physica status solidi b

phys. stat. sol. (b) **243**, No. 2, 378–381 (2006) / **DOI** 10.1002/pssb.200562705

Electric transport and coding sequences of DNA molecules

C. T. Shih[*, 1, 2]

[1] Department of Physics, Tunghai University, Taichung, Taiwan
[2] Physics Division, National Center for Theoretical Sciences, Hsinchu, Taiwan

Received 21 August 2005, accepted 28 September 2005
Published online 12 December 2005

PACS 72.80.Le, 87.14.Gg, 87.15.Aa, 87.15.Cc

A tight-binding model is used to investigate the relation between the electric transport properties and the sequences of protein-coding regions of complete genomes. The sequence-dependent transmission coefficient for some particular propagation length w_G is closely related to the positions of genes on DNA. w_G is the "characteristic migration length" for each species.

The conductance of DNA molecules is one of the central problems of biophysics. Despite of the academic interest on the issue, DNA is one of the most promising candidates which may serve as the building block of molecular electronics. There have been many experimental results from different measurements for the last few years. Yet the results seem to be controversial and the problem is still unresolved [1]. The experimental results are amazingly diverse and almost cover all possibilities, ranged from insulating [2], semiconducting [3], Ohmic [4, 5], and even induced superconductivity [6]. The diversity of the experimental results comes from the setup of the measurements and the preparation of the DNA samples. One of the critical factors influencing the results is the contact of the DNA and electrodes [2, 7–9]. And the different nucleotide sequences of the DNA molecules used in the experiments also diversify the results because the transport properties are sequence-dependent.

Aside from the physical properties, the statistical features of the symbolic sequences of DNA have also been intensely studied during the past years [10–15]. The previous works are mainly focused on the correlations and linguistic properties of the symbols A, T, C, and G. These properties of the symbolic sequences are the results of evolution, and the underlying driving forces are the principles of physics and chemistry. On the other hand, the correlation of sequences will influence the physical and chemical properties, such as the relation between the mechanical properties and sequences of DNA [16]. Thus it is reasonable to conjecture that the sequence-dependent electric properties can play some role during the evolution process in nature. For example, it is postulated that there may be proteins which can locate the DNA damage by detect the long-range electron migration properties [17, 18]. In this paper, the relation between electric transport properties and the gene-coding/nocoding parts of genomic sequences will be discussed.

The simplest effective tight-binding Hamiltonian for a hole propagating in the DNA chain can be written as [19, 20]

$$H = \sum_n \varepsilon_n c_n^\dagger c_n + t_0 \sum_n (c_n^\dagger c_{n+1} + \text{h.c.}) , \tag{1}$$

* e-mail: ctshih@thu.edu.tw

where each lattice point represents a nucleotide of the chain. $c_n^\dagger$ (c_n) is the creation (destruction) operator of a hole at the n-th site. ε_n is the potential energy at the n-th site, which is determined by the ionization potential of the corresponding nuecleotide. ε_n equals to 8.24 eV, 9.14 eV, 8.87 eV, and 7.75 eV for n = A, T, C, and G, repectively [21]. The DNA molecule is assumed to be connected between two semi-infinite electrodes with energy $\varepsilon_m = \varepsilon_G = 7.75\,\text{eV}$. The hopping integral $t_0 = 1\,\text{eV}$ [19]. Note that $n \in ([-\infty, 1]$ and $n \in [N+1, \infty]$ are for electrodes and $n \in [2, N]$ are for nucleotides.

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

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

where

$$M_n = \begin{pmatrix} \dfrac{E-\varepsilon_n}{t_0} & -1 \\ 1 & 0 \end{pmatrix}, \tag{3}$$

where E is the energy of the injected hole. In electrodes, the wave functions are simply plane waves and the dispersion of the hole is $\varepsilon_m + 2\cos k$. So the range of possible E is $[\varepsilon_m - 2, \varepsilon_m + 2] = [5.75\,\text{eV}, 9.75\,\text{eV}]$. Matching the wave functions at the contacts ($n = 1$ and $n = N$) we get the transmission coefficient

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

The transmission of several sequences of complete genomes $S = (s_1, s_2, \ldots, s_{N_{\text{tot}}})$ is studied. Since the total length N_{tot} of the complete genome is usually much longer than the distance which holes can migrate along the DNA chain even for the smallest N_{tot} for viruses, we won't measure the transmission through the whole chain but only shorter segments instead. 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 $1 \le i \le N_{\text{tot}} - w + 1$ from S. Starting from $i = 1$ and sliding the window, we can get the "transmission sequence" $T_w(E, i)$ of the $S_{i,w}$ for all i, which depends on the energy of the injected hole E, the starting position of the segment i, and the propagation length w. For further analysis of the whole genome sequences, we integrate $T_w(E, i)$ in an energy interval $[E, E + \Delta E]$:

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

In the remaining of this paper, the results for $E = 5.75\,\text{eV}$ and $\Delta E = 4\,\text{eV}$ are discussed. That is, the transmission coefficient is integrated in the whole bandwidth. $\bar{T}_w(E, \Delta E, i)$ will be shorten as $\bar{T}_w(i)$ for this $(E, \Delta E)$ in the following text.

Since the transport properties are related to the mechanism of DNA damage repairing, there could be correlation between the locations of genes and the integrated transmission $\bar{T}_w(i)$. In Fig. 1, $\bar{T}_{240}(i)$ and the coding regions are compared for part of the sequence of the third chromosome of Saccharomyces cerevisiae (bakery yeast, accession number = NC.001135 for GenBank [22]). It is obvious 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 define a binary "coding sequence" $G(i) = 1$ (0) if the i-th nucleotide was in the protein-coding (noncoding) region, and then normalize $G(i)$ and

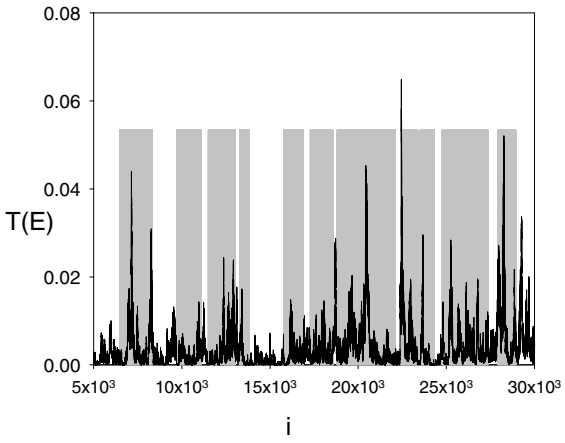

Fig. 1 Comparison of $\bar{T}_{240}(i)$ (line) and the coding regions (shaded area) of the 5000-th to 30000-th nucleotides of the third chromosome of yeast.

$\bar{T}_w(i)$ in the following way

$$G'(i) = G(i) - \frac{1}{N}\sum_j G(j)\,,$$

$$g(i) = \frac{G'(i)}{\sqrt{\sum_j (G'(j))^2}}\,,$$

and

$$\bar{T}'_w(i) = \bar{T}_w(i) - \frac{1}{N}\sum_j \bar{T}_w(i)\,,$$

$$t_w(i) = \frac{\bar{T}'_w(i)}{\sqrt{\sum_j (\bar{T}'_w(i))^2}}\,. \tag{6}$$

The overlap between these two normalized sequences is defined as [23, 24]

$$\Omega(w) = \sum_i g(i)\, t_w(i)\,. \tag{7}$$

Several $\Omega(w)$ functions for different genomes are shown in Fig. 2. It can be seen that there is maximal positive or negative overlap $\Omega_{\max}$ at some "characteristic migration length" w_G for each genome.

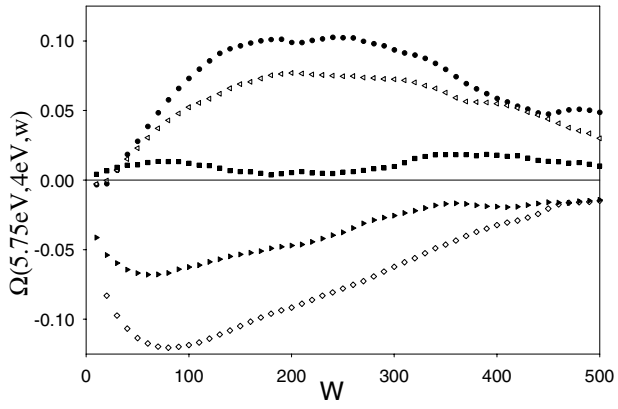

Fig. 2 $\Omega(w)$ for several genomes: chromosomes III (full circles) and VIII (open triangles) of yeast, acinetobacter sp. ADP1 (open diamonds), and chlamydia trachomatis D/UW-3/CX (full triangles). Full squares are for the randomized sequence of yeast chromosome VIII (see text).

The overlaps of coding and transmission sequences for yeast chromosomes III ($(w_G, \Omega_{\max}) = (240, 0.103)$) and VIII (NC.001140, $(w_G, \Omega_{\max}) = (200, 0.077)$) are positive, which means the coding regions have larger conductances. On the other hand, the overlaps for the two microbial genomes, acinetobacter sp. ADP1 (NC.005966, $(w_G, \Omega_{\max}) = (80, -0.129)$) and chlamydia trachomatis D/UW-3/CX (NC.000117, $(w_G, \Omega_{\max}) = (50, -0.075)$) are both negative, which means the coding regions have smaller conductance.

To ensure that $\Omega(w)$ shown above are physically and biologically meaningful, we compare the results with a random sequence. The sequence of yeast chromosome VIII is radomized with the percentages of A, T, C, G unchanged, then the overlap between $\overline{T}_w(i)$ of this randomized sequence and $G(i)$ of yeast chromosome VIII is calculated and shown in Fig. 2 (full squares). It is clear that this overlap is about one order of magnitude smaller then the real sequences. So $\Omega_{\max}$ and w_G are not artifacts, but intrinsic properties of genomes from the above comparison.

In summary, the correlation between the transport properties and the positions of genes is studied in this paper. There are two characteristic values $\Omega_{\max}$ and w_G for each genome. These two values may provide information for taxonomy or the mechanism of evolution.

Acknowledgements This research is supported by the National Science Council in Taiwan under the Grant No. 93-2112-M-029-001. Part of the calculations are performed in the IBM P690 and PC clusters in the National Center for High-performance Computing in Taiwan, and the PC clusters of the Department of Physics and Department of Computer Science and Engineering of Tunghai University, Taiwan. The author is grateful for their help.

References

[1] R. G. Endres, D. L. Cox, and R. R. P. Singh, Rev. Mod. Phys. **76**, 195 (2004).
[2] Y. Zhang, R. H. Austin, J. Kraeft, E. C. Cox, and N. P. Ong, Phys. Rev. Lett. **89**, 198102 (2002).
[3] D. Porath, A. Bezryadin, S. de Vries, and C. Dekker, Nature (London) **403**, 635 (2000).
[4] P. Tran, B. Alavi, and G. Gruner, Phys. Rev. Lett. **85**, 1564 (2000).
[5] K.-H. Yoo, D. H. Ha, J.-O. Lee, J. W. Park, J. Kim, J. J. Kim, H.-Y. Lee, T. Kawai, and H. Y. Choi, Phys. Rev. Lett. **87**, 198102 (2001).
[6] A. Y. Kasumov, M. Kociak, S. Gueron, B. Reulet, and V. T. Volkov, Science **291**, 280 (2001).
[7] H. Hartzell, B. Melord, D. Asare, H. Chen, J. J. Heremans, and V. Sughomonian, Appl. Phys. Lett. **82**, 4800 (2003).
[8] A. J. Storm, J. van Noort, S. de Vries, and C. Dekker, Appl. Phys. Lett. **79**, 3881 (2001).
[9] E. Maciá, F. Triozon, and S. Roche, Phys. Rev. B **71**, 113106 (2005).
[10] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992).
[11] 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).
[12] W. Li, Comput. Chem. (UK) **21**, 257 (1997).
[13] D. Holste, O. Weiss, I. Große, and H. Herzel, J. Mol. Evol. **51**, 353 (2000).
[14] T.-H. Hsu and S.-L. Nyeo, Phys. Rev. E **67**, 051911 (2003).
[15] P. W. Messer, P. F. Arndt, and M. Lässig, Phys. Rev. E **94**, 138103 (2005).
[16] C. Vaillant, B. Audit, C. Thermes, and A. Arnéodo, Phys. Rev. E **67**, 032901 (2003).
[17] S. R. Rajski, B. A. Jackson, and J. K. Barton, Mutat. Res. **447**, 49 (2000).
[18] E. Yavin, A. K. Boal, E. D. A. Stemp, E. M. Boon, A. L. Livingston, V. L. O'Shea, S. S. David, and J. K. Barton, Proc. Natl. Acad. Sci. **102**, 3546 (2005).
[19] S. Roche, Phys. Rev. Lett. **91**, 108101 (2003).
[20] S. Roche, D. Bicout, E. Maciá, and E. Kats, Phys. Rev. Lett. **91**, 228101 (2003).
[21] H. Sugiyama, and I. Saito, J. Am. Chem. Soc. **118**, 7063 (1996).
[22] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler, Nucl. Acids Res. **32** (Database issue), D23-6 (2004).
[23] C. T. Shih, Z. Y. Su, J. F. Gwan, H. C. Lee, B. L. Hao, and C. H. Hsieh, Phys. Rev. Lett. **84**, 386 (2000).
[24] C. T. Shih, Z. Y. Su, J. F. Gwan, B. L. Hao, C. H. Hsieh, J. L. Lo, and H. C. Lee, Phys. Rev. E **65**, 41923 (2002).

