Novel DNA sequence representations

SIR — Recent contributions from Lathe and Findlay[1] and Hayashi and Munkata[2] are manifestations of a widespread concern that the conventional (letter-sequence) representations of long DNA sequences are unsuitable for purposes other than elementary information storage. With regard to comprehensibility, such data appear to be analogous to the atomic coordinate tables seen in the specialized journals of X-ray crystallographers. Data in such form cannot be remembered, compared to other similar data, or scanned for certain features without some computer manipulations.

Lathe and Findlay's suggested sequence representation is based on four vertical lines of unequal lengths, each assigned to one of the four nucleotide bases[1]. (This idea appears to be related to the more general computer-based approach used by Daniels *et al.*[3].) For relatively short sequences the method creates attractive mnemonics capable of reflecting certain symmetry features of the sequence. However, for long DNA sequences of several thousand nucleotides this method would be limited by the same problems as those of the letter-sequence method — a compromise between excessive figure size and legibility, lack of indication of global sequence characteristics, and the impossibility of resolution adjustment. The conventional letter-sequence method seems adequate for short DNA where the conveyance of the short-range details of the information is the most important factor. Thus, the suitability of any alternative method must be tested not only on short but also on long DNA sequences.

We have proposed a graphic representation method[4-6] based on the idea of mapping the information content of a DNA sequence into a three-dimensional space curve such that the shape of the curve represents the sequence of nucleotides. A short space vector of characteristic direction is assigned to each of the four nucleotides (Fig. 1) and the curve is assembled from such vectors by joining them head-to-tail in the order of the nucleotides in the DNA sequence. The front and the side views of these 'H-curves' (or their stereoscopically viewed three-dimensional version) can provide not only short-range detail but can also reflect global nucleotide-distribution trends. For instance, DNA regions rich in certain nucleotides will result in H-curve regions shifting in characteristic directions. An important property of these curves is that they can be generated or viewed at different degrees of resolution. At high resolution the individual nucleotides of the sequence are observable but at lower resolution, (when curves are condensed to a relatively short vertical length) the regional variations of nucleotide-distribution patterns along the DNA sequence are emphasized. This latter property of our curves makes them particularly suitable for representing very long DNA sequences. (We recently published the evaluation of the H-curve of the λ genome sequence of 48,502 nucleotides[7].)

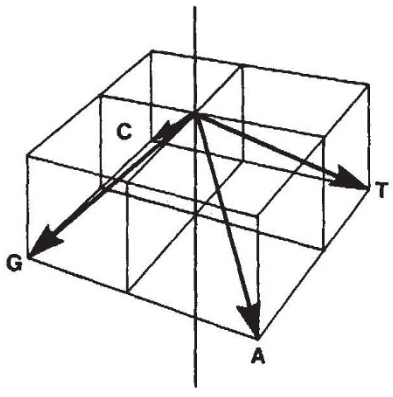

Fig. 1 Perspective illustration of the four small elemental space vectors used to depict the four nucleotides in the H-curve method of DNA sequence representation. The H curve for a DNA sequence is constructed by a head-to-tail assembly of these vectors according the nucleotide sequence of the DNA.

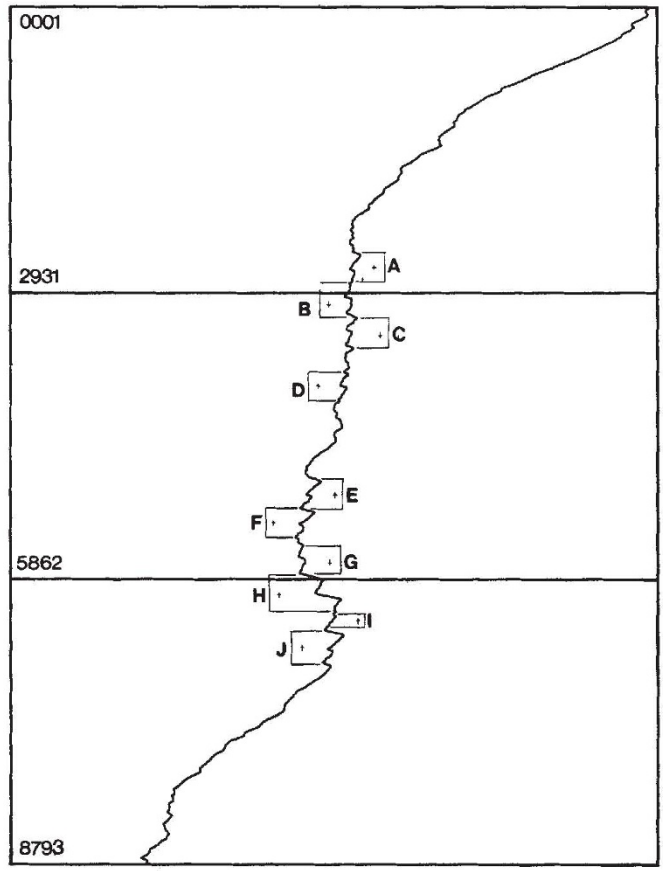

Fig. 2 The H curve of an 8,793-nucleotide-long human DNA sequence (an expressed isotype (5β) of the β-tubulin gene cluster[8]). The projection of the curve shown is the front view with the elemental vectors (here unresolvable) for the G and C nucleotides pointing to the left and those for the A and T nucleotides to the right. The vertical central portion of the curve corresponds to an unusual intron containing a cluster of ten 300-nucleotide-long *Alu* sequences (sections marked with letters). The top of the curve is the 5′ end. The vertical reference lines are 2,931 nucleotides apart and the smoothing factor[5] (*w*) was 10.

The curve shown in Fig. 2 illustrates the utility of our representation method for the display and location of characteristic repeating elements on a lengthy eukaryotic DNA sequence.

EUGENE HAMORI
Department of Biochemistry, Tulane University, School of Medicine, New Orleans, Louisiana 70112, USA

1. Lathe, R. & Findlay, R. *Nature* **311**, 610 (1984).
2. Hayashi, K. & Munakata, N. *Nature* **310**, 96 (1984).
3. Davidson, N. & Szybalski, W. in *The Bacteriophage Lambda* (ed. Hershey, A. D.) 62-70 (Cold Spring Harbor Laboratory, New York, 1971).
4. Hamori, E. *Fedn Proc.* **40**, 1647 (1981)
5. Hamori, E. & Ruskin, J. *J. biol. Chem.* **258**, 1318-1327 (1983).
6. Hamori, E. *Fedn Proc.* **42**, 2263 (1983).
7. Hamori, E. *Gene analyt. Techq.* **42**, 2263 (1983).
8. Lee, M. G.-S., Loomis, C. & Cowan, N. J. *Nucleic Acids Res.* **12**, 5823-5836 (1984).

TATHE AND FINDLAY REPLY — Sequences of the four letters G, A, T(U) and C, are particularly intractable to visual analysis. An ideal format would (1) reveal local features such as recognition sites for enzymes interacting with nucleic acids; (2) reveal large-scale base distribution patterns, and (3) be machine-readable. It is improbable that a single method can meet all three requirements.

The methods proposed by Hamori[1] and ourselves[2] are quite different. We are nevertheless struck by the parallels between the two formats (angle-vector H-curve and line-extension) and the two methods of