Simpler DNA sequence representations

SIR—The method of DNA visualization presented by Hamori[1,2] admits of a simpler representation which may be more generally useful. In Hamori's technique, each of the four nucleotides is represented by a vector in three-dimensional space having a characteristic, but variable, orientation. DNA structure then may be viewed from any chosen perspective in two-dimensional plots on computer graphical monitors. Different aspects of structure are examined by varying the geometric relationships among the four elemental vectors.

I suggest that a particular, canonical choice of vectors has certain advantages, both for exploring structure and for comparing sequences. Figure 1 illustrates an intronic sequence based on this approach, wherein successive guanine (G) residues are represented by unit vectors in the positive x-axis direction, complementary cytosine (C) residues correspond to negative x-axis unit vectors, and thymine (T) and adenine (A) residues correspond to unit vectors in the positive and negative y-axis directions respectively.

With this choice, all sequences can be represented in two dimensions in a unique manner, and differences in sequence structure become immediately apparent. In Fig. 2, which compares two actin genes, the regions of divergence of the exonic sequences are clear.

Furthermore, this method of embedding DNA sequences in a two-dimensional metric space permits the consistent application of several standard concepts based on distance measures. Two are worth mentioning. Plots of cumulative Manhattan distances from a site of origin (for example, the initiation codon) against position number can be useful in pinpointing similarities in structure of sequences coding for the same type of product; this ideally should be done after alignment for maximum sequence homology.

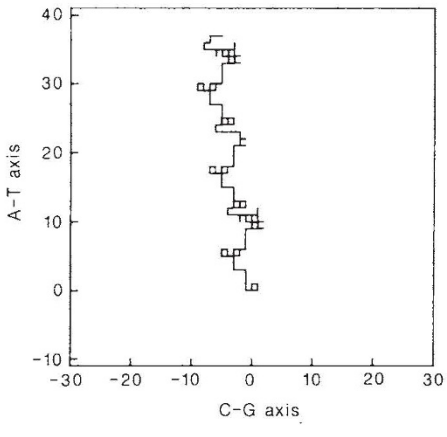

Fig. 1 Graphical representation of the 186-base pair (bp) nucleotide sequence of human BK papovavirus BK enhancer [3], showing three 12-bp repeats. The sequence begins at the bottom, at coordinate position [0,0], and reads (5′ to 3′) towards the top. Successive G, C, T and A nucleotides are depicted as coordinate shifts of one unit in the $+x$, $-x$, $+y$ and $-y$ directions, respectively.

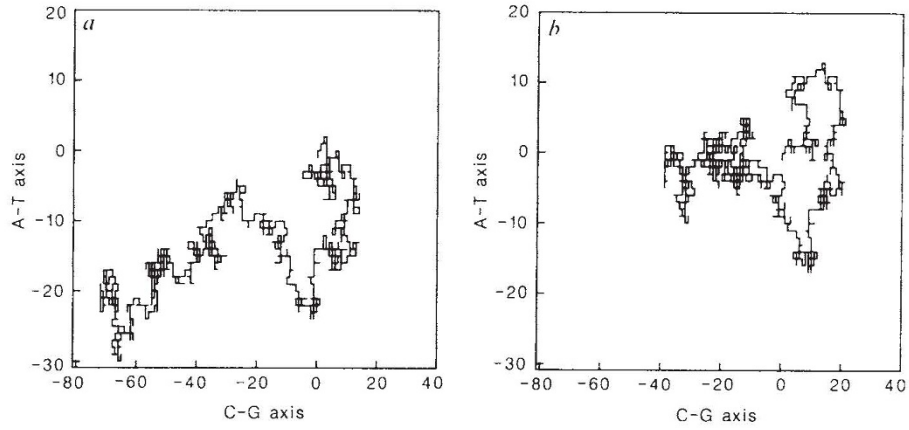

Fig. 2 Nucleotide sequence representations of two actin genes from *Drosophila melanogaster* (Canton S strain). The single intron has been removed from each, leaving a 1,131-bp exonic sequence. Data were obtained from the GenBank library. *a*, Cytological locus 79B; *b*, cytological locus 88F.

Finally, the general properties of classes of sequences (for example, actin genes from different groups of organisms, different types of immunoreceptors) can be examined by calculating the fractal dimension, in this space, of each sequence. This is simply the ratio of the logarithm of the total Manhattan path length to the logarithm of the distance of the end-point of the sequence from its origin: $\log [n(A) + n(G) + n(T) + n(C)] / \log [|n(G) / n(C)| + |n(T) - n(A)|]$, where $n(X)$ is the number of occurences of nucleotide X in the sequence.

M. A. GATES

Department of Zoology,
University of Toronto,
Toronto, Ontario, Canada M5S 1A1

1. Hamori, E. & Ruskin J. *J. biol. Chem.* **258,** 1318-1327 (1983).
2. Hamori, E. *Nature* **314,** 585 (1985).
3. Rosenthal, N. *et al. Science* **222,** 749-755 (1983).

Similarity of vaccinia 28K, v-*erb-B* and EGF receptors

SIR—Vaccinia virus, a member of the Poxviridae, has a large, double-stranded DNA (about 187,000 base pairs) and replicates in the host-cell cytoplasm. The genome contains more than 100 genes, some of which probably code for the enzymes involved in the viral transcription process. Recently, similarities were found between the sequences of the vaccinia 19K protein, EGF, transforming growth factor type 1, and several other mammalian proteins[1-3].

Using the FASTP rapid similarity search[4] and other programs of the Protein Identification Resource system, we have now found unusual sequence similarity (Figs 1,2) between the vaccinia virus 28K proteins a major late protein of unknown function, synthesized after DNA replication and the carboxyl ends of the v-*erb-B* transforming protein of avian erythroblastosis virus[6] and, to a lesser extent, the related human epidermal growth factor (EGF) receptor percursor[7].

When we used the ALIGN program to evaluate these similarities, the 28K protein and a portion (345–584) of the carboxy-terminal domain of the v-*erb-B* protein produced an alignment score of 4.9 s.d. (mutation data scoring matrix with 6 added to each term, penalty of 6 for a break in either sequence, 300 random

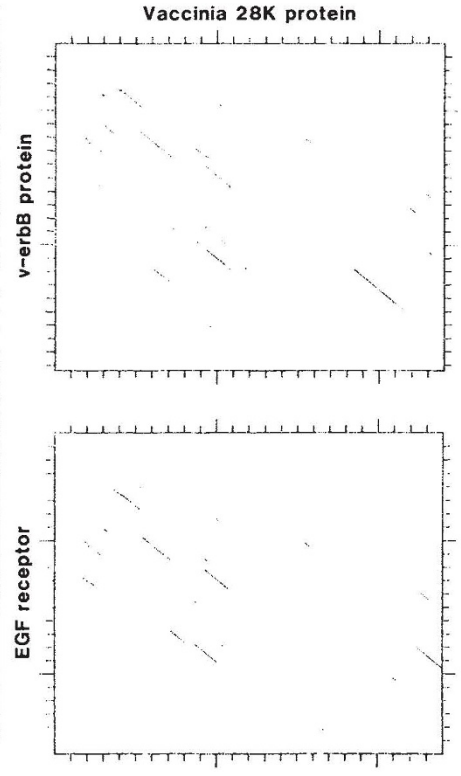

Fig. 1 Graphic matrix plots of the vaccinia 28K protein, the v-*erb-B* protein (350–604), and the EGF receptor precursor (920–1,170). Each point represents a score of 20 or more over a span of 25 amino acids using the mutation data scoring matrix.