. Many of the sequences were found to contain long open-reading frames that potentially encode helicase. Thus, the repetitive patterns in these genes might have a dual role to play. They could function similar to telomeric repeats in extending the life of a cell line. They could also function as important protein domains that are responsible for the helicase function. The Y′-elements contain some highly conserved domains of repeats. One which such domain identified as Pfam-B_59 in the PFAM database shows a unique bar (Figure 20) compared to the other Y′-elements. The helicases that showed bars are

Chromosome 4: YRF1-1/YDR543W
(Bar: 1530000–1530500 bp)
Chromosome 5: YRF1-2/YER190W
(Bar: 574900–575400 bp)
Chromosome 7: YRF1-3/YGR296W
(Bar: 1089000–1089400 bp)
Chromosome 12: YRF1-4/YLR466W
(Bar: 1069500–1070000 bp)
Chromosome 12: YRF1-5/YLR467W
(Bar: 1076250–1076750 bp)
Chromosome 14: YRF1-6/YNL339C
(Bar: 1600–2000 bp)
Chromosome 16: YRF1-7/YPL283C
(Bar: 1500–2000 bp).

Figure 21 shows helicases YRF1-1, YRF1-2, YRF1-3, YRF1-4, and YRF1-5. Part of the conserved domain is seen as a bar.

A large number of other subtelomeric genes show exactly the same bars with the same frequency and color characteristics. The genes are annotated as hypothetical ORFs with unknown functions. The proteins produced from these genes are found to have the same conserved Pfam-B_59 domain. The bar patterns found through spectrogram visualization support the hypothesis that the ORFs have similar functions to the Y′-elements.

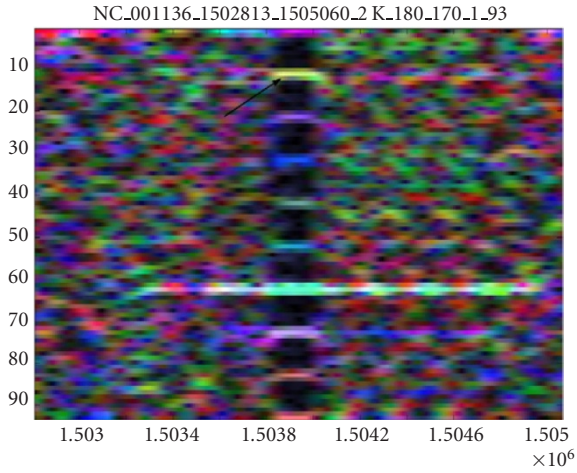

Figure 22: Spectrogram showing shaft in FIT1 gene. The arrow highlights period 18, showing an intensity corresponding to a repeat of 6 amino acids.

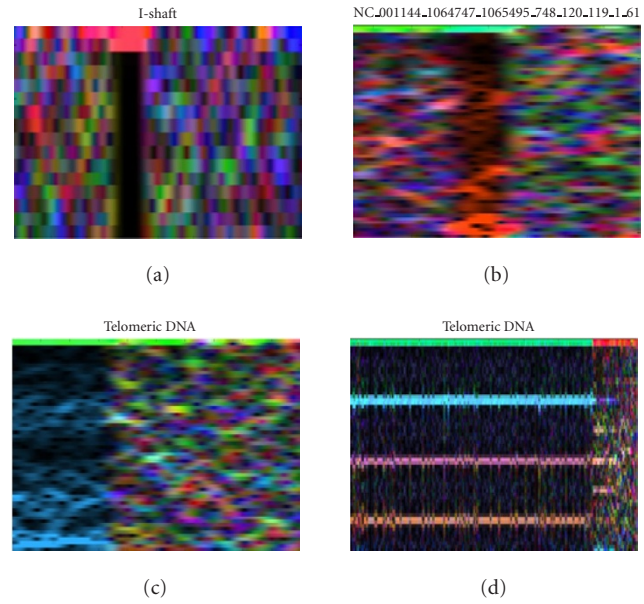

Figure 23: Four spectrograms showing very simple regions. (a) and (b) correspond to simple (1 and 3 bp) repeats in intergenic regions, while (c) and (d) show subtelomeric DNA found at the end of chromosomes.

A number of yeast cell wall glycoproteins such as PIR1, PIR3, HSP150, and TIR1 are characterized by the presence of tandem repeats of a region of 18 to 19 residues. The core region is highly conserved and has a consensus pattern of "SQ [IV] [STGNH] DSQ [LIV] Q [AIV] [STA]." The genomic DNA sequences of these proteins show prominent and characteristic bars whose frequency pattern represents the dominant periodicities. These bars are visually distinct in color and frequency pattern from the Y′-elements.

Some bars show the structural significance of protein in the cell. In yeast, the protein HKR1 coded on chromosome 4 is a cell surface protein that may regulate cell wall beta-glucan synthesis. A region of the gene shows strong bars at a number of relevant frequencies reflecting corresponding periodicities in the protein as well as the DNA sequences. The domain in the protein sequence is made up of 12 repeats of a 28 amino acid sequence, namely, "S [AV] [P] VAVSSTYTSSPSAPAAISSTYTSSP." It was predicted to have a beta-helix supersecondary structure with a high score by the Betawrap algorithm. The gene YIL169C in *S. cerevisiae* shows strong bars that correspond to a serine-rich domain in the protein. This domain extends through amino acids 92–154 and is identified as a potential T-SNARE coiled-coil domain.

3.2.3. Shafts and their structural significance

The shaft shown in Figure 22 is part of the FIT1 gene. It corresponds to a domain of repeats of 6 amino acids, namely, "SSAVET." The shaft shows a bright band at frequency 11, marked by an arrow. The remaining bars are all harmonics of this fundamental periodicity. As the DFT window size was 180 for this spectrogram, a frequency of 11 corresponds to a periodicity of 18 in the DNA sequence and a periodicity of 6 in the protein sequence. This protein domain is predicted with high probability as a large alpha helix by GCG-Pepstruct. Spectrogram analysis of genes CYC8 and GAL11 also show shafts with a prominent periodicity of 6 nucleotides. This translates to a periodicity of 2 amino acids in the protein. Rightly so, they represent QA repeats that form large alpha helices in both proteins.

Many shafts also represent low complexity, high flexibility regions made of GOR turns in the respective proteins. Gene YLR114C has a DENN (differentially expressed in neoplastic versus normal cells) domain. Part of this domain is a high flexibility region of *D* repeats. This region corresponds to a shaft.

Finally, found in the yeast genome were the simplest patterns possible. Some examples of very simple patterns are shown in Figure 23. Very simple repeats of a single to a few nucleotides create simple spectrograms with bright and dark regions. The simplest pattern possible is a dark vertical bar corresponding to a constant nucleotide sequence (e.g., TTTTTTT . . .). These patterns may correspond to subtelomeric DNA or to simple structures in protein-coding regions. Very simple patterns are useful because they serve as visual markers when navigating the genome.

3.2.4. An unannotated pattern

We observed a bar (with many strong periodicities) and a shaft in the region of 12500–13000 nucleotides of *S. cerevisiae* chromosome 1 (Figure 24). Except for this one pattern, every occurrence of quilts, bars, and shafts in the yeast genome was found to correspond to a gene. This region also shows a dominant 3 bp periodicity (the codon frequency). It is sandwiched between 2 genes (12047-12427 and 13364-13744). We found this region to be unannotated in the GenBank and other major databases. Based on these observations, we believe that the region might correspond to a missed gene or pseudogene.

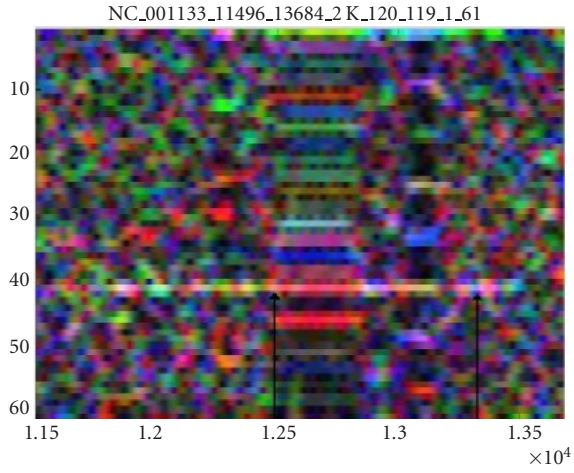

Figure 24: Spectrogram showing an unannotated pattern believed to correspond to a gene or pseudogene. The left arrow marks the end of a predicted gene. The right arrow marks the beginning of another predicted gene.

4. DISCUSSION AND CONCLUSIONS

We employed the short time Fourier transform (STFT) to create color spectrograms of the genomes of various organisms after developing a software tool allowing for easy visual navigation of the genomes via the spectrogram. Spectrograms were created for many different organisms of varying complexity, and we believe that the method can effectively identify any unusual patterns in any genome. Various structures and periodicities were found along all lengths of the chromosome, from a single gene to an entire chromosome. Important periodicities ranged from 0 to 300. We learned that there were no complex patterns in the phage genome and the number of complex patterns increased in frequency with the complexity of the organism. The higher organisms also showed more complex patterns per gene.

Periodicities from 0 to 300 were located and highlighted. We found periodicities relevant to the structure of DNA as well as periodicities involved in protein coding. Periodicities relevant to DNA structure included those concerning telomere structure, protein coding in DNA, DNA helical folding, DNA nucleosome binding, and DNA nucleosome superstructure. One of the characteristics of spectrogram color was that it correlated to Giemsa staining in human chromosomes, thus providing visual information regarding relative nucleotide content, including GC content. Minisatellites were easily visualized as well as the complexity of their constituent repeat pattern.

Patterns of quilts, bars, and shafts were also found on the sequence scale of individual genes. Although bars and shafts were restricted to protein-coding regions in the yeast genome, the same was not true for the higher organisms. In *C. elegans* and humans, some patterns extended into the introns of genes and many were also present in intergenic regions. Patterns were useful in associating homology between various proteins. They were also found to have biological significance, particularly in describing the structure of cell surface proteins. Many classes of cell surface proteins are known and within these classes, there also exist many variants. Cell surface proteins are involved in pathology, pharmacology, and cell signaling. Spectrogram analysis seems particularly well suited for the analysis of this important class of proteins.

A significant challenge in bioinformatics is finding sensible ways to manage the quantity and complexity of information in the genome. Spectrogram analysis of genomes exposes both sequence and frequency information on many scales of magnitude and therefore provides an almost unique visualization of DNA on any magnitude scale. We believe that, based on visual similarity of pattern type such as prominent periodicities and color, this method of frequency analysis is useful as a visualization tool. We found the tool to be particularly useful when used along with public databases and genome browsers. Spectrogram visualization gives a region of DNA a unique visual signature that is useful in quickly recognizing an area of interest. Though spectrograms are much more dynamic, they provide a road map similar to cytological maps used with the fruit fly. Further, this unique visual signature can also be used as a heuristic method of classifying domains in DNA protein-coding regions. Finally, the spectrogram gives insight regarding the physical structure of DNA in which a sequence of interest is embedded. Thus, DNA color spectrograms place sequences of interest in a much-needed larger context.

In summary, we used DNA color spectrograms to find biologically relevant patterns in the genomes of various organisms, some of which relate to DNA structure or protein coding. Similar patterns in different parts of various genomes were found to have similar functions. Various patterns included strong genome-wide periodicities and structures such as microsatellites, minisatellites, quilts, bars, and shafts. We believe that spectrogram analysis will be a useful tool in understanding the DNA structure, identifying protein domains, and predicting function and structure, as well as a discovery tool for novel DNA regions of potential biological significance.

ACKNOWLEDGMENT

Appreciation is expressed to Rick Thompson who introduced the terms “quilts,” “shafts,” and “bars,” and to Chris Fidyk who wrote the original software, implementing spectrogram development.

REFERENCES

[1] D. Anastassiou, “Frequency-domain analysis of biomolecular sequences,” *Bioinformatics*, vol. 16, no. 12, pp. 1073–1081, 2000.

[2] D. Anastassiou, “Genomic signal processing,” *IEEE Signal Processing Magazine*, vol. 18, no. 4, pp. 8–20, 2001.

[3] J. C. Shepherd, “Periodic correlations in DNA sequences and evidence suggesting their evolutionary origin in a comma-less genetic code,” *J. Mol. Evol.*, vol. 17, no. 2, pp. 94–102, 1981.

[4] J. C. Shepherd, “Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and

its possible evolutionary justification," *Proc. Natl. Acad. Sci. USA*, vol. 78, no. 3, pp. 1596–1600, 1981.
[5] J. C. Shepherd, "From primeval message to present-day gene," *Cold Spring Harb. Symp. Quant. Biol.*, vol. 47, Pt 2, pp. 1099–1108, 1983.
[6] J. W. Fickett, "Recognition of protein coding regions in DNA sequences," *Nucl. Acids. Res.*, vol. 10, pp. 5303–5318, 1982.
[7] D. Rhodes and A. Klug, "Helical periodicity of DNA determined by enzyme digestion," *Nature (London)*, vol. 286, pp. 573–578, August 1980.
[8] G. P. Lomonossoff, P. J. Butler, and A. Klug, "Sequence-dependent variation in the conformation of DNA," *J. Mol. Biol.*, vol. 149, pp. 745–760, July 1981.
[9] A. Klug, L. C. Lutter, and D. Rhodes, "Helical periodicity of DNA on and off the nucleosome as probed by nucleases," *Cold Spring Harb. Symp. Quant. Biol.*, vol. 47, pp. 285–292, 1983.
[10] L. J. Peck and J. C. Wang, "Sequence dependence of the helical repeat of DNA in solution," *Nature*, vol. 292, pp. 375–378, July 1981.
[11] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, Garland Publishing, New York, USA, 4th edition, Chapter 4, 2002.
[12] Y. Niimura and T. Gogobori, "*In silico* chromosome staining: Reconstruction of Giemsa bands from the whole human genome sequence," *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 2, pp. 797–802, 2002.
[13] A. Bird, "CpG islands as gene markers in the vertebrate nucleus," *Trends in Genetics*, vol. 3, pp. 342–347, 1987.
[14] S. Subramanian, V. M. Madgula, R. George, et al., "Triplet repeats in human genome: distribution and their association with genes and other genomic regions," *Bioinformatics*, vol. 19, no. 5, pp. 549–552, 2003.
[15] S. Subramanian, R. K. Mishra, and L. Singh, "Genome-wide analysis of Bkm sequences (GATA repeats): predominant association with sex chromosomes and potential role in chromatin organization and function," *Bioinformatics*, vol. 19, no. 6, pp. 681–685, 2003.
[16] K. B. Murray, D. Gorse, and J. Thornton, "Wavelet transforms for the characterization and detection of repeating motifs," *J. Mol. Biol.*, vol. 316, no. 2, pp. 341–363, 2002.
[17] Y. Nakamura, M. Leppert, P. O'Connell, et al., "Variable number of tandem repeat (VNTR) markers for human gene mapping science," *Science*, vol. 235, no. 4796, pp. 1616–1622, 1987.
[18] M. Bony, D. Thines-Sempoux, P. Barre, and B. Blondin, "Localization and cell surface anchoring of the Saccharomyces cerevisiae flocculation protein Flo1p," *Journal of Bacteriology*, vol. 179, no. 15, pp. 4929–4936, 1997.
[19] M. Yamada, N. Hayatsu, A. Matsuura, and F. Ishikawa, "Y′-Help1, a DNA helicase encoded by the yeast subtelomeric Y′ element, is induced in survivors defective for telomerase," *J. Biol. Chem.*, vol. 273, no. 50, pp. 33360–33366, 1998.

David Sussillo received his B.S. degree in computer science from Carnegie Mellon University in 1999 and his M.S. degree in electrical engineering from Columbia University in 2003. He is currently pursuing his Ph.D. degree in the Doctoral Program for Neurobiology and Behavior at Columbia University. His current research interests include signal processing of genomic signals, vision processing in the primary visual cortex, and computer applications in biomedical research.

Anshul Kundaje received his B.S. degree from Veermata Jijabai Technological Institute (VJTI), the University of Mumbai in 2001 and M.S. degree from Columbia University in 2002, both in electrical engineering. Presently, he is pursuing a Ph.D. degree in computer science at Columbia University. His research focus is computational biology, specifically applying machine learning and signal processing techniques to solving hard biological problems. His prime interest is in reverse engineering of genetic and protein networks using multiple sources of biological data such as mRNA expression, time-series, sequence, and protein data.

Dimitris Anastassiou is a Professor and Director of Columbia's Genomic Information Systems Laboratory at Columbia University. He received the Ph.D. degree in electrical engineering from the University of California, Berkeley, in 1979. From 1979 to 1983, he was a research staff member at the IBM Thomas J. Watson Research Center, Yorktown Heights, NY. Since 1983, he has been with the Department of Electrical Engineering, Columbia University. He is an IEEE Fellow, the recipient of an IBM Outstanding Innovation Award, a National Science Foundation Presidential Young Investigator Award, and a Columbia University Great Teacher Award. His previous research interests have been in the area of digital signal processing and information theory with emphasis on the digital representation of multimedia signals, with contributions to the international digital television coding standard, MPEG-2. He is the founder and previous Director of Columbia University's Image and Advanced Television Laboratory. His research is now exclusively focused on applying his expertise in engineering to the emerging field of computational biology.