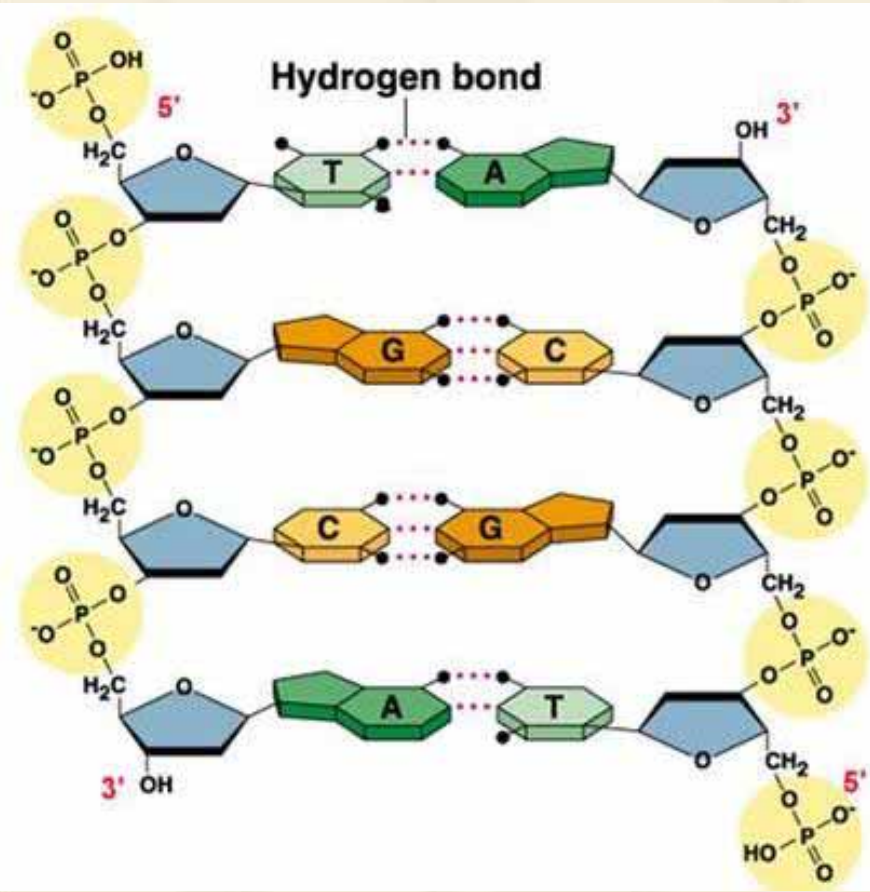
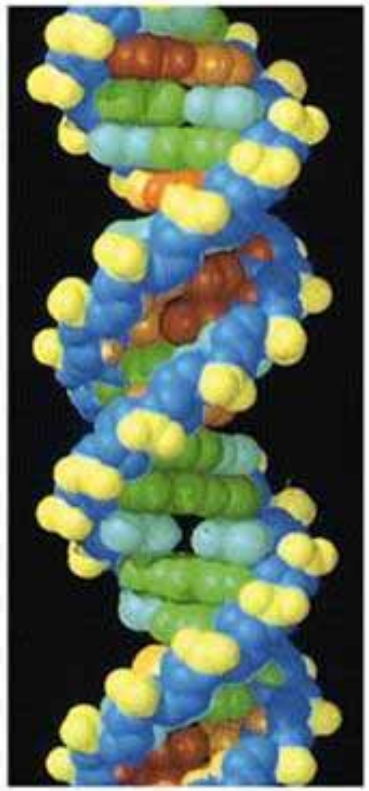
Long Range Correlation and Possible Electron Conduction through DNA Sequences

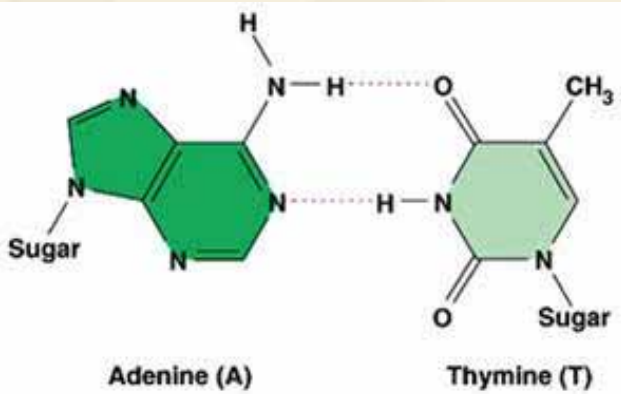
Sheng-Cheng Wang, Ping-Cheng Li*, and Hsen-Che Tseng
Department of Physics, National Chung-Hsing University

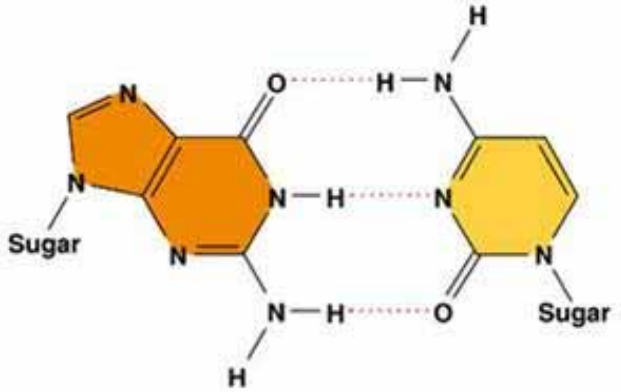
1. **Introduction**
2. **Long range correlation measurement in a DNA sequence**
3. **Charge Transfer in DNA**
4. **Electronic wave functions in a DNA sequence**
5. **Conclusions**

1. Introduction









2. Long range correlation measurement in a DNA sequence

Adapting C.K. Peng's convention **[R1]** we assigned

$$\begin{cases} u(i) = +1 \text{ for purines} & (A, G) \\ u(i) = -1 \text{ for pyrimidines} & (C, T) \end{cases}$$

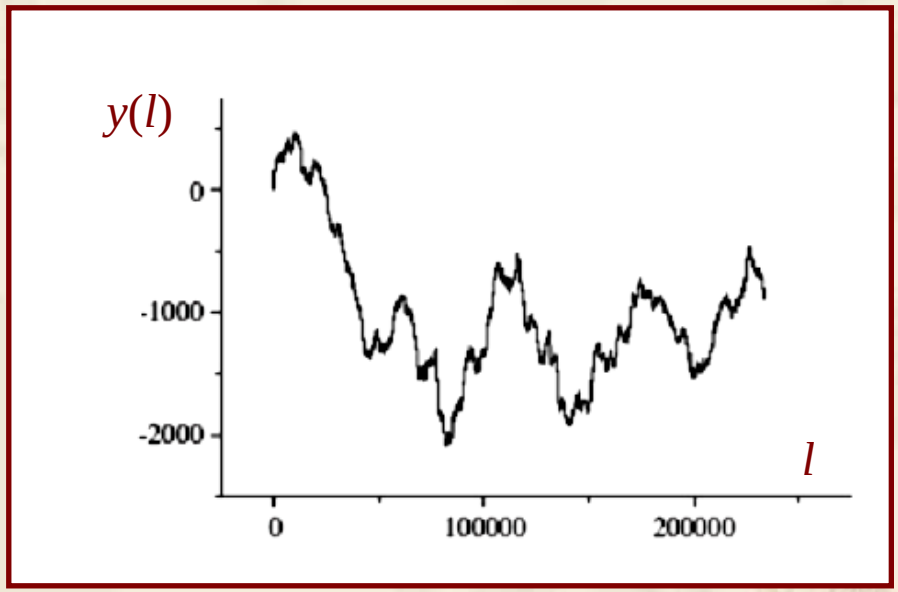
A DNA sequence mapped in this way resembles a walker staggers along a one dimensional path randomly and whereupon the second moment of the fluctuations of the random walk can be calculated.

The DNA sequence **"AGGCTTGAACTTAGGATTCG….."** is mapped into **"1, 1, 1, -1, -1 ,-1, ,1, 1, 1, -1, -1, -1, 1, 1, 1, 1, -1, -1, -1, 1,…."**.

Define the “net displacement” y as:

$$y(l) = \sum_{i=1}^{l} u(i) \qquad (1)$$

we may get the following landscape structure:



We may then calculate the root-mean-square fluctuations of various nucleotide distances l as:

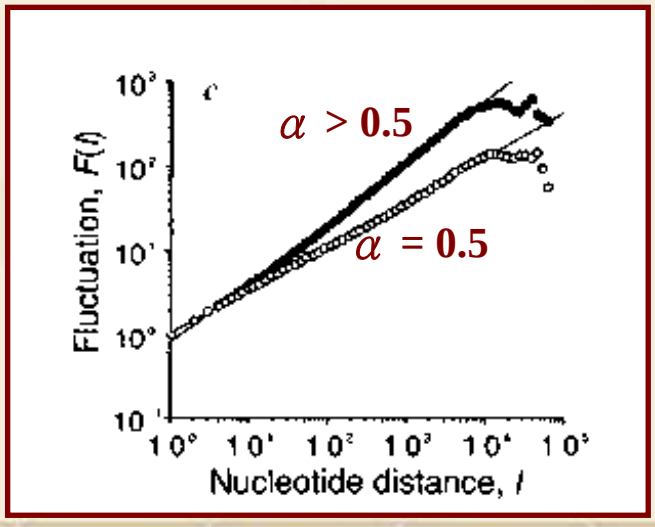
$$F(l) = \left[\left\langle [\Delta y(l)]^2 \right\rangle - \left\langle \Delta y(l) \right\rangle^2 \right]^{1/2} \qquad (2)$$

where the quantity $\Delta y(l)$ is defined as

$$\Delta y(l) = y(l_0 + l) - y(l_0) \qquad (3)$$

When $F(l)$ is plotted against l the diagram is seen to be linear on the logarithmic scale.

$$F(l) \sim l^{\alpha} \qquad (4)$$



$\alpha > 0.5$

There shows long range correlation in a sequence.

$\alpha = 0.5$

There shows a random sequence without correlations

- As it was suggested by Stephan Roche *et al*. **[R2] Hurst's analysis** was argued to be more reliable in the determination of the precise rescaling coefficients.

- **Hurst's analysis** manages to diminish "patchiness" structure [1,2] in a much more sophisticate fashion.

With a given sequence of size N the net displacement $x(n)$ of the random walker after n steps in a row is

$$x(n) = \sum_{i=1}^{n} u(i) \qquad 1 \le n \le N \qquad (5)$$

The net displacement difference between the two occasions when the random walker is either on the position m or on $(m+k)$ is defined as

$$\Delta x(m,k) = x(m+k) - x(m) \qquad (6)$$

Rescaled variables $X(m, k)$ can thus be defined as the following

$$X(m,k) = \Delta x(m,k) - \frac{k}{n}\Delta x(m,n) \quad \text{, where } 1 \le k \le n. \qquad (7)$$

By doing such kind of an overall subtraction a general trend of asymmetrical concentration of purines and pyrimidines is greatly diminished. The next step is to find the maximum and minimum values of the rescaled value $X(m, k)$ within the range $1 \leq k \leq n$ and calculate the difference, $S(m, n)$, between them.

$$S(m,n) = \max_{1\le k\le n}[X(m,k)] - \min_{1\le k\le n}[X(m,k)] \quad (8)$$

Average values obtained from these differences, $S(m, n)$ for $1 \leq m \leq N–n$, is calculated for different window length n.

$$\langle S(n)\rangle = \sum_{m=1}^{N-n} \frac{S(m,n)}{N-n} \quad (9)$$

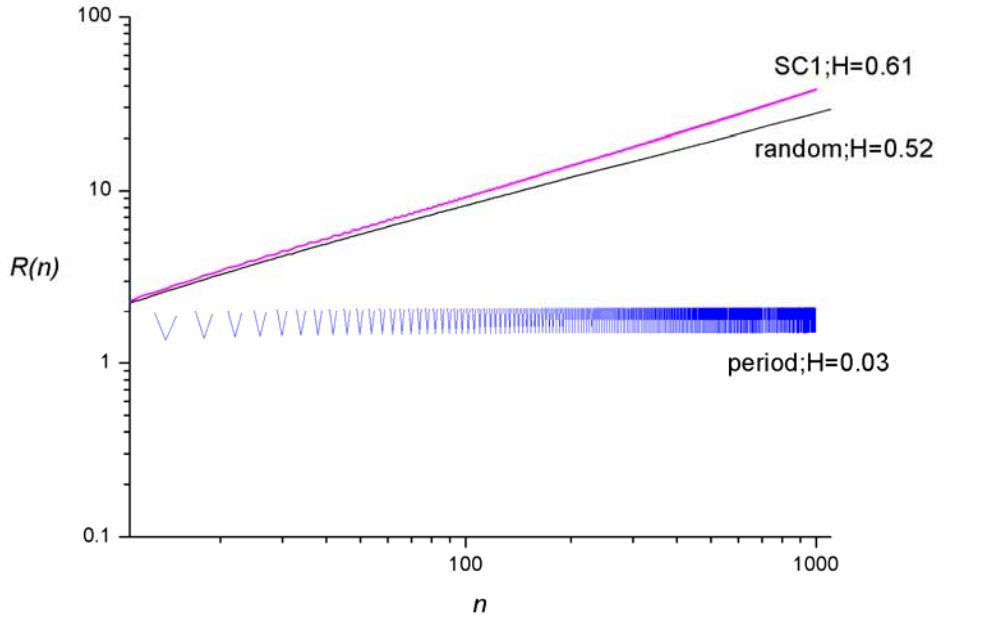
Finally the Hurst's exponent H is defined in the equation (9),

$$R(n) = \frac{\langle S(n)\rangle}{\sigma(n)} \propto n^{H} \quad (10)$$

where σ^2 (n) is the standard deviation of $u(i)$ over steps of length n and $R(n)$ is known as the rescaled range function.

Sequences generated from the ordinary Brownian motion show H = 0.5, a typical value for sequences without long-range correlations. Whereas H > 0.5 is expected for a long range correlated sequence.

- **H = 0.5** a random sequence without long range correlation.
- **H > 0.5** a sequence with long range correlation.

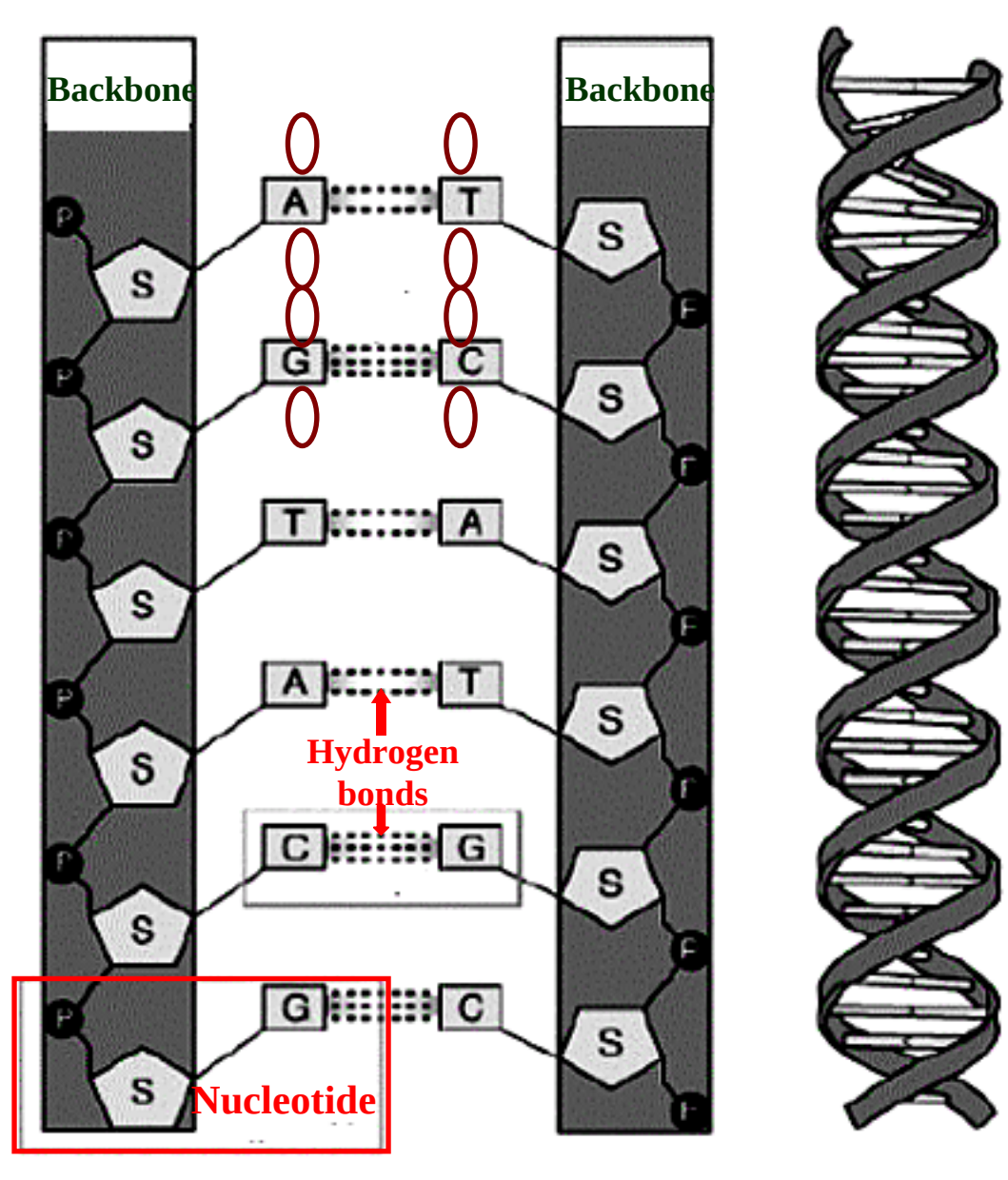


Rescaled range function *R(n)* versus *n*. The measured Hurst exponents are 0.61, 0.52, and 0.03 respectively for sequences taken from the first chromosome of *Saccharomyces cerevisiae*, a random sequence and a periodic array. Both the random sequence and the periodic sequence are of the length 100000 bp. In this figure the curve for *Saccharomyces cerevisiae* is drawn in pink color, the curve for the random sequence in black color, and the curve for the periodic sequence in blue color.

Number of S.C. chromosome	Length (bp)	Hurst exponent	DFA exponent
1	230208	0.61	0.67
2	813178	0.62	0.66
3	316616	0.60	0.67
4	1531916	0.63	0.67
5	576869	0.60	0.66
6	270148	0.62	0.69
7	1090946	0.63	0.67
8	562642	0.60	0.67
9	439885	0.59	0.68
10	745666	0.61	0.66
11	666454	0.63	0.68
12	1078174	0.61	0.67
13	924429	0.63	0.67
14	784331	0.61	0.68
15	1091287	0.61	0.67
16	948062	0.60	0.66

Hurst exponents and DFA exponents calculated from S.C. chromosome.

3. Charge Transfer in DNA



The primary structure of a DNA duplex with four nucleotides. The elliptical loops show the overlap of the π bonds along the base stacking direction.

Factors that affects charge transfer through a DNA chain:

1. Mechanical stress when a DNA chain is fastened between two electrodes for measurements.
2. The order of complexities in a sequence.
3. Local density distributions of different nucleotides.
4. The presence of neighboring water molecules and counter ions.
5. Temperature.
6. Many others.

Our application with the tight-binding model:

1. In a sense we are not able to construct a sensible model of physics at this stage in order to explain the real physical conduction mechanism but rather we are here to provide *various presentations of data analysis* according to the **tight-binding model** which has been extensively utilized for many authors. **[R3, R4, R5]**.
2. This type of research should be regarded as *a gateway to the real physics*.
3. Many averaged quantities obtained through the application of the tight-binding model can in principle be regarded as *characteristic quantities of various DNA sequences*.

Tight-binding model

The simplest **effective Hamiltonian** describing the propagation of a hole in the DNA chain is **[R3, R6, R7]** :

$$H = \sum_n \varepsilon_n |n\rangle\langle n| + t\sum_n \left[|n\rangle\langle n+1| + |n\rangle\langle n+1| \right] \qquad (11)$$

1. We made the same choice with reference **[R3]** for the hole site energies ε_n which are the energies required to excite an electron from corresponding nucleotide bases , ε_A **= 8.24 eV**, ε_T **= 9.14 eV**, ε_C **= 8.87 eV**, and ε_G **= 7.75 eV** (*A* = adenine, *T* = thymine, *C* = cytosine, and *G* = guanine).

2. The hopping integral ***t***, simulating the $\pi - \pi$ stacking between adjacent nucleotides and is the hopping probability amplitude between the neighboring sites, is taken to be **1 eV** in all calculations.

By proper rearrangement the Schrödinger equation can then be transformed into the following equation.

$$\psi_n \varepsilon_n + t\psi_{n+1} + t\psi_{n-1} - E\psi_n = 0 \qquad (12)$$

The function ψ_n is the projection of a eigenstate wave function ψ on the *n*th site and can be written as $\psi_n = \langle n|\psi\rangle$

These projected functions are related through the following equation:

$$\begin{pmatrix} \psi_{n+2} \\ \psi_{n+1} \end{pmatrix} = M_n \begin{pmatrix} \psi_{n+1} \\ \psi_n \end{pmatrix} = M_n \cdots M_1 \begin{pmatrix} \psi_1 \\ \psi_0 \end{pmatrix} \qquad (13)$$

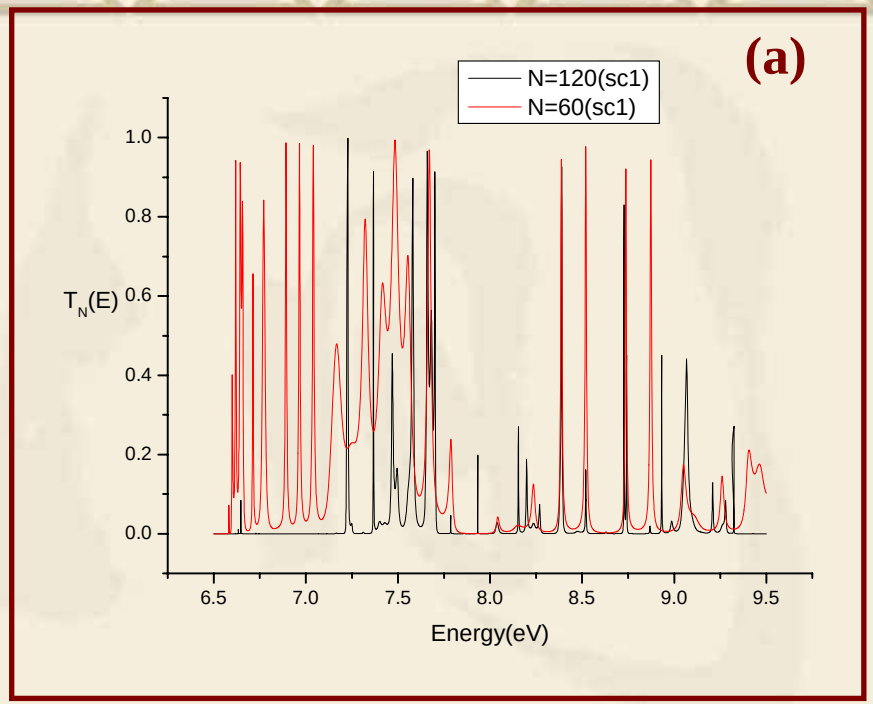
where $$M_n = \begin{bmatrix} \dfrac{E-\varepsilon_n}{t} & -1 \\ 1 & 0 \end{bmatrix} \qquad (14)$$

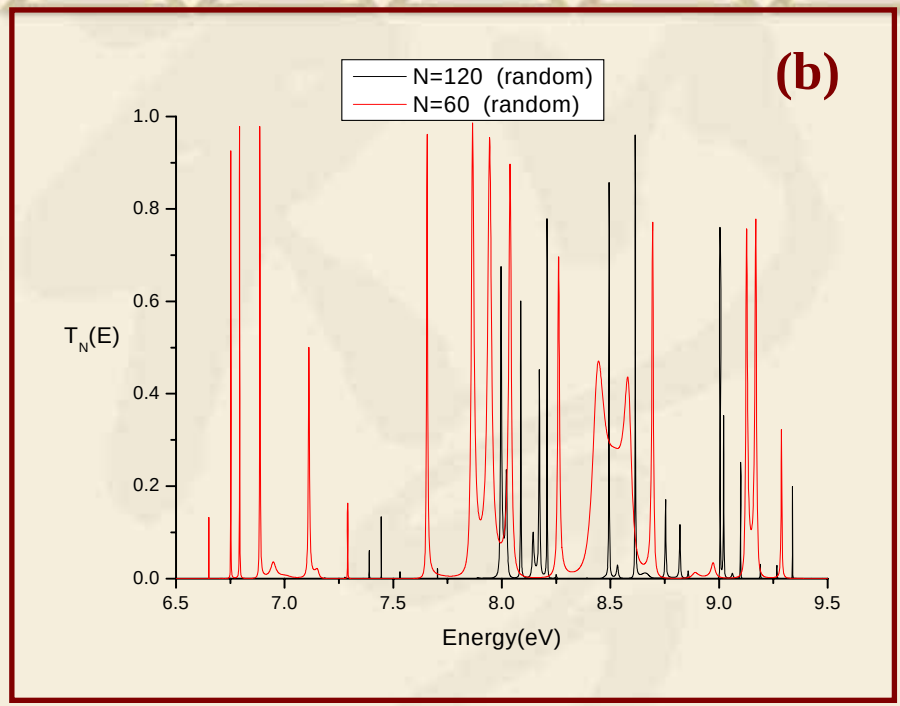
The transmission coefficient $T_N(E)$ which shows the probability of tunneling electrons through the *N*-site DNA chain is defined as

$$T_N(E) = [4 - \frac{(E-\varepsilon_m)^2}{t^2}] / \{ -\frac{(E-\varepsilon_m)^2}{t^2}(P_{12}P_{21}+1) + \frac{(E-\varepsilon_m)}{t}(P_{11}-P_{22})(P_{12}-P_{21}) + \sum_{i,j=1,2} P_{ij}^2 + 2 \} \quad (15)$$

with $P = M_N M_{N-1} \cdot \cdot \cdot M_1$ and M_n's are 2×2 matrices as defined in the equation (14).

The energy ε_m is the boundary energy assumed on the edge connecting points of two electrodes. It is taken to be the ionization energy of the guanine base, $\varepsilon_m = \varepsilon_G$.





(a) Transmission coefficients $T_N(E)$ as the function of energy E are calculated from two sequence chains taken from the first chromosome of *Saccharomyces cerevisiae*. Their lengths are N =**60** and **120**.

(b) Transmission coefficients $T_N(E)$ as the function of energy E calculated from *a random sequence* with N = **60** and **120** are shown here for comparison.

For each sequence of length L we calculate $T_N(E,i)$ for every piece of segment chain with width N and starting from the site i. Calculations were carried out from i = 1 to (L-N+1). In a sense the whole DNA sequence on a chromosome is taken to be a complete statistical ensemble. Afterward an integrated value, $S[T_N(E, i)]$, over an energy range ***E* = 5.75~9.75 eV** (ΔE = 4 eV) is computed **[R5]**.

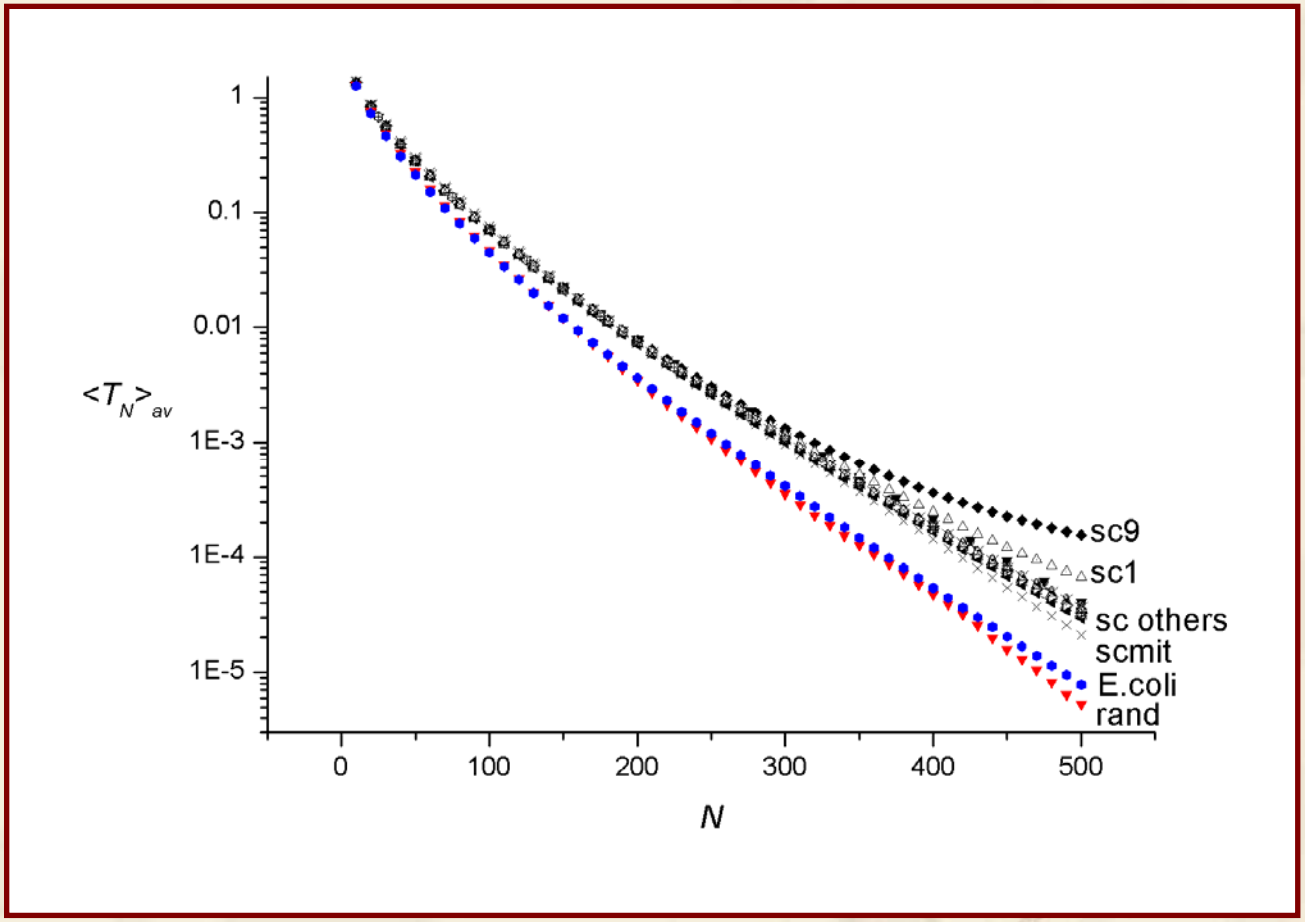
$$S[T_N(E,i)] = \int_{5.75}^{9.75} T_N(E,i)dE \qquad (16)$$

An averaged characteristic quantity can thus be defined as

$$\langle T_N \rangle_{av} = \frac{1}{M}\sum_{i=1}^{M} S[T_N(E,i)] \qquad (17)$$

where $M = L - N + 1$

The value $\langle T_N \rangle_{av}$ defined in the Eq. (17) is *an aggregated average quantity* over a well-defined energy range for a complete DNA sequence from a chromosome.



Averaged quantity $\langle T_N \rangle_{av}$ calculated from 16 chromosomes of *Saccharomyces cerevisiae* are shown in the figure. Much better conductivity of all sixteen chromosomal DNA chains are observed over a wide range of N when compared with values calculated from a random sequence.

4. Electronic wave functions in a DNA sequence

Charge conduction in DNA chains based on the tight binding model is shown a promising approach for discussion. It is then profitable to reveal in some *detail patterns of electronic wavefunctions* which might provide some indications or clues for further interpretations.

In order to simplify our calculation we take the same strategy as adopted in the **[R4]**. Site energies ε_n are assigned **0.5 eV** for purines and **-0.5 eV** for pyrimidines. This kind of oversimplification trims away detailed energy variations on all sites along a sequence chain but retain the feature of complexities on sequential order of two different types of nucleotide bases.

ε_n **= 0.5 eV** for purines **A** and **G**

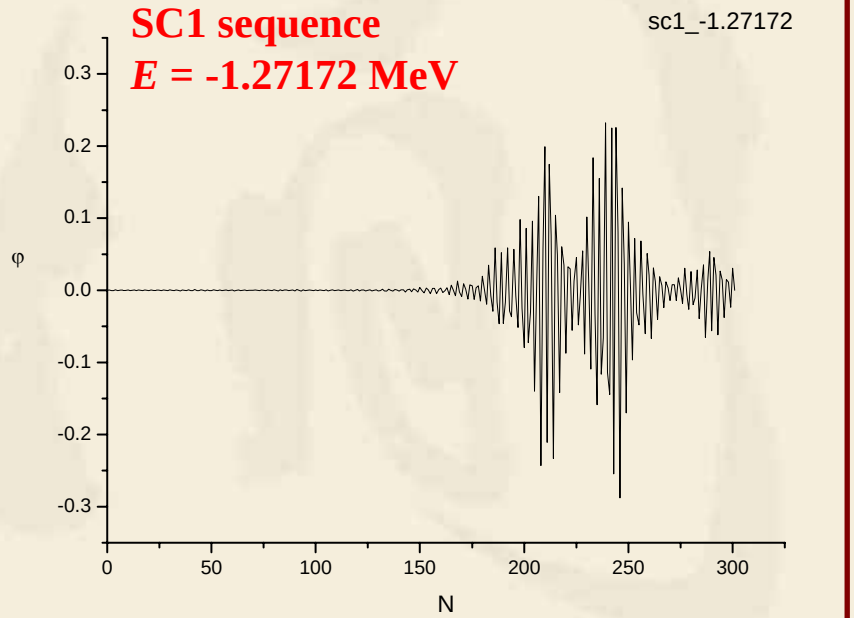
ε_n **= - 0.5 eV** for pyrimidines **C** and **T**

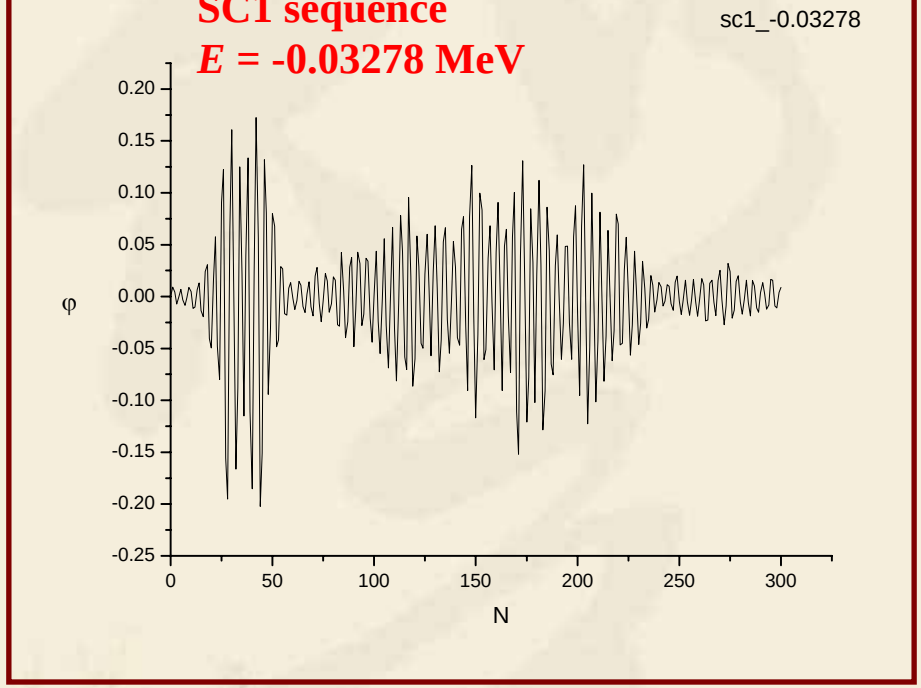
In order to calculate electronic wave functions in a DNA sequence of length N bp we transform the equation (5) into a matrix equation

$$\begin{pmatrix} \varepsilon_1 & 1 & 0 & \cdots & 0 \\ 1 & \varepsilon_2 & \ddots & \ddots & \vdots \\ 0 & 1 & \ddots & 1 & 0 \\ \vdots & \ddots & \ddots & \varepsilon_{N-1} & 1 \\ 0 & \cdots & 0 & 1 & \varepsilon_N \end{pmatrix} \begin{pmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_{N-1} \\ \psi_N \end{pmatrix} = \begin{pmatrix} E_1 & 0 & \cdots & \cdots & 0 \\ 0 & E_2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & E_{N-1} & 0 \\ 0 & \cdots & \cdots & 0 & E_N \end{pmatrix} \begin{pmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_{N-1} \\ \psi_N \end{pmatrix} \quad (18)$$

Electrons are assumed to be bound in the sequence so that $\psi_0 = \psi_{N+1} = 0$ on two boundary end sites are set.

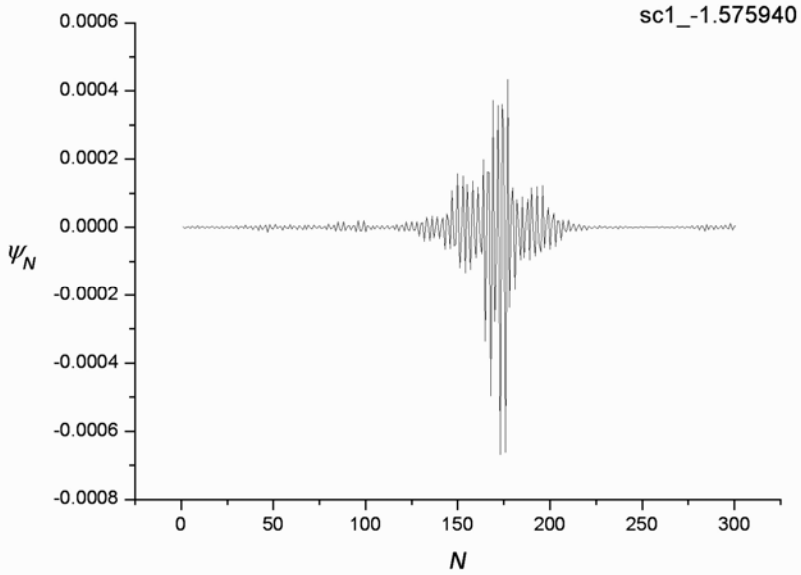
All component values of the wave function ψ on all sites for a sequence chain with length N are thus obtained.

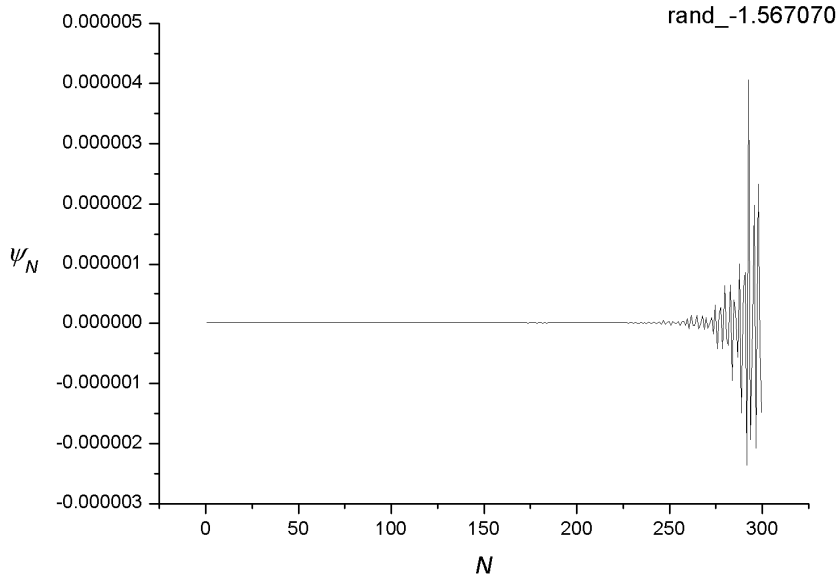




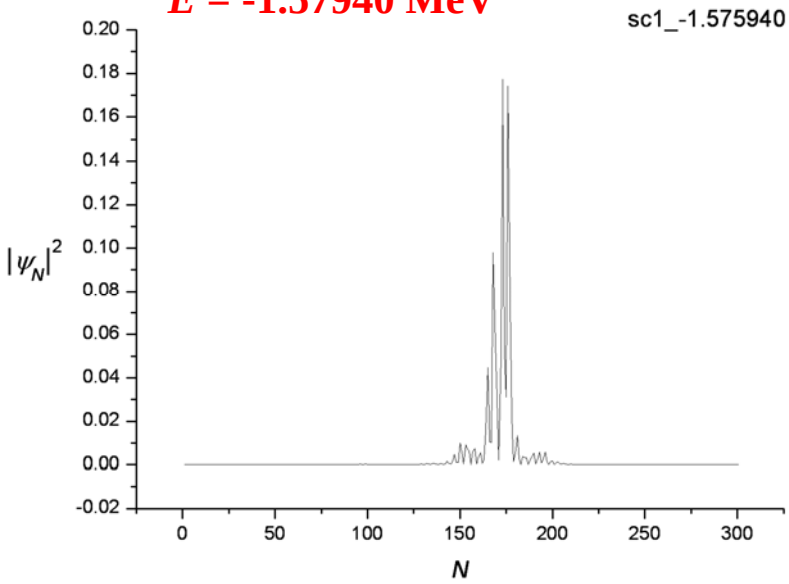
Example patterns of electronic eigenstate wave functions ψ_N for two different eigenenergies, ***E* = -1.271 eV** and ***E* = -0.033 eV**, are shown in this figure.

A sequence chain of length $N = 300$ is taken from the first chromosome for the calculation of ψ_N

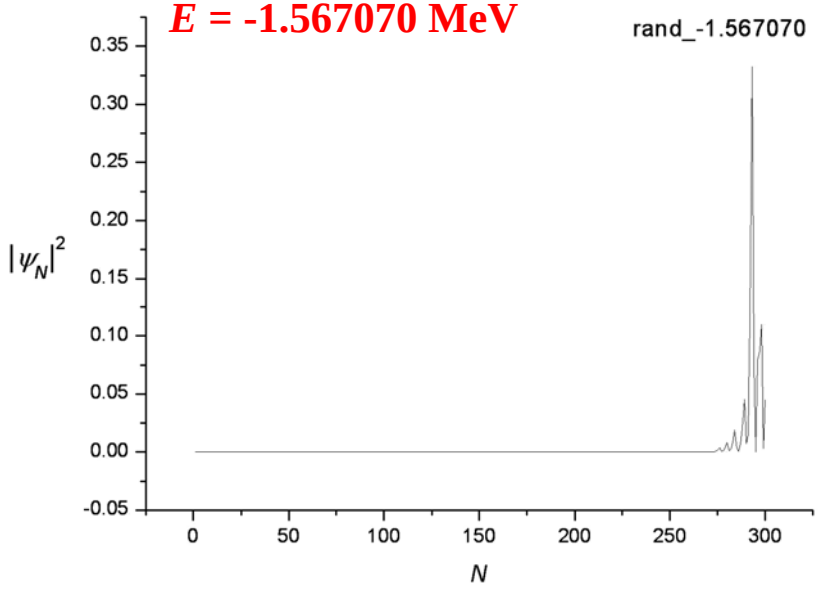




SC1 sequence
***E* = -1.57940 MeV**



random sequence
***E* = -1.567070 MeV**



Eigenstate wave function patterns are seen quite sensitive to the eigenenergy.

By computing the absolute square values of ψ_N on all sites we can have the electronic probability distribution

$$P(n) = \left|\psi_n\right|^2 \qquad (19)$$

which characterizes charge conduction through the chain.

Appearances of these probability distributions are so diverse that averaged characteristic quantities are needed for comparisons of segment chains from different DNA sequences.

For each probability distribution obtained with an eigenenergy E we define a mean radius

$$\left\langle r_P \right\rangle_E = \sum_{n=1}^{N} nP(n,E) \qquad (20)$$

to show an overall central position of a distribution function $P(n,E)$.

This quantity, $P(n,E)$, is then used to calculate *the second moment of the distribution* $\boldsymbol{D^2(E)}$.

$$D^2(E) = \sum_{n=1}^{N} P(n,E)\left(n - \langle r_P \rangle_E\right)^2 \qquad (21)$$

$\boldsymbol{D^2(E)}$ is the measure of distribution dispersion which characterizes the conductivity behavior along the chain on an eigenstate of energy E.

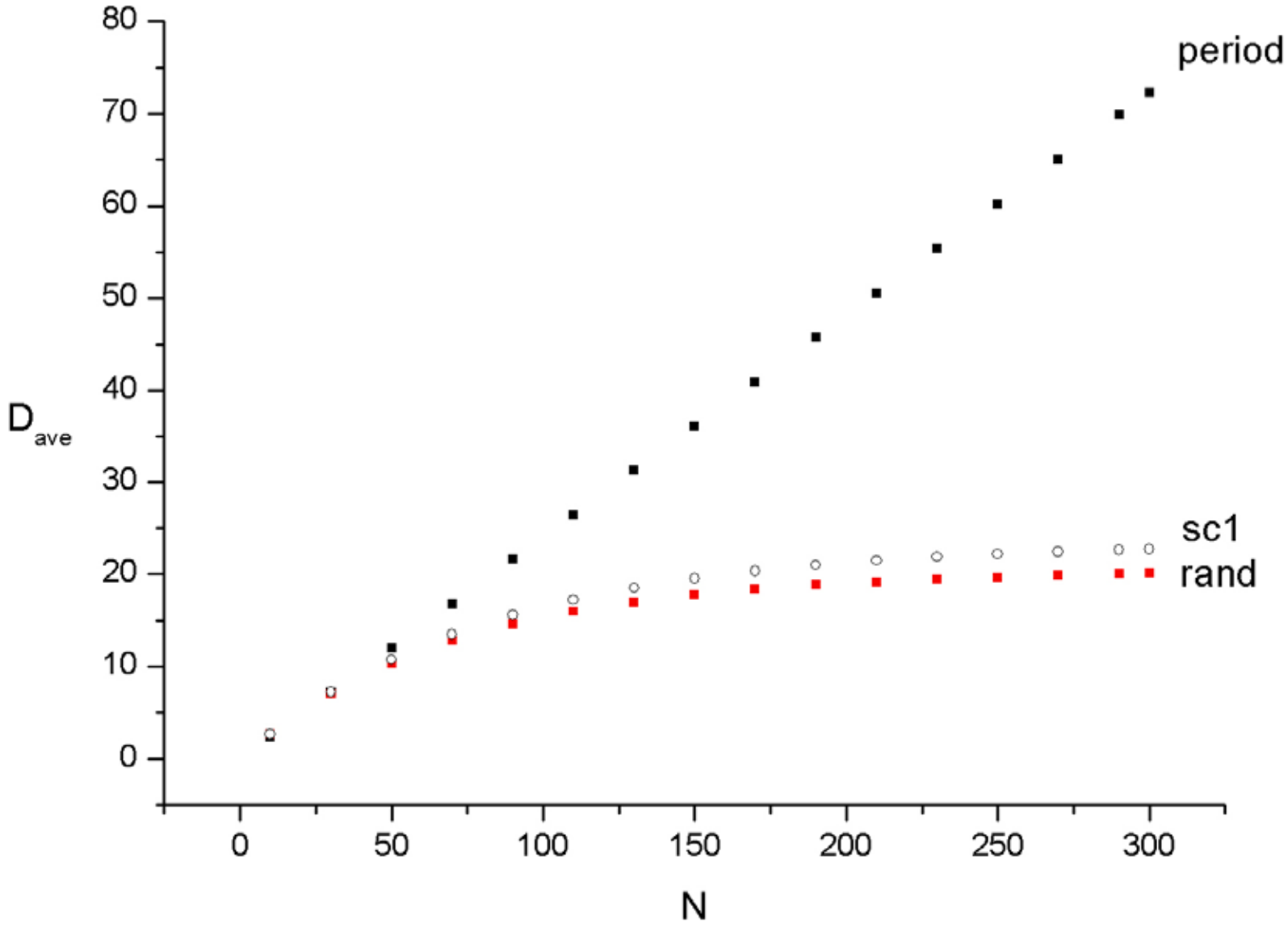
More dispersed distributions denote more delocalized wave functions and implies better ability of electronic conductions.

The square root values of the second moment $\boldsymbol{D^2(E)}$ are averaged through all eigenenergies and all segment chains from a complete DNA sequence to define an averaged value $\boldsymbol{D_{ave}}$.

Subsequently values of standard deviation σ for all chromosomes are also calculated.

Data	D_{ave} [bp]	σ [bp]
S.C.1	22.81	14.80
S.C.2	23.46	15.21
S.C.3	22.50	14.65
S.C.4	23.27	15.07
S.C.5	22.87	14.74
S.C.6	23.32	15.14
S.C.7	23.22	15.13
S.C.8	22.75	14.49
S.C.9	23.35	15.37
S.C.10	22.98	14.94
S.C.11	23.26	15.00
S.C.12	23.38	15.14
S.C.13	23.19	14.95
S.C.14	23.18	14.92
S.C.15	23.13	14.88
S.C.16	23.15	14.96
Random	20.14	13.17
Period	72.50	14.15

List of averaged second moment D_{ave} and its standard deviation (σ) are shown here. According to the data shown in this table long range correlated sequences from living cells indeed bear slightly better ability of charge conduction than a random sequence. A periodic sequence is expected to be the best conductor in this case.


80
70
60
50
40
30
20
10
0
D_{ave}
0
50
100
150
200
250
300
N
period
sc1
rand

5. Conclusions

1. It is profitable to design as many as possible methodologies to probe DNA's for further understandings on life phenomena as well as the natural evolution of life forms on the earth.

2. Base upon the *Hurst exponent analysis* and the *detrend fluctuation analysis* **(DFA)** we have confirmed long range correlations on DNA sequences from the *Saccharomyces cerevisiae* genome.

3. Base upon the *transmission coefficient analysis* and *study of overall averaged dispersion of wave functions* we conclude that DNA sequences from the *Saccharomyces cerevisiae* genome do indeed manifest better ability of charge conduction statistically when compared with a random sequence.

4. Data presented in this talk are from real DNA sequences that survived from harsh natural evolution process. Charge conduction ability is strongly affected by order arrangement of four different types of nucleotide bases. Better charge conductivity promises better damage recognition efficiency and may be better chance of survival through cruel natural environment.

References:

[R1] C.K. Peng *et al*. Nature (London) **356**, (1992) 168.

[R2] Stephan Roche, Dominique Bicout, Enrique Maciá, and Efim Kats, Phys. Rev. Lett. **91**, (2003) 228101.

[R3] Stephan Roche, Dominique Bicout, Enrique Maciá, and Efim Kats, Phys. Rev. Lett. **91**, (2003) 228101.

[R4] Pedro Carpena, Pedro Bernaola-Galvain, Plamen Ch. Ivanov, and H. Eugene Stanley, Nature **418**, (2002) 955.

[R5] C.T. Shih, Phys. Rev. E **74**, (2006) 010903.

[R6] Y. A. Berlin, A. L. Burin, and M. A. Ratner, Superlattices Microstruct. **28**, 241 (2000).

[R7] A.Voityuk *et al.*, J. Chem. Phys. **114**, 5614 (2001).