Frequency spectra of DNA sequences: application to a human bladder cancer gene

Clifford A Pickover

Remote Information Access Systems Group, Computer Sciences Department, IBM Thomas J Watson Research Center, PO Box 218, Yorktown Heights, NY 10598, USA

Two useful ways of describing base content and periodicity for nucleic acid sequences are the spectrogram and 3D power spectrum, representations similar to those frequently used in the field of digital signal processing. A description of a vector graphics facility for coordinated computation and display of such functions is presented. The interactive nature of its user interface and the variety of input parameters available to the user greatly facilitate the characterization of a particular nucleic acid sequence. In this paper, calculations are performed for a human bladder oncogene.

Keywords: nucleic acid sequences, base content, periodicity, spectrogram, 3D power spectrum, human bladder oncogene

received 12 December 1983, revised 5 March 1984

Many double-stranded DNA properties are correlated with the DNA nucleotide base composition. In general if P is some observed property of a DNA duplex (such as optical activity, melting temperature, or density), then: $P = f(P_{at}, P_{gc}, \chi_{gc})$, where P_{at} and P_{gc} are intrinsic properties of adenine-thymine and guanine-cytosine pairs, and χ_{gc} is the mole fraction of *G-C* in the DNA[1]. For example, it has been shown empirically that for a given salt concentration and pH, the melting temperature of a double-stranded oligonucleotide. *Tm*, is given by: $Tm(°C) = 69.3 + 41X_{gc}$. The melting temperature *Tm* is sufficiently sensitive to base composition that local fluctuation in X_{gc} will produce local regions with varying *Tm*. Because the coupling of these regions is not infinitely strong, individual regions melt independently from one another. For instance, denaturation diagrams of rDNA show regular patterns of stable (high *G-C* content) spacers followed by unstable regions(low *G-C* content), and each is transcribed into a 40S precursor rRNA[2]. It has been shown that the melting transition of synthetic DNAs with a regularly alternating sequence is quite sharp. For a discussion on nearest-neighbour interactions see Cantor and Schimmel[1].

This report is intended to introduce briefly a computer graphics characterization of nucleic acid sequences which is sensitive to regularities of the pattern of nucleotide bases. This method allows simultaneous analysis of two variables as a function of position in the DNA sequence in a manner analogous to a frequency-amplitude-time graph used in electronic signal processing.

DESCRIPTION OF GRAPHICS SYSTEM

A useful way of describing *G-C* content and periodicity for a particular nucleic acid sequence is the 3D power spectrum and spectrogram (amplitude versus frequency versus position in sequence). In 1947, the search for an effective way to display and examine speech waveforms in natural speech led to the development of the 'sound spectrograph' at Bell Laboratories. This spectrograph, which is in common use today, displays a trivariate representation of speech energy, with abscissal time and ordinal frequency; relative intensity is indicated by darkness on the graph. For topographic power spectra, an alternate trivariate representation, hillocks are indicative of relative intensity. In this work, a research system for the computation of digital spectrograms and topographic spectral distribution functions (power spectra) of nucleic acid sequences was developed (see Pickover[3]) for the use of a similar system to represent protein breathing motions). The user console consists of a vector graphics display (Tectronix 618) and a standard CRT terminal (IBM 3277 GA). Since many displays (eg the Tectronix 618) and hardcopy devices are bilevel, ie produce just two intensity levels, halftoning for the spectrogram was accomplished by the ordered dither technique[4]. The DNA-graphics system also may be run with colour options on an IBM 3279 terminal. The support software is implemented in PL/I[5]. Data entry can be accomplished using voice input if desired. The user of the system need only speak the names of the four bases into the voice recognition system in order for the bases to be entered into the file; such a system is useful for non-typists. The interactive nature of its user interface and the variety of input parameters available to the user (eg scaling, DNA base-value assignment, and window size and overlap) greatly help the characterization of a particular DNA or RNA sequence.

COMPUTATIONAL FRAMEWORK

Traditional time-series analysis techniques are used in the calculations necessary for the graphics displays[6,7,8].

The system requires as input a file containing a listing of all the nucleic acid bases (*G*, *C*, *A*, *T*) of the sequence to be analyzed. The user assigns a numerical value to each of the bases, thereby converting the string of characters to a digitized waveform. The user also specifies a data window (ie, how many contiguous bases will be characterized in each strip of the output display) and a window overlap (how many bases the adjacent windows have in common). If the user wishes to focus on smaller regions of the DNA and is not concerned with low frequency patterns, a small window may be chosen. If the user wants a more global characterization of the patterns within the DNA sequence, a large window is chosen. The overlap parameter makes it easier to capture the spatial dynamics of the features of interest. In both the 3D power spectrum and spectrogram, the mean window value is subtracted, and Hamming cosine tapers[9] are applied to each window of data prior to fast Fourier transformation. Hamming functions taper the data window to zero at each end. For a description of the use of Hamming window in reducing artifacts in power spectral density estimates, see Koopmans[6]. In the 3D plot, a filter is used to smooth the display output data. For a discussion of data smoothing techniques, see Bevington[10].

SIMPLIFIED EXAMPLE FOR HYPOTHETICAL DNA SEQUENCE

Hillocks in the 3D spectrum and darkness on the spectrogram, indicate prominent periodicities in the input sequence; the more common the periodicity, the greater the intensity. The absolute intensity scale of the plot can be altered by a user-specified constant. The following is an example for a simple hypothetical DNA sequence. Assume that the input sequence contains 500 bases with the following pattern *GGGGGAAAAAGGGGGAAAAA* . . . Let the nucleotide translation values be as follows: $G = 1$, $A = -1$. The program would subsequently convert this sequence to a square waveform with a period of 10 base units. As a result, the output spectrum would contain a peak centered on the '10' position on the *x*-axis (period rather than frequency labels are used on the output display to facilitate interpretation). If the window size were 100 and the overlap were 0, the program would automatically walk along the sequence and represent every adjacent 100 bases with a peak at 10. The final 3D plot would contain five strips, and the plot would resemble a ridge. At the other extreme, an entirely random sequence of bases would not give rise to any conspicuous features in the output display, since no base frequency is more common than another. The computation used to create the graphics indicates pattern regularity. A random sequence has no regularity.

EXAMPLE APPLICATION TO CANCER GENE

An example of the output of the graphics system for an actual DNA sequence is presented in Figure 1. The calculation was performed for a human bladder oncogene[11]. 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

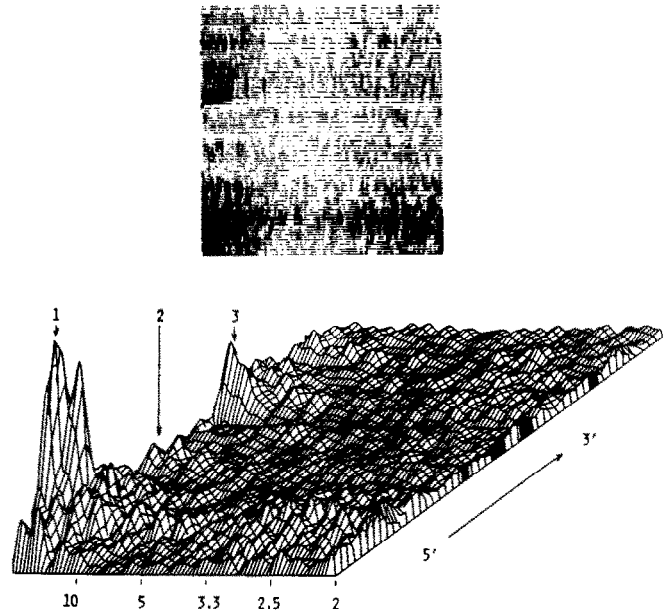

Figure 1. 3D spectral distribution function (amplitude versus frequency versus position in sequence) computed for the DNA sequence of a human bladder cancer gene. The DNA sequence may be thought of as running from front to back (5' to 3' end). Four exons (1670–1779), (2047–2226), (2381–2540), (3238–3354) are indicated by shaded side regions. Frequency of G-C *to* A-T *changes are represented on the x-axis (low frequency on the left). Period values are actually labelled to make interpretation easier. The three peaks (1, 2, 3) indicated by arrows are discussed in the text. The spectrogram (inset), an alternate representation of the same data, portrays position in sequence (ordinate) versus frequency (abscissa). Amplitude is indicated by darkness on the plot*

bladder carcinoma oncogene is derived from a sequence of similar structure present in the normal human genome.

The number of DNA parameters which can be visualized by this method is large. One can envision searching for periodicities in purines vs pyrimidines, charge characteristics of the bases, base-size, and a host of other physical or biologically relevant parameters. Each base may be assigned a different number, and the numbers need not be whole. In the example presented in this paper, triply bonded bases (*GC*) are differentiated from doubly bonded bases (*AT*) by assigning the nucleotide input values as follows: $G = 1$, $C = 1$, $A = -1$, $T = -1$. In order to emphasize regions of high *G-C* or *A-T* content, these values are assigned only if there are several adjacent bases in the same category (in this example five values were required otherwise a value of 0 is sent to the program). In this mapping, the 'DNA waveform' is often flat, with waveform excursions occurring at areas rich in either base-category. The requirement of five equivalent bases is arbitrary and meant only to suggest the variety of parameters that may be experimented with. What is being viewed, in this example, are periodicities in the pattern of the above two categories fluctuating along the sequence. Apart from visualizing periodicity, *GC*

and *AT* content are also a factor in the plot. If there is no major clustering of adjacent bases in the same category, there can be no periodicity. For Figure 1, the window size is 350 bases with an overlap of 275.

In Figure 1, the 4170-base DNA sequence may be thought of as running from front to back (5′ to 3′ end) of the figure. Four exons (1670–1779), (2047–2226), (2381–2540), (3238–3354) are indicated by shaded regions. Frequency of *G-C* to *A-T* changes is represented on the *x*-axis. Several prominent features can be seen on the map, and, interestingly, these features correspond to biologically important areas of the DNA sequence. The largest peak (1), occurring roughly between bases 590 and 900, corresponds with the sequence between two *Xma* III sites. When this sequence is deleted it reduces drastically the transforming activity of the oncogene, indicating the crucial role played by this non-coding sequence[11]. In addition, comparison of the nucleotide sequence of the entire coding region of the oncogene with that of its normal homolog reveals that in the stretch of 1683 bases which constitute the four exons and three introns there are only 2 base changes (at 1704 and 2720), only one of which is in a coding region. Peak 2, in a coding region, and Peak 3, in a non-coding region, correspond to these two spots. Peak 2 appears to be a 'hot spot' for point mutations[11].

Each of the peaks lies predominantly in the low frequency domain of the display, with most of the intensity associated with periods greater than 10 bases. The largest period which may be displayed with the window size used for this example is 375, and it appears from Figure 1 that there is intensity at low frequency patterns of the order of 10–100 bases long.

CONCLUSIONS

Both the spectrogram and 3D power spectrum present nucleic acid sequence data in a way which can be interpreted visually by the molecular geneticist. This facilitates a variety of nucleotide sequence characterizations. The interactive nature of the research station allows for the rapid generation of these functions using a variety of input data scaling and windowing parameters. A report like this can only be viewed as introductory due to the large 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 periodicities in the DNA sequence by having the program produce many DNA maps by automatically iterating through a large number of input parameters. In this way, the program may suggest important features and parameters to the human analyst which would not even be considered otherwise. The correlation of the resulting features with biological relevance would be the next area of study necessary. It is hoped that the spectrographic methods presented here will provide a useful tool for future representations of nucleic acid sequences.

REFERENCES

1 **Cantor, C and Schimmel, P** *Biophys. Chem. Part III* W H Freeman and Company, San Francisco, CA, USA (1980)
2 **Wensink, P and Brown, D** *J. Mol. Biol.* No 60 (1971)
3 **Pickover, C** *Science* No 223 (1984)
4 **Foley, J and Van Dam, A** *Fundamentals of interactive computer graphics* Addison-Wesley, MA, USA (1982)
5 **Hughes, J** *PL/I programming* John Wiley, NY, USA (1973)
6 **Koopmans, L** *The spectral analysis of time series* Academic Press, NY, USA (1974)
7 **Dixon, W** *Biomedical computing programs* University of California Press, CA, USA (1977)
8 **Bendat, J and Piersol, R** *Measurement and analysis of random data* John Wiley, NY, USA (1966)
9 **Oppenheim, A and Schafer, R** *Digital signal processing* Prentice-Hall, NJ, USA (1975)
10 **Bevington, P** *Data reduction and error analysis for the physical sciences* McGraw-Hill, NY, USA (1969)
11 **Reddy, E** *Science* No 220 (1983)