# Computer-drawn faces characterizing nucleic acid sequences

**Clifford A Pickover**

IBM Thomas J Watson Research Center, Yorktown Heights, NY 10598, USA

*A brief introduction to a rather unorthodox computer graphics characterization of DNA sequences is presented. This visual method of data reduction is accomplished by computer-drawn faces which function as multivariate representations sensitive to regularities and irregularities of the statistical properties of the sequence of bases. Various graphical methods of representing multivariate data using icons, or symbols, have been discussed previously. The system presented here is special in that it has as its primary focus the rapid characterization of a multidimensional data series using an interactive graphics system with a variety of controlling parameters.*

*Keywords: nucleic acid sequence, graphical icon generation, displaying multivariate data*

received 11 March 1984, accepted 5 June 1984

Computer graphics has become increasingly useful in the representation and interpretation of multidimensional data with complex relationships. Pseudocolour, animation, 3D figures and a variety of shading schemes are among the techniques used to reveal relations not easily visible from simple correlations based on 2D linear theories.

Various graphical methods of representing multivariate data using icons, or symbols, have been discussed previously[1-4]. In general, $n$ data parameters are each mapped into a figure with $n$ features, each feature varying in size or shape according to the point's coordinate in that dimension. One particularly novel method of representing multivariate data has been presented by Chernoff[1]. The data sample variables are mapped to facial characteristics; thus, each multivariate observation is visualized as a computer-drawn face. Such faces have been shown to be more reliable and more memorable than other tested icons[2] and allow the human analyst to grasp many of the essential regularities and irregularities in the data. This aspect of the graphical point displays capitalizes on the feature integration abilities of the human visual system, particularly at higher levels of cognitive processing[2].

The objective of this research was the development of a vector–graphics system which allows the rapid generation of faces characterizing a dataset which fluctuates through time or space. In this paper, faces are used to represent statistical properties of the sequence of bases in the DNA of a human bladder cancer gene[5,6]. The interactive nature of the system's user interface, and the automatic iteration of the algorithms using a user-determined data window size, greatly facilitate the characterization of a particular data sequence. Though the need for realistic faces is probably not great owing to the ability of humans to caricaturize, an attempt has been made to make a more plausible face than some previous algorithms.

## DESCRIPTION OF THE SYSTEM

The user console consists of a vector graphics display (Tektronix 618) and a standard CRT terminal (IBM 3277 GA)[7]. The support software is implemented in PL/I[8]. In order to use the system for a DNA sequence, a file containing a listing of the nucleic acid bases (G, C, A, T) is required. The user of the system need only speak the names of the four bases into a voice recognition system in order for the bases to be entered into the file. Such a system is especially useful for non-typists. The user specifies a window size (how many contiguous bases will be represented in each face of the output display; eg, 100 bases), and the program then automatically passes through the sequence, drawing one face for each window. The specific features in the DNA sequence which are catalogued by the faces are discussed in the following section. The windows may overlap by a specified number of bases. If the user desires 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 the current applications, ten facial parameters, $F(1,2,3,4,5,6,7,8,9,10)$ are used, and each facial characteristic has ten settings, $S(1,2,3,4,5,6,7,8,9,10)$, providing for 10 billion possible different faces. The controlled features are: head eccentricity, eye eccentricity, pupil size, eyebrow slant, nose size, mouth shape, eye spacing, eye size, mouth length, and degree of mouth opening. The mouth is constructed using parabolic interpolation routines and the other features are derived from circles, lines and ellipses. A middle-setting face $S(5,5,5,5,5,5,5,5,5,5)$ is shown below.

## EXAMPLE APPLICATION TO CANCER GENE

Many double-stranded DNA properties are correlated with the DNA nucleotide base composition[9]. For example, the melting temperature $T_m$ is sufficiently sensitive to base composition that local fluctuation in the base composition will produce local regions with varying $T_m$. Because the coupling of these regions is not infinitely strong, individual regions melt independently from one another. The simplest way to take a specific sequence into account is to assume that all effects are dominated by nearest-neighbour interactions. In this paper, the occurrences of the 10 posssible neighbour pairs (GC, GA, GT, GG, CA, CT, CC, TA, TT and AA) are individually summed within a data window, and the deviation of the sums from the expected (random) value causes deviations of the facial parameters from their middle positions. The number of possible DNA characteristics that can be visualized by this method is very large, and the cataloguing of nucleotide pairs serves only as one illustrative example of the use of computer-drawn faces.

The computer-drawn faces were calculated and displayed for a human bladder oncogene[6]. Oncogenes have been detected in tumours 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 of similar structure present in the normal human genome.

An example of the output of the graphics system is presented in Figure 1. The 4100-base DNA sequence may be thought of as running from face 1 (upper left) to face 100 (lower right). Four exons, or coding regions, (1670–1779), (2047–2226), (2381–2540), (3238–3354) are indicated by the unnumbered enclosing boxes. Several biologically relevant regions are indicated in the figure. Region 1, encompassed by the box with the label '1', roughly corresponds with the sequence between two Xma III sites which, when deleted,

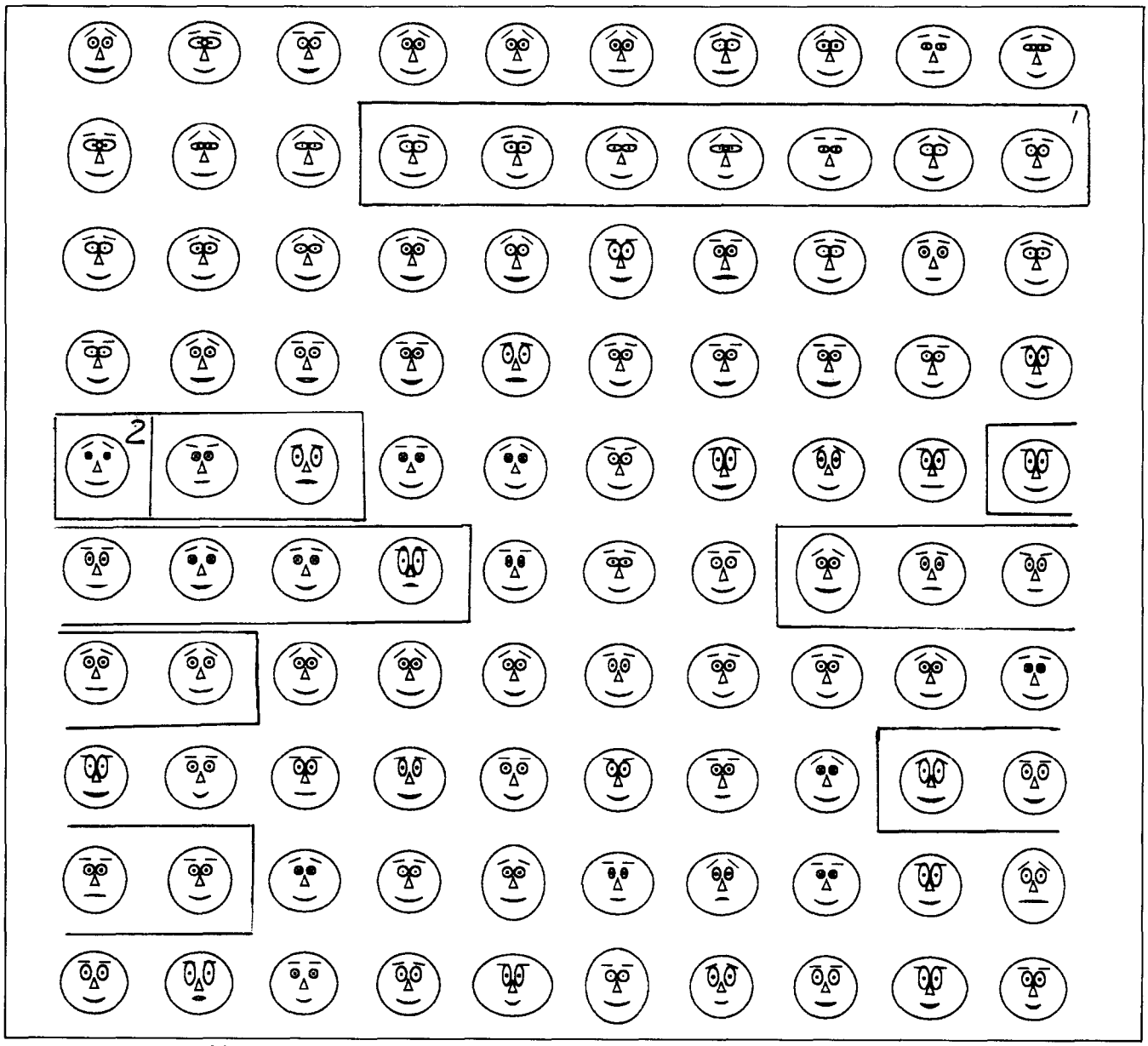


*Figure 1. Sequence of faces representing a cancer gene. The 4100-base DNA sequence may be thought of as running from face 1 (upper left) to face 100 (lower right). Each face represents 41 bases, and the data windows do not overlap. Biologically important regions are indicated by boxes (see text)*

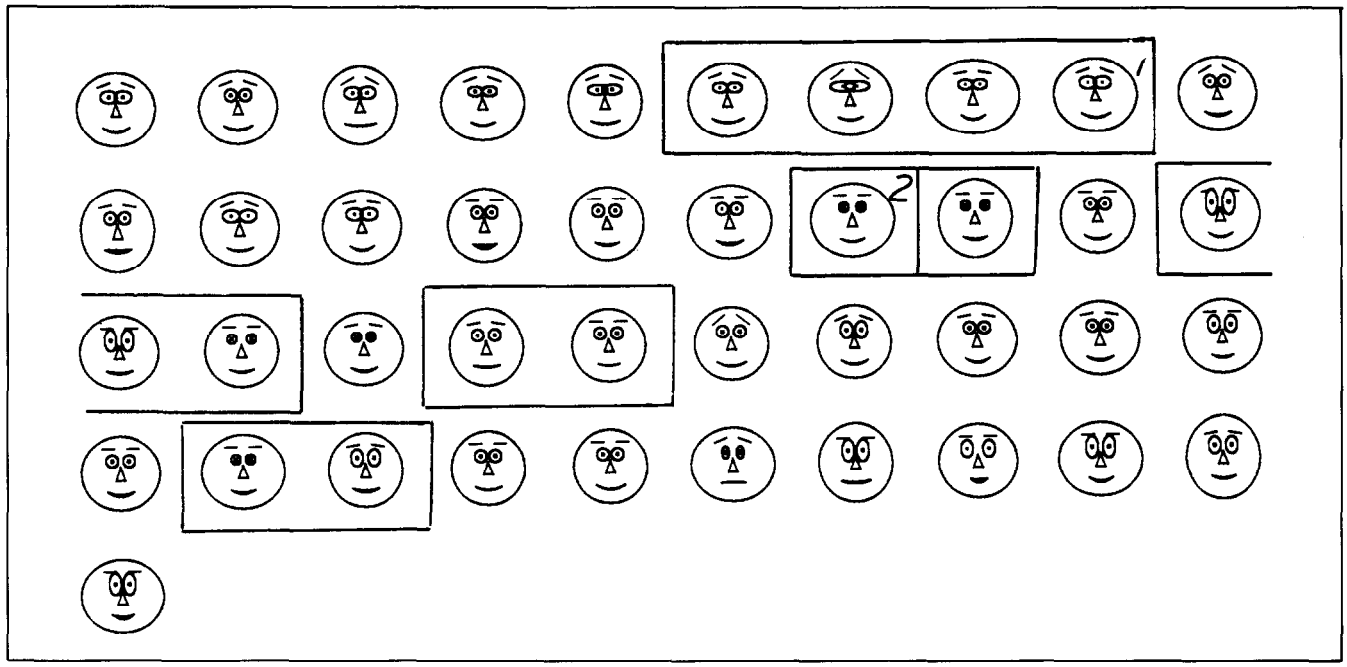

*Figure 2. Sequence of faces representing a cancer gene, using a larger data window than that in Figure 1. The 4100-base DNA sequence may be thought of as running from face 1 (upper left) to face 41 (bottom). Each face represents 100 bases, and the data windows do not overlap. Biologically important regions are indicated by boxes (see text)*

drastically reduces the transforming activity of the oncogene. Thus this non-coding sequence, occurring between bases 590 and 900, plays a crucial role in gene function[6]. Comparison of the nucleotide sequence of the entire coding region of the oncogene with that of its normal homologue reveals only one base change (at 1704) in a coding region. This base is contained in the sequence characterized by Box 2. Box 2 appears to be a 'hot spot' for point mutations[6]. Another representation of the same gene, using a larger data window of 100 bases, is shown in Figure 2. Below is one face representing the entire 4100-base sequence.

## CONCLUSIONS

Graphics are generally limited to a finite number of dimensions, requiring that many multivariate data problems be reduced to fewer dimensions before analysis. It has been shown previously that icons are often useful in allowing a user to detect and comprehend important phenomena and perhaps for communicating major conclusions to others. Various techniques of icon generation have been proposed. What makes the graphics system described in this paper special is its primary focus on the rapid characterization of a dynamic sequence of data using an interactive graphics system with a variety of controlling parameters.

The computer-drawn faces, such as those presented in Figures 1 and 2, allow one to detect substantial changes in the base composition along a gene sequence. A massive amount of data is condensed into a small number of faces. Whether or not noticeable changes occurring in the faces correlate reliably with changes in biological function could provide the basis for some extremely important future research. It is left to the readers to judge for themselves if there appear to be conspicuous facial characteristics in the icons in Box 1 and in the 'hot spot' Box 2, with the feature-mapping used in this paper. In addition, it would be valuable to determine whether the average face for the entire sequence differs substantially from faces calculated for other types of genes. If reliable differences are found it may be possible to use such icons as 'fingerprints' for gene class.

Potential novel mappings of data to the faces can be envisioned. For instance, all of the face parameter settings may contribute somewhat to the calculation of each individual facial-characteristic setting $S_i$, each $S$ having a characteristic weighting factor:

$$S_I = \alpha_1 S_1 + \alpha_2 S_2 + \alpha_3 S_3 + \alpha_4 S_4 + \alpha_5 S_5 + \ldots$$

where $\alpha_1 \gg \alpha_n$.

In summary, the icons presented in this disclosure provide a qualitative awareness of complex DNA sequence trends, and may subsequently guide the researcher in applying more traditional statistical calculations. They point out regularities and irregularities in the DNA sequence which are not easy to observe using conventional techniques. In the bladder cancer gene, one can, at a glance, note changes in characteristics of the DNA sequence by observing the series of computer-drawn faces. The interactive nature of the research station allows for the rapid generation of these functions using several input parameters. A report such as this can only be viewed as introductory owing 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 statistical features of the DNA sequence by having the program produce many DNA faces by automatically iterating through a large number of input parameters. In this way, the program may suggest to the human analyst important features and parameters

which would not even be considered otherwise. The reliable correlation of resultant features with biological relevance would be the next necessary area of study. It is hoped that the multivariate representation presented here will provide a useful tool for future representations of nucleic acid sequences.

## REFERENCES


1 **Chernoff, H** *J. Amer. Stat. Assoc.* Vol 68 (1973) p 361
2 **Wilkinson, L** in *Conference proceedings: human factors in computer systems* Gaithersburg, Maryland, USA (1982) p 202
3 **Pickover, C** *J. Educ. Tech. Syst.* Vol 13 (1985) p 185
4 **Pickover, C** *Computers & Graph* (1985) to be published
5 **Pickover, C** 'Frequency spectra of DNA sequences: application to a human bladder cancer gene' *J. Mol. Graph.* Vol 2 No 2 (June 1984) pp 50–52
6 **Reddy, E** *Science* Vol 220 (1983) p 1061
7 **Pickover, C** *Science* Vol 223 (1984) p 181
8 **Hughes, J** *PL/I Programming* John Wiley and Sons, New York (1975)
9 **Cantor, C and Schimmel, P** *Biophysical chemistry, part III* W H Freeman and Company, San Francisco, USA (1980) pp 1150–1165


**CORRIGENDUM**

In the September 1984 issue, the colour plates for the paper 'VENUS — a program to display protein structure using raster colour graphics' by Y Iga and N Yasuoka (page CP2) were incorrectly positioned. The left-hand picture should have been placed above the caption headed Colour plate 2, the centre picture above the caption headed Colour plate 3 and the right hand picture above the caption headed Colour plate 1.