6. Sen, D. & Gilbert, W. *Nature* **334,** 364–366 (1988).
7. Jin, R., Breslauer, K. J., Jones, R. A. & Gaffney, B. L. *Science* **250,** 543–546 (1990).
8. Guschlbauer, W., Chantot, J.-F. & Thiele, D. *J. biomolec. Struct. Dyn.* **8,** 491–511 (1990).
9. Boelens, R., Scheek, R. M., Dijkstra, K. & Kaptein, R. *J. magn. Reson.* **62,** 378–386 (1985).
10. Feigon, J., Wang, A. H.-J., van der Marel, G. A., van Boom, J. H. & Rich, A. *Nucleic Acids Res.* **12,** 1243–1263 (1984).
11. Patel, D. J., Kozlowski, S. A., Nordheim, A. & Rich, A. *Proc. natn. Acad. Sci. U.S.A.* **79,** 1413–1417 (1982).
12. Oka, Y. & Thomas, C. A. Jr *Nucleic Acids Res.* **15,** 8877–8898 (1987).
13. Kintanar, A., Klevit, R. E. & Reid, B. R. *Nucleic Acids Res.* **15,** 5845–5862 (1987).
14. Feigon, J. *et al.* in *DNA Structures* (eds Lilley, D. M. J. & Dahlberg, J. E.) (Academic, Orlando, in the press).
15. Sklenář, V. & Bax, A. *J. magn. Reson.* **74,** 469–479 (1987).
16. Kumar, A., Ernst, R. R. & Wüthrich, K. *Biochem. biophys. Res. Commun.* **95,** 1–6 (1980).
17. Marion, D. & Bax, A. *J. magn. Reson.* **80,** 528–533 (1988).
18. Davis, D. G. & Bax, A. *J. Am. chem. Soc.* **107,** 2820–2821 (1985).
19. Wüthrich, K. *NMR of Proteins and Nucleic Acids* (Wiley, New York, 1986).
20. States, D. J., Haberkorn, R. A. & Ruben, D. J. *J. magn. Res.* **48,** 286–292 (1982).
21. Borzo, M. & Laszlo, P. *C. hebd. Séanc. Acad. Sci., Paris* **287,** 475–478 (1978).
22. Pinnavaia, T. J. *et al., J. Am. chem. Soc.* **100,** 3625–3627 (1978).
23. de Leeuw, F. A. A. M. & Altona, C. *J. comput. Chem.* **4,** 428–437 (1983).
24. Brünger, A. T. *X-PLOR (Version 2.1) Manual* 1–291 (Yale Univ., New Haven, CT, 1990).
25. Nilges, M., Habazettl, J., Brünger, A. T. & Holak, T. A. *J. molec. Biol.* **219,** 499–510 (1991).
26. Wang, Y., Jin, R., Gaffney, B., Jones, R. A. & Breslauer, K. J. *Nucl. Acids Res.* **19,** 4619–4622 (1991).
27. Macaya, R. F., Schultze, P. & Feigon, J. *J. Am. chem. Soc.* **114,** 781–783 (1992).

ACKNOWLEDGEMENTS. We thank A. Lipanov for suggesting that we study the Oxy-3.5 sequence, K. Koshlap and S. Malek for synthesizing and purifying the DNA and E. Wang for discussion. This work was supported by the NSF Presidential Young Investigator Award with matching funds from AmGen, DuPont/Merck Pharmaceuticals, Monsanto and Sterling Drug to J.F. F.W.S. was supported in part by the NIH.

Long-range correlations in nucleotide sequences

C.-K. Peng*, S. V. Buldyrev*, A. L. Goldberger†, S. Havlin*‡, F. Sciortino*, M. Simons†§ & H. E. Stanley*

* Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215, USA
† Cardiovascular Division, Harvard Medical School, Beth Israel Hospital, Boston, Massachusetts 02215, USA
‡ Physical Sciences Laboratory, Division of Computer Research and Technology, National Institutes of Health, Bethesda, Maryland 20892, USA
§ Biology Department, MIT, Cambridge, Massachusetts 02139, USA

DNA SEQUENCES have been analysed using models, such as an *n*-step Markov chain, that incorporate the possibility of short-range nucleotide correlations[1]. We propose here a method for studying the stochastic properties of nucleotide sequences by constructing a 1:1 map of the nucleotide sequence onto a walk, which we term a 'DNA walk'. We then use the mapping to provide a quantitative measure of the correlation between nucleotides over long distances along the DNA chain. Thus we uncover in the nucleotide sequence a remarkably long-range power law correlation that implies a new scale-invariant property of DNA. We find such long-range correlations in intron-containing genes and in nontranscribed regulatory DNA sequences, but not in complementary DNA sequences or intron-less genes.

For the conventional one-dimensional random walk model, a walker moves either up ($u(i)=+1$) or down ($u(i)=-1$) one unit length (u) for each step i of the walk[2]. For the case of an uncorrelated walk, the direction of each step is independent of the previous steps. For the case of a correlated random walk, the direction of each step depends on the history ('memory') of the walker. The DNA walk is defined by the rule that the walker steps up ($u(i)=+1$) if a pyrimidine occurs at a position a linear distance i along the DNA chain, whereas the walker steps down ($u(i)=-1$) if a purine occurs at position i. Does such a walk display only short-range correlations (as in an n-step Markov chain) or long-range correlations (as in critical phenomena and other scale-free 'fractal' phenomena).

This DNA walk provides a graphical representation for each gene and permits the degree of correlation in the nucleotide sequence to be directly visualized, as in Fig. 1. Figure 1 naturally motivates a quantification of this correlation by calculating the 'net displacement' (y) of the walker after l steps, which is the sum of the unit steps $u(i)$ for each step i,

$$y(l) \equiv \sum_{i=1}^{l} u(i). \tag{1}$$

An important statistical quantity characterizing any walk[2] is the root mean square fluctuation $F(l)$ about the average of the displacement; $F(l)$ is defined in terms of the difference between the average of the square and the square of the average,

$$F^2(l) \equiv \overline{[\Delta y(l) - \overline{\Delta y(l)}]^2} = \overline{[\Delta y(l)]^2} - [\overline{\Delta y(l)}]^2, \tag{2a}$$

of a quantity $\Delta y(l)$ defined by

$$\Delta y(l) \equiv y(l_0+l) - y(l_0). \tag{2b}$$

Here the bars indicate an average over all positions l_0 in the gene. Operationally, this is equivalent to (1) taking a set of calipers set for a fixed distance l, (2) moving the beginning point sequentially from $l_0=1$ to $l_0=2$ and so on (3) calculating the quantity $\Delta y(l)$ (and its square) for each value of l_0, and (4) averaging all of the calculated quantities to obtain equation (2a).

The mean-square fluctuation is related to the auto-correlation function $C(l) = \overline{u(l_0)u(l_0+l)} - [\overline{u(l_0)}]^2$ through the relation[3] $F^2(l) = \sum_{i=1}^{l}\sum_{j=1}^{l} C(j-i)$. Because F is related to C by a summation, the fluctuation in F is substantially reduced compared with the fluctuation in C. Hence measurement of F leads to a more reliable characterization of the DNA walk than measurement of C.

The calculation of $F(l)$ can distinguish three possible types of behaviour. (1) If the nucleotide sequence were random, then $C(l)$ would be zero on average (except $C(0)=1$), so $F(l) \sim l^{1/2}$ (as expected for a normal random walk). (2) If there were a local correlation extending up to a characteristic range R (such as in Markov chains), then $C(l) \sim \exp(-l/R)$; nonetheless the asymptotic behaviour $F(l) \sim l^{1/2}$ would be unchanged from the purely random case[3]. (3) If there is no characteristic length (that is, if the correlation were 'infinite-range'), then the scaling property of $C(l)$ would not be exponential, but would instead be a power law function, and the fluctuations will also be described by a power law

$$F(l) \sim l^{\alpha} \tag{3}$$

with $\alpha \neq \frac{1}{2}$.

Figure 1a shows a typical example of an intron-containing gene. The DNA walk has an obviously very jagged contour which corresponds to long-range correlations. Our calculation of $F(l)$ for this gene is shown in Fig. 2a. That the data are linear over three decades on this double logarithmic plot confirms that $F(l) \sim l^{\alpha}$. We calculated a least-squares fit and find a straight line with slope $\alpha = 0.67 \pm 0.01$.

Figure 1b shows the DNA walk for the cDNA sequence of this same gene, whereas Fig. 1c shows the data for a typical intron-less gene. In contrast to Fig. 1a, these intron-free sequences have less jagged contours, suggesting a shorter range correlation. In almost all the intron-less sequences studied, purine-rich regions (compared with the average concentration over the entire strand) alternate with pyrimidine-rich regions, corresponding to the 'up-hill' and 'down-hill' portions of the DNA walk. To take into account the fact that the concentrations of purines and pyrimidines are not constant throughout the single-strand nucleotide sequence, we partitioned each DNA walk representation into segments demarcated by the global maximum ('max') and minimum ('min') displacements and analysed the fluctuation in each segment. Figure 2a also shows the data for the cDNA, and a least-squares fit gives a straight line with slope $\alpha = 0.49 \pm 0.01$. To ensure we did not introduce a bias by this 'min-max partitioning', we also used the same partitioning procedure to analyse the full gene and find no change in the exponent α.

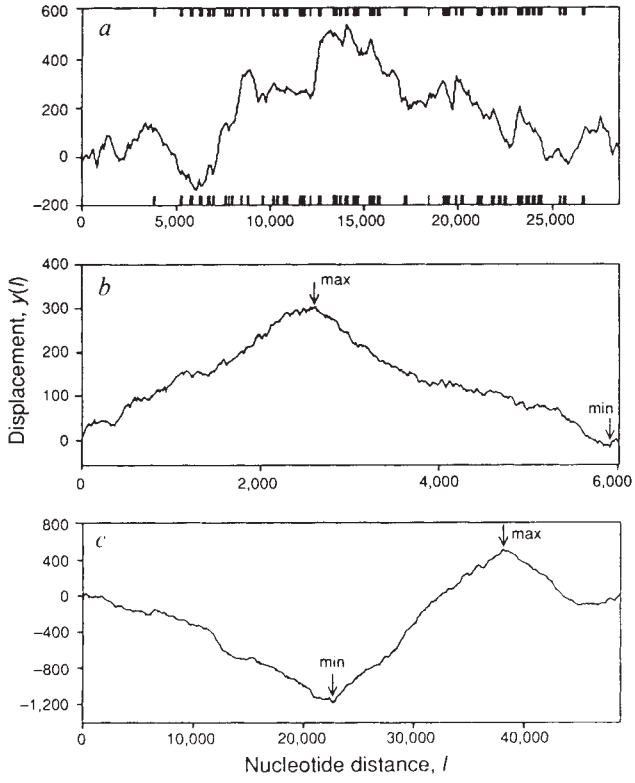

FIG. 1 The DNA walk representations of intron-rich human β-cardiac myosin heavy-chain gene sequence (*a*) its cDNA (*b*), and the intron-less bacteriophage λ DNA sequence (*c*). Note the more complex fluctuations for the intron-containing gene in *a* compared with the intron-less sequences in *b* and *c*. Heavy bars denote coding regions of the gene. So that the graphical representation was not affected by the global differences in concentration between purines and pyrimidines, DNA walk representations were plotted so that the end point has the same vertical displacement as the starting point (for the statistical analysis, we use the original definitions (1)–(3), without any adjustment of vertical displacement). The minimum (min) and maximum (max) points on the landscape are denoted by arrows, and their application in the analysis is described in the text. For almost all intron-less genes and cDNA sequences studied, there appear regions with one strand bias, followed by regions of a different strand bias. The fluctuation on either side of the overall strand bias is random, a fact that is plausible by visual inspection of the DNA walk representations.

To see if this scaling behaviour is universal, we applied our analysis to a large number of representative genomic and cDNA sequences across the phylogenetic spectrum, some of which are shown in Table 1. The myosin heavy-chain family is particularly useful because it encompasses a number of long genomic and cDNA sequences for species ranging from yeast to humans. We also analysed other sequences encoding a variety of other proteins as well as regulatory DNA sequences. We discovered that long-range correlations ($\alpha > 1/2$) are characteristic of intron-containing genes and nontranscribed genomic regulatory elements (group A). Thus, the average value (± 2 s.e.m.) of α for the first 14 entries of Table 1 is 0.61 ± 0.03. By contrast, for cDNA sequences and genes without introns, we find that $\alpha = 1/2$ indicating no long-range correlation (group B). In keeping with this, the average value of α for the last 10 entries of Table 1 is 0.50 ± 0.01. This significant difference in the value of α for the two groups of nucleotide sequences is confirmed in Fig. 2*b*, where we average the actual fluctuations over the two sets of entries in Table 1. Thus, the calculation of $F(l)$ for the DNA walk representation provides a new, quantitative method to distinguish genes with multiple introns from intron-less genes and cDNAs based solely on their statistical properties.

To determine if the long-range correlations we observe could be due to long repetitive nucleotide sequences, we tested whether the correlations would persist when the nucleotides blocks of size L were shuffled. Specifically, we divided the nucleotide sequence into many small segments (each of length $L = 100$) and then randomized the purine–pyrimidine order in each segment. We find, as expected, that the long-range correlation does persist. This calculation shows that the long-range correlation is a 'global' property of the nucleotide organization and is independent of the details of any repeating sequences. The calculation is equivalent to a 'renormalization' of the sequence in which each individual segment is replaced by the corresponding uniform concentration of nucleotides.

The four nucleotides (adenine, guanine, thymine and cytosine) were analysed using the DNA walk and we observed long-range correlations in intron-containing genes. We find the most robust fits to power law scaling by computing the DNA walk for nucleotide class (purine versus pyrimidine) rather than for any specific nucleotide.

To test if our results could be artefacts of a strand bias (that is, that over some distance one strand is more purine-rich), we studied many 'artificial' nucleotide sequences with and without strand bias. We find the same exponent α for both cases; the effect of a strand bias is to give a non-zero value for the quantity $\overline{\Delta y(l)}$, but this quantity is subtracted in the actual calculations of the mean deviation from the average in equation (2*a*). We

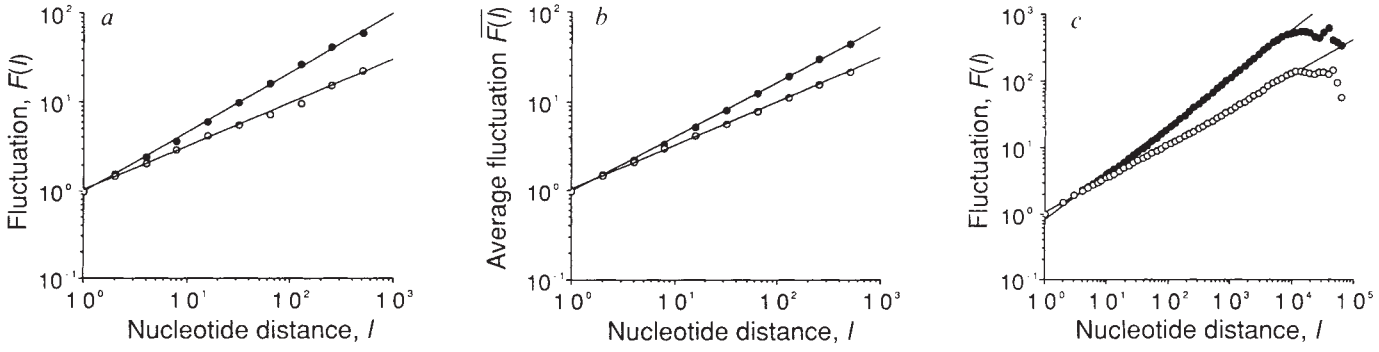

FIG. 2 Double logarithmic plots of *a*, the mean square fluctuation function $F(l)$ as a function of the linear distance l along the DNA chain for the human β-cardiac myosin heavy chain gene (●, $\alpha \approx 0.67$) and its cDNA (○, $\alpha \approx 0.49$) *and b*, the average of $F(l)$ over the entries in groups A (●, $\alpha \approx 0.62$) and B (○, $\alpha \approx 0.49$) of Table 1. The difference in the slopes is statistically significant, consistent with the possibility of long-range correlations in group A and short-range correlations in group B. For clarity, the data points shown are separated by intervals of integer powers of 2. *c*, A correlation of even longer range for the intron-containing human β-globin chromosomal region (73,376 nucleotides; ●, $\alpha \approx 0.71$). Shown for comparison is the '1-step' standard Markov chain analysis of the same nucleotide sequence, displaying the expected exponent $\alpha = 1/2$ (○, $\alpha \approx 0.52$).

TABLE 1 Summary of the correlation analysis of 24 sequences selected across the phylogenetic spectrum.

Sequence	Code	Comments	Length (nt)	% Introns	α
	GROUP A				
Adenovirus type 2	ADBCG	Intron-cont. Virus	35,937	n.d.	0.56
Caenorhabditis elegans MHC gene 1	CELMYO1A	Gene	12,241	51	0.61
C. elegans MHC gene 2	CELMYO2A	Gene	10,780	44	0.54
C elegans MHC gene 3	CELMYO3A	Gene	11,621	49	0.61
C. elegans MHC unc 54 gene	CELMYUNC	Gene	9,000	25	0.58
Chicken c-*myb* oncogene	CHKMYB15	Gene (5'-end)	8,200	92	0.60
Chicken embryonic MHC	CHKMYHE	Gene	31,111	78	0.65
Drosophila melanogaster MHC	DROMHC	Gene	22,663	72	0.56
Goat β-globin	GOTGLOBE	Gene*	10,194	n/a	0.58
Human β-globin	HHUMHBB	Chromosomal region	73,326	n/a	0.71
Human metallothionein	HUMMETIA	Gene	2,941	91	0.61
Human α-cardiac MHC	HUMMHCAG1	Gene (N-terminal)	2,366	72	0.65
Human β-cardiac MHC	HUMBMYH7	Gene	28,438	73	0.67
Rat embryonic skeletal MHC	RATMHCG	Gene	25,759	76	0.63
				$\alpha_{mean} \pm (2$ s.e.m.$) = 0.61 \pm 0.03$	
	GROUP B				
Bacteriophage λ	LAMCG	Intronless Virus	48,502	0	0.53
Chicken c-*myb* oncogene	CHKCMYBR	cDNA	2,218	0	0.50
Chicken nonmuscle MHC	CHKMYHN	cDNA	7,003	0	0.47
Dictiostelium discodium MHC	DDIMYHC	cDNA	6,681	0	0.49
Drosophila melanogaster MHC	DROMYONMA	cDNA	6,338	0	0.47
Human β-cardiac MHC	HUMBMYH7CD	cDNA	6,008	0	0.49
Human dystrophin	HUMDYS	cDNA	13,957	0	0.53
Human embryonic MHC	HUMSMHCE	cDNA (partial)	3,382	0	0.51
Human mitochondrion	HUMMT	Intronless Gene	16,569	0	0.49
Yeast MHC	SCMYO1G†	Intronless Gene	6,108	0	0.50
				$\alpha_{mean} \pm (2$ s.e.m.$) = 0.50 \pm 0.01$	

Sequences are divided into two groups on the basis of their intron content; within each group the sequences are ordered alphabetically. Note that $\alpha > 0.5$ implies the existence of long-range correlations, whereas $\alpha \approx 0.5$ implies only short-range correlations. The second column (code) lists the GenBank names (unless specified otherwise).

nt, Number of nucleotides per sequence.
*, eta-globin activating region (nontranscribed DNA).
n.d., no data; exon/intron map not fully known.
n/a, not applicable; nontranscribed DNA regions.
MHC, myosin heavy chain.
†, EMBL name.

also find that the only way to obtain a deviation from $\alpha = 1/2$ is to introduce a long-range correlation; for example, studies of an artificial sequence characterized by long-range correlation with correlation parameter of 0.61 yield a graph of the form of Fig. 2*a*, with a slope of exactly 0.61.

To confirm that the nucleotide correlations are truly long range, we now discuss the breakdown of linearity that must ultimately occur in graphs such as Fig. 2*a* and *b*. We do not show the graphs for distances larger than 1,000 because the statistical error increases (not because the correlation vanishes, which would be indicated by the slope changing to the value $\alpha = 1/2$). An increase in statistical error when there are less than roughly 10 independent data sets is usually found for analysis of this sort. For example, if a gene has 10,000 nucleotides, then we obtain only 10 independent sets of data when the calipers are separated by a distance of 1,000. Indeed, the 'fall-off' in the straight line behaviour is typical of all fractal analysis and is also found for artificial sequences of correlated numbers. We find that genes with more nucleotides possess still longer regions of linear (power law) behaviour than the three decades of linearity shown in Fig. 2*a* and *b*. For example, Fig. 2*c* is the correlation graph for the human β-globin chromosomal region, which has 73,376 nucleotides; note that the linearity extends to roughly 7,000 nucleotides. For comparison, Fig. 2*c* also shows the standard Markov chain analysis, which corroborates the classical result $\alpha = 1/2$.

We have thus introduced a new method to display correlations in the sequence of nucleotides, and defined a quantitative measure of the degree of correlation which is derived from a random walk representation. We find that the nucleotide sequence in intron-containing genes is highly correlated, and that the correlation is remarkably long range, indeed, nucleotides thousands of base pairs distant are correlated. Moreover, the quantitative scaling of the correlation is of the power law form observed in numerous phenomena having a self-similar or 'fractal' origin. A previous report of long-range correlations in DNA did not provide sufficient evidence to establish this fact[4]. Finally, we note that such long-range correlations are generally associated with the existence of a nonequilibrium dynamic process[2,5–7]. Interestingly, cDNAs do not have this property and seem to exist in an equilibrium state. Our finding of long-range correlations in intron-containing genes seems to be independent of the particular gene or the encoded protein. It is observed in genes as disparate as myosin heavy chain, β-globin and adenovirus (Table 1). The functional (and structural) role of introns remains uncertain (for review see ref. 8). Although our data do not resolve the 'intron-late'[9] versus 'intron-early'[10,11] controversy about gene evolution, they do reveal intriguing fractal properties of genome organization that need to be accounted for by any such theory. □

Received 12 August; accepted 3 December 1991.

1. Tavaré, S. & Giddings, B. W. in *Mathematical Methods for DNA Sequences* (ed. Waterman M. S.) 117–132 (CRC Press, Boca Raton, 1989).
2. Montroll, E. W. & Shlesinger, M. F. in *Nonequilibrium Phenomena II. From Stochastics to Hydrodynamics* (eds Lebowitz, J. L. & Montroll, E. W.) 1–121 (North-Holland, Amsterdam, 1984).
3. Stanley, H. E. *Introduction to Phase Transitions and Critical Phenomena* 120–121 (Oxford University Press, London, 1971).
4. Li, W. *Santa Fe Institute Technical Report* No. SFI-91-02 (1991).
5. Dutta, P. & Horn, P. M. *Rev. mod. Phys.* **53,** 497–516 (1981).
6. Bak, P., Tang, C. & Wiesenfeld, K. *Phys. Rev. Lett.* **59,** 381–384 (1987).
7. Shlesinger, M. F. *Rev. phys. Chem.* **39,** 269–290 (1988).
8. Doolittle, W. F. in *Intervening Sequences in Evolution and Development* (eds Stone, E & Schwartz, R.) 42–62 (Oxford University Press, New York, 1990).
9. Gilbert, W. *Nature* **271,** 501 (1978).
10. Darnell, J. E. Jr *Science* **202,** 1257–1260 (1978).
11. Doolittle, W. F. *Nature* **272,** 581–582 (1978).

ACKNOWLEDGEMENTS. We thank J. Hausdorff for contributions in the initial stages of this project, C. DeLisi, L. Liebovitch, R. D. Rosenberg, M. Schwartz and R. Voss for discussions, N. Shworak for metallothionein sequences and Y. W. He, G. Huber, B. K. Lee, R. Nossal, D. R. Rigney, A. Parsegian and P. Trunfio for comments. Partial support was provided to A.L.G. by the G. Harold and Leila Y. Mathers Charitable Foundation, the National Heart, Lung and Blood Institute and NASA, to M.S. by the American Heart Association, and to H.E.S., F.S., S.B. and G.K.P. by the Office of Naval Research and the National Science Foundation.