Patchiness and Correlations in DNA Sequences

Samuel Karlin, Volker Brendel

Science, New Series, Volume 259, Issue 5095 (Jan. 29, 1993), 677-680.

REFERENCES AND NOTES

1. P. W. Postma and J. Lengler, *Microbiol. Rev.* **49**, 232 (1985); M. H. Saier, Jr., *ibid.* **53**, 109 (1989); N. D. Meadow, D. K. Fox, S. Roseman, *Annu. Rev. Biochem.* **59**, 497 (1990); S. Roseman and N. D. Meadow, *J. Biol. Chem.* **265**, 2993 (1990).
2. M. J. Novotny *et al.*, *J. Bacteriol.* **162**, 810 (1985).
3. P. W. Postma *et al.*, *ibid.* **158**, 351 (1984).
4. C. M. Anderson, R. E. Stenkamp, R. C. McDonald, T. A. Steitz, *J. Mol. Biol.* **123**, 207 (1978).
5. K. M. Flaherty *et al.*, *Nature* **346**, 623 (1990).
6. W. Kabsch *et al.*, *ibid.* **347**, 37 (1990).
7. Domains of GK consist of IA, residues 1 to 35, 49 to 83, 165 to 173, and 221 to 253; IB, residues 36 to 48 and 82 to 164; IIA, residues 254 to 306 and 373 to 472; IIB, residues 307 to 372. Unique additions are IC, residues 174 to 220, and IIC, residues 456 to 501.
8. K. M. Flaherty *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 5041 (1991).
9. K.-U. Frohlich, K. D. Entian, D. Mecke, *Gene* **36**, 105 (1985). Residues numbers for hexokinase are quoted for the translated gene, not the crystallographic sequence.
10. D. W. Pettigrew and D. C. Thomas, unpublished results.
11. J. K. DeRiel and H. Paulus, *Biochemistry* **17**, 5134 (1978); *ibid.*, p. 5141; *ibid.*, p. 5146.
12. P. J. Bjorkman *et al.*, *Nature* **329**, 506 (1987); R. St. Charles, D. A. Walz, B. F. P. Edwards, *J. Biol. Chem.* **264**, 2092 (1989); G. M. Clore, E. Apella, M. Yamada, K. Matsushima, A. M. Gronenborn, *ibid.*, p. 18907; J. H. Hurley *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 8635 (1989); K. Valegard *et al.*, *Nature* **345**, 36 (1990).
13. B. Lee and F. M. Richards, *J. Mol. Biol.* **55**, 379 (1971). Probe radius was 1.4 Å.
14. W. S. Bennett, Jr. and T. A. Steitz, *ibid.* **140**, 183 (1980).
15. J. W. Thorner and H. Paulus, *J. Biol. Chem.* **248**, 3922 (1973).
16. The superposition included residues 5 to 25, 48 to 68, 77 to 82, 164 to 174, 240 to 256, 260 to 266, 268 to 276, 302 to 319, 381 to 402, 407 to 436, and 437 to 454 of GK, and residues 5 to 25, 116 to 136, 142 to 147, 156 to 166, 170 to 186, 195 to 201, 204 to 212, 221 to 238, 306 to 327, 335 to 364, and 367 to 384 of HSC70.
17. D. Worthylake *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 10382 (1991).
18. D.-I. Liao *et al.*, *Biochemistry* **30**, 9583 (1991).
19. J. G. Pelton *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 3479 (1991).
20. N. D. Meadow and S. Roseman, *J. Biol. Chem.* **257**, 14526 (1982).
21. N. D. Meadow *et al.*, *J. Biol. Chem.* **261**, 13504 (1986).
22. In contact 2, residues 1 to 11 of III^{Glc} bind to residues 125 to 129, 195, 258 to 259, 274 to 281, and 399 to 401 of the GK subunit of the second tetramer. In contact 3, residues 27 to 29, 55 to 56, and 155 to 161 of III^{Glc} interact with residues 183, 195 to 197, 219, 277, and 298 to 300 of the same GK subunit as for contact 2.
23. D. R. Davies, E. A. Padlan, S. Sheriff, *Annu. Rev. Biochem.* **59**, 439 (1990); J. Janin and C. Chothia, *J. Biol. Chem.* **265**, 16027 (1990).
24. J. Reizer *et al.*, *J. Biol. Chem.* **267**, 9158 (1992).
25. K. A. Presper *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 4052 (1989).
26. R. E. Klevit and E. B. Waygood, *Biochemistry* **25**, 7774 (1986).
27. O. Herzberg *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 2499 (1992).
28. M. Ikura *et al.*, *Science* **256**, 632 (1992); W. E. Meador *et al.*, *ibid.* **257**, 1251 (1992).
29. J. G. Pelton, D. A. Torchia, N. D. Meadow, S. Roseman, *Biochemistry* **31**, 5215 (1992).
30. ———, *Protein Science*, in press.
31. J. H. Hurley, A. M. Dean, P. E. Thorsness, D. E. Koshland, Jr., R. M. Stroud, *J. Biol. Chem.* **265**, 3599 (1990); J. H. Hurley, A. M. Dean, J. L. Sohl, D. E. Koshland, Jr., R. M. Stroud, *Science* **249**, 1012 (1990); A. M. Dean and D. E. Koshland, Jr., *ibid.*, p. 1044.
32. GK and III^{Glc} were mixed at 1:1 stochiometry to yield a final protein concentration of approximately 20 mg/ml. A 5-μl drop of the well solution consisting of 0.5 to 0.8 M sodium acetate, 100 mM MES (pH 6.0) was mixed with 5 μl of the protein mixture. Crystals of up to 1.0 mm along the largest dimension were obtained in 1 week at room temperature. The space group is *I*222 with a = 123.4, b = 124.3, c = 133.6 Å. The packing parameter [B. W. Matthews, *J. Mol. Biol.* **33**, 491 (1968)] is V_m = 3.5 Å^3/D for one monomer each of GK and III^{Glc} in the asymmetric unit. The crystals diffracted to 2.4 Å. Heavy atom derivatization was performed in a storage solution of 16% (w/v) PEG 1550, 40 mM Pipes (pH 7.0). A 3-day soak in a 1:10 dilution of a saturated solution of *cis*-platinum(II) diamine dichloride (cis-Pt) and a 2-week soak of a 1:100 dilution of a saturated solution of $HgCl_2$ in the presence of 50 mM 2-mercaptoethylamine yielded useful derivatives. "ADP" refers to crystals soaked in storage solution containing 5 mM ADP, 5 mM $MgCl_2$, and 5 mM glycerol. Protein was purified as described [H. R. Faber, D. W. Pettigrew, S. J. Remington, *J. Mol. Biol.* **207**, 637 (1989); J. G. Pelton, D. A. Torchia, N. D. Meadow, C.-Y. Wong, S. Roseman, *Biochemistry* **30**, 10043 (1992)].
33. R. Hamlin, *Methods Enzymol.* **114**, 416 (1985).
34. A. J. Howard *et al.*, *ibid.*, p. 452.
35. M. G. Rossmann, *J. Appl. Crystallogr.* **12**, 225 (1979); M. F. Schmid *et al.*, *Acta Crystallogr. Sect. A* **37**, 701 (1981).
36. T. C. Terwilliger and D. Eisenberg, *Acta Crystallogr. Sect. A* **39**, 813 (1983).
37. B. C. Wang, *Methods Enzymol.* **115**, 90 (1985).
38. T. A. Jones, *ibid.*, p. 157.
39. A. T. Brünger, J. Kuriyan, M. Karplus, *Science* **235**, 458 (1987); A. T. Brunger, M. Karplus, G. A. Petsko, *Acta Crystallogr. Sect. A* **45**, 50 (1989).
40. In the first round of refinement, the partial model (75% of all atoms) was energy-minimized, annealed from 6000 K to 300 K, and minimized again. After group B-factor refinement, the R factor dropped from 51 to 32% for all data from 6.0 to 2.8 Å. After greater than 90% of all atoms had been modeled in electron density from the final solvent-flattened MIR map, refinement was continued with all data from 5.0 to 2.6 Å. Four rounds of manual rebuilding and simulated annealing from 2000 K to 300 K and two rounds of rebuilding with conventional conjugate gradient minimization led to the current model. A glycerol molecule was located in an $F_{obs} - F_{calc}$ Fourier synthesis following the third round of simulated annealing refinement. The final model contained 650 amino acid residues, one glycerol molecule, and no water molecules. About 2% of the main chain torsion angle (ϕ, ψ) combinations deviated from allowed regions of the Ramachandran diagram by more than 20°. Residues 1 to 3, 230 to 236, and 500 to 501 of GK and residues 12 to 18 of III^{Glc} as well as 18 side chains of GK could not be located in the electron density and have been omitted from the model. The ADP molecule was placed in a difference electron density map and the model refined.

41. We thank K. Kallio for technical assistance, R. Huber and his colleagues for calling our attention to their derivitization method [H. W. Hoeffken *et al.*, *J. Mol. Biol.* **204**, 417 (1988)]; D. B. McKay for providing complete coordinates of HSC70; and B. W. Matthews, F. W. Dahlquist, and W. Tulip for helpful discussions. Supported by the National Institutes of Health grants GM 42618-01A1 (S.J.R.), American Cancer Society ACS 1296 (S.J.R.), 5-R37 GM38759 (S.R.), a grant from the Lucille P. Markey Charitable Trust, and a postdoctoral fellowship from the American Cancer Society (J.H.H.). Crystallographic refinement was carried out in part at the Pittsburgh Supercomputing Center under grant DMB900038P (J.H.H. and B. W. Matthews). Coordinates have been deposited in the Protein Data Bank.

8 July 1992; accepted 6 November 1992

Patchiness and Correlations in DNA Sequences

Samuel Karlin and Volker Brendel

The highly nonrandom character of genomic DNA can confound attempts at modeling DNA sequence variation by standard stochastic processes (including random walk or fractal models). In particular, the mosaic character of DNA consisting of patches of different composition can fully account for apparent long-range correlations in DNA.

Genomic global and local compositional heterogeneity is widely recognized. The many examples of DNA heterogeneity in existence include isochore compartments [regions dominated by either G + C or A + T as determined by thermal-melting studies or density-gradient centrifugation; for example, the G + C– and A + T–rich halves of the bacteriophage lambda genome (*1*); see (*2*) for examples in mammalian species]; mobile insertion elements [such as Alu in human, Ty in yeast, and IS in *Escherichia coli* (*3*)]; characteristic satellite centromeric tandem repeats [such as the 171–base pair human alpha satellite DNA (*4*)]; characteristic telomeric sequences [such as AGGGTT tandem repeats in humans (*5*)]; HTF islands [vertebrate DNA sequences that occur generally upstream of genes and are abundant with nonmethylated CpG (*6*)]; repetitive extragenic palindromes (REPs) in *E. coli* and *Salmonella typhimurium* (*7*); repeat induced point (RIP) mutation in certain fungi (*8*); recombinational hot spots [such as chi elements in *E. coli* (*9*)]; universal underrepresentation of the dinucleotide TpA (*10*); suppression of the dinucleotide CpG in vertebrates (*11*); the pervasive rarity of the tetranucleotide CTAG (*12*); GNN periodicity in coding sequences (*13*); and methyltransferase modifications (*14*).

Thus, genome organization is complex and variegated. In general, genomic sequences are not homogeneous on any scale. For example, eukaryotic sequences are often endowed with tandem repeats accruing

Department of Mathematics, Stanford University, Stanford, CA 94305.

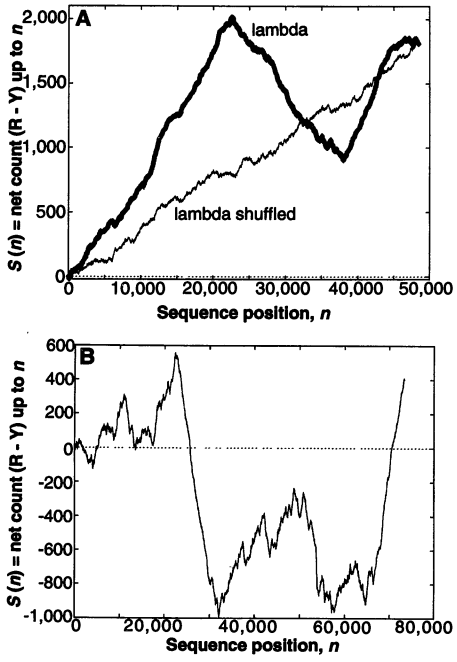

Fig. 1. (**A**) Random walk plot of the bacteriophage lambda sequence (thick line) and a shuffled sequence of the same composition (thin line). For each occurrence of R (purine) the graph moves up one unit, and for each occurrence of Y (pyrimidine) the graph moves down one unit. Thus $S(n)$ is the net count of R minus Y from the beginning of the sequence up to position n. (**B**) Analogous random walk plot for the human beta-globin region (GenBank/EMBL file name HUMHBB).

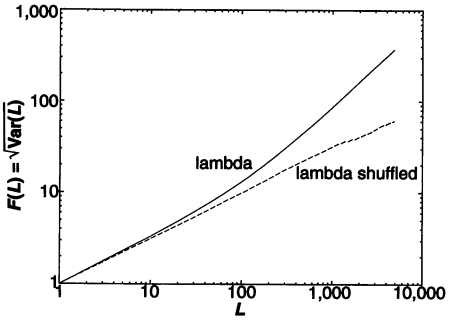

Fig. 2. Fluctuation plot for the bacteriophage lambda sequence (solid line) and a shuffled sequence of the same composition (broken line). The function $F(L)$ is the estimated standard deviation of R minus Y counts over all windows of size L along the sequence (see text).

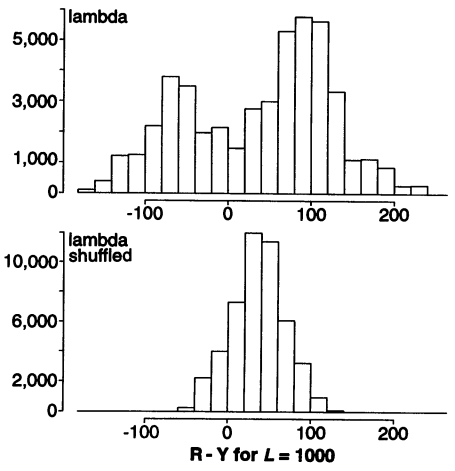

Fig. 3. Histograms of R minus Y counts in windows of size L = 1000 for bacteriophage lambda and a shuffled sequence of the same composition. The bimodal distribution for lambda has a mean of 38 and a standard deviation of 86. The shuffled sequence has a similar mean of 37 but a much smaller standard deviation of 32, close to the expected value of $\sqrt{1000}$.

from polymerase slippage or unequal crossing-over and with distant direct and inverted repeats promoted in part by transposition, translocation, recombination, amplification, and excision recurrences. Many genomic sequences exhibit polymorphisms, strain variation, DNA inversions, and rearrangements reflecting a state of flux.

These phenomena raise fundamental and methodological questions. Can DNA sequence variation be reasonably modeled by stochastic processes of a tractable genre? In this context, there has recently been intense discussion about the existence and nature of long-range correlations within DNA sequences (*15–18*). One approach has been to study DNA sequence variation by means of a random walk generated by an incremental variable that associates to position i the value $X(i) = 1$ or -1, depending on whether the ith nucleotide of the sequence is based on R (purine) or Y (pyrimidine), respectively (Fig. 1) (*15*). Let

$$S(n) = \sum_{i=1}^{n} X(i),\ n = 1,2,\ldots,N \quad (1)$$

be the cumulative variable of the random walk, and let us denote the cumulation over a segment defined by a sliding window of length L by

$$S_k(L) = S(k + L - 1) - S(k - 1) = \sum_{i=k}^{k+L-1} X(i) \quad (2)$$

where k is the position where the window begins. The analysis in (*15*) implicitly assumes that the distribution of nucleotides is stationary (stationarity); in particular, it is assumed that the probability distribution of $S_k(L)$ is independent of the sequence position k. For each window size L, the sample variance of $S_k(L)$, $\mathrm{Var}(L)$, gives a measure of sequence variation. Its value is determined from the data moving along the sequence ($k = 1,2,\ldots,N - L + 1$). Under the condition of stationarity, $\mathrm{Var}(L)$ is an estimate of the auto-covariance function

$$C(L) = \sum_{i=1}^{L}\sum_{j=1}^{L} \mathrm{Cov}\,[X(i),X(j)] \quad (3)$$

where $\mathrm{Cov}[X(i),X(j)]$ denotes the covariance of $X(i)$ and $X(j)$. Here, $\mathrm{Cov}[X(i), X(i + r)] = \mathrm{Cov}[X(1), X(1 + r)] = C_r$ for all $i = 1,2,\ldots$, and $r = 1,2,\ldots$, and thus

$$C(L) = LC_0 + 2\sum_{r=1}^{L-1} (L - r)C_r \quad (4)$$

For a stationary stochastic process with asymptotically strongly independent increments, C_r decreases to zero rapidly as $r \to \infty$; for example, $C_r \simeq c\sigma^r$ where $c > 0$ and $0 < \sigma < 1$. In this case, $C(L)$ grows with order L. In contrast, for a process with asymptotically weakly independent increments (implying weak long-range correlations), C_r decreases to zero at a slow rate; for example, $C_r \simeq cr^{\alpha-2}$ where $c > 0$ and $1 < \alpha < 2$. In this case, $C(L)$ grows with order L^α (*19, 20*).

Recent papers (*15, 16*) proffer the asymptotically weakly independent stationary process as a model to describe apparent long-range dependencies inherent to many DNA sequences. However, the assumption of stochastic stationarity is problematic in view of the great degree of local and global heterogeneity, as indicated above. We show here that most DNA sequence variation can be explained by compositional patchiness and does not involve the higher order organization implied by long-range correlations [see also (*18*)].

The effect of genomic patchiness on variance and correlation assessments can be illustrated by analysis of the sequence of the bacteriophage lambda and a corresponding random sequence obtained by shuffling the original sequence. Figure 1A displays the random walk plot for both sequences. The shuffled sequence attains the overall excess of purines over pyrimidines observed for the lambda sequence with an approximately steady increase throughout, as would be expected for a random walk with drift (such as for independent tosses of a biased coin). The log-log plot of $F(L) = \sqrt{\mathrm{Var}(L)}$ as a function of L (Fig. 2) for the shuffled sequence gives the expected line with slope ~0.5. The corresponding plot for lambda, however, reveals a marked deviation from a straight line with the values of $F(L)$ exceeding the ones expected for a homogeneous random sequence. This larger variance is most easily explained by the extant patchiness of the lambda sequence (see theory below). To illustrate the patchiness, Fig. 3 gives the histograms of the R minus Y counts in windows of length L = 1000 for lambda and for the shuffled sequence. For lambda, one clearly sees a bimodal histo-

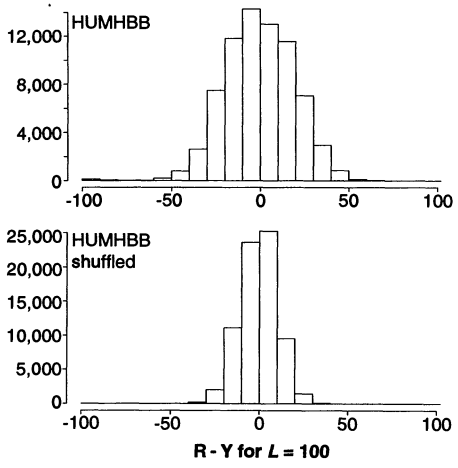

Fig. 4. Histograms of R minus Y counts in windows of size $L = 100$ for the human beta-globin region (HUMHBB) and a shuffled sequence of the same composition. Both distributions have a mean of about 0, but the human sequence displays a standard deviation which is larger by a factor of 2 (19.5 versus 10.3).

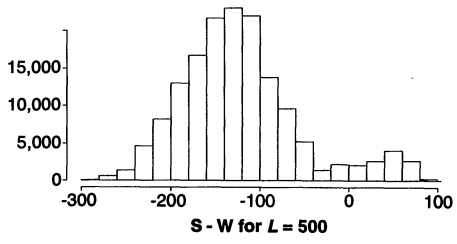

Fig. 5. Histograms of S (G or C) minus W (A or T) counts in windows of size $L = 500$ for the 155844 nucleotide sequence of the tobacco chloroplast genome (GenBank/EMBL file name CHNTXX). The mean is −121.5. The standard deviation is 67.2, about threefold larger than expected for a random sequence of the same composition.

gram with peaks corresponding to pyrimidine- and purine-rich windows that are derived from the pyrimidine- and purine-rich segments of the sequence. This distribution has a variance that is considerably larger than the variance for the unimodal, bell-shaped distribution associated with the shuffled sequence. Similar results hold for other values of L, leading to the discrepancy between the natural and the shuffled sequences that is evident in Fig. 2 (see Figs. 1B and 4 for a eukaryotic example).

We shall argue that such a discrepancy follows, in general, from a theoretical analysis of sequences modulated by a multipatch model. Suppose that in a first approximation, the sequence is composed of a mosaic of patches of distinct underlying compositions [that is, the distribution of the increments $X(i)$ in each patch type is homogeneous but differs between patches]. The variance of $S_k(L)$ over all windows, irrespective of patch type, can be obtained by classical partitioning according to patch type (analysis of variance within and between groups) as the sum of the within-patch variance (average of the variances over the different patch types) plus the between-patch variance (variance of the average segmental values in the different patch types) (*21*). Formally,

$$\text{Var } S(L) = E[\text{Var}\{S_k(L) \mid (k,k+L-1) \text{ in patch } P\}] + \text{Var}E[S_k(L) \mid (k,k+L-1) \text{ in patch } P] \quad (5)$$

where E denotes the appropriate expectation and $(k,k+L-1)$ refers to the nucleotides in sequence positions k to $k+L-1$. The variances within each patch are about $c_P L$, where the positive constants c_P depend on the patch type P. Thus, averaging over the different patches gives a within-patch variance estimate of the order L. The variance of all the different segmental patch values (the between-patch variance) is of the order dL^2, where the nonnegative constant d is zero only when the underlying distributions of all patches are the same (*22*). In total, the overall variance of $S_k(L)$ for a sequence that comprises at least two distinct patches is given by $cL + dL^2$. Accordingly, depending on the complexity of the DNA sequence (the degree of local inhomogeneity and global patchiness), the variance plot is more accurately represented by a curve of the form $cL + dL^2$, which might be fitted by L^α (for some α, $1 \leq \alpha \leq 2$) over short intervals with α varying from interval to interval. Here L has to be small compared to the sequence length; otherwise, the patchiness will be smoothed out by averaging over the window size. More generally, the contribution of the dL^2 term will tend to be large when L is small compared to the typical patch size, but will be reduced for larger L.

An intrinsic difficulty with the approach taken in (*15*) is that the plot of $\log F(L)$ against $\log L$ is not a straight line over the whole range of L values (*23*) (Fig. 2), contrary to the assumption of stationarity. Given that compositional heterogeneity occurs on all scales, attempts at breaking up a sequence into more homogeneous segments (*15, 16*) cannot succeed, unless the segments are chosen to be so small that the whole notion of long-range correlations evaporates. Observations analogous to those reported here hold, to varying degrees in our data analyses, for most sequences, including both the intron-containing genes and the cDNAs analyzed in (*15*). Moreover, the pervasive phenomenon of patchiness also occurs in other alphabets, as in, for example, S (G or C) versus W (A or T) or G versus non-G. A typical example is given in Fig. 5. The multimodal character of the distribution is evident. Similar distributions for the S − W segmental counts are obtained for other sequences, including the long herpes-virus genomes, the human beta-globin region, and yeast chromosome III. The pervasive patchiness of DNA in all alphabets precludes modeling DNA sequences by a stationary process and, in particular, by a process having long-range dependence. From this perspective, systematic long-range correlations in DNA sequences are doubtful.

Biological phenomena are generally replete with variability, which is at the core of evolutionary developments and is also operating at the molecular level [to explain the extant variability, see the discussion between the protagonists (*24*) and antagonists (*25*) of the neutral theory of molecular evolution]. Mathematical and statistical models can, at best, help in providing benchmarks for interacting with biological data and in interpreting experimental results (*26*). Components of variance analyses such as described here might assist in assessing sources of molecular sequence variation.

REFERENCES AND NOTES

1. R. B. Inman, *J. Mol. Biol.* **18**, 464 (1966).
2. G. Bernardi *et al.*, *Science* **228**, 953 (1985); G. Bernardi, D. Mouchirond, C. Bautier, G. Bernardi, *J. Mol. Evol.* **28**, 7 (1988).
3. D. E. Berg and M. M. Howe, *Mobile DNA* (American Society for Microbiology, Washington, DC, 1989).
4. H. F. Willard and J. S. Waye, *Trends Genet.* **3**, 192 (1987).
5. E. H. Blackburn, *Nature* **350**, 569 (1991).
6. A. P. Bird, *ibid.* **321**, 209 (1986).
7. E. Gilson, W. Saurin, D. Perrin, S. Bachellier, M. Hofnung, *Nucleic Acids Res.* **19**, 1375 (1991).
8. E. U. Selker, *Annu. Rev. Genet.* **24**, 579 (1990).
9. S. Krawiec and M. Riley, *Microbiol. Rev.* **54**, 502 (1990).
10. C. Burge, A. Campbell, S. Karlin, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 1358 (1992).
11. J. Josse, A. D. Kaiser, A. Kornberg, *J. Biol. Chem.* **236**, 864 (1961).
12. S. Karlin, C. Burge, A. Campbell, *Nucleic Acids Res.* **20**, 1363 (1992).
13. J. W. Ficket, *ibid.* **10**, 5303 (1982).
14. M. Nelson and M. McClelland, *ibid.* **19**, 2045 (1991).
15. C.-K. Peng *et al.*, *Nature* **356**, 168 (1992).
16. W. Li and K. Kaneko, *Europhys. Lett.* **17**, 655 (1992); R. Voss, *Phys. Rev. Lett.* **68**, 3805 (1992).
17. J. Maddox, *Nature* **358**, 103 (1992); I. Amato, *Science* **257**, 747 (1992).
18. S. Nee, *Nature* **357**, 450 (1992).
19. B. B. Mandelbrot and J. W. Van Ness, *SIAM Rev.* **10**, 422 (1968); M. S. Taqqu, *Stochastic Processes Appl.* **7**, 55 (1978).
20. The theory of processes with stationary increments satisfying weak long-range dependencies is intimately related to self-similar stochastic processes [invariant in distribution under judicious scaling of time and space; specifically, if Z_t is such a process, then Z_{at}, for $a > 0$, is distributed like $a^\alpha Z_t$ for suitable $\alpha > 0$; an example would be fractional Brownian motion; see Y. G. Sinai, *Theor. Probab. Appl.* **21**, 64 (1976) and references in (*16*)]. Self-similar processes are of interest in theoretical physics, in hydrology, and in studies of $1/f$ noises [D. Wolf, Ed., *Noise in Physical Systems* (Springer-Verlag, New York, 1978); M. Cas-

sandro and G. Jona-Lasinio, *Adv. Phys.* **27**, 913 (1978)]. The mathematics of self-similar phenomena is elegant, tantalizing, and currently popular [see, for example, recent books on fractal sets and processes, chaotic behavior, and nonlinear dynamical systems, such as G. L. Baker and J. P. Gollub, *Chaotic Dynamics* (Cambridge Univ. Press, Cambridge, 1990); D. Ruelle, *Chance and Chaos* (Princeton Univ. Press, Princeton, NJ, 1991)]. However, attempts at fitting fractal models to various phenomena—such as geologic formations, market fluctuations, cardiac arrhythmias, meteorological anomalies, and fluid turbulence—are tenuous and controversial [see, for example, I. Amato, *Science* **256**, 1763 (1992)].

21. The representation of the total variance of a heterogeneous population that is composed of subpopulations as a sum of within- and between-subpopulation variance is at the core of ANOVA (analysis of variance) methods; see, for example, R. R. Sokal and F. J. Rohlf, *Biometry: The Principles and Practice of Statistics in Biological Research* (Freeman, New York, ed. 2, 1981), pp. 198–205; M. Fisz, *Probability Theory and Mathematical Statistics* (Wiley, New York, ed. 3, 1963), pp. 512–520; C. R. Rao and J. Kleffe, *Estimation of Variance Components and Applications* (Elsevier, New York, 1988).

22. In the simplest case, let the sequence consist of two patch types and assume that any particular window is equally likely to fall into either patch type. The observed variables are then

$$S_{jk}(L) = \sum_{i=k}^{k+L-1} X_j(i)$$

where $X_j(i) = 1$ with probability p_j and -1 with probability $q_j = 1 - p_j$, and $j = 1$ or 2 depending on which patch type the window starting at position k belongs to. The expectation and variance of S_{jk} are $E[S_{jk}] = L(p_j - q_j)$, and $\mathrm{Var}[S_{jk}] = L\,4\,p_jq_j$. Thus, the average within-patch variance is given by $L\,2\,(p_1q_1 + p_2q_2)$, and the between-patch variance is given by

$$\frac{1}{2}L^2[(p_1 - q_1)^2 + (p_2 - q_2)^2] - \frac{L^2}{4}[(p_1 - q_1) + (p_2 - q_2)]^2 = L^2(p_1 - p_2)^2$$

Only in the case $p_1 = p_2$ does the L^2 term vanish.

23. V. V. Prabhu and J.-M. Claverie, *Nature* **359**, 782 (1992).
24. M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983).
25. J. H. Gillespie, *The Causes of Molecular Evolution* (Oxford Univ. Press, Oxford, 1992).
26. S. Karlin and V. Brendel, *Science* **257**, 39 (1992).
27. This work was supported in part by NIH grants HG00335-04 and GM10452-29 and NSF grant DMS86-06244.

16 September 1992; accepted 13 November 1992

The Skipping of Constitutive Exons in Vivo Induced by Nonsense Mutations

Harry C. Dietz,* David Valle, Clair A. Francomano, Raymond J. Kendzior, Jr., Reed E. Pyeritz, Garry R. Cutting

Nonsense mutations create a premature signal for the termination of translation of messenger RNA. Such mutations have been observed to cause a severe reduction in the amount of mutant allele transcript or to generate a peptide truncated at the carboxyl end. Analysis of fibrillin transcript from a patient with Marfan syndrome revealed the skipping of a constitutive exon containing a nonsense mutation. Similar results were observed for two nonsense mutations in the gene encoding ornithine δ-aminotransferase from patients with gyrate atrophy. All genomic DNA sequences flanking these exons that are known to influence RNA splicing were unaltered, which suggests that nonsense mutations can alter splice site selection in vivo.

The fibrillin gene (*FBN1*) encodes a 350-kD glycoprotein component of the extracellular microfibril (*1*). Mutations in this gene cause Marfan syndrome, a systemic disorder of connective tissue with manifestations in the ocular, skeletal, and cardiovascular systems (*2–4*). Although the genomic organization of the gene has not been determined, approximately 7 of 10 kb of *FBN1* cDNA have been cloned and sequenced (*5*). The characterization of *FBN1* defects that cause Marfan syndrome resulted in the identification in one patient of an allele with a 66-nucleotide (nt) deletion in mature mRNA resulting from in-frame skipping of an entire exon. This patient had no family history of Marfan syndrome but had classic and severe manifestations of the disorder. The only identified sequence variation unique to this patient was a T → G transversion at position +26 of the skipped exon, which resulted in a premature TAG termination codon. Exon skipping restored the open reading frame of the mutant transcript. Alternative splicing was not observed in either parent of the patient or in 70 unrelated individuals. We subsequently observed two recurrences of this phenomenon in patients with different nonsense mutations in the gene (*OAT*) for ornithine δ-aminotransferase (OAT), a nuclear-encoded mitochondrial matrix enzyme. Mutations in *OAT* cause gyrate atrophy (GA), an autosomal recessive, slowly progressive chorioretinal degeneration leading to blindness in middle age. We propose that these nonsense mutations induce the skipping of constitutive exons in vivo.

A 540-nt fragment of *FBN1* cDNA was amplified by the polymerase chain reaction (PCR) and subjected to single-strand conformation polymorphism analysis. An abnormally migrating band unique to a sample from a patient with classic Marfan syndrome (patient MS-7) was observed (*6*). Reamplification of DNA recovered from the abnormal band showed a heteroduplex that contained a wild-type product (540 bp) and a smaller (474 bp) product. Direct sequencing of each template demonstrated a 66-bp deletion in the smaller product (Fig. 1) that encompassed the 3′ region of one of five eight-cysteine domains in *FBN1* that are homologous to a motif found in transforming growth factor–β1 binding protein (*5*). This cysteine-rich domain may participate in protein-to-protein interactions.

The sequencing of PCR-amplified genomic DNA revealed that the deleted 66-bp region of cDNA represents an entire exon (hereafter referred to as exon B) (Fig. 2A). To determine the basis for exon skipping, we amplified by PCR a 3.0-kb region of genomic DNA from MS-7, his parents, and an unrelated and unaffected individual (Co1). The resulting product spanned the 3′ end of the upstream exon (A), the skipped exon (B), and two downstream exons (C and D). With the exception of the central 2.0 kb of the intron following exon C (intron C), this region was sequenced for all four individuals (Fig. 2B). A wild-type sequence was observed for all of the cis-acting elements known to influence RNA splicing, including the 3′ and 5′ splice sites and the predicted branchpoint flanking exon B (*7–9*) in all samples. Patient MS-7's sample, however, had a unique mutation: a T → G substitution in one allele at position +26 in exon B (Fig. 2C). The corresponding amino acid alteration is a substitution of a termination codon (X) for tyrosine (Y) at codon 1215 (Y1215X) in the characterized coding sequence for *FBN1* (*5*). A Cvn I restriction site is created by this alteration and can be used to screen DNA for mutation Y1215X (*6*). None of 55 unaffected control subjects and 52 unrelated probands with Marfan syndrome carried this defect.

Another single base substitution was identified at position −64 of intron A. Patient MS-7 was heterozygous A/G, his mother was A/A, his father was G/G, and Col was A/G. This polymorphism occurs away from the conventional placement for a branchpoint sequence (positions −18 to −40) and within a context without homol-

H. C. Dietz, C. A. Francomano, R. J. Kendzior, Jr., R. E. Pyeritz, G. R. Cutting, Departments of Pediatrics and Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD 21205.
D. Valle, Howard Hughes Medical Institute, Department of Pediatrics, The Johns Hopkins University School of Medicine, Baltimore, MD 21205.
*To whom correspondence should be addressed.