Violation of the single-parameter scaling hypothesis in human chromosome 22 with charge transfer models

Ai-Min Guo and Shi-Jie Xiong
Department of Physics and National Laboratory of Solid State Microstructures, Nanjing University, Nanjing 210093, China
(Received 1 April 2008; revised manuscript received 15 December 2008; published 24 April 2009)

We investigate transport properties of DNA sequences in human chromosome 22 and compare the results with those of a random artificial DNA sequence based on the single- and double-stranded charge transfer models. The statistical quantities, including the Hurst exponent, the distribution of Lyapunov exponent (LE), the central moments, and the scaling parameter, are numerically calculated by using the transfer-matrix approach. It is found that the existence of satellite DNA segments in human chromosome 22 could result in deviations from usual Gaussian distribution of LE. Our results suggest that the presence of the satellite DNA segments, together with the long-range correlations and the base-pairing correlations could lead to the violation of single-parameter scaling hypothesis which holds for the random artificial DNA sequence although the behaviors of the averaged LEs for both DNA sequences are similar. This provides a viewpoint to analyze differences between the genomic DNA sequences and the nonliving random ones on the basis of localization properties of wave functions in the sequences.

DOI: 10.1103/PhysRevE.79.041924 PACS number(s): 87.14.G−, 72.15.Rn, 05.40.−a

I. INTRODUCTION

The study of statistical patterns in DNA molecules has attracted considerable attention among the physics and biology communities [1–4], owing to its envisioned impact on nuclear metabolism regulation and on evolution of life. Scale-invariance properties in genomic sequences containing thousands of nucleotides are extensively investigated, especially based on mapping DNA sequences into numerical quantities and calculating the autocorrelation function. The obtained results may be used to characterize and graphically represent the genetic information stored in DNA [2]. These algorithms include the power spectrum analysis [5,12], wavelet transform technique [6,7], detrended fluctuation analysis (DFA) [8,9], and entropy-based approach [10,11]. Additionally, the long-range correlations in DNA sequences are believed to play a crucial role in the positioning of nucleosomes [12] and in promoting the long-range charge migration [13–16]. While a characteristic sequence dependence of charge transport is important, the statistical properties of quantities related to the charge transport in DNA may provide deeper understanding of functioning of some DNA segments or the whole sequence in biological processes, such as DNA mutations [17], and should deserve particular concern.

The DNA molecules, which are made up from four nucleotides guanine (G), adenine (A), thymine (T), and cytosine (C), can be viewed as a one-dimensional (1D) chain. In this way the motion of electrons (holes) in single-stranded DNA (ssDNA) can be incorporated into an effective tight-binding Hamiltonian [18],

$$\mathcal{H} = \sum_i \varepsilon_i c_i^\dagger c_i - \sum_i t_{i,i+1}(c_i^\dagger c_{i+1} + c_{i+1}^\dagger c_i), \tag{1}$$

where each lattice site represents a nucleobase of the chain. Here $c_i^\dagger$ (c_i) is the creation (annihilation) operator of a hole at the ith site, ε_i is the on-site energy, and $t_{i,i+1}$ is the nearest-neighbor hopping integral.

For 1D disordered systems the single-parameter scaling (SPS) analysis is widely applied to describe the statistics of transport quantities, including the conductance and the Lyapunov exponent (LE). For instance, the SPS hypothesis suggests that the probability distribution of the LE is Gaussian, and in the regime of strong localization where the system size is larger than the localization length, the whole distribution can be determined by only one parameter, and the scaling parameter τ satisfies

$$\tau = \sigma^2 n/\gamma = 1, \tag{2}$$

with σ and γ being the standard deviation and the mean value of the distribution, respectively, and n being the system size. The numerical calculations of low-dimensional Anderson models exhibited excellent agreement with SPS [19–22], while other studies reported the violation of SPS in systems with extremely strong or weak disorder [23,24]. The issue of validity of SPS relation (2) in genomic or artificial DNA sequences is important but, to our knowledge, has so far been rarely addressed theoretically.

The main objective of this paper is to perform an analysis of transport properties of DNA sequences in the first completely sequenced human chromosome 22 (Chr22) based on the single- and double-stranded charge transfer models [25], and to illustrate whether the SPS hypothesis is still valid for description of the transport properties in genomic sequences. To this purpose, we numerically investigate the statistics of DNA segments in the Chr22 sequence, taking into account both the single- and double-stranded structures, and compare the results with those of a random GATC sequence with identical probability 1/4 for each nucleobase by using the strong-weak (SW) mapping rule and the transfer-matrix approach. The statistical quantities, such as the Hurst exponent, the LE distribution, the central moments, and the scaling parameter, are calculated with a large ensemble of realizations. For genomic sequence, each realization is a segment with length n randomly and uniformly selected from the re-

alistic Chr22 sequence. For a DNA segment, the length n is in units of base-pair spacing (3.4 Å). We come to the conclusion that the statistics of transport properties of the Chr22 sequence are obviously different from that of the random artificial DNA sequence, and the underlying physics is related to the presence of the satellite DNA segments and the long-range correlations in the Chr22 sequence. And the base-pairing correlations can also play a significant role in the statistical properties of both DNA sequences.

The paper is organized as follows. In Sec. II, the Hurst exponent is introduced to measure the nucleotide correlations in the Chr22 and random DNA sequences. In Sec. III the statistical properties of the LE of ssDNA are studied, while in Sec. IV the statistical properties of the LE of double-stranded DNA (dsDNA) are investigated. Finally, the results are summarized in Sec. V.

II. HURST EXPONENT

Recently, the Hurst exponent is introduced to measure the nucleotide correlations in DNA sequences [7,14]. The statistical method is to construct a DNA walk based on a mapping rule. Here the SW rule is employed [4,8], i.e., the walker steps up $[u(j)=+1]$ if a guanine or cytosine (G, C) occurs at the jth position, whereas the walker steps down $[u(j)=-1]$ if an adenine or thymine (A, T) occurs at the jth position. After n steps the net displacement of the DNA walk is $D(n)=\Sigma_{j=1}^{n}u(j)$, and the variance of the walk is $\sigma^2(n)=1-D^2(n)/n^2$. Following the prescription of Hurst's analysis the rescaled variables are defined as $X(k,n)=D(k)-kD(n)/n$, and the range $S(n)$ of the DNA walk is expressed as $S(n)=\max[X(k,n)]-\min[X(k,n)]$ for $1\leq k\leq n$. To avoid spurious effects due to a particular configuration, the rescaled range function $R(n)$ is written as

$$R(n)=\langle S(n)\rangle/\langle\sigma(n)\rangle. \tag{3}$$

The Hurst exponent H is then defined as the slope of linear fitting of $\ln[R(n)]$ versus $\ln(n)$. For uncorrelated random walk the rescaled range function reads $R(n)=\sqrt{\pi n/2}-1$ with $H=0.5$. It is expected that the Hurst exponent should be different for different length scales of the DNA sequence [14,16,26]. In this perspective, H can be obtained through

$$H(n)=\{\ln[R(n)]-\ln[R(n-h)]\}/[\ln(n)-\ln(n-h)], \tag{4}$$

with h being the step length.

In Fig. 1, we plot the rescaled range function and the Hurst exponent for the Chr22 and random DNA sequences, as a function of the sequence length. It becomes clear that the random DNA sequence is uncorrelated and exactly follows the power-law behavior with $H=0.5$ albeit of small fluctuations (Fig. 1, inset). While for the Chr22 sequence, it has long-range correlations characterized with a persistent behavior ($H>0.5$), and H is strongly length scale dependent and exhibits heterogeneities of several characteristic lengths that is consistent with previous calculations using the DFA [9,27]. We note that the Hurst exponent will be somewhat underestimated for relatively short DNA sequences [28], and the actual correlations should be stronger for longer DNA sequences. Therefore, one may expect that the statistical properties of the LEs will be different between the Chr22 and random DNA sequences, and SPS relation (2) could be violated for the former but retained for the latter.

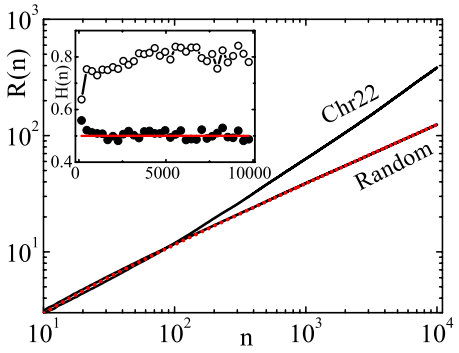

FIG. 1. (Color online) Rescaled range function $R(n)$ versus length n. The red dotted line, which almost coincides with the curve for the random DNA sequence, refers to $R(n)=\sqrt{\pi n/2}-1$. Inset: the Hurst exponent $H(n)$ versus n for the Chr22 (○) and random (●) DNA sequences. Here we adopt $h=300$. The red solid line corresponds to $H=0.5$. The results are averaged over 5000 realizations.

III. STATISTICAL PROPERTIES OF SINGLE-STRANDED DNA

Now we turn to calculate the statistical properties of the LEs of ssChr22 and single-stranded random (ssRandom) DNA molecules. The on-site energies are taken as $\varepsilon_G=7.75$ eV, $\varepsilon_A=8.24$ eV, $\varepsilon_T=9.14$ eV, and $\varepsilon_C=8.87$ eV throughout the paper [29], while the coupling is set to $t=1$ eV. We emphasize that for ssDNA the results presented in the following also hold for more realistic coupling [30,31], e.g., $t=0.4$ and 0.1 eV. For ssDNA the LE, as the inverse of the localization length l_{loc}, is written as

$$\gamma_n(E)=\lim_{n\to\infty}\frac{1}{2n}\ln[\mathrm{Tr}(M_nM_n^{\dagger})], \tag{5}$$

with the transfer matrix

$$M_n=\prod_{i=n}^{1}\begin{pmatrix}\frac{\varepsilon_i-E}{t} & -1\\ 1 & 0\end{pmatrix}.$$

In Fig. 2, we show the energy-dependent LEs for the ssChr22 and ssRandom DNA sequences. From the calculation,

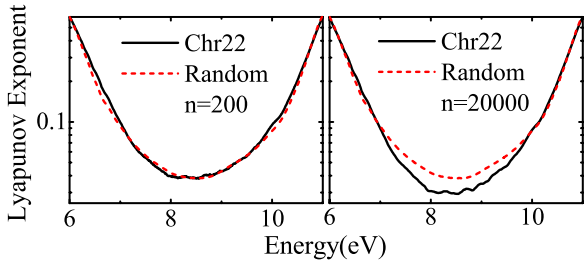

FIG. 2. (Color online) Energy-dependent LEs for the ssRandom and ssChr22 sequences with $n=200$ and 20 000. The results are averaged over 2×10^5 realizations.

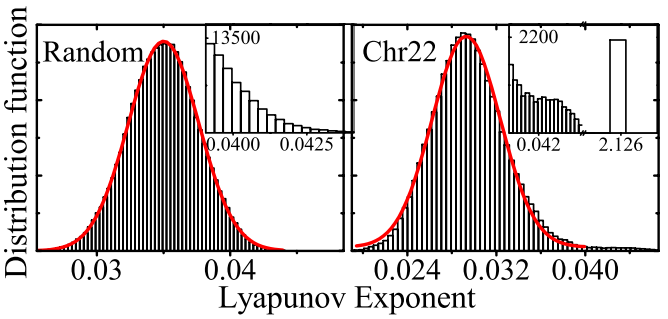

FIG. 3. (Color online) Distribution functions of the LEs for the ssRandom and ssChr22 sequences with $n=5000$. The red solid lines are fitting curves of the Gaussian distribution (the fitting curve of the ssChr22 sequence is for the DNA segments of which $\gamma_n \leq 0.04$), with the mean value and the variance determined by the numerical data. The insets show the enlarged views in the region from 0.0391 to 0.0441 (left panel) and from 0.0388 to 2.13 (right panel). The number of samples is 10^6.

the LEs of both ssDNA sequences do not scale with the size of the system and remain a constant for a certain energy, indicating the localization of states over the whole energy spectrum. Note that for $n=200$ the LE of the ssChr22 sequence is almost identical to that of the ssRandom sequence, whereas for $n=20000$ the LE of the ssChr22 sequence is slightly smaller due to its stronger long-range correlations in larger length scales as mentioned above. However, such correlations in the Chr22 sequence are not sufficient to support the localization-delocalization transition, and all states remain localized with the maximum localization length $l_{\mathrm{loc}} \sim 30$.

In what follows we will discuss the statistical properties of the LEs of DNA sequences. To avoid spurious effects our results are determined over an ensemble of 10^6 realizations. For the Chr22 sequence, the samples are randomly and uniformly selected from the representative segment entitled NT_{011520} [25]. Here we focus on the energy level at the band center [20–23], i.e., $E=8.5$ eV.

In Fig. 3, we present the statistical distribution of the LEs and the fitting Gaussian curves for the ssChr22 and ssRandom DNA sequences (we stress that the fitting Gaussian curve of the ssChr22 sequence is for the DNA segments of which $\gamma_n \leq 0.04$). As we can see, there exists a very long tail (the maximum of γ_n reaches 2.13) in the LE distribution of the ssChr22 sequence (Fig. 3, right panel, inset). These satellite DNA segments of which the LEs are far from the mean value γ, approximately locating in the region from the 71 74 070th to the 72 02 601th nucleobase of the NT_{011520} segment, will drastically affect the statistical properties of the LEs as discussed below and should implicate some special biological information. In order to quantitatively measure the deviation of the statistics from the Gaussian distribution, the corresponding central moments,

$$\mu_k = \langle (\gamma_n - \langle \gamma_n \rangle)^k \rangle, \tag{6}$$

are calculated. To quantitatively describe the deviation from the Gaussian distribution, we adopt the normalized central moments defined as $\tilde{\mu}_k = \mu_k / \mu_2^{k/2}$. For a Gaussian distribution $\tilde{\mu}_k = (k-1)!!$ when k is even and $\tilde{\mu}_k = 0$ when k is odd. In Table I we list the second central moment μ_2, the skewness κ, and the normalized central moments $\tilde{\mu}_k$ for the ssDNA and dsDNA sequences. It clearly appears that the normalized central moments with even orders for the ssRandom sequence comply with $\tilde{\mu}_k = (k-1)!!$ within the numerical accuracy and the skewness is -0.0172, suggesting a Gaussian distribution as expected. The normalized central moments of the ssChr22 sequence, however, are much larger and increase with increasing the order, and the skewness is 19.8, exhibiting a significant deviation from a Gaussian distribution. This is mainly attributed to the satellite DNA segments because the skewness is reduced to 2.04 by removing them from the Chr22 sequence. The LE distribution of the wssChr22 sequence, which denotes the genomic sequence by removing the satellite DNA segments from the Chr22 sequence, also exhibits a remarkable deviation from a Gaussian distribution.

To further illustrate how the LE distribution is deviated to Gaussian statistics, in Fig. 4 we plot the normalized central moments as a function of the sequence length. For the ssRandom sequence, the second central moment roughly follows the power-law relation $\mu_2 \propto n^{-0.989}$ (Fig. 4, top panel, inset), consistent with the central limit theorem. And the normalized central moments $\tilde{\mu}_4$ and $\tilde{\mu}_6$ are length independent, and are approximated to $\tilde{\mu}_4 \approx 3.00$, $\tilde{\mu}_6 \approx 15.0$, i.e., $\mu_4 \approx 3.00\mu_2^2$ and $\mu_6 \approx 15.0\mu_2^3$. This further suggests a good Gaussian distribution for the ssRandom sequence. In contrast, the second central moment of the ssChr22 sequence is

TABLE I. The second central moment μ_2, the skewness κ, and the normalized central moments $\tilde{\mu}_k$ for the ssDNA and dsDNA sequences with the parameters used in Figs. 3 and 7, respectively. The error bars are about 1 or 2 orders of magnitude smaller than the size of the normalized central moments for the DNA sequences. Here wssChr22 (wdsChr22) stands for the genomic sequence by removing the satellite DNA segments from the Chr22 sequence.

	μ_2	$\tilde{\mu}_3(\kappa)$	$\tilde{\mu}_4$	$\tilde{\mu}_5$	$\tilde{\mu}_6$	$\tilde{\mu}_7$	$\tilde{\mu}_8$
ssChr22	1.05×10^{-2}	1.98×10^{1}	3.98×10^{2}	8.03×10^{3}	1.63×10^{5}	3.30×10^{6}	6.71×10^{7}
dsChr22	8.00×10^{-4}	-1.61×10^{1}	3.03×10^{2}	-5.73×10^{3}	1.09×10^{5}	-2.06×10^{6}	3.93×10^{7}
wssChr22	1.54×10^{-5}	2.04	1.67×10^{1}	1.27×10^{2}	1.14×10^{3}	1.06×10^{4}	1.03×10^{5}
wdsChr22	1.03×10^{-4}	5.14×10^{-2}	3.05	1.28	1.91×10^{1}	3.42×10^{1}	2.53×10^{2}
ssRandom	7.04×10^{-6}	-1.72×10^{-2}	3.01	-2.12×10^{-1}	1.51×10^{1}	-2.06	1.05×10^{2}
dsRandom	9.67×10^{-6}	1.68×10^{-2}	3.02	2.54×10^{-1}	1.54×10^{1}	3.94	1.12×10^{2}

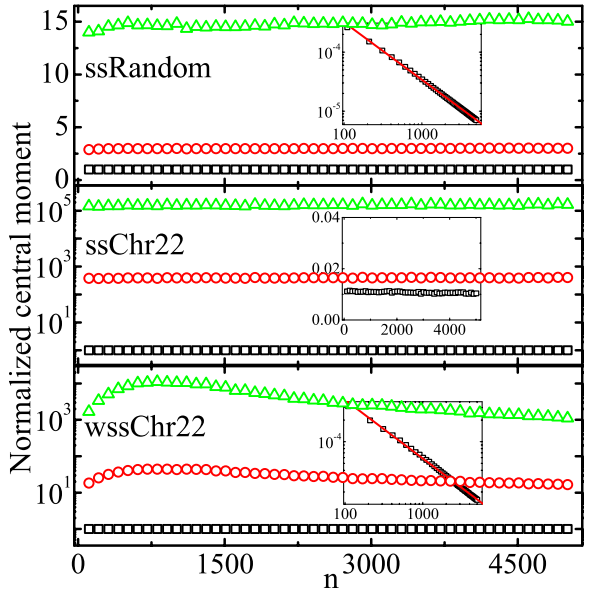

FIG. 4. (Color online) First three normalized central moments $\tilde{\mu}_k$ with even orders versus length n for the ssRandom, ssChr22, and wssChr22 sequences. The symbols from lower to upper (□, ○, △) correspond to $\tilde{\mu}_2$, $\tilde{\mu}_4$, and $\tilde{\mu}_6$, respectively. Inset: μ_2 versus n for the ssRandom, ssChr22, and wssChr22 sequences. The red solid lines in top and bottom panels are linear fitting curves with slope −0.989 and −0.815, respectively.

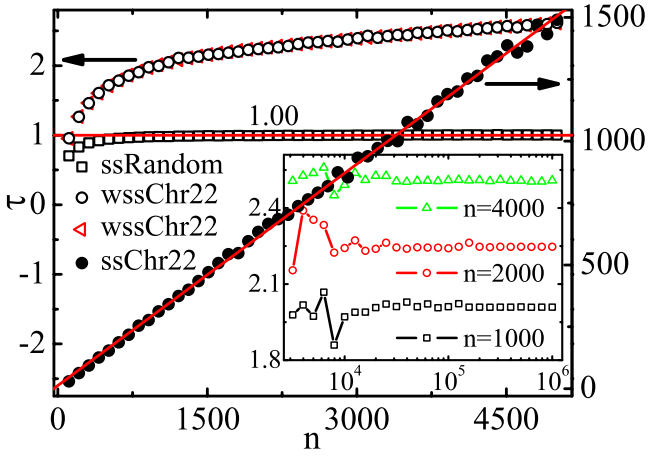

FIG. 5. (Color online) Scaling parameter τ versus length n for the ssRandom, ssChr22, and wssChr22 sequences. The symbol ◁ corresponds to the scaling behavior of the wssChr22 sequence by employing the sliding window strategy to extract the DNA samples from the Chr22 sequence. The red solid lines are fitting curves (see text). The number of samples is 10^6. Inset: τ versus the number of samples for the wssChr22 sequence by using the sliding window strategy. The symbols □, ○, and △ correspond to the scaling behavior of the wssChr22 sequence with the length $n=1000$, 2000, and 4000, respectively.

irrespective of the length (Fig. 4, middle panel, inset), and we roughly have $\tilde{\mu}_4 \approx 394$ and $\tilde{\mu}_6 \approx 1.58 \times 10^5$. This is mainly related to the satellite DNA segments because the dependence of the second central moment of the wssChr22 sequence on the length can be approximately expressed as $\mu_2 \propto n^{-0.815}$ (Fig. 4, bottom panel, inset), and the normalized central moments $\tilde{\mu}_4$ and $\tilde{\mu}_6$ become much smaller although they seem to rely on the length. These imply that the LE distributions of the ssChr22 and wssChr22 sequences are non-Gaussian. Accordingly, the scaling parameter, which relates the mean value and the variance of the LE to the system size, is nearly length independent for the ssRandom sequence and is proportional to the length for the ssChr22 sequence, as illustrated in Fig. 5. It can be seen that the scaling parameter of the ssRandom sequence satisfies $\tau \approx 1.00$ for $n \gg l_{\mathrm{loc}}$, in good agreement with the SPS hypothesis. On the contrary, the scaling parameter of the ssChr22 sequence is proportional to the length with slope 0.299, suggesting the violation of the SPS hypothesis. In the absence of the satellite DNA segments, the scaling parameter of the wssChr22 sequence is much smaller, and its dependence on the length can be written as $\tau \propto n^{0.177}$, indicating the violation of the SPS hypothesis. These results seem to suggest that the satellite DNA segments and the long-range correlations play an important role in the statistics of transport properties in the Chr22 sequence. We have also calculated the scaling behavior of the wssChr22 sequence by using the sliding window strategy to extract the DNA samples from the Chr22 sequence. As we can see, there is no evident difference between the scaling behavior of the wssChr22 sequence by using the two different strategies when the number of samples is large enough, e.g., 10^6 (Fig. 5, main frame). The scaling parameter will fluctuate around a fixed value when the number of samples is not very large and will approach the fixed value when the number of samples is sufficiently large (Fig. 5, inset).

IV. STATISTICAL PROPERTIES OF DOUBLE-STRANDED DNA

A straightforward generalization of model (1), accounting for the double-stranded structure of DNA molecules, is the two-leg ladder model [32–39]. The intrachain and interchain couplings are set to $t=0.4$ eV and $\lambda_{\mathrm{GC}}=\lambda_{\mathrm{AT}}=0.9$ eV [31], respectively. For dsDNA the LE can be calculated by using the standard method of the Gram-Schmidt reorthonormalization, after every ten steps of multiplication of transfer matrices [40]. Here the smallest LE is considered since it is the most physically significant quantity and its inverse is the localization length.

The energy-dependent LEs for double-stranded Chr22 (dsChr22) and dsRandom DNA sequences are plotted in Fig. 6. It clearly shows that the LEs of both dsDNA sequences never vanish within the whole energy spectrum for $t=0.4$ eV and $\lambda=0.9$ eV (Fig. 6, main frame) as well as for more realistic coupling (Fig. 6, inset), consistent with previous works [33,38]. From the calculation, the LEs of both dsDNA sequences are also length independent. Therefore, one can make conclusion that the intrinsic DNA correlations, arising from the base-pairing (G pairs with C, while A pairs with T), are not sufficient to delocalize electronic states in long DNA sequences even if they are associated with the additional nucleotide correlations in one strand. However, the base-pairing correlations in DNA molecules will somewhat enhance the localization length. The maximum localization length of the dsRandom sequence is $l_{\mathrm{loc}} \sim 54$, which

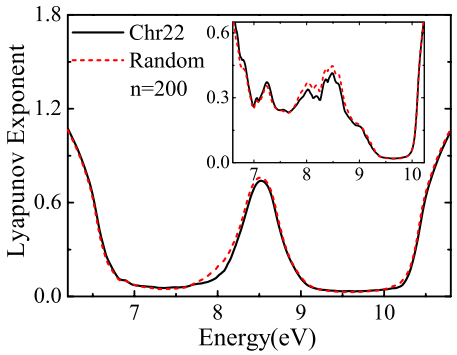

FIG. 6. (Color online) Energy-dependence of LEs for the dsRandom and dsChr22 sequences with $n=200$ for $t=0.4$ eV, $\lambda_{GC}=\lambda_{AT}=0.9$ eV (main frame) and for more realistic coupling $t=0.4$ eV, $\lambda_{GC}=0.9$ eV, and $\lambda_{AT}=0.34$ eV [30,31] (inset). The results are averaged over 2×10^5 realizations. The LEs of the dsChr22 and dsRandom sequences are also length independent within the numerical accuracy.

is comparable to the size of samples used in several experiments [41,42].

In Fig. 7, we show the statistical distribution of the LEs and the fitting Gaussian curves for the dsChr22 and dsRandom sequences. Because of the satellite DNA segments mentioned above, there also exists a long tail (the minimum of γ_n reaches 0.155) in the LE distribution of the dsChr22 sequence (Fig. 7, inset). Notwithstanding, on the one hand, we note that the position of the tail is changed to locate in the left side of the LE distribution, that is due to the base-pairing correlations. On the other hand, the influence of the tail will be weaker on the statistics of the LE for the dsChr22 sequence since the difference of the LE of the satellite DNA segments from the mean value γ is much smaller than that of the ssChr22 sequence. For instance, Fig. 7 suggests that the deviation of the LE distribution to Gaussian statistics observed for the dsChr22 sequence is rather small, and the normalized central moments are much smaller as compared with that of the ssChr22 sequence (Table I). However, from Table I it clearly appears that the absolute value of the normalized central moments of the dsChr22 sequence satisfies $|\tilde{\mu}_k| \gg (k-1)!!$, and its skewness is -16.1, suggesting a strong deviation from the Gaussian statistics. This is mainly

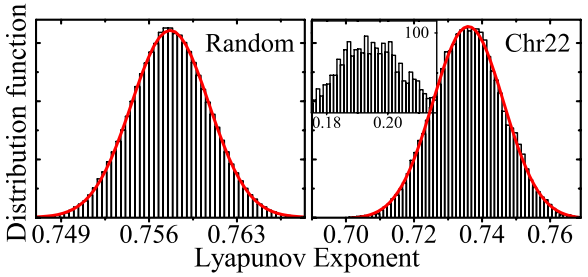

FIG. 7. (Color online) Distribution functions of the LEs for the dsRandom and dsChr22 sequences with $n=5000$. The red solid lines are fitting curves of the Gaussian distribution, with the mean value and the variance determined by the numerical data. The inset shows the enlarged view in the region from 0.175 to 0.216 for the dsChr22 sequence. The number of samples is 10^6.

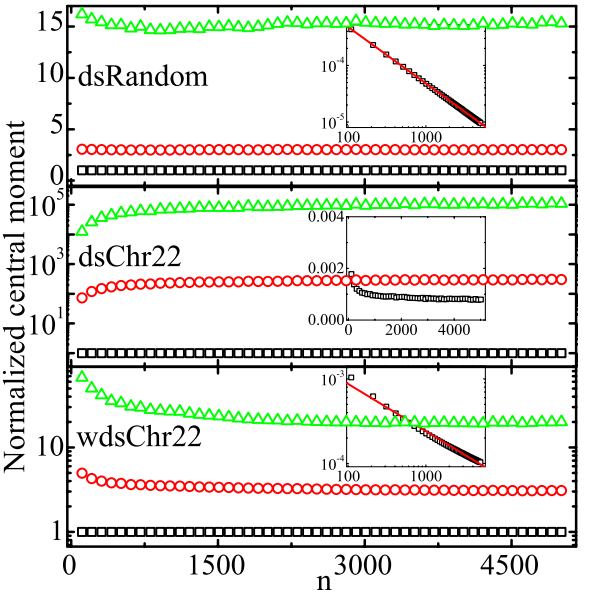

FIG. 8. (Color online) First three normalized central moments $\tilde{\mu}_k$ with even orders versus length n for the dsRandom, dsChr22, and wdsChr22 sequences. The symbols from lower to upper (□, ○, △) correspond to $\tilde{\mu}_2$, $\tilde{\mu}_4$, and $\tilde{\mu}_6$, respectively. Inset: μ_2 versus n for the dsRandom, dsChr22, and wdsChr22 sequences. The red solid lines in top and bottom panels are linear fitting curves with slope -1.00 and -0.559, respectively.

attributed to the satellite DNA segments because the LE distribution of the wdsChr22 sequence, which denotes the genomic sequence by removing the satellite DNA segments from the Chr22 sequence, seems to be Gaussian-type. While for the dsRandom sequence, the normalized central moments with even orders comply with the relation $\tilde{\mu}_k=(k-1)!!$ within the numerical accuracy, and the skewness is 0.0168, suggesting a Gaussian distribution.

The normalized central moments, as a function of the sequence length, are reported in Fig. 8. It can be seen that the second central moment exactly follows the power-law relation $\mu_2 \propto n^{-1.00}$ for the dsRandom sequence (Fig. 8, top panel, inset), and the normalized central moments satisfy $\tilde{\mu}_4 \approx 3.00$ and $\tilde{\mu}_6 \approx 15.0$, and thus $\mu_k \propto n^{-k/2}$ for $k=2$, 4 and 6, further suggesting a good Gaussian distribution. While for the dsChr22 sequence, the second central moment will slowly decrease with increasing the length and saturate at large length scale (Fig. 8, middle panel, inset), and we approximately have $\tilde{\mu}_4 \approx 290$ and $\mu_6 \approx 9.86\times10^4$. In the absence of the satellite DNA segments, the dependence of the second central moment on the length observed for the wdsChr22 sequence can be approximated to $\mu_2 \propto n^{-0.559}$ (Fig. 8, bottom panel, inset), and the normalized central moments $\tilde{\mu}_k$ decrease with increasing the length although their value is very close to $(k-1)!!$ when the length is considerably large. This indicates that the LE distributions of the dsChr22 and wdsChr22 sequences are non-Gaussian. Accordingly, the scaling parameter is irrespective of the length for the dsRandom sequence but is proportional to the length for the dsChr22 sequence, as illustrated in Fig. 9. By inspecting Fig. 9, the scaling parameter of the dsRandom sequence is approximately a constant estimated to $\tau \approx 0.0635$, remarkably

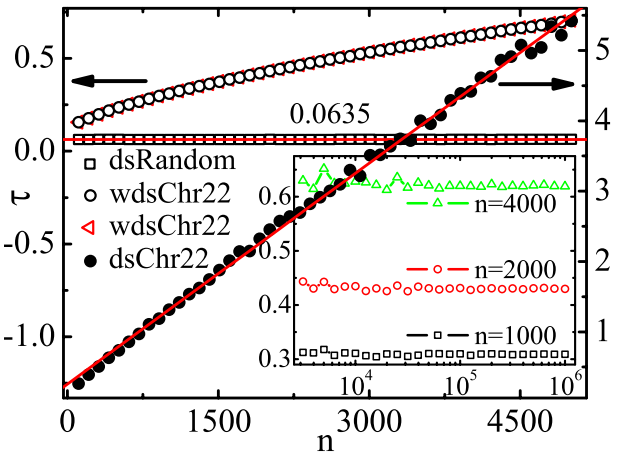

FIG. 9. (Color online) Scaling parameter τ versus length n for the dsRandom, dsChr22, and wdsChr22 sequences. The symbol $\triangleleft$ corresponds to the scaling behavior of the wdsChr22 sequence by using the sliding window strategy. The red solid lines are fitting curves (see text). The number of samples is 10^6. Inset: τ versus the number of samples for the wdsChr22 sequence by using the sliding window strategy. The symbols $\square$, $\bigcirc$, and $\triangle$ correspond to the scaling behavior of the wdsChr22 sequence with the length $n=1000$, 2000, and 4000, respectively. It can be seen that the difference between the scaling behavior of the wdsChr22 sequence by employing the two different strategies vanishes when the number of samples is 10^6 (main frame). The scaling parameter will fluctuate around a fixed value when the number of samples is not very large and will approach the fixed value when the number of samples is sufficiently large (inset). This is the same as discussed in the case of the ssDNA sequences.

deviating from the SPS hypothesis. This is mainly attributed to the double-stranded structure and the base-pairing correlations and to the fact that the localization length becomes comparable with the base-pair spacing (Fig. 7, left panel), since in the case of completely random double chain, the scaling parameter is about $\tau \approx 0.849$ if the mean value of the LE is $\gamma \approx 0.388$ and $\tau \approx 1.01$ if $\gamma \approx 0.0314$ (data not shown). While for the dsChr22 sequence, the scaling parameter increases linearly with the length with slope 1.05×10^{-2} and obviously deviates from the SPS hypothesis. The scaling parameter of the wdsChr22 sequence is much smaller and its dependence on the length can be written as $\tau \propto n^{0.441}$, suggesting a strong deviation from the SPS hypothesis. These results seem to suggest that the satellite DNA segments and the base-pairing correlations play an important role in the transport properties of the dsDNA sequences.

V. CONCLUSIONS

In summary, we numerically investigate the transport properties of the Chr22 and random DNA sequences by calculating the Hurst exponent, the LE distribution, the central moments, and the scaling parameter in case of both single- and double-stranded structures, and examine the validity of the SPS hypothesis to describe the transport properties of the genomic sequence. It is found that the random DNA sequence follows the behavior $R(n)=\sqrt{\pi n/2}-1$, and the Chr22 sequence exhibits long-range correlations and the Hurst exponent is strongly length scale dependent. In the case of single-stranded structure, the LE distribution of the random sequence is Gaussian and the scaling parameter satisfies $\tau \sim 1$ in the regime of strong localization, in accordance with the SPS. For the Chr22 sequence, the LE distribution is non-Gaussian and the scaling parameter is proportional to the sequence length with slope 0.299, suggesting the violation of the SPS. This is mainly attributed to the presence of the satellite DNA segments which approximately locate in the region from the 71 74 070th to the 72 02 601th nucleobase of the NT_{011520} segment in the Chr22 sequence. In the case of double-stranded structure, a similar behavior is observed for the random sequence except that the scaling parameter is estimated to be 0.0635. This significant deviation from the SPS hypothesis and the difference from the ssRandom sequence are not only attributed to the base-pairing correlations but also to the fact that the localization length becomes comparable with the base-pair spacing. While for the Chr22 sequence, the LE distribution is also non-Gaussian, and the position of the tail is changed to locate in the left side of the LE distribution and the slope of the $\tau-n$ curve is changed to be 1.05×10^{-2}. This is mainly due to the presence of the satellite DNA segments. In both cases the central moments of the random sequence decrease exponentially with increasing the length with a decay exponent $k/2$ for μ_k $(k=2,4,6)$, whereas that of the Chr22 sequence appear to be length independent. Due to the critical effects of the satellite DNA segments on the transport properties of the Chr22 sequence, a more systematic study should, however, be undertaken to further illustrate their role in biological processes. The method can also be employed to study the statistics of transport properties in other genomic sequences with thousands of nucleotides on the basis of quantum-mechanical wave functions.

ACKNOWLEDGMENTS

This work was supported by the State Key Programs for Basic Research of China (Grant Nos. 2005CB623605 and 2006CB921803), and by National Foundation of Natural Science in China under Grant Nos. 10474033 and 60676056.

[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] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[3] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).
[4] P. Bernaola-Galván, P. Carpena, R. Román-Roldán, and J. L. Oliver, Gene **300**, 105 (2002).
[5] W. Li and D. Holste, Phys. Rev. E **71**, 041910 (2005).
[6] A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995).
[7] B. Audit, C. Vaillant, A. Arneodo, Y. d'Aubenton-Carafa, and C. Thermes, J. Mol. Biol. **316**, 903 (2002).
[8] S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C. K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **51**, 5084 (1995).
[9] P. Carpena, P. Bernaola-Galván, A. V. Coronado, M. Hackenberg, and J. L. Oliver, Phys. Rev. E **75**, 032903 (2007).
[10] P. Bernaola-Galván, I. Grosse, P. Carpena, J. L. Oliver, R. Román-Roldán, and H. E. Stanley, Phys. Rev. Lett. **85**, 1342 (2000).
[11] D. Holste, I. Grosse, and H. Herzel, Phys. Rev. E **64**, 041917 (2001).
[12] C. Vaillant, B. Audit, and A. Arneodo, Phys. Rev. Lett. **99**, 218103 (2007).
[13] P. Carpena, P. Bernaola-Galván, P. Ch. Ivanov, and H. E. Stanley, Nature (London) **418**, 955 (2002).
[14] S. Roche, D. Bicout, E. Maciá, and E. Kats, Phys. Rev. Lett. **91**, 228101 (2003).
[15] E. L. Albuquerque, M. S. Vasconcelos, M. L. Lyra, and F. A. B. F. de Moura, Phys. Rev. E **71**, 021910 (2005).
[16] A. M. Guo, Phys. Rev. E **75**, 061915 (2007).
[17] C. T. Shih, S. Roche, and R. A. Römer, Phys. Rev. Lett. **100**, 018105 (2008).
[18] Y. A. Berlin, A. L. Burin, and M. A. Ratner, Chem. Phys. **275**, 61 (2002).
[19] P. W. Anderson, D. J. Thouless, E. Abrahams, and D. S. Fisher, Phys. Rev. B **22**, 3519 (1980).
[20] L. I. Deych, A. A. Lisyansky, and B. L. Altshuler, Phys. Rev. B **64**, 224202 (2001).
[21] K. Slevin, Y. Asada, and L. I. Deych, Phys. Rev. B **70**, 054201 (2004).
[22] Y. Y. Zhang and S. J. Xiong, Phys. Rev. B **72**, 132202 (2005).
[23] H. Schomerus and M. Titov, Phys. Rev. B **67**, 100201(R) (2003).
[24] L. I. Deych, M. V. Erementchouk, A. A. Lisyansky, and B. L. Altshuler, Phys. Rev. Lett. **91**, 096601 (2003).
[25] The sequence is retrieved from the National Center for Biotechnology Information (NCBI) build 36.2 human genome assembly. The human chromosome 22 contains about 3.49×10^7 nucleotides, the largest segment NT_{011520} about 2.33×10^7 nucleotides, and the second largest segment about 4.25×10^6 nucleotides.
[26] B. Audit, C. Thermes, C. Vaillant, Y. d'Aubenton-Carafa, J. F. Muzy, and A. Arneodo, Phys. Rev. Lett. **86**, 2471 (2001).
[27] G. M. Viswanathan, S. V. Buldyrev, S. Havlin, and H. E. Stanley, Biophys. J. **72**, 866 (1997).
[28] A. V. Coronado and P. Carpena, J. Biol. Phys. **31**, 121 (2005).
[29] H. Sugiyama and I. Saito, J. Am. Chem. Soc. **118**, 7063 (1996).
[30] A. A. Voityuk, J. Jortner, M. Bixon, and N. Rösch, J. Chem. Phys. **114**, 5614 (2001).
[31] Y. J. Yan and H. Y. Zhang, J. Theor. Comput. Chem. **1**, 225 (2002).
[32] K. Iguchi, Int. J. Mod. Phys. B **11**, 2405 (1997).
[33] A. Sedrakyan and F. Domínguez-Adame, Phys. Rev. Lett. **96**, 059703 (2006).
[34] X. F. Wang and T. Chakraborty, Phys. Rev. Lett. **97**, 106602 (2006).
[35] R. Gutiérrez, S. Mohapatra, H. Cohen, D. Porath, and G. Cuniberti, Phys. Rev. B **74**, 235105 (2006).
[36] E. Maciá, Phys. Rev. B **75**, 035130 (2007).
[37] A. V. Malyshev, Phys. Rev. Lett. **98**, 096801 (2007).
[38] V. M. K. Bagci and A. A. Krokhin, Phys. Rev. B **76**, 134202 (2007).
[39] A. M. Guo and S. J. Xiong, Phys. Lett. A **372**, 3259 (2008).
[40] A. MacKinnon and B. Kramer, Z. Phys. B: Condens. Matter **53**, 1 (1983).
[41] D. Porath, A. Bezryadin, S. de Vries, and C. Dekker, Nature (London) **403**, 635 (2000).
[42] H. Cohen, C. Nogues, R. Naaman, and D. Porath, Proc. Natl. Acad. Sci. U.S.A. **102**, 11589 (2005).