observed, the long-range correlation is more prominent. In Voss's paper, the only two categories (phage and bacteria) where the white-noise component dominates contain no introns, consistent with our explanation that long-range correlations are perhaps due to the repetitive patterns that are observed most often in introns.

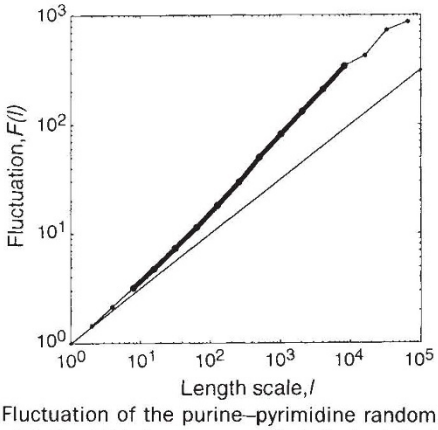
Wentian Li
Cold Spring Harbor Laboratory,
PO Box 100,
Cold Spring Harbor, New York 11724, USA
Kunihiko Kaneko
Department of Pure and Applied Science
University of Tokyo
Komaba, Meguro
Tokyo 153 Japan

SIR — Peng *et al.*[1] reported apparent long-range correlations within various intron-containing gene sequences up to length scales of about 10 kb, but not in intronless genes. Voss[2] later reported such correlations, manifested as spectral density estimates that obey the $1/f^{\beta}$ law, where $\beta \sim 0.8$, in DNA sequences from various organisms. Our analysis of the first completely sequenced eukaryotic chromosome[3] (*Saccharomyces cerevisiae* chromosome III, 315 kb), unlike that of Prabhu and Claverie[4], supports those findings and extends the length scales over which correlations are found. Moreover, we provide a simple statistical test for the significance of such correlations based on the analysis of variance.

The complete yeast chromosome III sequence (EMBL accession number X59720) encompasses roughly 182 putative coding regions, of which only 3% contain introns. It also contains substantial stretches of presumably noncoding DNA in gene-flanking regions. We have not concerned ourselves here with the question of identifying intronless sequences (see ref. 4), but ask whether there are indeed long-range correlations within the entire chromosome.

Following the method of Peng *et al.*, we have plotted the fluctuation $F(l)$, that is, the root mean square (r.m.s.) of the differences between the purine and pyrimidine frequencies in all regions of length l, for $l = 2, 4, 8$ and so on. We find that the fluctuation grows as l^{α}, with $\alpha \sim 0.7$ when the length scale l is greater than 10 (see figure). (Although α here is related to β above, there is no explicit connection between them, and the terminology we use follows that in the references cited herein.) The slope of this plot (α) remains statistically significantly greater than 0.5 (the expectation for an uncorrelated series) up to about $l = 8$ kb. Such higher than expected slopes indicate correlations in the purine/pyrimidine ratio in regions of length l. The longest sequence analysed in ref. 1, the 73-kb human β-globin region, showed a significantly elevated slope only up to about l=1 kb, although the apparent high slope extends further.

The statistical significance can be assessed by analogy to standard statistical analysis of variance. Each point in the figure is obtained by first dividing the series into $2m$ non-overlapping regions of length l and computing the fluctuation, $F(l)$. By joining the neighbouring pairs of such regions, we create m new regions of length $2l$, and compute the fluctuation for these longer regions, $F(2l)$. The purine/pyrimidine differences in the new, larger regions have a fluctuation or variance related to the 'between sum-of-squares' of the analysis of variance, whereas the fluctuation in the original regions is related to the 'total sum of squares'. The statistical significance of any excess variance in the larger



Fluctuation of the purine–pyrimidine random walk in yeast chromosome III plotted against length scale. Intervals with significantly larger than expected slope (P<0.01) are shown with a heavy line. The straight line has the expected slope, α=0.5, for an uncorrelated series.

regions (the main-effect) can be obtained by comparing

$$\frac{(m-1)\,F^2(2l)/2}{(2m-1)\,F^2(l) - (m-1)\,F^2(2l)/2}$$

to the critical value of Fisher's F-distribution with m and m–1 degrees of freedom.

The correlations we observe could result from long regions of shifted base composition. As Nee[5] shows, such regions can result in slopes greater than 0.5 on the fluctuation plot. However, unlike Nee, we believe that the observations of Peng *et al.* do imply a lack of independence of the purine/pyrimidine occurrences, at long ranges. Because the purines and pyrimidines tend to be loosely clustered in the DNA sequence, knowledge that a particular base within a cluster is a purine augments the probability that another, possibly distant base in the same cluster is also a purine. That is, the bases within a cluster or region are not acting independently. Such regions are well known. For example, CpG islands up to ~150 bases in length, or multiple repetitive elements arranged in clusters extending up to about 10 kb are found in other genomes. Thus, some patterns have already been observed on length scales similar to that of our finding. Voss found long-range correlations in the 229 kb cytomegalovirus genome (our test gives significance out to l = 16 kb), but the biology of viral genomes is known to differ markedly from that of nuclear chromosomes (for example, the replication cycle does not require interaction with a centromere for segregation). Our finding of significant correlation in a complete, functional chromosome at scales eight times larger than in the β-globin gene region supports the idea that such correlations are a general property of eukaryotic DNA, and not necessarily limited to isolated examples. Repeating the analysis, counting the number of weakly pairing (A or T) versus strongly pairing (C or G) nucleotides or counting any single nucleotide (A, T, C or G), showed similar results.

It will be interesting to see if statistically significant correlations are observed at still larger scales (100 kb, 1 Mb), when still longer DNA sequences are determined. Although the mechanisms giving rise to these long-range patterns or correlations remain unexplained, such mechanisms are possibly involved in the functioning, regulation and perhaps even of evolution of the genome.

Peter J. Munson
Ronald C. Taylor
George S. Michaels
Division of Computer Research and Technology,
National Institutes of Health,
Bethesda, Maryland 20892, USA

1. Peng, C.-K. *et al. Nature* **356**, 168–170 (1992).
2. Voss, R. F. *Phys. Rev. Lett.* **68**, 3805–3808 (1992).
3. Oliver, S. G. *et al. Nature* **357**, 38–46 (1992).
4. Prabhu, V. V. & Claverie, J.-M. *Nature* **357**, 782 (1992).
5. Nee, S. *Nature* **357**, 450 (1992).

Gas correction

SIR — Perfect gases (see J. Maddox *Nature* **359**, 669; 1992) are modelled as comprising point particles having mass but not volume, billiard-ball or otherwise. If they evinced volume, they would be imperfect, and instead of $pV = nRT$, we should have to replace V by, for example, the van der Waals ($V - nb$), where b represents the sum of the particle volumes in a mole of the gas; or by more elaborate devices, as listed by Maddox.

David R. Rosseinsky
Department of Chemistry,
The University,
Exeter EX4 4QD, UK