Long Range Correlations in DNA: Scaling Properties and Charge Transfer Efficiency

Stephan Roche,[1] Dominique Bicout,[2] Enrique Maciá,[3] and Efim Kats[2,4]
[1]*Commissariat à l'Énergie Atomique, DSM/DRFMC/SPSMS, 17 avenue des Martyrs, 38054 Grenoble, France*
[2]*Laue-Langevin Institute, 6 rue Jules Horowitz, BP 156, F-38042, Grenoble, France*
[3]*Departamento de Fisica de Materiales, Facultad de Fisicas, Universidad Complutense, E-28040 Madrid, Spain*
[4]*L. D. Landau Institute for Theoretical Physics, RAS, Moscow, Russia*
(Received 12 June 2003; published 26 November 2003)

We address the relation between long-range correlations and charge transfer efficiency in aperiodic artificial or genomic DNA sequences. Coherent charge transfer through the highest occupied molecular orbital states of the guanine nucleotide is studied using the transmission approach, and the focus is on how the sequence-dependent backscattering profile can be inferred from correlations between base pairs.

DOI: 10.1103/PhysRevLett.91.228101 PACS numbers: 87.14.Gg, 72.20.Ee, 72.80.Le

During the past few years, the nature of *long range correlations* in DNA sequences has been the subject of intense debate [1–3]. Scale invariant properties in complex genomic sequences with thousands of nucleotides have been investigated, in particular, with wavelet analysis [2], and have been argued to play a crucial role in gene regulation and cell division. Besides, among the many physical, chemical, or biological phenomena that might be inferred from sequence correlations, charge transfer properties deserve particular concern. Indeed, a precise understanding of DNA-mediated charge migration would have a strong impact on the description of damage recognition process and protein binding, or on engineering biological processes [4,5]. The π-stacked array of DNA base pairs (bp) (made up from nucleotides: guanine G, adenine A, cytosine C, thymine T) provides an extended path to convey long-range charge transport although dynamical motions of base pairs, or energetic sequence-dependent heterogeneities, are expected to reduce long-range efficiency. Photoexcitation experiments have unveiled that charge excitations can be transmitted between metallointercalators, preferentially through the guanine highest occupied molecular orbitals (G-HOMO) of the DNA bridge [5,6]. Such experiments and mesoscopic transport measurements on single artificial or genomic DNA sequences contacted between metallic electrodes have also been the subject of controversial debate [7]. While accurate determination of absolute values of conductivity is important, characteristic sequence dependences of charge transport could provide valuable clues to mechanisms and biological functions of transport. Such an issue has up to now been poorly addressed experimentally and theoretically. In that perspective, the possible role of long-range correlations on electronic delocalization has been recently anticipated [8]. In this Letter, the electronic transport properties are proven to be critically related to the nature and range of correlations.

Rescaling coefficients have been introduced as a useful measure of correlations in DNA sequences [1]. It relies on the evaluation of the second moment of the fluctuations of sequence composition. The statistical method consists of constructing a mapping of the nucleotide sequence onto a walk. A DNA walk is initiated from the first to the last nucleotide of the sequence with the rule that the walker steps down $[v(i) = -1]$ if a purine (A, G) occurs at position i along the sequence, whereas the walker steps up $[v(i) = +1]$ if a pyrimidine (T, C) occurs at position i. Given a nucleotide sequence of size N, the net displacement $x(n)$ of the nucleotide walker after n steps is $x(n) = \sum_{i=1}^{n} v(i)$, $1 \le n \le N$. Recently, Hurst's analysis [9] was argued to be more reliable for determining the precise rescaling coefficients [10]. We thus follow the prescription of Hurst's analysis to construct adjusted variables as $X(m, k) = \Delta x(m, k) - \frac{k}{n} \Delta x(m, n)$, $1 \le k \le n$, and define the range $S(m, n)$ for random walks of lengths n as $S(m, n) = \max_{1 \le k \le n}[X(m, k)] - \min_{1 \le k \le n}[X(m, k)]$. Now, the rescaled range function $R(n)$ is defined as [9]

$$R(n) = \frac{\langle S(n) \rangle}{\sigma(n)} \propto n^H, \tag{1}$$

where $\langle S(n) \rangle = \sum_{m=1}^{N-n} S(m, n)/(N - n)$ and $\sigma^2(n)$ is the standard deviation of $v(i)$ over walks of lengths n, and averaged over the entire sequence. The Hurst exponent H of the process is then defined through the scaling in Eq. (1). Interestingly, for short-ranged correlated random walk the exact result for the rescaled range function reads $R(n) = \sqrt{[\pi n/2]} - 1$ [9,11]. Thus, $H = 1/2$ for the ordinary Brownian motion. The existence of power-law behaviors suggests that there is no characteristic length scale associated with properties under consideration. It is clear at the first glance that DNA sequences are unlikely fully characterized by a single scaling exponent. One expects that the scaling behavior be different for different length scales of the sequence; i.e., the rescaling exponent is itself a function of the length scale n. In the case where a characteristic size n_c can be defined, one may postulate that $R(n)$ is still described by the power law in Eq. (1), but with a scale dependent rescaling exponents $H(n)$ such that $H(n) = H_1$ for $1 \le n < n_c$ and $H(n) = H_2$ for $n \ge n_c$.

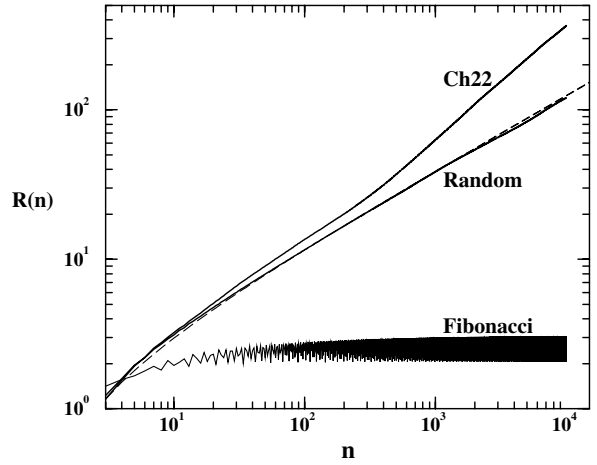

FIG. 1. Rescaled range function $R(n)$ versus n. Dashed line corresponds to $\sqrt{[\pi n/2]} - 1$.

In our study, we consider three sequences: a DNA sequence of the first completely sequenced human chromosome 22 (Ch22) containing about 33.4×10^6 nucleotides entitled NT_{011520} retrieved from the National Center for Biotechnology Information, a random DNA sequence (where A, T, C, G are evenly chosen probability 1/4) and a Fibonacci polyGC quasiperiodic sequence constructed starting from a G nucleotide as seed and following the inflation rule $G \rightarrow GC$ and $C \rightarrow G$. This gives successively G, GC, GCG, GCGGC, GCGGCGCG, GCGGCGCGGCGGC, ..., for sequences of length 1, 2, 3, 5, 8, 13, ..., respectively, such that its characteristic self-similar order introduces correlations on a broad scale range. The ratio [number of G]/[number of C] approaches the golden mean value $(1 + \sqrt{5})/2 \simeq 1.618$ in the limit of an infinite sequence. The random and Fibonacci sequences are used as prototypes of short-range (or uncorrelated) and strongly correlated systems, respectively.

The computed functions $R(n)$ for the three sequences described above are reported in Fig. 1 and values of H are summarized in Table I. It clearly appears from these calculations that the random sequence is indeed uncorrelated following the $\sqrt{[\pi n/2]}$ law, whereas the Fibonacci sequence is strongly correlated with a "ballistic behavior" and correlations in the Ch22 sequence exhibit a power-law behavior with a scaling exponent depending on the length scale. The Ch22 sequence has long-range correlations characterized by Hurst exponents greater than 1/2 (see Table I). Given the huge amount of nucleotides of the Ch22 sequence, the physically relevant question seems rather to address to which extent charge transport can be efficient through the G-HOMO, in comparison with uncorrelated random or quasiperiodic sequences. To have some elements of response, we now turn to the examination of charge transfer properties in these sequences. To this end, we consider an effective tight-binding Hamiltonian describing the energetics of a hole located at nucleotide site n [12,13],

TABLE I. Hurst exponents calculated from data in Fig. 1.

Sequence	N	Purines	$H(n_c = 300)$ H_1	H_2
Ch22	182 617	91 029	0.60	0.75
Random	182 617	91 118	0.50	0.50
Fibonacci	46 368	28 657	0.085	0.011

$$\mathcal{H} = \sum_n \varepsilon_n c_n^\dagger c_n - \sum_n t_0 (c_n^\dagger c_{n+1} + \text{H.c.}), \qquad (2)$$

where $c_n^\dagger$ (c_n) is the creation (annihilation) operator of a hole at site n. The hole site energies ε_n are chosen according to the ionization potentials of respective bases [13], $\varepsilon_A = 8.24$ eV, $\varepsilon_T = 9.14$ eV, $\varepsilon_C = 8.87$ eV, and $\varepsilon_G = 7.75$ eV, while the hopping integral, simulating the $\pi - \pi$ stacking between adjacent nucleotides, is taken as $t_0 = 1$ eV. The DNA sequences are further assumed to be connected to two semi-infinite electrodes whose energies ε_m are adjusted to simulate a resonance with the G-HOMO energy level, $\varepsilon_m = \varepsilon_G$, and with hopping integrals such that $t_m = t_0$. Note that *ab initio* studies suggest that $t_0 \sim 0.1$–0.4 eV [13], but the choice $t_m/t_0 = 1$ reduces backscattering of holes at the contact electrodes and allows for a larger accessible transmission spectrum and a better characterization of DNA's intrinsic conduction [12]. Sites comprised between $[-\infty, 0] \cup [N + 1, +\infty]$ belong to the leads, whereas sites $i = 1, N$ are associated with the sequence of size N under study. The transmission coefficients are computed using the transfer matrix formalism in which the time independent Schrödinger equation is projected into a localized basis by properly accounting for the boundary conditions [14]. Let ψ_n denote the wave function with energy E at site n, we obtain from Eq. (2) the recurrent equation, $\binom{\psi_{N+2}}{\psi_{N+1}} = M_N \binom{\psi_{N+1}}{\psi_N} = M_N \cdots M_1 \binom{\psi_1}{\psi_0}$, where M_n is a 2×2 matrix with elements $M_n(1,1) = (E - \varepsilon_n)/t_{n+1}$, $M_n(1,2) = -t_n/t_{n+1}$, $M_n(2,1) = 1$, and $M_n(2,2) = 0$. The transmission coefficient $T_N(E)$ that gives the fraction of tunneling electrons transmitted through the N-site DNA is related to the Landauer resistance as $(h/2e^2)[1 - T_N(E)]/T_N(E)$, where $h/2e^2$ is the quantum resistance and [14]

$$T_N(E) = \left[4 - \frac{(E-\varepsilon_m)^2}{t_m^2}\right] \Big/ \left\{ -\frac{(E-\varepsilon_m)^2}{t_m^2} (\mathcal{P}_{12}\mathcal{P}_{21} + 1) + \frac{(E-\varepsilon_m)}{t_m} (\mathcal{P}_{11} - \mathcal{P}_{22})(\mathcal{P}_{12} - \mathcal{P}_{21}) + \sum_{i,j=1,2} \mathcal{P}_{ij}^2 + 2 \right\} \qquad (3)$$

with $\mathcal{P} = M_N M_{N-1} \ldots M_1$. For a given energy, $T_N(E)$ reflects the level of backscattering events in the hole transport through the sequence. As metallic leads are adjusted to the G-HOMO energy level, the hole transport will experience a sequence-dependent contribution of backscattering according to the distribution of C, T, and A potential barriers over

the length scale of the sequence. To compare transmission properties of different chains, the behavior of the Lyapunov coefficient, $\gamma_N(E) = \frac{1}{2N}\ln(T_N(E))$, is also calculated. $\gamma_N(E)$ has been extensively investigated to sort out the main features of complex localization patterns [15,16]. For systems with uncorrelated disorder, $\gamma_N(E)$ provides the localization length $\xi(E) = 1/[\lim_{N\to\infty}\gamma_N(E)]$. In the presence of scale invariance properties, the underlying structure of $\gamma_N(E)$ reflects the self-similarity of the spectrum [16].

Following our analysis on correlations, the $T_N(E)$ for the three sequences of Table I have been computed, varying the sequence length. The random and Fibonacci quasiperiodic based sequences are generated starting from the first nucleotide of the sequence up to N bp, while the Ch22-based sequences are constructed by starting from the bp = 15 000 of the full Ch22 sequence and then extracting the first N bp, namely AGGGCATCGCTAACGAGGTCGCCGTCCACA GCATCGCTATCGAGGACACCACACCGTCCA for $N = 60$ bp. Figures 2 and 3 present the comparison of $T_N(E)$ between the quasiperiodic and Ch22 sequences and between uncorrelated random DNA and Ch22 sequences, respectively, with the same number of bp. Lyapunov coefficients for quasiperiodic and Ch22-based sequences are also displayed in Fig. 4.

General trends of Figs. 2 and 3 are that $T_N(E)$ is characterized by an energy spectrum of resonant peaks with high transmission. As the sequence length increases, fewer states will present good transmittivity, due to the progressive fragmentation of the spectrum, although several peaks with high transmission remain at certain energy values, and new ones may appear. For Fibonacci and Ch22-based sequences, these resonant energies are robust enough to persist against backscattering effects due to interspersed bases along the sequence. This point is illustrated in Figs. 2 and 3 where one observes that Fibonacci (Ch22-based sequences) of 180 bp (360 bp) exhibit states with better transmission properties than those present in a 60 bp (300 bp) long sequence. In addition, $\gamma_N(E)$ shown in Fig. 4 illustrates intrinsic properties of the two correlated sequences albeit of different nature. Indeed, the series of main elliptic bumps found in the Fibonacci sequence with 60 bp are reproduced in the 480 bp sequence, which present additional features associated with the partitioning of spectrum. While self-similarity fully characterizes the quasiperiodic sequence, the scaling properties in Ch22 rely on a totally different kind of long-range correlations, with no hints of self-similar patterns.

In contrast, the fragmentation of the spectrum strongly affects the transmittivity of the uncorrelated random sequence. All resonant states (when any) are evenly affected and the corresponding transmission decreases as the sequence length gets longer. From a statistical analysis over many random sequences, it clearly appears that Ch22-based sequences exhibit a much higher charge transfer efficiency over much longer distances in comparison with uncorrelated random sequences.

Nevertheless, to improve our understanding and gain some physical insights about characteristic features exhibited by these sequences, we now focus on quasiperiodic sequences since it has been shown that the global structure of the electronic spectrum of such chains can be obtained in practice by considering very short periodic approximants to infinite quasiperiodic chains [16]. These sequences are characterized by long-range correlations that manifest themselves on electronic properties in terms of power-law localization of eigenstates in the thermodynamic limit or power-law increase of Landauer resistance in finite samples [16]. For this

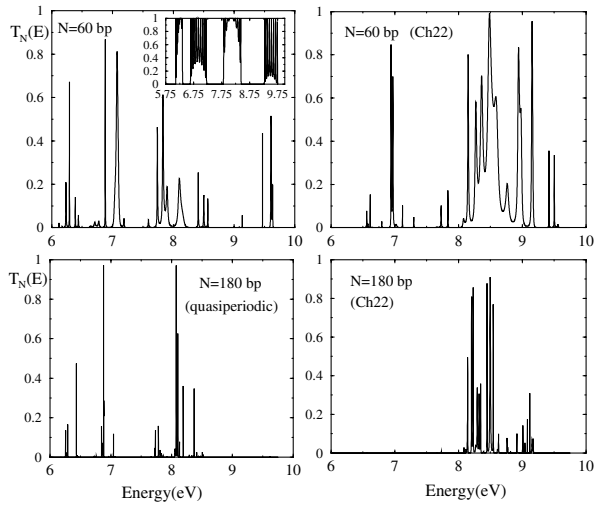

FIG. 2. Transmission coefficient for Fibonacci polyGC quasiperiodic (left frames) and Ch22-based sequences (right frames). Inset: $T_N(E)$ in Eq. (4) for a periodic approximant of length $N = 50$ bp.

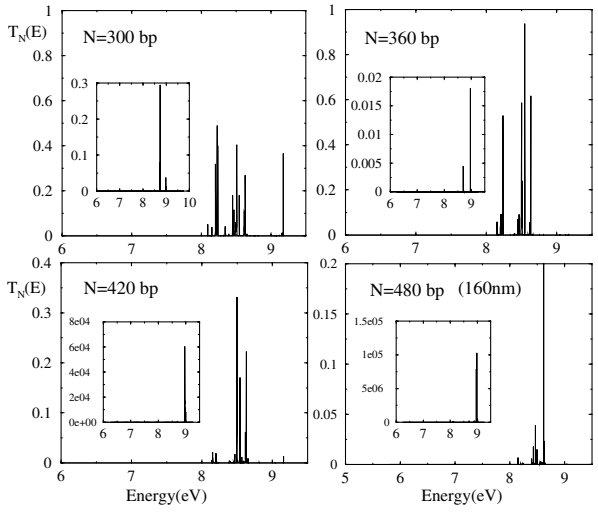

FIG. 3. Transmission coefficient for Ch22-based sequences (main frames) and typical results (over about 50 sequences) for uncorrelated DNA random chains (insets) with the same number of nucleotides.

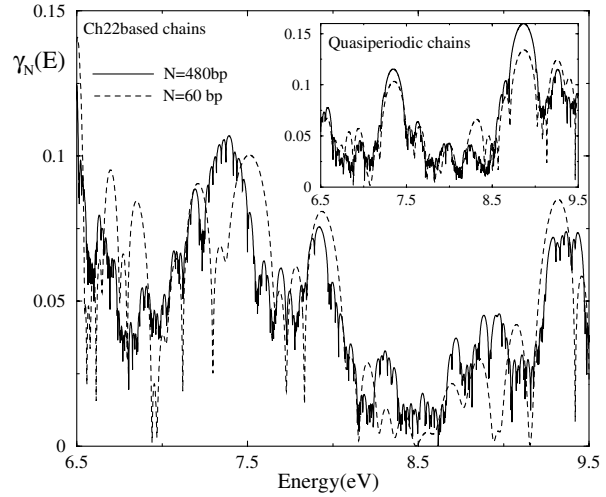

FIG. 4. Lyapunov coefficient for Ch22-based (main frame) and Fibonacci polyGC quasiperiodic sequences (inset).

purpose, we consider a periodic approximant whose unit cell is GCGGC. The corresponding dispersion relation of this approximant is given by, $2t_0^5\cos(5q) = (E-\varepsilon_G)^3 \times (E-\varepsilon_C)^2 - t_0^2(E-\varepsilon_G)(E-\varepsilon_C)(5E-4\varepsilon_G-\varepsilon_C) + t_0^4(5E-3\varepsilon_G-2\varepsilon_C)$. The energy spectrum of the GCGGC approximant is composed of three broad bands (of bandwidth $\simeq 0.5$–0.6 eV) centered at the energies $E_2 = 6.915$ eV, $E_3 = 8.143$ eV, and $E_4 = 9.527$ eV, plus two narrower bands (of bandwidth $\simeq 0.25$ eV) located at the edges of the spectrum at $E_1 = 6.191$ eV and $E_5 = 10.213$ eV. These analytical results allow us to properly assign the different resonant peaks appearing in the spectrum of the transmission coefficient (shown in the inset of Fig. 2) with respect to the four main subbands of the spectral window [5.75, 9.75 eV]. States belonging to the broader central bands around $E_2 = 6.915$ eV and $E_3 = 8.143$ eV turn out to be very robust to the progressive fragmentation of the energy spectrum. Accordingly, one is tempted to conclude from the simple inspection of Fig. 2 (left frames) that these states should exhibit good transport properties even in the thermodynamic limit. To further substantiate such an assertion, we consider in addition the transmission coefficient corresponding to the GCGGC approximant,

$$T_N(E) = [1 + q(x,y)U^2_{(N/5)-1}(w)]^{-1}, \qquad (4)$$

where $x = (E-\varepsilon_C)/2t_0$, $y = (E-\varepsilon_G)/2t_0$, $w = 16x^2y^3 - 16xy^2 - 4yx^2 + 3y + 2x$, the $U_{n-1}(w)$ is a Chebyshev polynomial of the second kind, and $q(x,y) \equiv A^2/(1-y^2) + B^2 - 1$ with $A \equiv -24xy^3 - 16x^2y^2 + 6xy + 2x^2 + 32x^2y^4 + 4y^4 + y^2$ and $B \equiv 32x^2y^3 - 8x^2y - 24xy^2 + 4y^3 + 3y + 2x$. The resonance condition then reads $q(x,y)U^2_{(N/5)-1}(w) = 0$, while the condition $q(x,y) \equiv 0$ yields $E_l = 4.317$ eV (which does not belong to the spectrum) and $E_u = 10.158$ eV (located near the center of the uppermost band, which is not included in our spectral window). On the other hand, the roots of the Chebyshev polynomial label a full transmission peak series according to the relationship $w = \cos(5k\pi/N)$ with $k = 0, \ldots, N$. This is illustrated in the inset of Fig. 2 (top left) where one observes oscillations in the energy dependence of the transmission curve for a sequence GCGGC with 10 units. By a deeper analysis, we find that Fibonacci quasiperiodic sequences as long as 160 nm, i.e., $\sim$450 bp, will still allow for nearly resonant transmission around two specific energies $E_2 \simeq 6.9$ eV and $E_3 \simeq 8.1$ eV.

In summary, long-range correlations in aperiodic DNA sequences seem to induce coherent charge transfer over longer length scales, when compared to uncorrelated sequences. Such a feature has been illustrated, in particular, in Chromosome 22-based sequences. Given that the nature of long-range correlations differs in coding versus noncoding regions of a given genomic DNA [3], a more systematic study of charge transport should, however, be undertaken.

[1] C. K. Peng *et al.*, Nature (London) **356**, 168 (1992); W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).
[2] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992); A. Arneodo *et al.*, Phys. Rev. Lett. **74**, 3293 (1995); B. Audit *et al.*, Phys. Rev. Lett. **86**, 2471 (2001).
[3] S. V. Buldyrev *et al.*, Phys. Rev. E **51**, 5084 (1995); H. Herzel and I. Grosse, Phys. Rev. E **55**, 800 (1997); D. Holste and I. Grosse, Phys. Rev. E **67**, 061913 (2003).
[4] E. Braun *et al.*, Nature (London) **391**, 775 (1999).
[5] C. Treadway, M. G. Hill, and J. K. Barton, Chem. Phys. **281**, 409 (2002).
[6] M. A. Ratner, Nature (London) **397**, 480 (1999); D. J. Bicout and E. Kats, Phys. Lett. A **300**, 479 (2002).
[7] M. Di Ventra and M. Zwolak, in *Encyclopedia of Nanoscience and Nanotechnology*, edited by H. S. Nalwa (American Scientific Publishers, San Francisco, 2003).
[8] P. Carpena *et al.*, Nature (London) **418**, 955 (2002); **421**, 764 (2003).
[9] H. E. Hurst, Proc. Am. Soc. Civ. Eng. **76**, 1 (1950).
[10] A. Montagnini *et al.*, Phys. Lett. A **244**, 237 (1998).
[11] W. Feller, Ann. Math. Stat. **22**, 427 (1951).
[12] Y. A. Berlin, A. L. Burin, and M. A. Ratner, Superlattices Microstruct. **28**, 241 (2000); S. Roche, Phys. Rev. Lett. **91**, 108101 (2003).
[13] H. Sugiyama and I. Saito, J. Am. Chem. Soc. **118**, 7063 (1996); A. Voityuk *et al.*, J. Chem. Phys. **114**, 5614 (2001).
[14] A. Stone *et al.*, Phys. Rev. B **24**, 5583 (1981).
[15] J. L. Pichard, J. Phys. C **19**, 1519 (1986).
[16] E. Maciá *et al.*, Phys. Rev. B **49**, 9503 (1994).