Mosaic organization of DNA nucleotides

C.-K. Peng,[1,2] S. V. Buldyrev,[1] S. Havlin,[1] M. Simons,[2,3] H. E. Stanley,[1] and A. L. Goldberger[2]

[1] *Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215*

[2] *Cardiovascular Division, Harvard Medical School, Beth Israel Hospital, Boston, Massachusetts 02215*

[3] *Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139*

(Received 20 October 1993)

Long-range power-law correlations have been reported recently for DNA sequences containing noncoding regions. We address the question of whether such correlations may be a trivial consequence of the known mosaic structure ("patchiness") of DNA. We analyze two classes of controls consisting of patchy nucleotide sequences generated by different algorithms—one without and one with long-range power-law correlations. Although both types of sequences are highly heterogeneous, they are quantitatively distinguishable by an alternative fluctuation analysis method that differentiates local patchiness from long-range correlations. Application of this analysis to selected DNA sequences demonstrates that patchiness is not sufficient to account for long-range correlation properties.

PACS number(s): 87.10.+e, 05.40.+j

Recently there have been several reports that certain DNA sequences may display long-range power-law correlations extending across more than 10^4 nucleotides [1–5]. Moreover, it appears that coding sequences do not display long-range correlations [1,2]. How to interpret these findings is unclear at present. It is known that DNA nucleotides form a mosaic comprised of "patches" (excess of one type of nucleotide) [6–10]. The possibility that such patchiness is related to the appearance of long-range power-law correlations has been raised by Nee [9] and by Karlin and Brendel [10]. Here we address this possibility by systematically studying distinct forms of mosaic structure.

A useful way of analyzing patchiness arising from the heterogeneous purine-pyrimidine content is the DNA walk [1], defined as follows: A one-dimensional walker dictated by the nucleotide sequence takes one step down when there is a purine [$u(i) = -1$] and one step up when there is a pyrimidine [$u(i) = 1$]. The displacement of the walker after n steps, $y(n)$, is defined as $y(n) \equiv \sum_{i=1}^{n} u(i)$, and is displayed on a graph of y vs n as in Fig. 1 [11].

Figure 1(a) shows a representative DNA walk for a class of artificial "control" sequences generated by stitching together subsequences that correspond to random walks with different nucleotide composition (strand bias) [12]. Figure 1(b) shows a representative DNA walk for a different class of sequences, generated by a model with a well-defined long-range power-law correlation [13]; the heterogeneous structure of this control sequence arises directly from the long-range correlation itself.

We find apparent patchiness in real DNA sequences—both in the noncoding and coding regions. Figure 1(c) displays the *E. coli* K12 genomic fragment (composed of more than 80% coding regions), while Fig. 1(d) shows the human *T*-cell receptor alpha/delta locus ($< 10\%$ coding).

The fundamental difference between the two control sequences is that the first does not possess long-range power-law correlations, while the second does. Therefore a appropriate scaling analysis of the correlation properties should be able to distinguish between them. In Ref.

[1], a "min-max" method was proposed to take into account the "nucleotide heterogeneity." A potential drawback of this method is that it requires the investigator to judge how many local maxima and minima of a landscape to utilize in the analysis. Here we present an alternative method— *"detrended fluctuation analysis" (DFA)*—that is independent of investigator input and permits the detection of long-range correlations embedded in a patchy landscape [such as Fig. 1(b)], and also avoids the spurious detection of apparent long-range correlations that are an artifact of patchiness [Fig. 1(a)].

The DFA method comprises the following steps.

(1) Divide the entire sequence of length N into N/ℓ nonoverlapping boxes, each containing ℓ nucleotides, and define the "local trend" in each box (proportional to the compositional bias in the box) to be the ordinate of a linear least-squares fit for the DNA walk displacement in that box [14].

(2) Define the "detrended walk," denoted by $y_\ell(n)$, as the difference between the original walk $y(n)$ and the local trend. Calculate the variance about the detrended walk for each box, and calculate the average of these variances over all the boxes of size ℓ, denoted $F_d^2(\ell)$ [15].

To illustrate the DFA method, we show in Fig. 2 a 1000-nucleotide subsequence of the DNA walk of bacteriophage λ [GenBank name: LAMCG, 48502 bp (bp denotes base pair)]. Figure 2(a) shows the local trends when this subsequence is partitioned into boxes of size $\ell = 100$ while Fig. 2(b) shows the local trends when the subsequence is partitioned into boxes of size $\ell = 200$. It is apparent by visual inspection that the variance increases with the box size. The dependence of variance on box size gives rise to the scaling properties of the fluctuations.

If only short-range correlations (or no correlations) exist in the nucleotide sequence, then the detrended DNA walk must have the statistical properties of a random walk (unbiased or biased) so $F_d(\ell) \sim \ell^{1/2}$; however, if there is long-range power-law correlation (i.e., no characteristic length scale), then $F_d(\ell) \sim \ell^\alpha$ with $\alpha \neq 1/2$ [16].
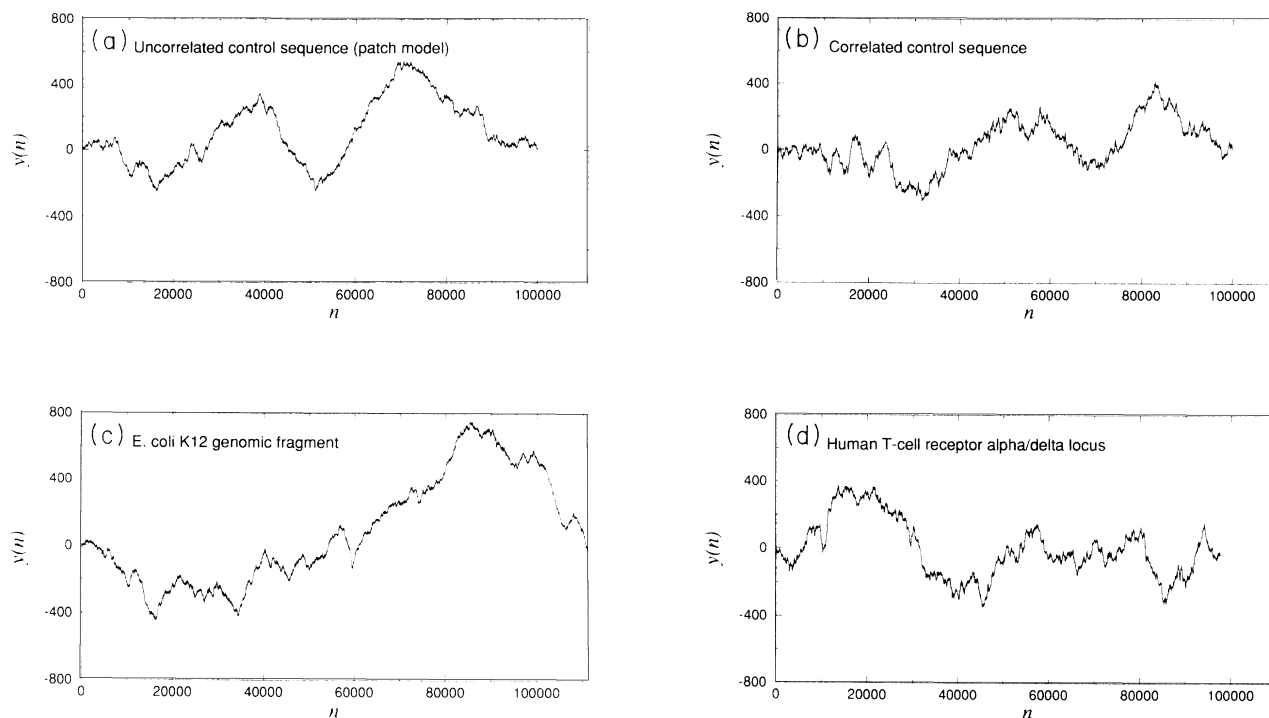
FIG. 1. (a) DNA walk for a control sequence obtained by stitching together biased random walks; the characteristic length for the patches is 2500. (b) DNA walk for a control sequence obtained from building a long-range correlation into a set of 100 000 "nucleotides" which are correlated with power-law exponent $\alpha = 0.61$ using the procedure of Ref. [13]. (c) DNA walk for a genomic fragment containing mostly coding regions [*E. coli* K12 genome, 0–2.4 min region, GenBank name: ECO110K, 111401 bp]. (d) DNA walk for a typical intron-containing chromosomal region of a comparable length (human *T*-cell receptor alpha/delta locus, GenBank name: HUMTCRADCV, 97634 bp). Large subregions ("patches") of uniform overall slope ("strand bias") reflect the mosaic structure. To facilitate the comparison of subtle fluctuations, each landscape is plotted so that the end point has the same vertical displacement as the starting point, i.e., the overall bias has been removed.

Figure 3 shows the results of applying DFA to the DNA walks displayed in Fig. 1. We note that the two types of control sequences are clearly distinguishable by their differing values of the scaling exponent: $\alpha = 0.51$ for the uncorrelated control while $\alpha = 0.61$ for the correlated control. Moreover, the *E. coli* chromosomal sequence of Fig. 1(c), composed primarily of *coding* regions, has the same exponent ($\alpha \sim 0.5$) as the uncorrelated patch model of Fig. 1(a). In contrast, the genomic sequence containing *noncoding* regions of Fig. 1(d) has the same exponent ($\alpha > 0.5$) as the correlated control sequence of Fig. 1(b) [17].

Figure 3 also demonstrates the fact that the DFA method is capable of identifying the characteristic length scale of the biased subregions of the uncorrelated control sequence of Fig. 1(a). The arrow indicates the crossover from $\alpha \approx 0.5$ to larger α values occurring at $\ell \equiv \ell_{\times} \approx 2500$, a value that coincides with the characteristic length scale "built into" the uncorrelated control sequence of Fig. 1(a) [12]. Moreover, the *E. coli* data also exhibit a crossover at roughly the same value of $\ell_{\times}$, suggesting the existence of a characteristic patch size in the *E. coli* nucleotide sequence (probably corresponding to the average length of a protein). We find similar crossover phenomena for other coding sequences. In contrast, no such crossover phenomenon occurs for the correlated control sequence of Fig. 1(b) or for the intron-
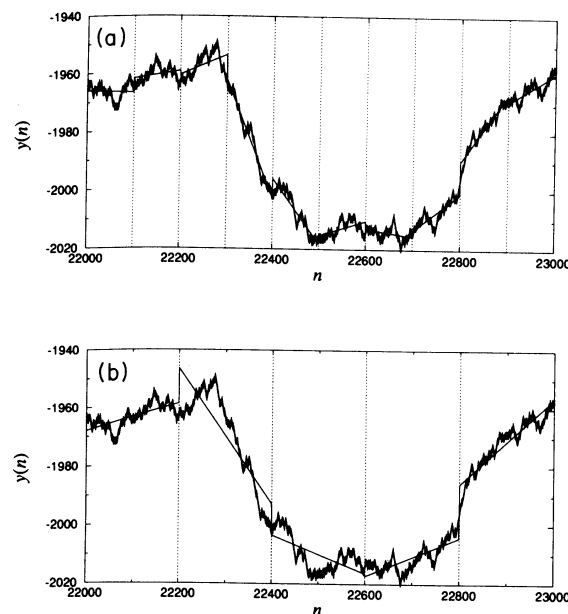

FIG. 2. DNA walk generated by a subsequence of the bacteriophage λ genome. The detrended fluctuation analysis (DFA) is applied in (a) to box size $\ell = 100$, and in (b) to box size $\ell = 200$. Shown in each box is the least-squares fit to the data in that box. One sees that the typical variance for a box in (a) is smaller than for a box in (b).
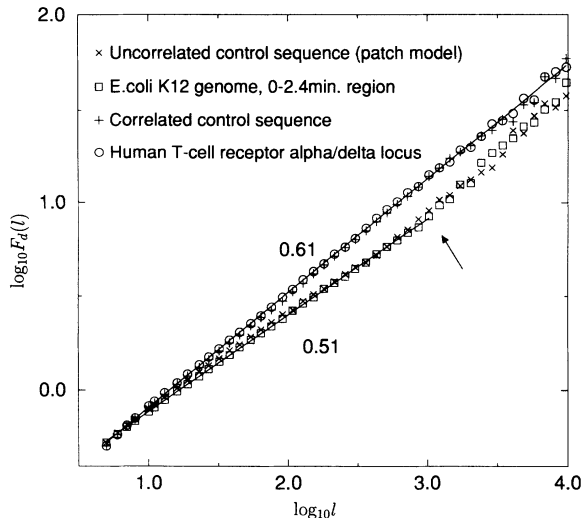
FIG. 3. DFA analysis of the four landscapes shown in Fig. 1. The uncorrelated biased random walk [Fig. 1(a)] ($\times$) is similar to the *E. coli* genomic coding fragment [Fig. 1(c)] ($\square$), while the correlated control sequence [Fig. 1(b)] ($+$) is quite similar to the highly noncoding human T-cell receptor alpha/delta locus [Fig. 1(d)] ($\circ$). The lower solid line, the best fit for *E. coli* data from $\ell = 4$ to 861, has slope 0.51. The upper solid line, the best fit for human data from $\ell = 4$ to 8192, has slope 0.61. The arrow denoting the crossover phenomenon is explained in the text.

containing DNA of Fig. 1(d), corresponding to the existence of correlations at all scales (i.e., no characteristic patch size).

We also utilize the DFA method to study the complete distribution (normalized histogram), $P(y_\ell)$, of the basic quantity $y_\ell(n)$. One quantitative measurement of this distribution is the standard deviation, $F_d(\ell)$. Figure 4 shows the histogram for the actual DNA and control sequences in Fig. 1. The histograms of the coding sequence and uncorrelated control sequence are virtually indistinguishable only when both are rescaled by the *correct* exponent $\alpha = 0.51$ [(Fig. 4(a)]. Moreover, the histograms of the sequence containing noncoding regions and of the correlated control sequence are virtually indistinguishable only when both are rescaled by the *correct* exponent $\alpha = 0.61$ which is appropriate for a correlated sequence [Fig. 4(b)]. This finding demonstrates that the exponent α describes the scaling properties of the entire distribution, not just the standard deviation $F_d(\ell)$.

The DFA method, therefore, allows us to unambiguously differentiate two properties that were previously difficult to distinguish: (i) the value of α, which measures the degree of long-range correlation (indicating the absence of a characteristic length scale), and (ii) the approximate value of $\ell_\times$, the crossover value, indicating the characteristic length scale of patches. Moreover, these results indicate that *patchiness by itself cannot account for long-range power-law correlations found in noncoding regions*—since the uncorrelated patch model does not lead to $\alpha > 1/2$. However, recent model studies demonstrate explicitly how one can find $\alpha > 1/2$ if the patches have no characteristic length scale [18].

A discrepancy between the work of Ref. [10] and the alternative analysis presented here is exemplified in Fig. 5, which reanalyzes the example studied in Ref. [10]: the complete genome for bacteriophage λ consisting primarily of coding sequences. The DNA walk for this sequence (inset) exhibits three prominent patches of different strand bias. The "uncorrected" (nondetrended) fluctuation analysis applied in Ref. [10] shows a slope close to 0.5 only for quite small values of ℓ ($\ell < 20$), and a slope crossing over to a value close to unity for large ℓ. Reference [10] attributed the long-range correlations reported in Refs. [1–4] to crossover phenomena of the sort exhibited in the example shown in the top curve of Fig. 5. However, DFA for the same DNA sequence (bottom curve) reveals that the scaling region with $\alpha \sim 0.5$ actually extends up to 3 decades, thus providing evidence that this heterogeneous sequence resembles the patch model *without* long-range correlations [18]. This crossover behavior, observed in many coding sequences, is readily dis-
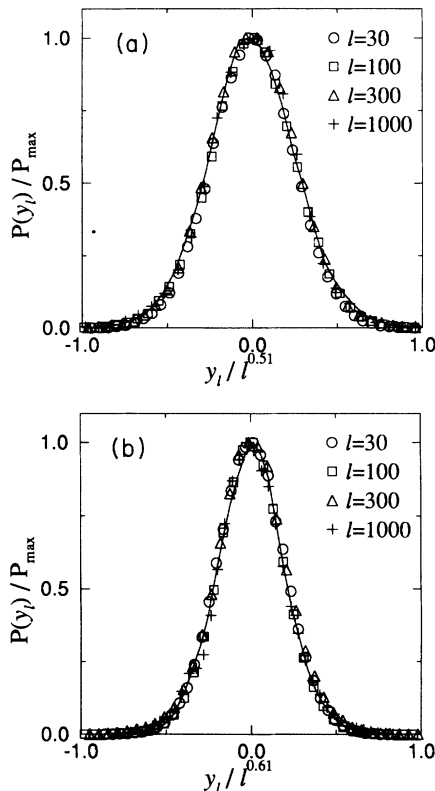

FIG. 4. Histogram of y_ℓ. We rescale the abscissa by ℓ^α and rescale the vertical variable, $P(y_\ell)$, such that the peaks of all histograms are the same ($= 1$). Shown are four different box sizes ℓ. (a) shows the genomic fragment of Fig. 1(c), plotted with $\alpha = 0.51$. Similarly, (b) shows the intron-containing gene of Fig. 1(d) for $\alpha = 0.61$. The solid line in (a) is for the uncorrelated control sequence of Fig. 1(a) with $\ell = 400$, while the solid line in (b) is for the correlated control sequence of Fig. 1(b) with $\ell = 400$. These histograms can be very well fitted by $P(y_\ell) \sim \exp(-y_\ell^\delta/c)$, where c is a constant and $\delta \approx 1.8$ (very close to Gaussian, for which $\delta = 2$). The form of the histogram provides important information for validating different models proposed recently for explaining long-range correlations in noncoding sequences [18,19].
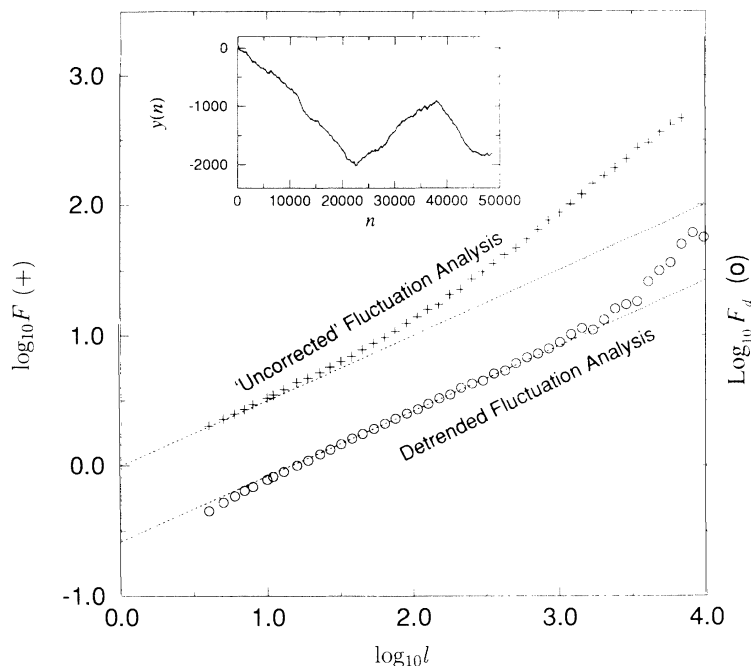
FIG. 5. Comparison of the fluctuation analysis used in Ref. [10] and the DFA presented here. The DNA sequence is for the complete genome of lambda phage, whose DNA walk appears in the inset. The two parallel dotted lines have slope 0.5. A best fit straight line to $F_d(\ell)$ for the interval $\ell = 4$ to 1024 has slope $\alpha = 0.51$.

tinguishable from true power-law correlations observed in noncoding sequences which usually exhibit constant α value (> 0.5) over several decades—more than 3 decades in the case of the data displayed in Fig. 3.

We wish to thank C. DeLisi, F. Sciortino, and T. Vicsek for valuable discussions, and P. Jensen, J. Kertész, W. Li, M. Matsa, and S. M. Ossadnik for a critical reading of the manuscript.

[1] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992).

[2] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).

[3] C. L. Berthelsen, J. A. Glazier, and M. H. Skolnick, Phys. Rev. A **45**, 8902 (1992).

[4] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992); **71**, 1777 (1993); S. V. Buldyrev, A. Goldberger, S. Havlin, C.-K. Peng, F. Sciortino, M. Simons, and H. E. Stanley, *ibid.* **71**, 1776 (1993).

[5] P. J. Munson. R. C. Taylor, and G. S. Michaels, Nature (London) **360**, 636 (1992).

[6] G. Bernardi, B. Olofsson, J. Filipski, M. Zerial, J. Salinas, F. Cuny, M. Meunier-Rotival, and F. Rodier, Science **228**, 953 (1985).

[7] G. A. Churchill, Bull. Math. Biol. **51**, 79 (1989).

[8] J. W. Fickett, D. C. Torney, and D. R. Wolf, Genomics **13**, 1056 (1992).

[9] S. Nee, Nature (London) **357**, 450 (1992).

[10] S. Karlin and V. Brendel, Science **259**, 677 (1993).

[11] Variants of this rule permit studying patchiness arising from other forms of nucleotide heterogeneity. For example, adenine heterogeneity can be studied by forming a walk with 3 steps up for each adenine and 1 step down for a nonadenine.

[12] To be precise, we randomly divide a sequence of length N into k subregions and assign a compositional bias (purine-pyrimidine concentration) to each subregion from a Gaussian distribution of width σ and mean 0.5. The sequences thus generated are patchy by construction and the patch size distribution displays a characteristic length given by N/k. In the case of Fig. 1(a), we divided the whole sequence (10^5 bp) into 40 subregions with length ranging from 68 bp to 11 679 bp. The characteristic length is $10^5/40 = 2500$ bp. The width of the bias distribution is $\sigma = 0.1$. As expected, the landscape of the typical control sequence from this biased random model is highly heterogeneous.

[13] To generate long-range correlated variables, we start with random uncorrelated variables, taken from a uniform distribution. We then form the Fourier transform and multiply by a power in q space. Finally, we Fourier transform back to real space and thereby obtain long-range correlated variables with a correlation function given by a power law. This method was proposed in S. Havlin, R. B. Selinger, M. Schwartz, H. E. Stanley, and A. Bunde, Phys. Rev. Lett. **61**, 1438 (1988).

[14] A more general procedure is to apply the same method except, instead of incrementing the box position by a distance ℓ between each measurement, to increment the position by a distance ℓ/S, where S is a sliding parameter that can be chosen to give optimal statistics.

[15] The precise definition of the *average* variance on length scale ℓ is $F_d^2(\ell) \equiv \frac{1}{N}\sum_{n=1}^{N} y_\ell^2(n)$.

[16] Note that if $\alpha < 1/2$, there exist long-range correlations that reflect an alternation of different nucleotide types, while if $\alpha > 1/2$ the long-range correlations reflect a persistence of the same nucleotide type. For the ideal case of a random sequence, $\alpha = 0.5$. However, since the length of the sequences is not infinite, we must take into account the statistical fluctuations for a finite size sample. A discussion of the effects of the finite length of DNA sequences analyzed appears in C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, M. Simons, and H. E. Stanley, Phys. Rev. E **47**, 3730 (1993). To measure the exponent α, we calculate the slope of $\log_{10} F_d(\ell)$ versus $\log_{10} \ell$ for $\ell \geq 4$, since for very small ℓ we expect deviation (introduced by the detrending algorithm) from power-law behavior. Numerical studies of various types of control sequences show that this deviation is negligible (within the error of statistical fluctuations) if we evaluate the exponent from $\ell \geq 4$.

[17] We have tested the DFA method extensively on DNA sequences from genomic and cDNA sequences which we and others had previously studied by other methods, as well as correlated and uncorrelated control sequences. In all cases, the DFA method gave results as convincing as the examples in the text.

[18] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **47**, 4514 (1993). The DFA method has been successfully used to study changes in fractal complexity with evolution in S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, H. E. Stanley, and M. Simons, Biophys. J. **65**, 2675 (1993).

[19] A. Yu. Grosberg, Y. Rabin, S. Havlin, and A. Nir, Biofisika (Russia) **26**, 1 (1993); Europhys. Lett. **23**, 373 (1993).