DNA vectorgrams: Representation of cancer genes as movements on a 2D cellular lattice

by Clifford A. Pickover

A brief introduction to a computer graphics characterization of cancer DNA sequences, as well as other biologically interesting sequences, is presented. The procedure described takes DNA sequences containing *n* bases and computes *n* two-dimensional real vectors. When displayed on a planar unit-cellular lattice, these characteristic patterns appear as a "DNA vectorgram," *C*(*n*). Several demonstration plots are provided which indicate that *C*(*n*) is sensitive to certain statistical properties of the sequence of bases and allows the human observer to visually detect some important sequence structural properties and patterns not easily captured by traditional methods. The system presented has as its primary focus the fast characterization of the progression of sequence data using an interactive graphics system with several controlling parameters.

Introduction

DNA contains the basic genetic information of all living cells. The sequences of bases of DNA (adenine, cytosine, guanine, and thymine—A, C, G, and T) may hold information concerning protein synthesis as well as a variety of regulatory signals. For example, specific A/T-rich regions are thought to be codes for beginning transcription. In addition, certain *specific* viral sequences elicit cancerous changes in cells in artificial media and in animals. Although the genes implicated in the development of cancer were first observed in work with viruses, many of these genes have now been found to be part of the normal cell's genome as well [1].

In addition to containing such regulatory codes and tumor-promoting codes [1], DNA base sequence and composition are often correlated with physical properties of the DNA. For example, the melting temperature is related to the mole fraction of triple-bonded G/C in the DNA, and the melting transition of synthetic DNAs with regularly alternating sequences is quite sharp [2]. An interesting and common feature of eucaryotic DNA is the presence of tandem as well as interspersed base-sequence repeats throughout the genome (for references, see [3]). These repeating units range in size from dinucleotide repeats to longer interspersed sequences, for example, large sequences

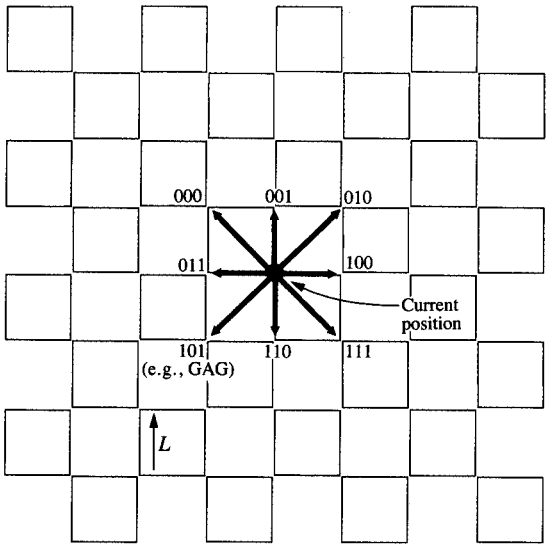


Figure 1

The mapping of the digit strings onto characteristic two-dimensional patterns traced out on a cellular lattice of cell length L.

known as "ALU sequences" found in higher organisms [3]. Finally, processes of DNA rearrangement and recombination and a variety of topological and conformational changes are all affected by the specific sequence of bases in DNA [4].

Fairly detailed comparisons between DNA sequences are useful and can be achieved by a variety of brute-force statistical computations [2, 5], but sometimes at a cost of the loss of an intuitive feeling for the structures. Differences between sequences may obscure the similarities. Even determining whether a particular sequence is *random* is curiously difficult. The best that can be done is to specify certain tests for types of randomness and then to call a sequence random to the degree that it passes them. For example, for DNA one can insist that each base occur with frequency 1/4. Of course, this does not test for the spatial progression of the bases—and permutations of bases taken two at a time, three at a time, $\cdots$, n at a time must also be checked. The importance of "randomness" in studying sequence data (and in understanding implications for evolution) is discussed in [6, 7]. The approach described in this paper provides a method for simply representing and comparing random and DNA sequences in such a way that several sequence features may be detected by the analyst's eye.

Among the methods available for biomolecule characterization (for both protein and nucleic-acid sequences), computer graphics is emerging as an important tool [8]. Since the characterization of DNA base content, periodicity and both long-range and nearest-neighbor sequence data are currently active areas of research [3, 9]. I introduce a computer graphics characterization of nucleic-acid sequences which is sensitive to the patterns in the progression of the bases. This method involves the conversion of the DNA sequence to binary data and subsequent mapping of the data to a two-dimensional pattern on a cellular lattice. I have previously discussed mapping of genetic information to a binary waveform—with analyses analogous to those used in electronic signal processing [9].

Motivation and method

• *Lattice transformation*

For the examples in this paper, triple-bonded bases (G/C) are differentiated from double-bonded bases (A/T) by assigning nucleotide input values as follows: G = 1, C = 1, A = 0, T = 0. Since the sequences generated by this means are strings of 0s and 1s, the human observer may find difficulty in distinguishing different sequences. A technique which has proved useful in overcoming this drawback involves the transformation of the digit strings into characteristic two-dimensional patterns traced out on a unit-cellular lattice. This approach was invented by D. H. Green, who applied it to shift registers of digital computers [10], and the simple conversion pattern I use follows that of Green, as shown in **Figure 1**. Three digits at a time are inspected and assigned a direction of movement over a cellular lattice. Therefore, each of the three-digit combinations causes a vector to be drawn from a point on the lattice to one of the eight points immediately adjacent, in accord with the coding system shown. This procedure is repeated using serial overlapping windows of length three, and therefore a pattern characteristic of the DNA sequence is drawn on the lattice. Three-digit windows are used because the subsequent eight directions are easily represented on a tightly packed 2D unit-cellular lattice (two-digit windows give four directions on a tightly packed lattice but yield patterns that are visually less rich), and because the genetic sequence is often organized in terms of triplets ("codons"). Other mapping schemes, however, can be imagined and yield useful patterns, as discussed in the Conclusions section.

• *Net movement*

When this approach is used, sequences with a predominance of repeating Gs or Cs, for example, show a net movement along the right lower diagonal. In general, sequences with high G/C content show a downward tendency. When the transformation diagrammed in Figure 1 is used—if for each combination of three bases found in the sequence there exists at some other region another combination which is the logical inverse (e.g., G and A interchanged; 010 vs. 101),

then the net movement is zero. Therefore, it is possible for the trace to return to the initial point. **Figure 2** is an example for a repeating sequence. As would be expected, various common short-sequence control elements in the DNA each give signature patterns (for examples of such control codes, such as the Pribnow box and CAT box, see [11]).

• *Random DNA strings*

In order to fully appreciate and utilize the DNA vectorgrams, it is necessary to digress and review the implications of vectorgram application to a random bit string. For a random walk on a plane we can estimate the most probable distance (R) traveled by a particle after N equal steps by

$$R = L\sqrt{N}, \tag{1}$$

where L is the length of each straight track walked [12]. For our lattice, L is not constant, due to the diagonals, and this formula therefore cannot be applied. In order to derive the appropriate equation, one may use either the results of many random-walk experiments on the lattice or probability theory. Dr. S. H. Biyani of IBM East Fishkill, in unpublished work, has produced a derivation based in part on the Central Limit Theorem [13] and on the fact that when the x and y components of a distance have a Gaussian distribution, the distance has a Rayleigh distribution. Both the random-walk experiments and the theoretical method yield

$$R = 1.085L\sqrt{N}, \tag{2}$$

where L is the lattice grid spacing. I will refer to R as the "expected" distance. When large numbers of random test DNAs were entered into the system, this approximation was found to be excellent. I introduce a lattice-persistence parameter ρ, useful for comparing DNA vectorgrams:

$$\rho = D/R, \tag{3}$$

where D is the actual measured distance between the DNA termini on the lattice, and R is the distance expected for a random sequence, given in Equation (2). For random sequences, $\rho \approx 1$.

• *DNA vectorgram plots*

In this graphics system, parameters such as the start and stop base number and the step size L can be entered by the user—thereby allowing magnification of various regions of interest. The starting point for all sequences is placed at the center of the plot, and circles with radius R are superimposed [Equation (2)] to facilitate comparison of plots and to suggest deviation from randomness. Since the sequences are all of different sizes, different scale factors (i.e., step sizes) were necessary to fit the vectorgram on the plot, and these are given in the figure captions. It should be noted that these figures represent snapshots of a temporal process whereby a bright light moves on the vector-graphics screen as it progresses along the sequence. This dynamic aspect helps to orient the viewer and facilitates the detection of

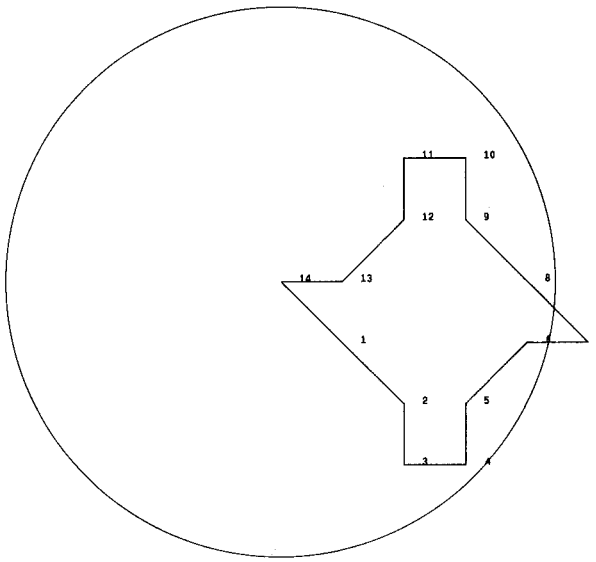


Figure 2

DNA vectorgram computed for the repeating sequence ...GGGGAAGAATACGAGGGGAA....

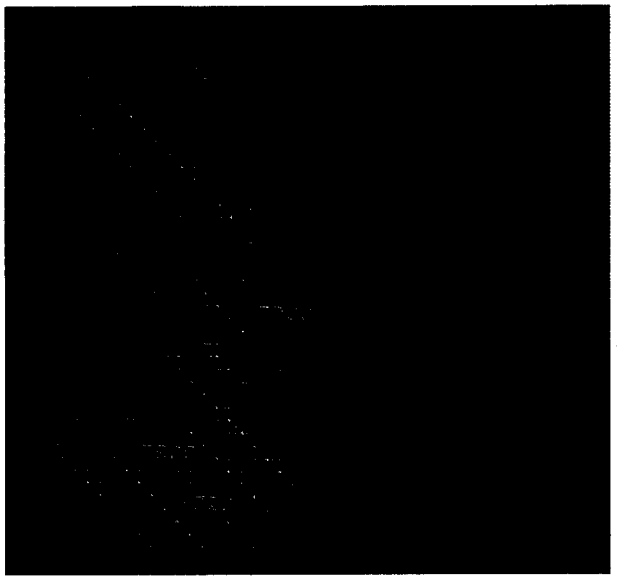
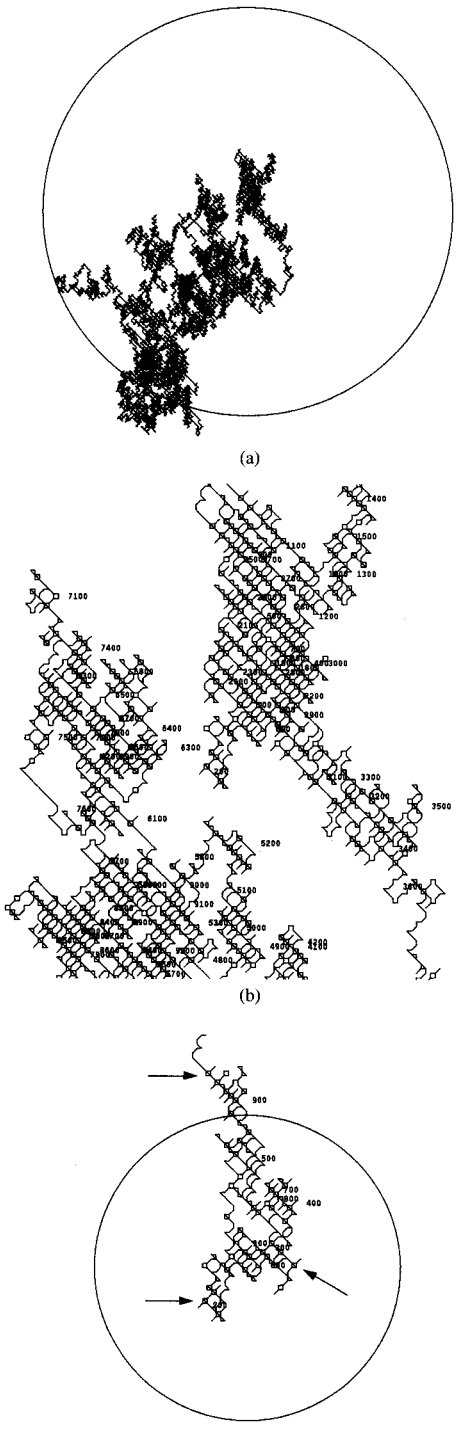


Figure 3

Color DNA vectorgram showing spatial evolution of pattern.



(a) Random nucleotide-base input sequence (number of bases for this plot = 30 000; grid size $L = 0.22$). (b) Magnification of a central region of Figure 4(a). (c) Magnification containing just the first 1000 random bases of Figure 4(a).

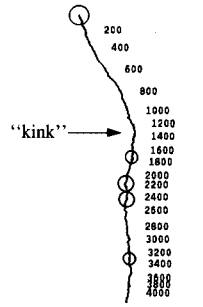
patterns. The user may also request labeling (numbering) of the sequences at specified intervals. The system reports such parameters as D, R, ρ, N, and α. The parameter α is defined to be the angle between the positive x-axis (the horizontal axis through the origin of the vectorgram) and a line drawn between the terminus and the origin. As can be seen, α is different for different DNAs, although it appears to be restricted to regions near 90° and 270° for the DNAs tested. It is also possible to plot a dot matrix representing a sequence (see below). Color options facilitate the localization of features of interest [e.g., distinguishing interons and exons (which are defined in the following section), and portraying the spatial evolution of patterns for overlapping regions of the vectorgram] and are suggested when color capability is available. **Figure 3** is an example random-DNA vectorgram where the color changes every 1000 bases (red, blue, green, magenta, yellow, cyan).

Demonstrations

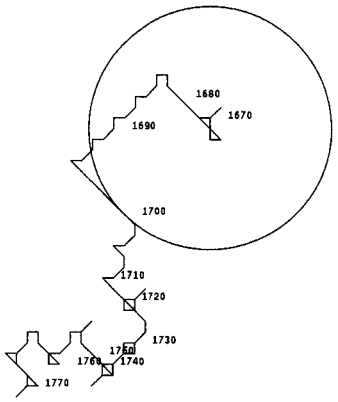
Figure 4(a) shows a DNA vectorgram for a random input sequence. This is useful for comparison with the DNA sequences to follow, which are visually far from random. As expected, its R-vector (shown as a circle with radius R) has roughly the same length as the actual net linear displacement D along the lattice. **Figure 4(b)** shows a magnification of a central region of the plot, and **Figure 4(c)** shows a magnification containing just the first 1000 random bases. Besides the fact that $\rho \approx 1$, note the characteristic closed bisected diamond shapes frequent in all these figures computed from random bit strings. "True" DNA (Figures 5, 7–11) apparently travels too "fast" and does not stay in a lattice region long enough to form these small bisected diamonds.

Figures 5 and 7–11 show DNA vectorgrams for several input sequences. Occasionally, regions of obvious periodicity and other structural features are pointed out to the reader by ellipses in the figures. An example of the output of the graphics system for a large DNA sequence is presented in **Figure 5**. The calculation was performed for a human bladder oncogene [10]. Oncogenes have been detected in tumors representative of each of the major forms of human cancer, and some have been shown to be able to induce malignant transformations in certain cell lines. This bladder carcinoma oncogene is derived from a sequence having similar structure present in the normal human genome. **Figure 6** is a representation of the same 4150 oncogene bases where the binary numbers are represented by dots (thus, e.g., 101011 = . . .). The sequence progresses from left to right and from bottom to top. Predictably, it is much easier to detect trends and patterns in the vectorgrams than by viewing this simple bitmap of base content *en masse*.

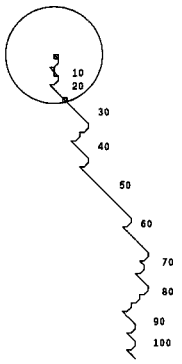
Figure 5(a) is notable for its extreme ρ parameter ($\rho = 23$). The vectorgram, far from being random, travels a mostly downward course indicating strings containing a



(a)



(b)



(c)

Figure 5

(a) Human bladder oncogene (cancer gene) ($\rho = 23$; $L = 0.023$). (b) A magnification of the first oncogene exon (109 bases) shown in Figure 5(a) ($\rho = 2.4$; $L = 1.75$). (c) Magnification of the first 109 (noncoding) bases of the oncogene shown in Figure 5(a) ($\rho = 7.1$; $L = 0.70$).

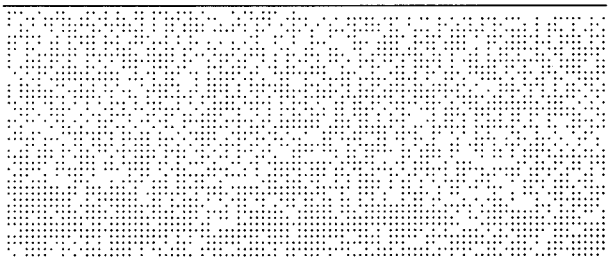


Figure 6

Bitmap for the human bladder cancer gene (1s indicated by dots).

predominance of 1s (011, 101, 111). The most prominent feature on the map is the "kink"—the global shift of the direction in $C(n)$—at about base 1350, and interestingly this feature corresponds to a biologically important area of the DNA sequence. Transcriptional control signals for the synthesis of RNA species are expected to be present in the 5′ noncoding sequence upstream from the initiator codon. [5′ and 3′ (see below) indicate the directionality of the DNA sequence and are derived from numbering of the carbons in the sugar molecules which make up the backbone. They indicate the polarity of the DNA strands.] Reddy [11] searched upstream from the initiator codon for sequences related to the Pribnow box (TATAAA). Two sets were found (1336–1341, 1415–1421), and the kink in the curve occurs precisely at this location separating control signals and "enhancer regions" [11] from the coding groups to follow.

In plants and animals, the vast majority of DNA is never translated to protein. These stretches of silent DNA are called "interons." The coding regions are "exons." In Figure 5(a), the 4150-base DNA sequence may be thought of as starting at the center of the paper and running downward (5′ to 3′ end). Four exons (1670–1779, 2047–2226, 2381–2540, 3238–3354) are encompassed by small circles. Figure 5(b) is a magnification of the first exon (109 bases), and for comparison, Figure 5(c) shows the first 109 bases of the sequence in a noncoding region. Note the different "look" of these two functionally different pieces of DNA.

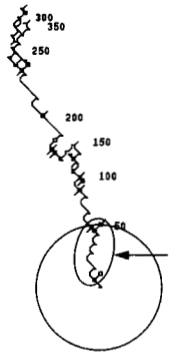


Figure 7

Human somatostatin I gene ($\rho = 2.8; L = 0.90$).

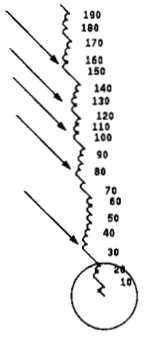


Figure 8

A/T-rich spacer region of *X. laevis* oocyte 5S DNA ($\rho = 7.7; L = 0.35$).

Figure 7 is computed for the human somatostatin I gene (somatostatin is a small neuropeptide and hormone found in the brain and intestine) [14] and has a trend opposite to that of the cancer gene ($\alpha \approx 90°$ instead of $\alpha \approx 270°$). This is due in part to a significant number of repeating A/Ts. The ellipse delimits an obvious repeating motif, and other repeats and long stretches can also be seen in the plot. **Figure 8** is computed for the A/T-rich spacer region of *Xenopus laevis* (African toad) oocyte 5S DNA [15]. The periodic runs of several Ts or As in a row (000), indicated by arrows, can be clearly seen. This sequence has the second greatest persistence length of those sequences tested ($\rho = 7.7$). The Harvey and Kirsten strains of murine sarcoma virus have been isolated from mouse tumors and shown to be capable of inducing sarcomas and leukemias in mice. **Figure 9** represents viral Harvey murine sarcoma DNA, and the vectorgram clearly shows a downward trend. Three structurally different regions are made noticeable by the vectorgram; the two terminal domains are separated by an intermediate segment with several base repeats of (111)-rich regions. The lower domain appears to be more random. **Figure 10** is computed for the Kirsten sarcoma virus; it shows an upward trend with several visually distinct (0)-regions. Other obvious periodicities are denoted by the ellipses. **Figure 11** is computed for an SV40 (simian virus) deletion mutant [16], and it contains an interesting handle (bases 170–200) with three repeating, almost identical hooks corresponding to tandem repeats in the sequence.

Conclusions

As a result of the proliferation of cancer and noncancer sequences in the DNA data base, which has been far greater than ever anticipated [17], it becomes useful to develop tools to help characterize both small- and very large-scale genetic information. In this paper, a procedure is described for taking DNA sequences containing n bases and computing n two-dimensional real vectors. When displayed on a planar unit-cellular lattice, these characteristic patterns appear as a "DNA vectorgram," $C(n)$. $C(n)$'s sensitivity to certain regularities and irregularities in the DNA sequence allows it to function as a pattern-recognition "device"; this permits the human observer to visually detect some important sequence structural properties and patterns not easily captured by traditional methods. An alternate method used to capture sequence periodicities is the power-spectrum approach [3, 9]. However, though this technique can be very illuminating, in many applications it does have certain significant drawbacks. For one, power spectra are phase-insensitive. One nontrivial consequence of this is that very orderly and random data can, in theory, give rise to similar spectra [18]. In contrast, vectorgrams employ a computationally simple and fast algorithm (no Fourier transform is required, as with conventional techniques) and assumptions and complicating factors (window size and windowing effects such as resolution problems and side-lobe distortion) are minimized. In addition, $C(n)$ does not have stringent requirements with regard to the number of input

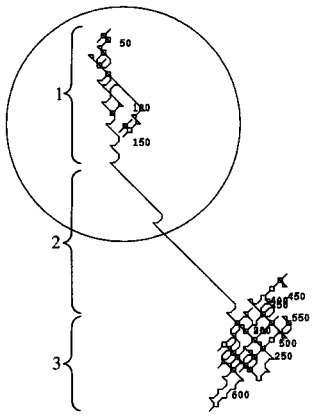


Figure 9

Viral Harvey murine sarcoma DNA ($\rho = 1.9; L = 0.70$).

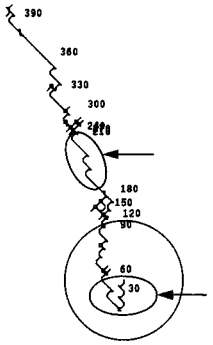


Figure 10

Kirsten sarcoma virus ($\rho = 4.7; L = 0.45$).

points, a constraint which often affects numerical stability, storage requirements, or execution time in traditional algorithms.

In conclusion, the different DNAs tested produce different-looking vectorgrams; some "travel" upward along the lattice, and others downward. As might be expected for DNA, $C(n)$ is in general not random. Randomness can be detected from the "look," the persistence length, and the presence of bisected diamonds in the lattice. The lattice-persistence length ρ is significantly greater than that determined for random DNA, and it can be as great as 22 for the oncogene or as small as 1.9 for the viral Harvey murine sarcoma DNA. As demonstrated by the examples, certain structural features are made evident by the vectorgrams. For example, interspersed repeats are manifested by repeating motifs on the lattice, as shown by the several "hooks" in the SV40 DNA or the periodic runs in the *X. laevis* DNA. Interestingly, α is different for different DNAs, although it appears to be restricted to regions near 90° and 270° for the DNAs tested. Why no DNAs point at 0° or 180° and whether this generalization holds when more DNAs are tested are questions for further studies. One possibility is already under investigation: In a private communication, Dr. Marek Kimmel of the Memorial Sloan-Kettering Cancer Institute has derived mathematical models which suggest that stochastic Markov correlations between adjacent bases in the DNA string (or any string) may account for the preferred upward and downward trends in the vectorgram.

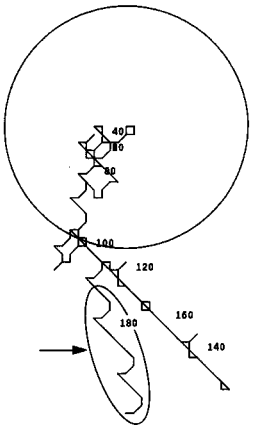


Figure 11

SV40 deletion mutant ($\rho = 2.0; L = 1.3$).

One limitation of the binary system used here is that it does not distinguish between sequences such as ATATATAT ··· and AAAAAAAA ···, which have different folding and

bending properties. Also, while certain important palindromes give rise to characteristic patterns, the binary method can obscure others. Both of these factors motivate the study of additional base assignments (see below).

$C(n)$ is useful in two ways: 1) it provides a qualitative and comparative measure for certain DNA periodicities and patterns, even for segments of DNA having significantly differing base content, and 2) it can be used to search with a single calculation for many pronounced structural features within one sequence. Since $C(n)$ presents nucleic-acid sequence data in a way which can be visually interpreted by the researcher, nucleotide-sequence characterizations are facilitated. The interactive nature of the research station allows for the rapid generation of these functions by using several parameters and magnifications.

This paper should be viewed as introductory because of the wide variety of DNA parameters which can potentially be visualized by this method. The exploration of this large parameter space provides a provocative area for future research. It may be possible to discover interesting properties and periodicities in the DNA sequence by having the program produce many vectorgrams by automatically iterating through a large number of input parameters and mappings. In this way, the program may suggest to the human analyst important features and parameters which would not even be considered otherwise. The correlation of resultant features with biological relevance would be the next necessary area of study.

Finally, other avenues for future research include the testing of other lattices (e.g., hexagonal), other window sizes (2, 4, 5, 6, $\cdots$), other base assignments (e.g., G = 1, C = 2, A = 3, T = 4), other sequences, and three-dimensional lattices. In addition, the extension of the lattice-movement representation to other disciplines is actively being investigated; present research includes applications to "clipped" (binary) speech waveforms and to the syllable patterns of Shakespeare. The vectorgrams may also be useful for distinguishing different classes of noise which have similar traditional spectra. Future studies would also include *variable* base assignment within a triplet window. For example, I have set G = 0 if G is in the first position of the triplet window, but 1 in the other two positions—with resultant interesting and conspicuous vectorgram features. It is hoped that the lattice application introduced here will provide a useful tool for future representations of nucleic acid sequences. Recent proposals to sequence the entire three-billion-base sequence of the human genome [17], advances in understanding viral sequences that induce cancer [1, 11] and in understanding the mechanisms by which oncogenes are activated in tumors [19], recent improved gene-mapping techniques (with accompanying proliferation of sequence data), and increasing commercial interest [20] motivate further assessment of the DNA vectorgram.

Acknowledgments

I would like to thank Dr. B. D. Silverman (IBM Yorktown) for supplying the on-line sequence data for some of the figures. I also thank Dr. S. H. Biyani (IBM East Fishkill) for useful discussions on probability theory, and Dr. Marek Kimmel (Memorial Sloan-Kettering Cancer Institute) for suggesting mathematical models describing vectorgram trends.

References

1. J. Bishop, "Oncogenes," *Sci. Amer.* **246,** 81–92 (March 1982).
2. C. Cantor and P. Schimmel, *Biophysical Chemistry, Part III*, W. H. Freeman and Co., San Francisco, 1980.
3. B. D. Silverman and R. Linsker, "A Measure of DNA Periodicity," *J. Theor. Biol.* **118,** 295–300 (1986).
4. S. Wasserman and N. Cozzarelli, "Biochemical Topology: Application to DNA Recombination and Replication," *Science* **232,** 951–955 (1986); G. Quigley, G. Ughetto, G. Van Der Marel, J. Van Boom, A. Wang, and A. Rich, "Non-Watson-Crick G–C and A–T Base Pairs in a DNA-Antibiotic Complex," *Science* **232,** 1255–1264 (1986).
5. P. Friedland and L. Kedes, "Discovering the Secrets of DNA," *Commun. ACM* **28,** 1164–1186 (1985).
6. "Eukaryotes, Prokaryotes: Who's First?" *Science News* **129,** 280 (1986).
7. R. Lewin, "Computer Genome Is Full of Junk DNA," *Science* **232,** 577–578 (1986).
8. C. Pickover, "Computer-Drawn Faces Characterizing Nucleic Acid Sequences," *J. Molec. Graph.* **2,** 107–110 (1984); C. Pickover, "Spectrographic Representations of Globular Protein Breathing Motions," *Science* **223,** 181 (1984); C. Pickover, "The Use of Random-Dot Displays in the Study of Biomolecular Conformation," *J. Molec. Graph.* **2,** 34 (1984).
9. C. Pickover, "Frequency Representations of DNA Sequences: Application to a Human Bladder Cancer Gene," *J. Molec. Graph.* **2,** 50 (1984).
10. D. H. Green, "Shift-Register Derived Patterns," *Cybernetic Serendipity*, J. Reichardt, Ed., Frederick Prager Publishers, New York, p. 99.
11. E. Reddy, "Nucleotide Sequence Analysis of the T24 Human Bladder Carcinoma Oncogene," *Science* **220,** 1061 (1983).
12. G. Gamow, "The Law of Disorder," *The Mystery of Matter*, L. Young, Ed., Oxford University Press, New York, 1965, pp. 409–427.
13. W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. II, John Wiley & Sons, Inc., New York, 1966, p. 253.
14. L. Shen and W. Rutter, "Sequence of the Human Somatostatin I Gene," *Science* **224,** 168–171 (1984).
15. N. Federoff and D. Brown, "*Xenopus Laevis* Oocyte 5S DNA," *Cell* **13,** 701–706 (1978).
16. C. Benoist and P. Chambon, "*In Vivo* Sequence Requirements of the SV40 Early Promoter Region," *Nature* **290,** 304 (1981).
17. R. Lewin, "Proposal to Sequence the Human Genome Stirs Debate," *Science* **232,** 1598–1599 (1986).
18. M. Martin and C. Pickover, "Short-Term Phase Characterization in Dynamic Signal Analysis," *IBM Tech. Disclosure Bull.* **27,** No. 11, 6769–6771 (April 1985); C. Pickover and A. Khorasani, "Fractal Characterization of Speech Waveform Graphs," *Comput. & Graph.* **10,** 51–61 (1986).
19. A. Wong, J. Ruppert, J. Eggleston, S. Hamilton, S. Baylin, and B. Vogelstein, "Gene Amplification of *c-myc* and *N-myc* in Small Cell Carcinoma of the Lung," *Science* **233,** 461–464 (1986).
20. K. Schneider, "Gene Mapping Is Improved," *The New York Times*, Section D2, Thursday, June 26, 1986.

Received June 3, 1986; accepted for publication August 4, 1986

Clifford A. Pickover *IBM Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598.* Dr. Pickover received a Ph.D. in 1982 from Yale University's Department of Molecular Biophysics and Biochemistry, where he conducted research on X-ray scattering and protein structure. He joined IBM at the T. J. Watson Research Center in October 1982 as a member of the speech synthesis group and later worked on design-automation workstations. Dr. Pickover has published papers in the fields of ecology, biochemistry, biophysics, computer graphics, computers in education. multidimensional data analysis, mathematics, art, speech analysis, and music. He has received an IBM Invention Achievement Award and has been nominated for inclusion in *Who's Who in the Frontiers of Science and Technology.* Currently, Dr. Pickover is a member of the symbolic layout and extraction group and is writing a book entitled *Computers, Pattern, Chaos and Beauty.*