PHYSICAL REVIEW E **74**, 010903(R) (2006)

Characteristic length scale of electric transport properties of genomes

C. T. Shih
Department of Physics, Tunghai University, Taichung, Taiwan

(Received 4 October 2005; published 25 July 2006)

A tight-binding model together with a statistical method are used to investigate the relation between the sequence-dependent electric transport properties and the sequences of protein-coding regions of complete genomes. A correlation parameter Ω is defined to analyze the relation. For some particular propagation length w_{max}, the transport behaviors of the coding and noncoding sequences are very different and the correlation reaches its maximal value Ω_{max}. w_{max} and Ω_{max} are characteristic values for each species. A possible reason for the difference between the features of transport properties in the coding and noncoding regions is the mechanism of DNA damage repair processes together with natural selection.

DOI: 10.1103/PhysRevE.74.010903 PACS number(s): 87.15.Aa, 87.14.Gg, 72.80.Le

The conductance of DNA molecules is one of the central problems of biophysics because it plays a critical role in biological systems. For example, it is postulated that there may be proteins that can locate DNA damage by detecting the long-range electron migration properties [1,2]. Also, DNA is a promising candidate which may serve as the building block of molecular electronics because of its sequence-dependent and self-assembly properties.

There have been many experimental results on the conductance of DNA from different measurements for the last few years. Yet the results are still highly controversial [3]. The experimental results cover almost all possibilities, ranged from insulating [4], semiconducting [5], Ohmic [6,7], and even induced superconductivity [8]. The diversity comes from the methods of measurement and the preparation of the DNA samples, contact between DNA and electrodes, and the nucleotide sequences of the DNA [9–13].

Aside from the electrical properties, the statistical features of the symbolic sequences of DNA have also been studied intensely during recent years [14–19]. Previous work was mainly focused on the correlations and linguistic properties of the symbols A, T, C, and G, which represent the four kinds of bases adenine, thymine, cytosine, and guanine of the nucleotides, respectively. The analyses also give some eccentric results. For example, the statistical behavior of intron-free coding sequences is similar to that of random sequences while the intron-rich or junk sequences have long-range correlations. One should note that the roots of these statistical properties of the symbolic sequences are the results of evolution, and the underlying driving forces are the principles of physics and chemistry. In the other direction, the correlation of sequences will influence the physical and chemical properties [20,21]. Thus it is reasonable to conjecture that the electric properties can play a critical role during the evolution process in nature by methods such as DNA damage repair processes [1,2]. In this paper, the relation between electric transport properties and the gene-coding (noncoding) parts of genomic sequences will be discussed.

The simplest effective tight-binding Hamiltonian for a hole propagating in the DNA chain is [22,23]

$$H = \sum_n \epsilon_n c_n^\dagger c_n + \sum_n t_{n,n+1}(c_n^\dagger c_{n+1} + \mathrm{H.c.}) \tag{1}$$

where each lattice point represents a nucleotide base of the chain for $n \in [2,N]$. The DNA molecule is assumed to be connected between two semi-infinite electrodes ($n \in (-\infty,1]$ and $n \in [N+1,\infty)$). $c_n^\dagger$ (c_n) is the creation (destruction) operator of a hole at the nth site. ϵ_n is the potential energy at the nth site, which is determined by the ionization potential of the corresponding nucleotide. $\epsilon_A=8.24$ eV, $\epsilon_T=9.14$ eV, $\epsilon_C=8.87$ eV, and $\epsilon_G=7.75$ eV from *ab initio* calculation in the gas phase [24]. In a biological environment they will be reduced [3,25,26]. The hopping integral $t_{n,n+1}=t_m$ and t_{DNA} for electrodes and nucleotides, respectively. t_{DNA} is assumed to be nucleotide independent for simplicity. Typical values of the hopping integral between adjacent bases range from 0.03 to 0.4 eV from first-principles calculation performed on a short segment with two or a few bases [24,27–30]. With the reduction of ϵ_n by water and ions in the environment [25], and the longer chain used in the model, the effective t_{DNA} will be larger ($\sim$0.4 eV) [31,32]. To reduce the backscattering effect at the contacts, even larger t_{DNA} (up to 1 eV) was also used in previous studies [22,23,33–35]. $t_m=1$ is the energy unit. t_{DNA} ranges from 0.4 to 1, and $\epsilon_n=8.24$, 9.14, 8.87, 7.75, and 7.75 for A, T, C, G, and electrodes (ϵ_m), respectively. Note that only the ratio $(\epsilon_n-\epsilon_m)/t_{DNA}$ is relevant for the tight-binding study.

The eigenstates of the Hamiltonian $|\Psi\rangle=\Sigma_n a_n|n\rangle$ ($|n\rangle$ represents the state where the hole is located at the nth site) can be solved exactly by using the transfer matrix method:

$$\begin{pmatrix} a_{N+2} \\ a_{N+1} \end{pmatrix} = M_{N+1}M_N \cdots M_1 \begin{pmatrix} a_1 \\ a_0 \end{pmatrix} \equiv P(N) \begin{pmatrix} a_1 \\ a_0 \end{pmatrix} \tag{2}$$

where

$$M_n = \begin{pmatrix} \dfrac{E-\epsilon_n}{t_{n,n+1}} & -\dfrac{t_{n-1,n}}{t_{n,n+1}} \\ 1 & 0 \end{pmatrix}. \tag{3}$$

E is the energy of the injected hole. In the electrodes, the wave functions are plane waves and the dispersion of the hole is $\epsilon_m+2t_m \cos k$. So the range of possible E is $[\epsilon_m-2t_m, \epsilon_m+2t_m]=[5.75,9.75]$. The transmission coefficient has the following form [36]:

$$T(E) = \frac{4 - \left(\frac{E-\epsilon_m}{t_0}\right)^2}{\sum_{i,j=1,2} P_{ij}^2 + 2 - \left(\frac{E-\epsilon_m}{t_0}\right)^2 P_{11}P_{22} + \left(\frac{E-\epsilon_m}{t_0}\right)(P_{11}-P_{22})(P_{12}-P_{21})}. \tag{4}$$

The transmission of several sequences of complete genomes $S=(s_1,s_2,\ldots,s_{N_{tot}})$ is studied ($s_i=A$, T, C, or G). Since the total length N_{tot} of the complete genome is usually much longer than the distance over which holes can migrate along the DNA chain even for the smallest N_{tot} for viruses, only shorter segments are measured instead of the whole chain. A "window" with width w is defined to extract a segment $S_{i,w}=(s_i,s_{i+1},\ldots,s_{i+w-1})$ for $1 \leqslant i \leqslant N_w = N_{tot}-w+1$ from S. Starting from $i=1$ and sliding the window, we can get the "transmission sequence" $T_w(E,i)$ of $S_{i,w}$ for all i, which depends on the energy of the injected hole E, the starting position i, and the propagation length w. For further analysis of the whole genome sequence, $T_w(E,i)$ is integrated in an energy interval $[E,E+\Delta E]$:

$$\bar{T}_w(E,\Delta E,i) = \int_E^{E+\Delta E} T_w(E',i)dE'. \tag{5}$$

In the remainder of the paper, the transmission is integrated for the whole bandwidth, that is, $E=5.75$ and $\Delta E=4$. These two values will be omitted in the related formulas for brevity. 300 base pairs at the two ends of the DNA chain will be omitted in the following analysis because the telomere sequences at the terminals usually have larger transmission (due to the periodicity) and will dominate some of the average properties. Thus $N_w = N_{tot}-w+1-2\times 300$.

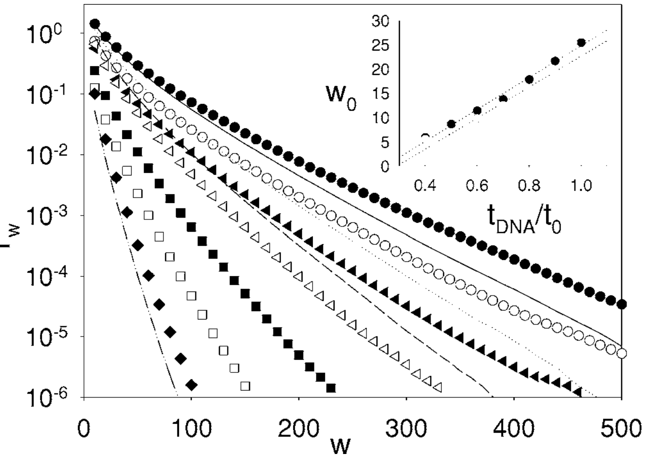

FIG. 1. T_w^{av} for Y3 with $t_{DNA}=1.0$ (full circles), 0.9 (open circles), 0.8 (full triangles), 0.7 (open triangles), 0.6 (full squares), 0.5 (open squares), and 0.4 (diamonds). Solid, dotted, dashed, and dash-dotted lines are for a random sequence R3 with $t_{DNA}=1.0$, 0.9, 0.8, and 0.4, respectively. The error bars are smaller than the size of the symbols. (Inset) Localization length w_0 of Y3 (full circles) and R3 (open circles) for each t_{DNA} (see text).

The averaged transmission $T_w^{av}=\frac{1}{N_w}\sum_i \bar{T}_w(i)$ versus propagation length w is plotted in Fig. 1 for the third chromosome of *Saccharomyces cerevisiae* (Y3) (bakery yeast, accession number=NC.001135 for GenBank [37]) with several values of t_{DNA}. T_w^{av} decreases exponentially with increasing w, which is consistent with the localization picture. The curves can be fitted by the function $T_w^{av}=ae^{-w/w_0}$. The inset of Fig. 1 shows the averaged localization length w_0 for each t_{DNA}. Note this is an averaged result of the complete genome, and the possibility of high conductance of some particular segments is not ruled out. Other important features are that $\bar{T}_w(i)$ decreases faster for smaller t_{DNA}, and w_0 is nearly proportional to t_{DNA}. The reason is that the backscattering is stronger for smaller t_{DNA}. T_w^{av} for a random sequence R3 with the same length and ratios of the four bases as Y3 are also shown in the lines of Fig. 1. It is clear that the transmission of the random sequence decreases faster (smaller w_0) than that of the natural genome due to the larger disorder. This result is consistent with Ref. [22].

Since the transport properties are related to the DNA damage repair mechanism, there could be correlation between the locations of genes and the corresponding $\bar{T}_w(i)$. In Fig. 2, $\bar{T}_{240}(i)$ and the coding regions are compared for part of the Y3 sequence. It seems that most of the sharp peaks of $\bar{T}_{240}(i)$ are located in the protein-coding region.

To check this correlation in a more quantitative way, I first

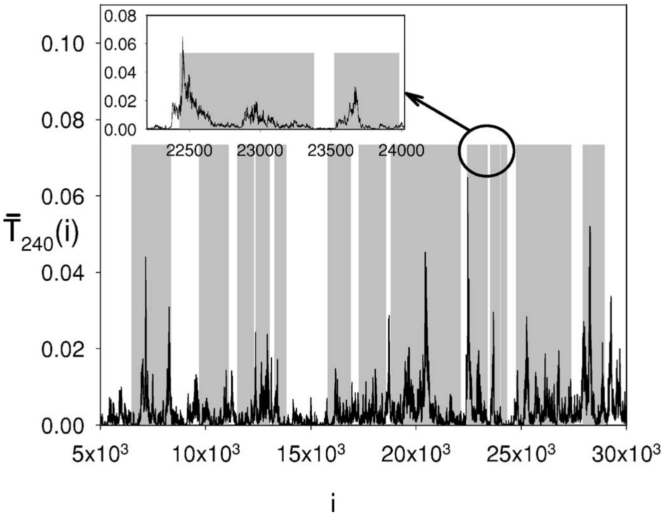

FIG. 2. Comparison of $\bar{T}_{240}(i)$ (line, $t_{DNA}=1$) and the coding regions (shaded area) of the range from the 5000th to the 30 000th nucleotide of Y3. (Inset) Enlarged plot from the 22 000th to the 24 000th nucleotide.

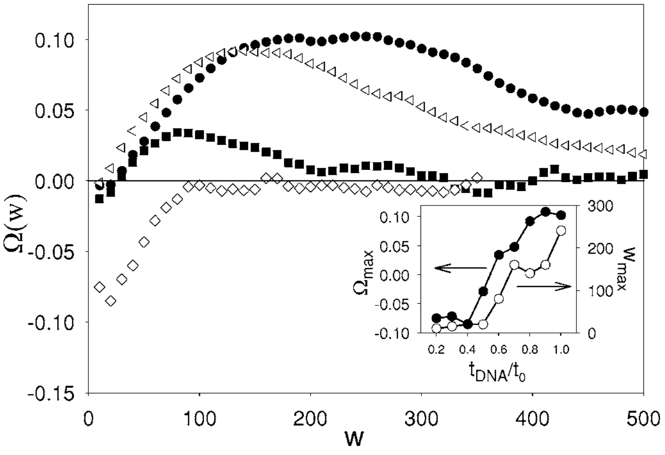

FIG. 3. $\Omega(w)$ for $t_{DNA}/t_0=1.0$ (circles), 0.8 (triangles), 0.6 (squares), and 0.4 (diamonds) of Y3. (Inset) Ω_{max} (full circles) and w_{max} (open circles) as functions of t_{DNA}.

define a binary "coding sequence" $G(i)=1(0)$ if the ith nucleotide was in the protein-coding (-noncoding) region, and then normalize $G(i)$ and $\bar{T}_w(i)$ in the following way:

$$G'(i)=G(i)-\frac{1}{N_w}\sum_j G(j), \quad g(i)=\frac{G'(i)}{\sqrt{\sum_j [G'(j)]^2}},$$

and

$$\bar{T}'_w(i)=\bar{T}_w(i)-\frac{1}{N_w}\sum_j \bar{T}_w(j), \quad t_w(i)=\frac{\bar{T}'_w(i)}{\sqrt{\sum_j [\bar{T}'_w(j)]^2}} \qquad (6)$$

The overlap between these two normalized sequences is defined as [38,39]

$$\Omega(w)=\sum_i g(i)t_w(i). \qquad (7)$$

In Fig. 3 $\Omega(w)$ for Y3 is shown for different t_{DNA}. For $t_{DNA}=1$, there is a maximum at $w_{max}=240$ with $\Omega_{max}=0.103$. Ω_{max} denotes the maximal absolute value of $\Omega(w)$ and can be positive or negative. The strong positive overlap implies that the holes can move more freely in the coding regions. As t_{DNA} decreases, both $\Omega_{\rm max}$ and w_{max} decrease. For $t_{DNA}\leqslant 0.5$, the overlap becomes negative which means the electronic conductance is poorer at the coding regions. Note that, although the values of Ω_{max} and w_{max} vary with the parameters used in the model, $G(i)$ and $\bar{T}_w(i)$ are correlated in general. That is, the transport properties in coding and noncoding regions are different.

Several $\Omega(w)$ with $t_{DNA}=1$ for different genomes are shown in Fig. 4. It can be seen that there is maximal positive or negative overlap Ω_{max} at some characteristic migration length w_{max} for each genome. $\Omega(w)$ for yeast chromosomes III, VIII, and X, and *Ureaplasma parvum* serovar 3 str. ATCC 700970 are positive, which means the coding regions have larger conductance. On the other hand, $\Omega(w)$ for *acinetobacter* sp. ADP1, *Deinococcus radiodurans* R1 chromosome II, and *Chlamydia trachomatis* D/UW-3/CX are negative, which means the coding regions have smaller conductance. (Ω_{max}, w_{max}) for these genomes are summarized in Table I.

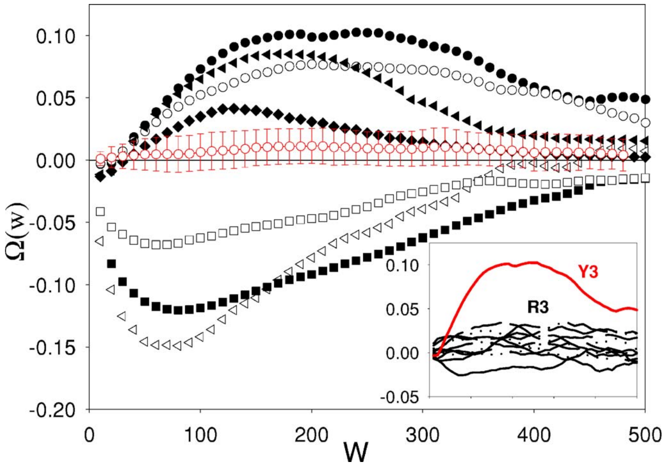

FIG. 4. (Color online) $\Omega(w)$ for several genomes: chromosomes III (full circles), VIII (open circles), and X (full triangles) of yeast, *Ureaplasma parvum* serovar 3 str. ATCC 700970 (full diamonds), *Acinetobacter* sp. ADP1 (full squares), *Deinococcus radiodurans* R1 chromosome II (open triangles), and *Chlamydia trachomatis* D/UW-3/CX (open squares). Red circles with error bars are averaged $\Omega(w)$ for R3, ten randomized sequences of yeast chromosome III (see text). (Inset) Comparison of $\Omega(w)$ for Y3 and the ten R3's.

To ensure that $\Omega(w)$ shown above are physically and biologically meaningful, we compare the results with random sequences. Ten sequences generated in the same way as R3 are analyzed and the averaged $\Omega(w)$ [overlap with the $g(i)$ of Y3] are shown in Fig. 4 (open circles with error bars). It is clear that its overlap is about one order of magnitude smaller then the real sequences. So Ω_{max} and w_{max} are not artifacts, but intrinsic properties of genomes from the above comparison.

One possible reason for the different w_{max} and Ω_{max} for different genomes is the content of A, T, C, and G, since that determines the shape and height of the potential barrier. But from the C+G percentage shown in Table I we see that the correlation between C+G percentage and (w_{max}, Ω_{max}) is not

TABLE I. Ω_{max} and w_{max} for the genomes studied in Fig. 4.

Genome	Access No.	C+G (%)	w_{max}	Ω_{max}
Yeast III	NC.001135	38.5	240	0.103
Yeast VIII	NC.001140	38.5	200	0.077
Yeast X	NC.001142	38.4	170	0.085
Ureaplasma parvum serovar 3 str. ATCC 700970	NC.002162	25.5	130	0.041
Acinetobacter sp. ADP1	NC.005966	40.4	80	−0.129
Deinococcus radiodurans R1 chromosome II	NC.001264	66.7	80	−0.149
Chlamydia trachomatis D/UW-3/CX	NC.000117	41.3	50	−0.075

evident. The R3's give completely different results from Y3 although they have exactly the same ATCG content. The correlation of the bases in the sequence may play a more important role than simply counting the ATCG ratios.

From the analysis above, it can be concluded that w_{max} is a characteristic length scale of the electric transport, which can identify the gene-coding regions, and Ω_{max} stands for the "sensibility" of this probing process.

The possible biological reason for these correlations is the mechanism of DNA damage repair processes. Since proteins use transport properties to probe the location of DNA damage [1,2], the transport of the coding areas should have particular features for the detecting process, while that of the noncoding regions is somewhat irrelevant.

Figure 4 shows two important features of Ω_{max}. First, each species has its characteristic values (w_{max},Ω_{max}). It can be postulated that the mechanisms detecting defects of DNA in different species are different due to the various biological and environmental features. Second, (w_{max},Ω_{max}) of the different chromosomes of the same species (yeast here) are very similar because they are in the *same* environment, and hence have the same DNA damage repair mechanism.

It should be noted that the model used in this study is an oversimplified one and the results may depend on the choice of the model. However, one of the most important properties can be extracted from this coarse-grained model—the coding regions have very different transport behavior from the noncoding parts at the characteristic length scale w_{max}. Each species has a different w_{max} depending on its environment. In the future, the model will be finer grained by introducing more realistic interactions like base-dependent hopping [40], sequence-dependent potentials [41], the effect of polarons [42,43], and charge-charge interactions [44]. Other types of models will also be studied to check the possible model dependence.

In summary, with this method combining the transfer matrix approach and symbolic sequence analysis, the correlation between the transport properties and the positions of genes is studied for complete genomes. There are two characteristic values Ω_{max} and w_{max} for each genome. These two values can provide information for taxonomy or the mechanism of evolution.

This research is supported by the National Science Council in Taiwan (Grant No. 93-2112-M-029-001-). Part of the calculations were performed at the National Center for High-Performance Computing in Taiwan. The author is grateful for the help.

[1] S. R. Rajski *et al.*, Mutat Res. **447**, 49 (2000).
[2] E. Yavin *et al.*, Proc. Natl. Acad. Sci. U.S.A. **102**, 3546 (2005).
[3] R. G. Endres *et al.*, Rev. Mod. Phys. **76**, 195 (2004).
[4] Y. Zhang *et al.*, Phys. Rev. Lett. **89**, 198102 (2002).
[5] D. Porath *et al.*, Nature (London) **403**, 635 (2000).
[6] P. Tran *et al.*, Phys. Rev. Lett. **85**, 1564 (2000).
[7] K.-H. Yoo *et al.*, Phys. Rev. Lett. **87**, 198102 (2001).
[8] A. Y. Kasumov *et al.*, Science **291**, 280 (2001).
[9] A. Y. Kasumov *et al.*, Appl. Phys. Lett. **84**, 1007 (2004).
[10] T. Heim *et al.*, Appl. Phys. Lett. **85**, 2637 (2004).
[11] H. Hartzell *et al.*, Appl. Phys. Lett. **82**, 4800 (2003).
[12] A. J. Storm *et al.*, Appl. Phys. Lett. **79**, 3881 (2001).
[13] E. Maciá *et al.*, Phys. Rev. B **71**, 113106 (2005).
[14] C.-K. Peng *et al.*, Nature (London) **356**, 168 (1992).
[15] S. V. Buldyrev *et al.*, Phys. Rev. E **51**, 5084 (1995).
[16] W. Li, Comput. Chem. (Oxford) **21**, 257 (1997).
[17] D. Holste *et al.*, J. Mol. Evol. **51**, 353 (2000).
[18] T.-H. Hsu *et al.*, Phys. Rev. E **67**, 051911 (2003).
[19] P. W. Messer *et al.*, Phys. Rev. Lett. **94**, 138103 (2005).
[20] C. Vaillant *et al.*, Phys. Rev. E **67**, 032901 (2003).
[21] E. Carlon *et al.*, Phys. Rev. Lett. **94**, 178101 (2005).
[22] S. Roche, Phys. Rev. Lett. **91**, 108101 (2003).
[23] S. Roche *et al.*, Phys. Rev. Lett. **91**, 228101 (2003).
[24] H. Sugiyama *et al.*, J. Am. Chem. Soc. **118**, 7063 (1996).
[25] E. B. Starikov, J. Photochem. Photobiol. C **3**, 147 (2002).
[26] E. Maciá, Nanotechnology **16**, S254 (2005).
[27] H. Zhang *et al.*, J. Chem. Phys. **117**, 4578 (2002).
[28] A. A. Voityuk *et al.*, J. Phys. Chem. B **104**, 9740 (2000); J. Chem. Phys. **114**, 5614 (2001).
[29] A. Troisi *et al.*, Chem. Phys. Lett. **344**, 509 (2001).
[30] F. C. Grozema *et al.*, ChemPhysChem **3**, 536 (2002).
[31] Y. A. Berlin *et al.*, Superlattices Microstruct. **28**, 241 (2000).
[32] G. Cuniberti *et al.*, Phys. Rev. B **65**, 241314(R) (2002).
[33] E. L. Albuquerque *et al.*, Phys. Rev. E **71**, 021910 (2005).
[34] W. Ren *et al.*, Phys. Rev. B **72**, 035456 (2005).
[35] R. A. Caetano *et al.*, Phys. Rev. Lett. **95**, 126601 (2005).
[36] E. Maciá, Phys. Rev. B **60**, 10032 (1999).
[37] D. A. Benson *et al.*, Nucleic Acids Res. **32**, D23 (2004).
[38] C. T. Shih *et al.*, Phys. Rev. Lett. **84**, 386 (2000).
[39] C. T. Shih *et al.*, Phys. Rev. E **65**, 041923 (2002).
[40] D. Klotsa *et al.*, Biophys. J. **89**, 2187 (2005).
[41] K. Senthilkumar *et al.*, J. Am. Chem. Soc. **125**, 13658 (2003).
[42] E. M. Conwell *et al.*, Proc. Natl. Acad. Sci. U.S.A. **97**, 4556 (2000); **102**, 8795 (2005).
[43] J. H. Wei *et al.*, Phys. Rev. B **72**, 064304 (2005).
[44] E. B. Starikov, Philos. Mag. Lett. **83**, 699 (2003).