CHAPTER 9

Power Law Correlations in DNA Sequences

Sergey V. Buldyrev*

Introduction

A wide variety of natural phenomena is characterized by power law behavior of their parameters. This type of behavior is also called scaling. The first observation of scaling probably goes back to Kepler[1] who empirically discovered that squares of the periods of planet revolution around the Sun scale as cubes of their orbits radii. This empirical law allowed Newton to discover his famous inverse-square law of gravity.

In the nineteenth century, it was realized that many physical phenomena, for example diffusion, can be described by partial differential equations. In turn, the solutions of these equations give rise to universal scaling laws. For example, the root mean square displacement of a diffusing particle scales as the square root of time.

In the twentieth century, power laws were found to describe various systems in the vicinity of critical points. These include not only systems of interacting particles such as liquids and magnets[2] but also purely geometric systems, such as random networks.[3] Scaling is also found to hold for polymeric systems, including both linear and branched polymers.[4] Since then, the list of systems characterized by power laws has grown rapidly including models of rough surfaces,[5] turbulence and earthquakes. Empirical power laws are found to characterize also many physiological, ecological, and socio-economic systems. These facts give rise to the increasingly appreciated "fractal geometry of nature".[6-15]

A major puzzle concerning genomes of eukaryotic organisms, is that the large percent of their DNA is not used to code proteins or RNA. In human genome, this "junk" DNA constitutes 97% of the total genome length which is equal to 3 billion nucleotides also called base-pairs (bp). The role of non-coding DNA is poorly understood. It seems that it evolves by its own laws not restricted by a specific biological function. These laws are based on probabilities of various mutations and as such resemble the laws governing other complex systems listed above. In this chapter, I will review the degree to which power laws can characterize fluctuating nucleotide content of the DNA sequences, see also a critical review of W. Li.[16]

The term "long range correlations" is often misunderstood, implying some mystical long-range interactions or information propagation in space. Therefore, I will start with a brief introduction in the theory of critical phenomena, in which this concept has been developed. An impatient reader can jump directly to section "Correlation Analysis of DNA Sequences".

*Sergey V. Buldyrev—Department of Physics, Yeshiva University, 500 West 185th Street, New York, New York 10033, U.S.A. Email: buldyrev@yu.edu

Power Laws, Scale-Free Networks and Genome Biology, edited by Eugene V. Koonin, Yuri I. Wolf and Georgy P. Karev. ©2006 Eurekah.com and Springer Science+Business Media.

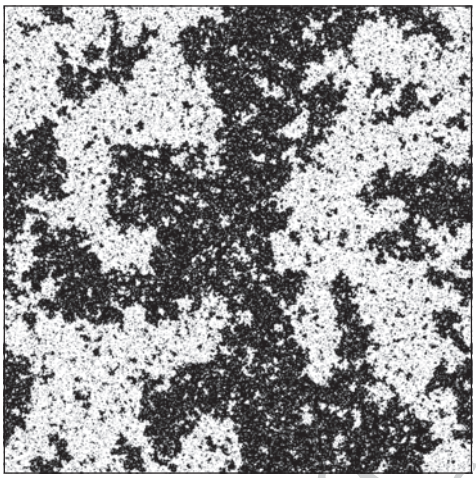

Figure 1. A snapshot of a two-dimensional system near its critical point. Black pixels represent gas particles. One can see density fluctuations of all different scales from a single particle to patches comparable with the entire system. This picture also represents an Ising magnetic near the Curie point, where black pixels are spins in positive orientation and white pixels are spins in negative orientation. The picture is obtained by computer simulations using the Metropolis algorithm at $T = T_c = 2.269185$.

Critical Phenomena and Long Range Correlations

One of the greatest advances in physics in the second half of twentieth century was the development of modern theory of critical phenomena.[2] The central paradigm of this theory is the importance of local fluctuations of the order parameter (Fig. 1). For a gas-liquid critical point, the order parameter is simply density. For the Curie point of a ferromagnetic, it is magnetization. Near the critical point T_c, the characteristic length scale ξ of the fluctuations, also known as the correlation length, grows according to a power law

$$\xi \sim |T - T_c|^{-\upsilon_c}. \tag{1}$$

The difference between the order parameters in the two phases (e.g., densities of gas and liquid) $\rho_l - \rho_g$ vanishes as the temperature approaches the critical point also according to a power law

$$\rho_l - \rho_g \sim (T_c - T)^{\beta_c}. \tag{2}$$

The positive quantities υ_c and β_c are called critical exponents. There are many other critical exponents α_c, γ_c, δ_c, η_c, etc., which characterize critical behavior of other parameters of the system.

The most spectacular manifestation of critical phenomena is critical opalescence. If one heats a closed transparent container filled by one third with water, the pressure inside it increases so that water and vapor remain at equilibrium: the water-vapor boundary is clearly visible and both phases are transparent. However, when the temperature approaches $T_c = 374$°C within 1°C, the phase boundary disappears, and the substance in the container becomes milky: the density fluctuations scatter light because their average size becomes larger than the wave length of light which is about half a micron. Thus the correlation length becomes more than thousand times larger than the average distance between molecules which is about 0.3 nanometers.

Since the fluctuations near the critical point become extremely large, the details of the interaction potential which acts on much smaller scales become irrelevant and hence all liquids

near the critical point have the same scaling behavior, i.e., they have exactly the same critical exponents, namely $\nu_c \approx 0.64$ and $\beta_c \approx 0.33$. Moreover, the theory predicts, that critical exponents are connected by several scaling relations, so that knowing any two exponents, for example ν_c and β_c one can predict the values of all the others. It turns out, that critical exponents depend only on dimensionality of space and some other major characteristics, such as dimensionality of spin orientations for magnetics. Thus, all variety of critical points can be classified by few universality classes so that all systems belonging to the same universality class have exactly the same values of critical exponents.

One of the simplest models for critical phenomena, the Ising model,[17] belongs to the same universality class as the liquid-gas critical point. We will discuss this model in greater detail, since it was first used by M. Ya. Azbel to describe possible correlations of nucleotides in the DNA.[18-20]

In the Ising model, atoms occupy sites on the d-dimensional lattice, for example on a square or a cubic lattice. In a one-dimensional system, atoms are placed equidistantly on a line. Each atom has a magnetic moment or spin, which may have only two orientations: up (s = +1) or down (s = -1). All pairs of spins occupying nearest neighboring sites interact with each other, so that they have a tendency to acquire the same orientation. The pair with the same orientations has negative potential energy $-\varepsilon$ while the pair with different orientations has positive potential energy $+\varepsilon$. Note that $\varepsilon < 0$ corresponds to the model of anti-ferromagnetic interactions. In addition, spins may interact with external magnetic field with energies $-h$ for positive spins and $+h$ for negative spins. It can be shown that this model is equivalent to the model of lattice gas, in which positive orientation of spins corresponds to the sites occupied by molecules, negative orientation indicates empty sites, two neighboring molecules attract with energy $-\varepsilon$, and the external field h corresponds to chemical potential which defines the average number of molecules in the system.

In 1973, M. Ya. Azbel[18] mapped a DNA sequence onto a one-dimensional Ising model by assigning positive spins s = +1 to strongly bonded pairs cytosine (C) and guanine (G) and negative spins s = -1 to weakly bonded pairs adenine (A) and thymine (T). (Complimentary base-pairs C and G located on the opposite strands of the DNA double helix are bonded by three hydrogen bonds, while A and T are boned only by two hydrogen bonds.)

One-Dimensional Ising Model

It is easy to solve the one-dimensional Ising model. According to the Boltzmann equation, the probability $p(U)$ to find a thermally equilibrated system in a state with certain potential energy is proportional to

$$p(U) \sim \exp(-U / k_B T), \tag{3}$$

where T is absolute temperature and k_B is Boltzmann constant. A striking simplicity of this equation is that it does not depend on any details of inter-atomic interaction and the details of motion of individual molecules. Once we know U and T, we can completely characterize our system in terms of the probability theory.

In the one-dimensional Ising model, a spin at position i can affect a spin at position i + 1 only through their direct interaction which is either $-\varepsilon$ if they orient the same way or $+\varepsilon$ if they orient in the opposite way. In the absence of magnetic field, the probabilities of these two orientations are proportional to $\exp(-U/k_B T)$, where $U = \pm\varepsilon$. Hence the probability of the same orientation is

$$p = \exp(\varepsilon / k_B T) / \left[\exp(\varepsilon / k_B T) + \exp(-\varepsilon / k_B T)\right] \equiv 1/(1+b), \tag{4}$$

where $b = \exp(-2\varepsilon/k_B T)$ and the probability of the opposite orientation is $q = 1 - p = b/(1 + b)$. Clearly, if T is small enough, b is also very small, and hence the probability for two neighboring spins to be in the same orientation is almost equal to one.

Do spins at a distant positions i and $(i + r)$ affect each other? To answer this question we must quantify this affect in mathematical terms. Two random variables $s(i)$ and $s(i + r)$ are called independent if the average of their product $\langle s(i)s(i + r)\rangle$ is equal to the product of their averages $\langle s(i)\rangle$ and $\langle s(i + r)\rangle$. Here and throughout the entire chapter $\langle \ldots \rangle$ denotes average taken over all possible positions i of the spins or nucleotide positions in a DNA sequence. The difference between these two quantities

$$\begin{aligned} C(r) &\equiv \langle s(i)s(i+r)\rangle - \langle s(i)\rangle\langle s(i+r)\rangle \\ &= \langle [s(i) - \langle s(i)\rangle][s(i+r) - \langle s(i+r)\rangle]\rangle \end{aligned} \tag{5}$$

characterizes the mutual dependence of two spins and is called correlation function. If $C(r) > 0$, the spins are correlated. If $C(r) < 0$, the spins are anti-correlated. Note that $C(0)$ coincides with the definition of variance of the variable $s(i)$. Note also that in general, for finite system of size L,$\langle s(i)\rangle \neq \langle s(i + r)\rangle$, because these two averages are taken over two different sets of positions i = 1,2...$L - r$ and $i + r = r + 1, r + 2, \ldots, L$. When r is comparable to L, this difference becomes substantial.

It can be easily shown (see next section) that for a one-dimensional Ising model the correlations decay exponentially $C(r) \sim \exp(-r/\xi)$ at any temperature. The inverse speed of the exponential decay ξ is identical to the correlation length. In the one-dimensional model, correlation length can diverge only if temperature approaches absolute zero. Thus the critical point for the one-dimensional model is $T_c = 0$.

In the next section we will show this by making a mathematical excursion into the theory of Markovian processes, which is a very useful tool in bioinformatics. This chapter may be omitted by a reader who does not want to go deep into mathematical details, but is useful for those whose goal is to apply mathematics in biology.

Markovian Processes

In order to compute correlation function, we will represent a sequence of spins in the Ising model as a Markovian process. Markovian processes are very important in bioinformatics, thus we briefly summarize their definition and properties.

A Markovian process[21] is defined as a process obeying the following rules. (i) A system at any time step t, can be in n possible states $\mathbf{e}_1, \mathbf{e}_2, \ldots \mathbf{e}_n$. (ii) The probability to find a system in a certain state at any time step depends only on its state at the previous time step. Thus to fully characterize a Markovian process, we must define a n x n set of transition probabilities p_{ij} which are the probabilities to find a system in a state $\mathbf{e}_i$ at time $t + 1$ provided that a time t it was in a state $\mathbf{e}_j$. Obviously, $\sum_{i=1}^{n} p_{ij} \equiv 1$. (iii) It is assumed that p_{ij} do not depend on time.

It is convenient to describe the behavior of a Markovian process in terms of vector algebra, so that the probabilities $p_i(t)$ to find a system in any of its n states at time t is an n-dimensional vector-column $\mathbf{p}(t)$. The sum of its components $p_i(t)$ is equal to unity. Accordingly, it is natural to arrange the transition probabilities p_{ij} into a n x n matrix $\mathbf{P}$. The j-th column of this matrix is the set of transitional probabilities p_{ij}. Using the rule of matrix multiplication combined with the law of probability multiplication for independent events, we can find

$$\mathbf{p}(t + r) = \mathbf{P}^r \mathbf{p}(t), \tag{6}$$

where $\mathbf{P}^r$ is the r–th power of matrix $\mathbf{P}$, which can be easily found once we determine eigenvectors $\mathbf{a}_i$ and eigenvalues λ_i of matrix $\mathbf{P}$. By definition, eigenvectors and eigenvalues satisfy a homogeneous system of linear equations

$$\mathbf{P}\mathbf{a}_i = \lambda_i \mathbf{a}_i. \tag{7}$$

which has a nontrivial solution only if its determinant is equal to zero. Accordingly, the eigenvalues satisfy an algebraic equation of n-th power which turns the determinant of the matrix $\mathbf{P} - \lambda\mathbf{I}$, where I is the unity matrix, into zero.

Once we have determined the eigenvectors and eigenvalues, we can write

$$\mathbf{P}^r = \mathbf{A}\Lambda^r\mathbf{A}^{-1}, \tag{8}$$

where Λ is the diagonal matrix formed by eigenvalues λ_i, and $\mathbf{A}$ is the matrix whose columns are eigenvectors $\mathbf{a}_i$.

Since the sum of elements in every column of matrix P is unity, the determinant of the matrix $\mathbf{P} - \mathrm{I}$ is equal to zero and one of the eigenvalues must be equal to unity: $\lambda_1 = 1$. The eigenvector $\mathbf{a}_1$, corresponding to this eigenvalue has a very special meaning. Its components yield the probabilities to find the system in each of its states for $r \to \infty$. We will show it in the following paragraph.

Except in some special degenerate cases, all the eigenvalues of a matrix are different. Assuming this, we can express the state of the system at time $t = r$ as a linear combination of the eigenvectors:

$$\mathbf{p}(t + r) = c_1\mathbf{a}_1 + c_2\lambda_2^r\mathbf{a}_2 + \ldots + c_n\lambda_n^r\mathbf{a}_n$$

where c_n are some coefficients, which can be obtained by multiplying the initial state of the system $\mathbf{p}(t)$ by matrix $\mathbf{A}^{-1}$. It can be easily seen from this equation that all eigenvalues must be less or equal to one: $|\lambda_i| \leq 1$. Indeed, if any $|\lambda_i| > 1$, the corresponding term in the above equation would diverge for $r \to \infty$, contradicting inequality $p_i(r) \leq 1$, which must be satisfied by the probabilities. Thus for all $i > 1$, $|\lambda_i| < 1$, and for any initial state of the system, we have $\lim_{r\to\infty}\mathbf{p}(r + t) = c_1\mathbf{a}_1$.

Thus, the average probability of finding the system in each of its states in a very long process is determined by the vector $c_1\mathbf{a}_1$, which can be readily found from the system of linear equations:

$$\mathbf{P}\mathbf{a}_1 = \mathbf{a}_1. \tag{9}$$

Since the determinant of this system is equal to zero, it has a nontrivial solution $c_1\mathbf{a}_1$, where c_1 is an arbitrary constant. Since the components of the vector $c_1\mathbf{a}_1$ have the meaning of the probabilities and, therefore, their sum must be equal to one, the coefficient c_1 must be the reciprocal of the sum of the elements of an arbitrary non-trivial solution $\mathbf{a}_1$ of Eq. (9).

The second-largest eigenvalue determines the decay of the correlations: $C(r) \sim \lambda_2^r = \exp(r \ln \lambda_2)$. By definition, the correlation length is the characteristic length of correlation decay which is determined by relation $C(r) \sim \exp(-r/\xi)$. Thus $\xi = 1/\ln(1/\lambda_2)$.

As an illustration of the Markovian formalism we can apply it to the one-dimensional Ising model. The matrix $\mathbf{P}$ in this case is simply

$$\mathbf{P} = \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix}, \tag{10}$$

where p is determined by Eq. (4). In order to find the eigenvalues, we must find the values of λ which turn the determinant of the matrix $\mathbf{P} - \lambda I$ into zero:

$$\begin{vmatrix} p-\lambda & 1-p \\ 1-p & p-\lambda \end{vmatrix} = 0. \tag{11}$$

This gives us a quadratic equation $(p-\lambda)^2-(1-p)^2=0$, with two roots $\lambda_1=1$, and $\lambda_2=2p-1$. The corresponding eigenvectors are

$$\mathbf{a}_1=\begin{pmatrix}\frac{1}{2}\\ \frac{1}{2}\end{pmatrix}\quad \mathbf{a}_2=\begin{pmatrix}-1\\ 1\end{pmatrix}. \tag{12}$$

Accordingly, we have

$$\mathbf{A}=\begin{pmatrix}\frac{1}{2} & -1\\ \frac{1}{2} & 1\end{pmatrix},\quad \mathbf{A}^{-1}=\begin{pmatrix}1 & 1\\ -\frac{1}{2} & \frac{1}{2}\end{pmatrix}. \tag{13}$$

and using Eq.(8),

$$\mathbf{P}^r=\begin{pmatrix}\frac{1+(2p-1)^r}{2} & \frac{1-(2p-1)^r}{2}\\ \frac{1-(2p-1)^r}{2} & \frac{1+(2p-1)^r}{2}\end{pmatrix}. \tag{14}$$

So we can see that the diagonal elements of this matrix, $p(r)=1/2+(2p-1)^r/2$ exponentially converge to 1/2 for $r\to\infty$. The speed of convergence determines the correlation length:

$$\xi=-1/\ln(2p-1)=1/\ln\left\{\left[\exp(2\varepsilon/k_BT)+1\right]/\left[\exp(2\varepsilon/k_BT)-1\right]\right\}. \tag{15}$$

For $T\to 0$ the correlation length diverges $\xi\approx\exp(2\varepsilon/k_BT)\to\infty$, while for $T\to\infty$ the correlation length approaches zero: $\xi\approx 1/\ln(k_BT/\varepsilon)\to 0$. For finite temperature the correlation length is finite. Hence for one-dimensional Ising model, there is no critical point at positive temperatures, however the absolute zero $T=T_c=0$ can be treated as a critical point because in its vicinity the correlation length diverges faster than any power. So one can identify exponent ν_c as being infinite.

The eigenvector $\mathbf{a}_1$ gives us equal probabilities for a spin to be in positive and negative orientations, thus the spontaneous magnetization being determined as $\langle s(t)\rangle=a_{11}-a_{21}=1/2-1/2=0$ remains zero for all temperatures. In order to compute correlation function, we must compute the average product $\langle s(t)s(t+r)\rangle$. With probability a_{11} the value $s(t)=1$. Given $s(t)=1$, the probabilities of $s(t+r)=1$ and $s(t+r)=-1$ are equal to the elements of the first column of matrix $\mathbf{P}^r$. Analogously for $s(t)=-1$, which occurs with probability a_{21}, the probabilities of $s(t+r)=1$ and $s(t+r)=-1$ are given by the elements in the second column of matrix $\mathbf{P}^r$. So

$$\langle s(t)s(t+r)\rangle=a_{11}\left[p_{11}(r)-p_{21}(r)\right]-a_{21}\left[p_{21}(r)-p_{22}(r)\right]=(2p-1)^r \tag{16}$$

and, therefore,

$$C(r)=(2p-1)^r. \tag{17}$$

Exponential versus Power Law Correlations

In the previous section, we see that the one-dimensional Ising model in the absence of magnetic field is equivalent to a two-state Markovian process. In general, it is clear that any one-dimensional model with short range interaction is equivalent to a Markovian process with a finite number of states, and for such a process correlations must decay exponentially as λ_2^r, where $\lambda_2<1$. Thus the correlation length must be finite and can diverge only for $T\to 0$. Intuitively, we can imagine a one dimensional model as a row of dancing people each holding hands with two neighbors: one is on the left and one is on the right. Once they are holding hands, the correlation can pass from one neighbor to the next. No matter how strong they are holding hands, there is a finite chance q that they will separate, and the correlation will stop.

The probability that the correlation spreads distance r is proportional to $(1 - q)^r \approx \exp(-qr)$, and hence the correlation length is finite and is inverse proportional to q.

In contrast, if the number of dimensions is larger than one, the interactions can propagate from point A to point B not only along a straight line from one neighbor to the next, but along an infinite number of possible paths connecting A and B. Accordingly, the correlation length can diverge for $T = T_c > 0$. Unfortunately, there are very few 2-dimensional models which can be solved exactly[17] and even those models have so complicated solutions that they are far beyond the scope of most physics textbooks. The most famous example of an exactly solvable 2-dimensional model is the Ising model, which was solved by Onsager in 1949. The solution is based on transfer matrices much more complicated than those we use in Section IV to solve the one dimensional model. It is much easier to simulate such a model on a computer and find an approximate numerical solution.

It can be shown that two-dimensional Ising model has a critical point at temperature $T_c = 2\varepsilon/\ln(1 + \sqrt{2})/k_B = 2.269\varepsilon/k_B$. At the vicinity of this temperature, the correlation function acquires a non-exponential behavior

$$C(r) \sim r^{-\eta} \exp(-r/\xi) \tag{18}$$

where $\eta = -1/4$ is a new critical exponent proposed by M. Fisher in 1964. The correlation length ξ diverges as $|T - T_c|^{-1}$, which means that $\nu_t = 1$. The spontaneous magnetization for $T < T_c$ is not equal to zero, but, for any sample, it can be either positive or negative.

The absolute value of spontaneous magnetization approaches zero for $T \to T_c$ as $(T - T_c)^{1/8}$, so $\beta_c = 1/8$. If the temperature increases above T_c (also known as Curie point), the sample loses its magnetization. This phenomenon can be observed by everyone in a kitchen-style experiment: take an arm from a compass, place it into the fire of the burner and keep it there until it starts to glow red. The Curie point for Iron is 700°C. Cool it and place it back into the compass. It does not show North any longer!

Figure 1 shows the results of a computer simulation of a two-dimensional Ising model on the $L \times L = 1024 \times 1024$ square lattice. The program is very simple. At any time step, a computer attempts to "mutate" a spin at a randomly chosen lattice site. It first computes the energy change ΔU in such a would be mutation. If $\Delta U \le 0$ the mutation always happens, if $\Delta U > 0$, it happens with probability $\exp(-\Delta U/k_B T)$. This algorithm invented by Metropolis in 1953,[22] leads to the Boltzmann distribution (3) of the probabilities to find a system in a state with total potential energy U. The proof of this fact is based on the theory of Markovian processes. Indeed, the set of Metropolis rules of flipping the spins can be represented as a transition matrix $\mathbf{P}$ with transition probabilities $p_{ij}\exp(-U_j/k_B T) = p_{ji}\exp(-U_i/k_B T)$, where U_i and U_j are the potential energies of the corresponding states. Obviously, vector $\mathbf{a}_1$ with components $a_{i1} = \exp(-U_i/k_B T)$ taken from the probability distribution (3) satisfies Eq. (9).

The system has periodic boundary conditions, so that pixels on the opposite edges of the system are in close proximity. In fact, the entire system can be viewed as a single line winded around the surface of a bagel. In such a system, site i has 4 neighbors $i + 1$, $i - 1$, $i + L$, and $i - L$, so the correlation can make really long jumps of length L and $-L$ along the line.

Black and white pixels show spins with positive and negative orientations respectively. One can see patches of irregular shapes and all possible sizes from very small, of one pixel size, to the giant one spanning the entire system. This scale-free property of patches is typical for systems with long range correlations with power law decay. Indeed, exponential decay of correlations $C(r) \sim \exp(-r/\xi)$ would imply a typical size ξ of patches so that the probability to find larger patches is exponentially small. The same picture can describe the behavior of gas particles near critical point. The molecules form clusters of all possible sizes which scatter light. Does this picture have anything to do with DNA?

It is well known that the DNA sequence has a mosaic structure[23] with patches of high concentration of strongly bonded CG base pairs alternating with patches of weakly bonded AT base pairs. These patches are called isochores and can span millions of base pairs. On a smaller scale of genes and exons, coding sequences have larger CG content than non-coding sequences. Finally there exist CpG islands of several hundred base pairs with high CG content.

May these patches have anything to do with Ising model? Of course DNA is not at thermal equilibrium and the concepts of temperature and potential energy cannot be applied to the study of its evolution. However, the evolution of DNA may be thought of as a Markovian process, similar to the Metropolis algorithm described above with mutation probabilities depending on the nature of neighboring nucleotides and on the pool of the surrounding nucleotides during replication process, which may be viewed as an external field or chemical potential.

There are several main objections to this idea:

1. **First objection**: the DNA chain is one dimensional. As we have seen above, long range correlations cannot exist in a one dimensional system.
   This objection can be easily overcome by the argument that the DNA molecule has an extremely complex three dimensional structure in which distant elements along the chain are in close geometrical proximity. Thus the correlation may propagate not only along the chain but may jump many steps ahead as in a toroidal Ising model shown in Figure 1. In 1993 Grosberg et al[24] proposed a model based on the distribution of loops in the polymer chain crumpled into a dense globular conformation. This simple model leads to the long range correlations decaying as a power law $r^{-\gamma}$, where $\gamma \approx 2/3$.
2. **Second objection**. The long-range correlations emerge only in the narrow vicinity of the critical point. Why in the biological system such as DNA, the probabilities of mutations are such that they correspond to the vicinity of the critical point?
   This objection is more difficult to overcome. However there are examples of simple models which drive themselves to the critical behavior. The most relevant example is a polymer chain in the solvent, in which the probability to find a monomer in a unit volume at distance r from a given monomer decays as $r^{1/\nu-3}$, where $\nu \approx 0.59$ is the correlation length exponent first determined by a Nobel prize winner P. Flory in 1949. In 1972, another Nobel prize winner P. G. de Gennes[4] mapped the problem of self-avoiding walks (which are believed to describe the behavior of polymers) to a model of a magnetic similar to an Ising model. He showed that the inverse polymer chain length $1/N$ is equivalent to the distance to the critical point $T - T_c$, and hence the correlation length ξ (which is equivalent to the radius of the polymer coil) grows as N^{ν}. A polymer chain has also a power law distribution of loops, determined by Des Cloiseaux.[25]
   In recent years, many models have been proposed that have a tendency of self organization (SOC) toward their critical points without any tuning of external parameters.[26,27] These models give rise to scaling, and produce sudden avalanche-like bursts of activity distributed according to a power law. Some SOC models are one-dimensional systems and have been applied to biological evolution.[28-30] Such models are of great interest and they might be relevant in studies of DNA sequences.
3. **Third objection**. Biological evolution is an extremely complex process which is governed by many different mechanisms acting at different length and time scales. The interplay of several characteristic length scales may lead to apparent power-law correlations, which thus lack universality of critical phenomena.[23]
   This objection is most probably correct. Indeed, the values of the correlation exponent are different for different species and change with distance r between the nucleotides (See section "Analysis of DNA Sequences"). Never the less, in the beginning of 1990s when the first long DNA sequences became publicly available, the idea to study them by correlation analysis attracted lot of attention.[31-41]

Correlation Analysis of DNA Sequences

Can correlation analysis be applied to DNA sequences? For a physicist or mathematician a DNA sequence looks like a text written in an unknown language, which is encoded in a 4-letter alphabet *A*, *C*, *G*, *T*. Each letter in this text corresponds to a DNA base pair. The first question one might ask is what is the overall fraction or frequency of each letter in this text. For example the frequency of letter "A", f_A is defined as $f_A = N_A/N$, where N_A is the number of letters "A" and N is the total length of the sequence.[42] This question is easy to answer, especially these days, when the total human genome is sequenced. In human genome, $f_A = f_T \approx 0.295$ and $f_G = f_C \approx 0.205$. Note that these numbers strongly depend on the organism under study. The second question one might ask: "Is there any apparent structure in this text, or it is indistinguishable from a text that would be typed by throwing a 4-sided dice?" (This dice can be made in the form of a Jewish toy, dreidel, with letters *A*, *C*, *G*, *T* on its sides which have slightly different surface areas, so that the probability of getting a letter on the top is equal to its frequency in the genome). For a text created by throwing such a dice, the events of getting any two letters at positions k and j are believed to be independent, so the probability of simultaneously getting letter X at position k and letter Y at position j is equal to the product $f_X f_Y$. If there is any structure in the text, the frequency $f_{XY}(r)$ of finding X at position k and Y at position $k + r$ will deviate significantly from the predicted value $f_X f_Y$.

To fully characterize all dependencies among four letters of the DNA alphabet one must compute 16 elements of dependence matrix $D_{XY}(r) = f_{XY}(r) - f_X f_Y$.[43] These dependence coefficients are equivalent to correlation functions used in the previous section to describe Ising model if the nucleotide sequence is replaced by a numerical sequence $s_x(k) = 1$ if nucleotide X is present at position k and $s_x(k) = 0$ otherwise:

$$D_{XY}(r) = \left\langle s_X(k) s_Y(r+k) \right\rangle - \left\langle s_X(k) \right\rangle \left\langle s_Y(k+r) \right\rangle, \tag{19}$$

where $\langle \ldots \rangle$ indicates the average over all k.

All other measures of correlations including nonlinear measures such as mutual information.[43-45] can be expressed via dependence coefficients. For example, one can introduce Purine-Pyrimidine (RY) correlation measure, in which any purine (A,G) is replaced by 1 and any pyrimidine (C,T) is replaced by -1. The numerical sequence for RY can be expressed as a linear combination of numerical sequences for each nucleotide $s_{RY} = s_A - s_C + s_G - s_T$. Accordingly,

$$\begin{aligned} C_{RY}(r) &= \left\langle s_{RY}(k), s_{RY}(k+r) \right\rangle - \left\langle s_{RY}(k) \right\rangle \left\langle s_{RY}(k+r) \right\rangle \\ &= D_{AA} + D_{CC} + D_{GG} + D_{TT} + 2\left(D_{AG} + D_{CT} - D_{GT} - D_{TT} - D_{AC} - D_{AT}\right). \end{aligned} \tag{20}$$

Analogously, one can introduce C_{SW}, ($S = C,G$; $W = A,T$) or C_{KM} ($K = A,C$; $M = G,T$) or any other correlation function based on a linear combination of the elementary measures s_A, s_C, s_G and s_T.[32,46,47] The coefficients of this linear combination can be presented in the form of a vector $\mathbf{m} = (m_A, m_C, m_G, m_T)$ which we will call a mapping rule. For example, for *RY* mapping rule, we define $\mathbf{m} = (1,-1,1,-1)$, and for *C* mapping rule we define $\mathbf{m} = (0,1,0,0)$. Accordingly, any correlation measure could be expressed as a quadratic form ($\mathbf{m} \cdot \mathbf{Dm}$), where $\mathbf{D}$ is the dependence matrix.

Definitely, some of these correlation measures such as C_{SW} are not zero for at least the size of the isochore i.e., a chromosomal region with high or low $C + G$ content. Isochores have a typical size of about 10^5 base pairs, so the correlations would be non-zero for at least $r \approx 10^5$.

A physicist whose goal is to understand some general principles of DNA organization may attempt to fit the behavior $D_{SW}(r)$ of by a power law function. A mathematical biologist[42,48-50] would rather try to characterize the size distribution of the isochores and their nucleotide content for various chromosomes and species and try to answer questions of

biological relevance rather than to measure some power law exponent, which has an ambiguous biological meaning and characterize isochores in a very indirect fashion.

In general, DNA is known for its complex mosaic structure,[23,42] with structural elements such as isochores, intergenic sequences, CpG islands, LINE(long interspersed elements) and SINE (short interspersed elements) repeats, genes, exons, introns, and tandem repeats.[51,52,53] Each of these structural elements has its different size distribution, nucleotide frequencies, and laws of molecular evolution, so the correlations in the DNA sequence have very complex structure, are different for different spices and can not be characterized by a universal power-law exponent, in a way it is observed in critical phenomena. Correlation studies by their nature involve averaging over large portions of a sequence, so they have a tendency to gloss over particular details. This is the main reason why they are not very popular in bioinformatics whose main tool is the search for sequence similarities[54] analogous to finding in an unknown language some already known words or names, which may shed some light on the meaning of their neighbors.

Never the less, characterization of correlations in DNA sequences has some intellectual merit and even practical importance for a biologist whose goal is to understand molecular evolution of DNA sequences.[55] There are several reasonable models of DNA evolution in which exact power-law correlations emerge.[56-59] The values of the exponents of these power laws depend on the parameters of the model, such as mutation rates and thus can be used to test certain assumptions of the models. These models are discussed in the three sections starting with section "Mutation Duplication Model of DNA Evolution".

Another problem with correlation studies, is that they can be affected by many characteristics of the system, for example sequence length. In order to avoid many potential pitfalls it is very important to understand basic properties of correlation measures and fine-tune them on the well known systems which can serve as golden standards. In the next sections we will introduce various correlation measures and illustrate their usage, applying them to the Ising model, whose correlation properties are well known. Again, an impatient reader may proceed to "Mutation Duplication Model of DNA Evolution".

Correlation Function

In the next four sections we will describe several methods of correlation analysis. To develop some intuition on their advantages and disadvantages we will apply them to the one-dimensional and two-dimensional Ising models, whose correlation properties are known theoretically.

The most straightforward analysis is the direct computation of the correlation function $C(r)$ defined in Eq. (5). Figure 2 shows the behavior of $\ln C(r)$ for the one-dimensional Ising model consisting of $L = 2^{16}$ spins for several values of T approaching zero. For small values of r, the graphs are straight lines with the slope equal to the inverse correlation length in complete agreement with Eq. (17). Figure 3 shows the behavior for the two-dimensional Ising model consisting of $L^2 = 2^8 \times 2^8$ spins above and below critical point. Figure 4 presents the corresponding snapshots of the system. The correlation length increases while temperature decreases toward $T_c \approx 2.27$ and then very quickly goes down again, as temperature continues to decrease.

This behavior may seem counterintuitive. Indeed, one can argue that correlations below T_c are so strong that the majority of spins acquire the same orientation. However, from a mathematical point of view, the majority of spins, say fraction p, has the same orientation. (In Fig. 4, T = 2.17, it is positive, but in other simulations, it may appear negative). White patches, indicating the negative orientation are small, isolated, and randomly distributed in the sample. These patches of the opposite orientation may be regarded as defects in the crystalline structure. Thus one can regard two spins at distant positions r and $r + k$ to be two independent random variables taking value 1 with probability p and value –1 with probability $1 - p$.

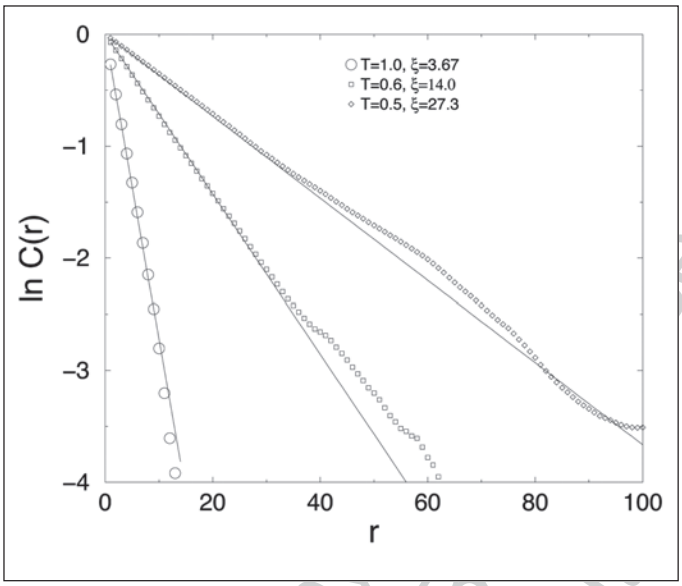

Figure 2. Logarithms of the correlation functions for the one-dimensional Ising model with $L = 2^{16}$ spins at three different temperatures $T = 1.0$ (°), $T = 0.6$ (□) and $T = 0.5$ (◊). The lines are drawn according to theoretical predictions of Eq. (17).

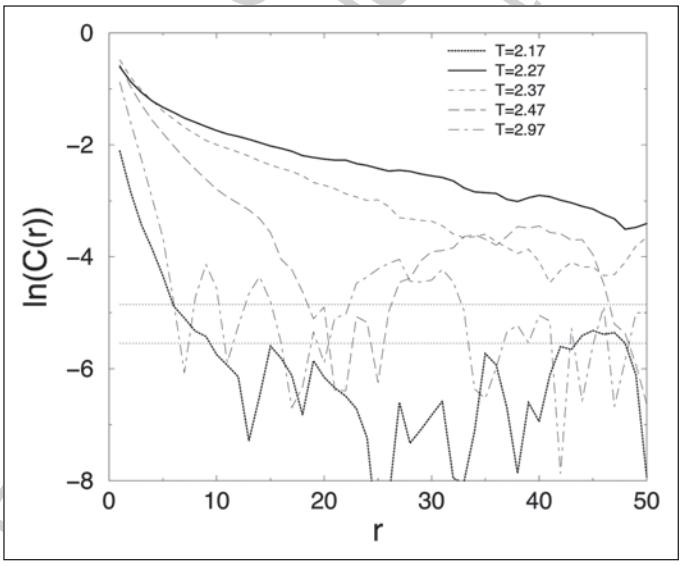

Figure 3. Logarithms of the correlation functions for the two-dimensional Ising model with $L = 2^8 \times 2^8$ spins at five different temperatures $T = 2.97$, $T = 2.47$, $T = 2.37$, $T = T_c = 2.27$, and $T = 2.17$. The straight horizontal lines show 68% and 95% confidence level for apparent correlations expected to be observed in an uncorrelated data of this length. Away from critical point, the behavior of correlations is well approximated by straight lines indicating exponential decay of correlations. The slopes of these lines are inverse proportional to the correlation length. Close to critical point, correlation length becomes extremely large.

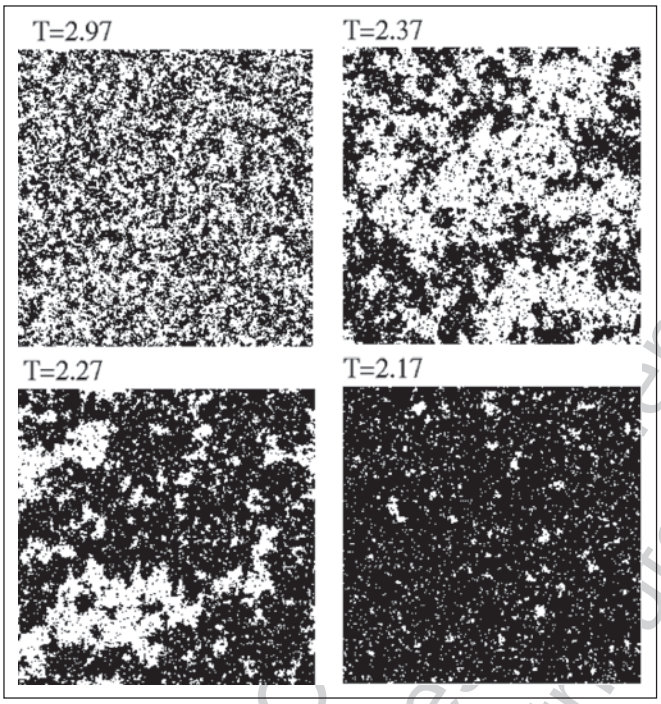

Figure 4. Snapshots of the Ising model far above the critical point $T = 2.97$, close to the critical point $T = 2.37$, at the critical point $T = 2.27$, and below the critical point $T = 2.17$. One can see that the patch sizes are the largest at the critical point.

Although $p >> 1 - p$, the average product $\langle s(k)s(r+k)\rangle$ of two independent variables $s(k)$ and $s(k + r)$ must be equal to the product of their averages $\langle s(k)\rangle\langle s(k + r)\rangle$, so the total correlation $C(r) = 0$. Note that $C(0) = 4p(1 - p)$, thus correlation function is small even for small r. Indeed, the graph corresponding to $T = 2.17$ starts at positions much below the graphs for $T \geq T_c$, for which $C(0) = 1$, since $p = 1/2$.

Note that calculations of $\ln C(r)$ become very inaccurate as $C(r)$ approaches zero. This is because the statistical error of calculating the correlation function becomes comparable with its value. Indeed, simple probabilistic analysis shows that for two independent variables x and y, the variable $(x - \langle x\rangle)(y - \langle y\rangle)$ has variance equal to the product of the variances of the variables x and y. When we compute correlation function, we average $\langle (x - \langle x\rangle)(y - \langle y\rangle)\rangle \equiv \langle s(k)s(r + k)\rangle - \langle s(k)\rangle\langle s(k + r)\rangle$ over $N = L \times L$ positions. In the best possible case, assuming all these measurements are independent, the standard error is the square root of variance divided by the square root of N. Since the variables $x \equiv s(k)$ and $y \equiv s(k + r)$ both have the variance $C(0) = 4p(1 - p)$, where p is the probability of a positive spin, we get this error $\sigma = 4p(1 - p)/\sqrt{N}$. Since for $T > T_c$ the probabilities of positive and negative spins are roughly equal, we have $\sigma = 1/256$. The horizontal lines indicate levels of σ, and 2σ corresponding to 68% and 95% confidence levels. Since in reality x and y are correlated, the number of independent measurements have to be divided by a factor proportional to ξ^d, where $d = 2$ is the dimensionality. The calculations of $C(r)$ become extremely inaccurate when we approach the critical point at which the correlation length diverges.

One can see that the values of the correlation function can be well approximated by the straight lines above the estimated standard error level, except for $T = 2.27$ and $T = 2.37$, when

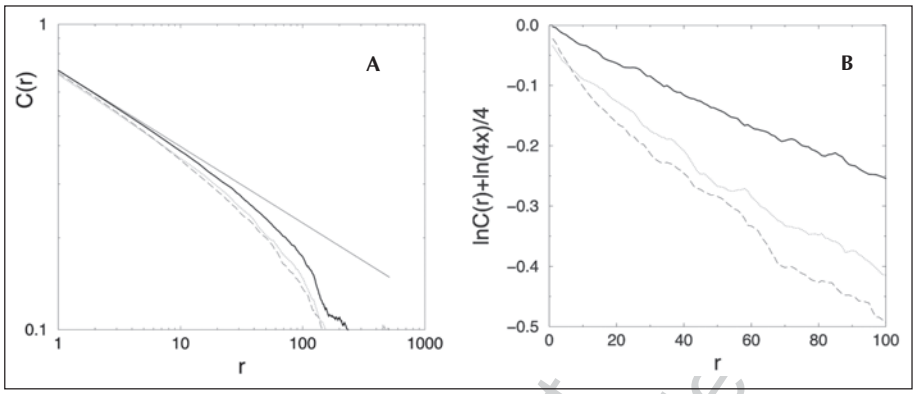

Figure 5. A) Double logarithmic plot of the correlation functions for the two-dimensional Ising model with $L = 2^{10} \times 2^{10}$ spins at critical temperature $T = T_c = 2.296$ for two realizations of the ferromagnetic model (dash and dotted lines) and the antiferromagnetic model (bold line). The straight line indicates theoretical fit $C(r) \sim r^{-1/4}/\sqrt{2}$. B) Logarithm of correlation function, multiplied by $r^{1/4}\sqrt{2}$. The linearity of the graph demonstrates exponential behavior of the corrected correlation function. The inverse slopes give values $\xi = 400$, $\xi = 270$, $\xi = 220$ comparable to half of the system size $L/2 = 512$.

the graphs start to bend upward as r decreases. This behavior may indicate a power law decay. To see this more clearly, we simulate a much larger system $L = 1024$ exactly at T_c. Figure 5A shows the behavior of correlation function on a double logarithmic plot. For small r the graph is approximately linear with slope –0.25 in agreement with the exact result for an infinite system $C(r) = r^{-1/4}/\sqrt{2}$.[17] However, one can see that the deviations rapidly increase with r and the agreement breaks down at about $r = 10$. Analyzing such a data, one can easily dismiss the possibility of power law correlations on the basis that their range is so small. In fact, this early deviation from the power law can be well explained by the finite size of the system $L = 1024$. Indeed, in a finite system, the correlation length cannot be larger than the radius of the system. In Figure 5B, we show that the correlation function can be well approximated by $C(r) \approx r^{-1/4} \exp(-r/\xi)/\sqrt{2}$, where ξ have different values comparable with the system radius ≈ 512. This example demonstrates difficulties associated with correct identification of power law correlations in a finite system.

It is illuminating to study also the anti-ferromagnetic Ising model, in which neighboring spins prefer to stay in the opposite direction, or be anti-correlated. At low temperatures, an anti-ferromagnetic system looks like a checker board. Mathematically, ferromagnetic and anti-ferromagnetic Ising models are identical, so that any configuration of the anti-ferromagnetic model corresponds to exactly one configuration of the ferromagnetic model which can be obtained by flipping all the spins according to a simple deterministic rule. Thus in both models, correlation length has the same finite value at any temperature, except at the critical point at which the correlation length in both models diverges. Nevertheless, the behaviors of correlation functions are totally different. In the anti-ferromagnetic case, correlation function is negative for all odd r and is positive for all even r (Fig. 6A.)

For $T > T_c$, the behavior of the absolute value of the correlation function is similar to that of the ferromagnetic model, both decaying exponentially with r, but below T_c in the anti-ferromagnetic case, the absolute values of correlations do not decay at all (Fig. 6B). However, if one average odd and even values of the correlation function, this averaged correlation function decays exponentially to zero as expected. This shows that the correlation length is finite and that there is no true long range correlations.

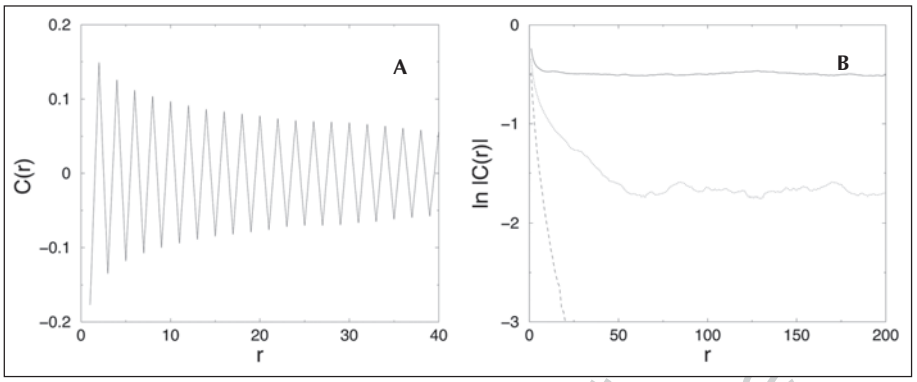

Figure 6. A) Correlation function for the two-dimensional anti-ferromagnetic Ising model. B) Absolute values of the correlation function below T_c (solid line), at T_c (dotted line) and above T_c, (dashed line).

The behavior of the anti-ferromagnetic model below critical temperature is similar to the behavior of coding sequences in DNA, which have a fixed reading frame[43] (see section "Models of Long Range Anti-Correlations'). For a totally uncorrelated sequence of codons in which codon frequencies are taken from real codon usage tables (i.e., no true long range correlations), certain correlation functions oscillate with period 3 with a fixed amplitude till the end of the reading frame. However, after averaging three successive values $C(r) + C(r + 1) + C(r + 2)$, the apparent correlations disappear.

Fourier Power Spectrum

In addition to large statistical errors in computation of $C(r)$, these calculations are also very slow, since the amount of operations is proportional to $r \times N$, where N is total number of points in the sample. An alternative way to study the correlations is to compute a power spectrum $S(f)$ which is the square of the absolute value of the Fourier transform of the function $s(k)$. This technique goes back to X-ray crystallography, in which the intensity of scattered X-rays at certain angle, appears to be a Fourier transform of the density correlation function in the sample under study.[60] It may also help to understand the Fourier transform technique in terms of a musical record. Imagine that $s(k)$ is a record of a melody. Now k is a continuum variable playing the role of time. Then $S(f)$ tells how much energy is carried by frequency (pitch) f. Unfortunately, applications of Fourier transform technique require substantial knowledge in mathematics involving complex numbers and trigonometry. In the following section, we give a brief review of the properties of Fourier transforms. Throughout this section we will use standard notations $i = \sqrt{-1}$ for imaginary unity and $\pi = 3.14159\ldots$. To simplify notations, we will also introduce an angular frequency $\omega = 2\pi f$.

Mathematically, the Fourier transform[61] of an infinitely long record is a result of an integral operator $\mathbf{F}$ acting on the function $s(x)$:

$$\tilde{s}(\omega) = \mathbf{F}s(\omega) = \int_{-\infty}^{\infty} e^{ix\omega} s(x)dx \equiv \int_{-\infty}^{\infty} \cos(\omega x)s(x)dx + i\int_{-\infty}^{\infty} \sin(\omega x)s(x)dx. \quad (21)$$

Since i is the imaginary unity, the result of a Fourier transform is a complex function $\tilde{s}(\omega) = a(f) + ib(\omega)$. The power spectrum $S(\omega)$ is defined as the square of the absolute value of the Fourier transform: $S(\omega) \equiv |\tilde{s}(\omega)|^2 \equiv \tilde{s}(\omega)\,\bar{\tilde{s}}(\omega)$, where $\bar{\tilde{s}}(\omega) = a(\omega) - ib(\omega)$ is a complex conjugate of $\tilde{s}(\omega)$. The signal $s(x)$ can be restored from $\tilde{s}(\omega)$ by the inverse Fourier transform

$$s(x) = \mathbf{F}^{-1}\tilde{s}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ix\omega}\tilde{s}(\omega)d\omega. \quad (22)$$

Fourier transforms have many interesting functional properties which make them a useful tool in data analysis. For example, $\mathbf{F}ds(x)/dx = -i\omega\,\tilde{s}\,(\omega)$ and $\mathbf{F}\int^{x} s(x)dx = i\,\tilde{s}\,(\omega)/\omega$. An important property of the Fourier transform is to turn a convolution of two functions into a product of their Fourier transforms:

$$\mathbf{F}\int_{-\infty}^{\infty} s_1(x)s_2(x+r)dx = \mathbf{F}s_1(\omega)\mathbf{F}s_2(-\omega). \tag{23}$$

Due to this property, the power spectrum of a function with zero average is equal to the Fourier transform of its autocorrelation function.

$$S(\omega) = \mathbf{F}C(r) \tag{24}$$

For example, if the correlations decay exponentially with correlation length ξ as for the one-dimensional Ising model or a one-step Markovian process, $C(r) = C(0)\exp(-r/\xi)$, we have

$$S(\omega) = 2C(0)\xi/(1+\omega^2\xi^2), \tag{25}$$

so the power spectrum is almost constant for low frequencies $\omega < 1/\xi$ and decays as $1/\omega^2$ for high frequencies $\omega >> 1/\xi$.

If the correlations decay as a power law (as at the critical point), $C(r) = |r|^{-\gamma}$, where $0 < \gamma < 1$, the power spectrum also decays as a power law $S(\omega) = c(\gamma)\omega^{-\beta}$, where

$$\beta = 1 - \gamma, \tag{26}$$

and $c(\gamma) = 2\cos[\frac{\pi}{2}(1-\gamma)]\Gamma(1-\gamma)$ does not depend on f. Here Γ is Euler's gamma-function.[61]

The case of approximately constant power spectrum is called white noise, since in this case all the frequencies carry the same energy (as in white light which is mixture of the colors of the rainbow corresponding to all different frequencies). The case $S(f) \sim 1/f^2$ is nicknamed "brown" noise since it describes Brownian motion and the case $S(f) \sim 1/f$ is called $1/f$-noise or "red" noise. The case $S(f) \sim 1/f^{\beta}$, with $0 < \beta < 1$corresponds to long range power-law correlations in the signal and is often called fractal noise. The power spectrum of the fractal noise looks like a straight line with slope $-\beta$ on a log-log plot.

In case of long range anti-correlations (as in the anti-ferromagnetic Ising model, Fig. 6) the correlation function oscillates with certain angular frequency ω_0. In this case, the behavior of the correlation function can be modeled as $C(r) \sim |r|^{-\gamma}\cos(\omega r)$. Analogous calculations[61] lead to $S(\omega) = c(\gamma)(|\omega_0 - \omega|^{-\beta} + |\omega_0 + \omega|^{-\beta})/2$. This expression is analytical at $\omega = 0$, but it has power law singularities at $\omega = \pm\omega_0$. Thus in case of anti-correlations, the graph of power spectrum does not look like a straight line on a simple log-log plot. One must plot $\ln P(\omega)$ versus $\ln|\omega_0 - \omega|$ in order to see a straight line with the slope $-\beta$.

If the correlation function decays for $r \to \infty$ faster than r^{-1}, its Fourier transform must be a continuous function limited for $f \to \infty$ and, therefore, cannot have singularity at any f. The log-log graph of such a function plotted against $f - f_0$ has zero slope in the limit $\ln|f - f_0| \to \pm\infty$, so one can conclude that $\beta = 0$ if $\gamma > 1$. If $\gamma = 1$, the Fourier transform may have logarithmic singularities, which also corresponds to zero slope $\beta = 0$.

Discrete Fourier Transform

In reality, however, we never deal with infinitely long time series. Usually we have a system of N equidistant measurements. In this case, a sequence of N measurements $s(k)$, $k = 0,1,...N-1$, can be regarded as vector $\mathbf{s}$ of the N-dimensional space. Accordingly, one can define a discrete Fourier transform,[62,63] of this vector not as an integral but as a sum

$$\tilde{\mathbf{s}} \equiv \mathbf{F}\mathbf{s} = \sum_{k=0}^{N-1} s(k)e^{2\pi ikq/N}, \tag{27}$$

which can also be regarded as a vector in N-dimensional space with components $\tilde{s}(q)$, q = 0,1,...$N-1$. The fractional quantity $f = q/N$ plays the role of frequency. As one can see, the discrete Fourier transform can be expressed in a matrix form $\tilde{\mathbf{s}} = \mathbf{F}\mathbf{s}$, where $\mathbf{F}$ is the matrix with elements $f_{kq} = \exp(2\pi ikq/N)$. Analogously, vector $\mathbf{s}$ can be restored by applying an inverse Fourier transform:

$$\mathbf{s} \equiv \mathbf{F}^{-1}\tilde{\mathbf{s}} = \frac{1}{N}\sum_{q=0}^{N-1}\tilde{s}(q)e^{-2\pi ikq/N}. \tag{28}$$

If one assumes that the sequence $s(k)$ is periodic, i.e., $s(k + N) = s(k)$, then the square of the discrete Fourier transform is proportional to the discrete Fourier transform of the correlation function as in case of the continuum Fourier transform.[62,63] Indeed, $|\tilde{s}(f)|^2 = \mathbf{F}\sum_{k=0}^{N-1} s(k)s(k+r)$.

It is natural to define the discrete power spectrum $S(f)$ to be exactly equal to the Fourier transform of the correlation function. Since the correlation function is defined as $C(r) = 1/N\sum_{k=0}^{N-1} s(k)s(k+r) - \langle s\rangle^2$, which involves division by N and subtraction of the average value, $S(f) \equiv |\tilde{s}(f)|^2/N$ for $f > 0$ and $S(0) = 0$, because $\tilde{s}(0) \equiv N\langle s(k)\rangle$.

The correlation function can be thus obtained as an inverse discrete Fourier transform of a power spectrum. Since frequencies $-q/N$ and $1-q/N$ are equivalent (due to 2π – periodicity of sines and cosines) and, for real signal, $\tilde{s}(-f)$ and $\tilde{s}(f)$ are complex conjugates, the values $S(q/N)$ and $S(1 - q/N)$ are equal to each other, so we can compute power spectra only up to the highest frequency $q/N = 1/2$.

If N is a natural power of two, $N = 2^n$, the discrete Fourier transform can be computed by a very efficient algorithm known as the Fast Fourier Transform (FFT).[62,63] The amount of operations in this algorithm grows linearly with N. This makes FFT a standard tool to analyze correlation properties of the time series.

Since the sequences we study are formed by random variables, the power spectra of such sequences are random variables themselves. Before proceeding further, it is important to calculate the power spectrum of a completely uncorrelated sequence of length N. As we have seen in section "Correlation Function", $C(0) > 0$ has the meaning of the average square amplitude (variance) of the original signal, while for $r > 0$, the values of $C(r)$ are Gaussian random variables with zero mean and standard deviation equal to $C(0)/\sqrt{N}$. Analogous conclusions can be made for $S(f)$. According to the central limit theorem,[21] the sum of N random uncorrelated variables $s(k)\exp(2i\pi kf)$ converges to a Gaussian distribution with mean equal to the sum of means and variance equal to the sum of variances of individual terms. Thus, we can conclude (after some algebra) that all $S(f)$ are identically distributed independent random variables with an exponential probability density $P(S(f)) = 1/[C(0)]\exp[-S(f)/C(0)]$. So the power spectrum of an uncorrected sequence has an extremely noisy graph. To reduce the noise one can average power spectra for many sequences, and the average value of the power spectrum will converge to a horizontal line $\langle S(f)\rangle = C(0)$ which is called the white noise level. An equivalent method is to average the values $S(f)$ for k neighboring frequencies $f, f + 1/N, f + 2/N,..., f + k/N$. Note that $\langle S(f)\rangle$ is equal to the Fourier transform of $\langle C(r)\rangle$, directly computed using Eq.(27), since as we see above, $\langle C(r)\rangle = 0$ for $r \neq 0$.

In the following, we will illustrate the usage of FFT computing power spectrum for a one- and two-dimensional Ising models near critical points.

Figure 7 shows the power spectrum for the one-dimensional Ising model consisting of $L = 2^{16}$ spins for $T = 0.5$ ($\xi = 27.3$), $T = 0.6$ ($\xi = 14.01$), $T = 1.0$ ($\xi = 3.67$). The power spectrum of the entire system for $N = L$ is very noisy so we show the running averages of the original data using window of 32 adjacent frequencies (gray fluctuating curves). The averages of 32 power spectra computed for 32 non-overlapping windows each of size $N = 2^{11}$ produce a very similar

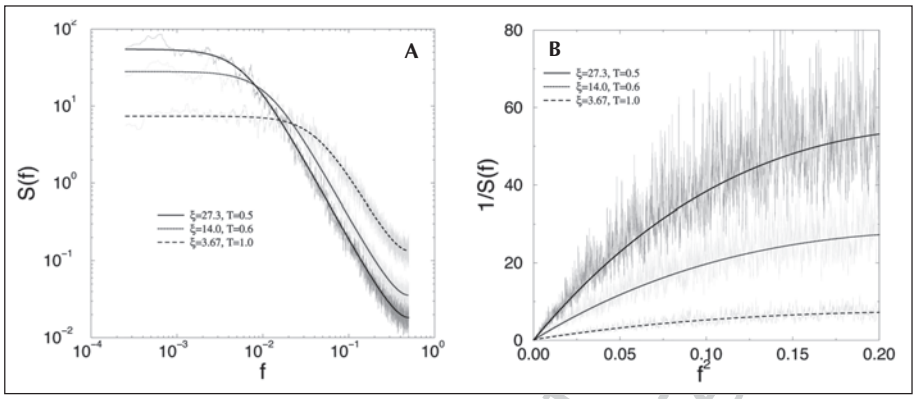

Figure 7. A) Power Spectrum of the one-dimensional Ising model with $L = 2^{16}$ spins for $T = 0.5$ ($\xi = 27.3$), $T = 0.6$ ($\xi = 14.0$), and $T = 1.0$ ($\xi = 3.67$). Smooth lines show analytical result Eq. (29). B) Inverse power spectrum of the same data plotted versus f^2. The slopes at $f^2 = 0$ are proportional to the values of the correlation length.

graph (not shown). The smooth bold lines represent exact discrete Fourier transform of the correlation function computed using Eqs. (4), (17) and (27)

$$S(f) = \frac{1-\lambda^2}{1+\lambda^2-2\lambda\cos(2\pi f)} \approx \frac{2\xi}{1+(2\pi\xi f)^2}, \tag{29}$$

where $\lambda = 2p - 1 = \exp(-1/\xi)$. These analytical results give excellent agreement with the numerical data. One way to estimate the correlation length is to measure a limit of $S(f)$ for $f \to 0$. This quantity can be applied to detect a characteristic patch size in the DNA sequence (see sections "Alternation of Nucleotide Frequencies" and "Models of Long Range Anti-Correlations"). Another, more accurate method[60] is to plot the inverse power spectrum $1/S(f)$ versus f^2 (Fig. 7B) and to measure the slope of this graph for $f^2 \to 0$. Indeed, according to Eq.(29), this slope is equal to $2\xi\pi^2$. These two methods give consistent results for exponentially decaying correlations, but technically speaking they measure two different properties of the power spectrum. In fact, the latter method gives the so called Debye persistence length $R^2 \sim \int_0^\infty C(r)r^2dr$, which is not the same as correlation length ξ, but is proportional to ξ for exponentially decreasing correlations, $C(r) \sim \exp(-r/\xi)$.

Figure 8A shows the power spectrum for a two-dimensional Ising model on a $L \times L = 2^{10} \times 2^{10}$ square lattice computed averaging power spectra for L horizontal rows each consisting of $N = L = 2^{10}$ points. The figure shows a remarkable straight line indicating long range power law correlations. However, the slope of the line $\beta = 0.86$ corresponds to $\gamma = 0.14$ which is almost two times smaller than the theoretical exact value $\gamma = \eta = 0.25$. The discrepancy shows that the power spectrum analysis of the finite system may often give inaccurate values of the correlation exponents.

Figure 8B shows a log-log plot of the power spectrum for a two-dimensional anti-ferromagnetic Ising model, plotted versus $1/2 - f$. The analysis in the previous section shows that since 1/2 is the frequency of the anti-ferromagnetic correlations, the power spectrum must have a power-law singularity in this point. Indeed, the graph gives an approximately straight line with slope $-\beta = -0.84$ similar to the case of ferromagnetic interactions.

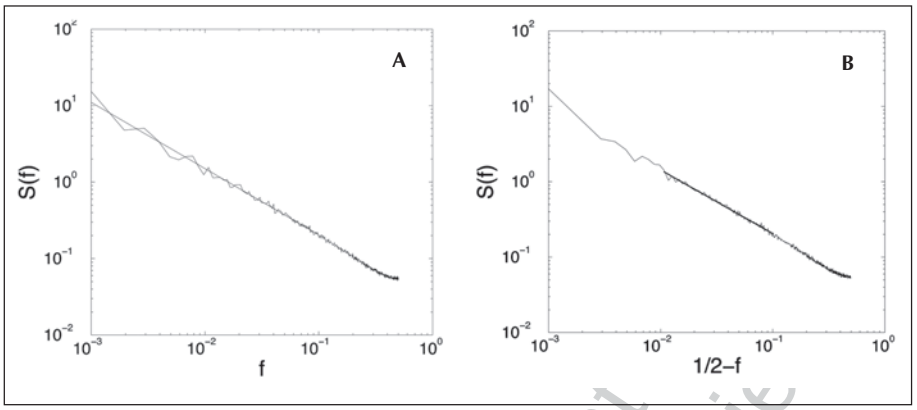

Figure 8. A) Power spectrum of the $2^{10} \times 2^{10}$ Ising model at the critical point. The slope of the straight line gives $\beta = 0.86$. B) Power spectrum of the $2^{10} \times 2^{10}$ anti-ferromagnetic Ising model at the critical point plotted versus $1/2 - f$. The slope of the straight line gives $\beta = 0.84$.

Detrended Fluctuation Analysis (DFA)

A somewhat more intuitive way to study correlations was proposed in the studies of the fluctuations of environmental records by Hurst in 1964.[7] This method is especially useful for short records. The idea is based on comparison of the behavior of the standard deviation of the record averaged over increasing periods of time with the analogous behavior for an uncorrelated record. According to the law of large numbers, the standard deviation of the averaged uncorrelated time series must decrease as the square root of the number of measurements in the averaging interval. This method naturally emerges when the goal is to determine an average value of a quantity (e.g., magnetization in the Ising model, or concentration of a certain nucleotide type in a DNA sequence) obtained in many successive measurements and to asses an error bar of this averaged value. Since the average is equal to the sum divided by the number of measurements, the same analysis can be performed in terms of the sum. In addition to its analytical merits, this method provides a useful graphical description of a time series which otherwise is difficult to see due to high frequency fluctuations.

A variant of Hurst analysis was developed in reference 64 under the name of detrended fluctuation analysis (DFA). The DFA method comprises the following steps:

1. For a numerical sequence $s(k)$, $k = 1,2,...L$ compute a running sum:

$$y(n) \equiv \sum_{k=1}^{n} s(k), \tag{30}$$

which can be represented graphically as a one dimensional landscape, (see Fig. 9A).

2. For any sliding observation box of length r which includes $r + 1$ values $y(k),y(k + 1),...y(k + r)$ define a linear function $y_k(x) = a_k + b_k x$ which provides the least square fit for these values, i.e., a_k and b_k are such that the sum of $r + 1$ squares

$$F_k^2(r) = \sum_{n=k}^{k+r} \left[y(n) - y_k(n) \right]^2 \tag{31}$$

has a minimal possible value $F_{k,\min}^2(r)$. Note that b_k has the meaning of the average value $\langle s(k) \rangle$ for this observation box, which is the local trend of the values $y(k)$. For a non-stationary sequence, the local average values $\langle s(k) \rangle$ can change with time. Since these trends are subtracted in each observation box, this analysis is called detrended. Note that $F_{k,\min}^2(1) \equiv 0$, so it is a trivial value which can be excluded from the analysis.

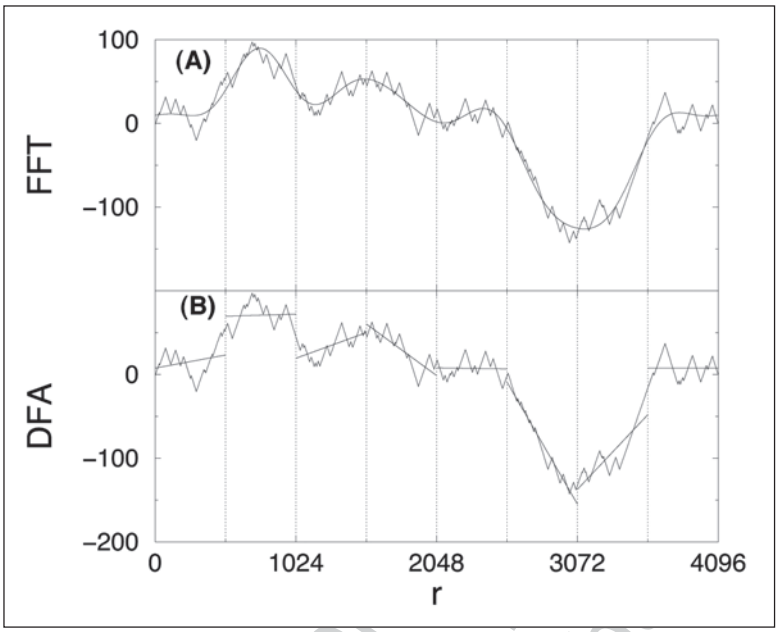

Figure 9. A) Low frequency Fourier approximation of the same landscape. All the frequencies $f > 1/r$ are removed. Fourier DFA computes the average square deviation of this approximation from the landscape. B) The detrended landscape $y_D(n)$ for the one-dimensional Ising model. Straight lines show least square linear fits obtained for different windows of size $r = 512$. Linear DFA computes the average square deviation of these fits from the landscape.

3. For $r > 1$, compute the average value of $F^2_{k,\min}(r)$ from $k = 1$ to $k = L - r$ and define the deternded fluctuation function as

$$F_D^2(r) \equiv \frac{1}{(L-r+1)(r-1)} \sum_{k=1}^{L-r} F_{k,\min}(r). \tag{32}$$

It can be shown, that for a long enough sequence $L \to \infty$ of uncorrelated values $s(k)$ (i.e., $C(r) = 0$ for $r > 1$) with finite mean and variance $C(0)$, we must have $F_D^2(r) \to (r+3)C(0)/15$. Thus the graph of $F_D(r)$ for such a sequence on a log-log plot is a straight line with slope the $\alpha = 1/2$ if plotted versus $r + 3$. Any deviation from the straight line behavior indicates the presence of correlations or anti-correlations. It can be also shown that for a sequence with long range power law correlations $C(r) \sim r^{-\gamma}$ for $0 < \gamma < 1$, the detrended fluctuation also grows as a power law $F_D(r) \sim r^{\alpha}$ as $r \to \infty$, where

$$\alpha = 1 - \gamma/2 > 1/2, \tag{33}$$

is called the Hurst exponent of the time series.

A Relation between DFA and Power Spectrum

There are many different ways to subtract local trends in Eq. (31).[65] One can subtract polynomials of various powers or linear combinations of sines and cosines of certain frequency instead of linear functions. All these different types of DFA have certain advantages and disadvantages. One way to subtract local trends is first to subtract a global trend and plot a sequence

$y_D(k) \equiv y(k) - ky(L)/L$. Next, compute a discrete Fourier transform with $N = L$ of this function $\tilde{y}(f) \propto \mathbf{F}y_D$ and subtract from the function $y_D(k)$ a low frequency approximation

$$y_r(k) = 1/L \sum_{|f|<1/r} \tilde{y}(f) \exp(-2\pi i f k),$$

(see Fig. 9B). A visual comparison of Figure 9A,B, suggests that these two procedures of subtracting local trends are equivalent. Thus we can define a Fourier detrended fluctuation as

$$F_{DF}^2(r) \equiv \frac{1}{L} \sum_{k=1}^{L} \left[y_D(k) - y_r(k) \right]^2. \tag{34}$$

According to Eq. (28), the residuals in the right hand side of Eq. (34) are equal to the high frequency part of the inverse Fourier transform:

$$y_D(k) - y_r(k) = 1/L \sum_{|f|\geq 1/r} \tilde{y}(f) \exp(-2\pi i f k).$$

The Fourier basis vectors are mutually orthogonal, i.e., $\sum_{k=1}^{L} \exp(2\pi i q k/L) \exp(-2\pi i q k/L) = L\delta_{pq}$, where $\delta_{pq} = 1$ if $p = q$ and $\delta_{pq} = 0$, otherwise. Thus, according to the L-dimensional analogy of the Pythagorean theorem, the square of the vector $\mathbf{y}_D(k) - \mathbf{y}_r(k)$ is equal to the sum of the squares of its orthogonal components and therefore,

$$F_{DF}^2(r) = 1/L^2 \sum_{|f|\geq 1/r}^{1/2} \left| \tilde{y}(f) \right|^2 = 1/L \sum_{|f|\geq 1/r}^{1/2} S_y(f). \tag{35}$$

The latter sum is nothing but the sum of all the high frequency components of the power spectrum $S_y(f)$ of the integrated signal.

Equation (35) allows us to derive the relation (33) between the exponents α and γ. Indeed, in continuum limit, this sum corresponds to the integral $\int_{f=1/r}^{\infty} S_y(f) df \sim \int_{f=1/r}^{\infty} S(f) f^{-2} df$, where $S(f)$ is the power spectrum of the original, non-integrated sequence $s(x)$ and the factor f^{-2} comes from the fact that the Fourier transform of the integrated sequence is proportional to the Fourier transform of the original sequence divided by f. As we see above (26), in case of power law correlations with exponent γ, we have $S(f) \sim f^{\gamma-1}$. Thus

$$F_{DF}^2(r) \sim \int_{f=1/r}^{\infty} S_y(f) df \sim \int_{f=1/r}^{\infty} S(f) f^{-2} df \sim (1/r)^{\gamma-1-2+1} = r^{2-\gamma}$$

If we assume that $F_{DF}(r) \sim F_D(r) = r^{\alpha}$ as visual inspection of Figure 9 suggests, we have $\alpha = 1 - \gamma/2$.

Figure 10A shows linear DFA and Fourier DFA for a one-dimensional Ising model on a double logarithmic plot. These two methods are graphically introduced in Figure 9. One can see a sharp transition from the correlated behavior for $r \approx \xi$ with slope $\alpha(r) > 1$ to an uncorrelated behavior for $r >> \xi$ with slope $\alpha(r) \approx 1/2$. The change of the slope can be also studied by plotting the local slope $\alpha(r)$ versus r (Fig. 10B). This graph shows that Fourier DFA can detect the correlation length more accurately than the linear DFA.

Figure 11 shows analogous plots for the two-dimensional Ising model with long range correlations $\gamma = 1/4$. One can see again that the Fourier DFA is more accurate in finding the correct value of the exponent $\alpha = 1 - \gamma/2 = 0.875$ than linear DFA.

In summary, we introduce three methods to study correlations: autocorrelation function $C(r)$, power spectrum $S(f)$, and DFA or Hurst analysis $F_D(r)$. For a signal with long range power law correlations $\gamma < 1$, all three quantities behave as power law:

$$\begin{aligned} C(r) &\sim r^{-\gamma} \qquad r \to \infty \\ S(f) &\sim f^{-\beta} \qquad f \to 0 \\ F_D(r) &\sim r^{\alpha} \qquad r \to \infty \end{aligned} \tag{36}$$

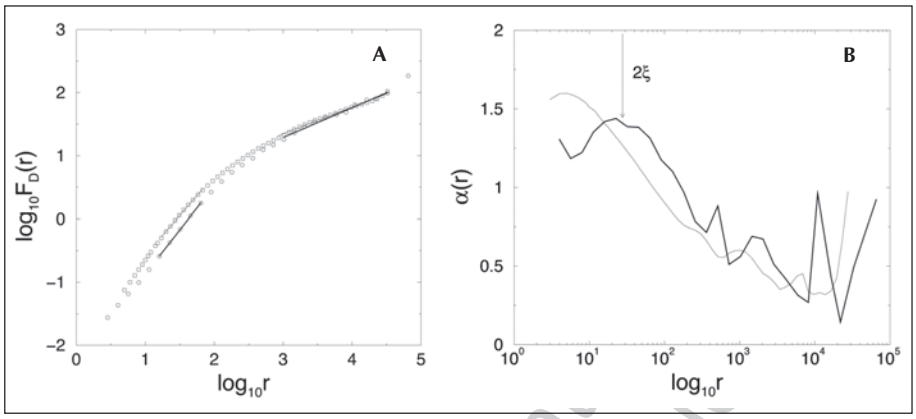

Figure 10. A) Linear detrended fluctuation (°) and Fourier detrended fluctuation (□) of the one dimensional Ising model for ($T = 0.6$, $L = 2^{16}$). The slopes of linear fits give local values of $\alpha = 1.24$ (thin line) and $\alpha = 1.38$ (bold line) for small $r \approx \xi = 14$ and $\alpha = 0.42$ (thin line), $\alpha = 0.47$ (bold line) for uncorrelated regime $r >> \xi$. B) The slope $\alpha(r)$ of the detrended fluctuations as function of r. Note that Fourier DFA gives a strong maximum at $r = 2\xi$ while linear DFA shows monotonic decay of α.

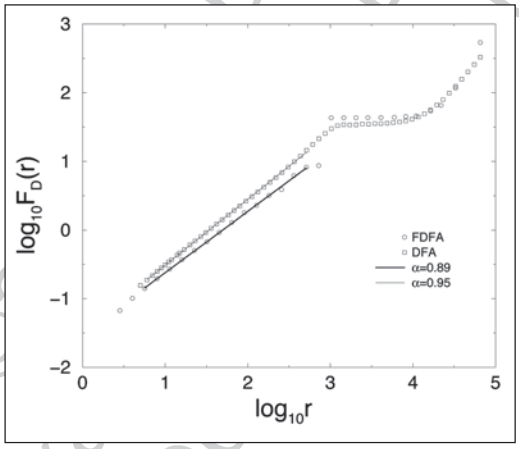

Figure 11. Linear detrended fluctuation (°) and Fourier detrended fluctuation (□) of the two dimensional Ising model for ($T = T_c$, $L = 2^{10}$). The slopes of linear fits give local values of $\alpha = 0.95$ (thin line) and $\alpha = 0.88$ (bold line) for small $r < L$. The steep jump in Fourier DFA at $L = 2^{10}$, indicates quasi-periodicity with period $L = 2^{10}$ due to the "bagel" geometry of the model.

where the exponents α, β, and γ are related via the following linear relations:

$$\begin{aligned} \beta &= \gamma - 1 \\ \alpha &= 1 - \gamma/2 \\ \alpha &= (\beta + 1)/2. \end{aligned} \tag{37}$$

If $\gamma > 1$, the exponents $\beta = 0$, $\alpha = 1/2$ are the same as for a short range correlated sequence with finite correlation length ξ.

Duplication-Mutation Model of DNA Evolution

In 1991, W. Li proposed a duplication-mutation model of DNA evolution which predicted long-range power law correlations among nucleotides.[56] As we see above, in a one dimensional system with finite range interactions, correlations must decay exponentially with distance. So in order to produce a power law decay of correlations, one must assume long-range interactions among nucleotides. In the model of W. Li, such interactions are provided by the fact that the time axes serves as an additional spatial dimension which connects distant segments of DNA developed from a single ancestor. The model is based on two assumptions both of which are well biologically motivated:

1. Every nucleotide can mutate with certain probability.
2. Every nucleotide can be duplicated or deleted with certain probability.

First phenomenon is known as point mutation which can be caused by random chemical reactions such as methylation.[51] Second phenomenon often happens in the process of cell division (mitosis and meiosis) when pairs of sister chromosomes exchange segments of their DNA (genetic crossover). If the exchanging segments are of identical length the duplication does not happen. However, if two segments differ in length by n nucleotides, the chromosome that acquires larger segment obtains an extra sequence of length n which is identical to its neighbor, while another chromosome loses this sequence. In many cases, duplications can be more evolutionary advantageous than deletions. This process leads to creation of large families of genes developed from the same ancestor. For simplicity we will start with a model similar to the original model of Li[56] which neglects deletions and deals only with duplication of a single nucleotide ($n = 1$). Next, we will discuss the implications of deletions. Schematically, this model can be illustrated by Figure 12. For simplicity, we assume only two types of nucleotides X and Y (say purine vs. pyrimidine or A vs. not A). Each level of the tree-like structure represents one step of the evolutionary process during which every nucleotide duplicates, a nucleotide X can mutate with probability p_Y into Y, and a nucleotide Y can mutate with probability p_X into X. This model can be illustrated by a "family" tree in which every nucleotide is connected to its parent in the previous generation and eventually to a single ancestor at the root of the tree.

After k duplication steps, this process will lead to a sequence of total 2^k nucleotides. The frequencies of nucleotides X and Y in this sequence can be computed using the theory of Markovian processes. Indeed, the sequence of mutations along any brunch of the tree connect-

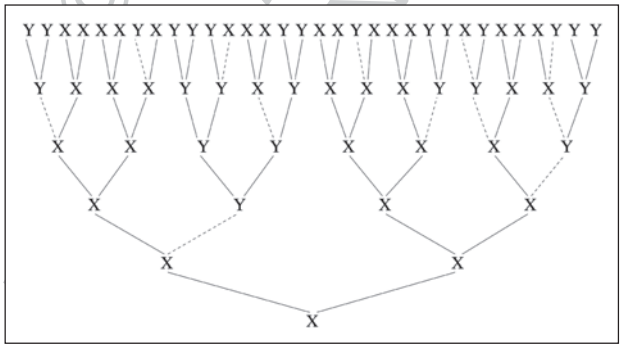

Figure 12. Mutation duplication model of W. Li.[56] At each time step, nucleotides or genes X and Y duplicate and may mutate with probability $P_X + P_Y \approx 1/12$. Mutations are indicated by dashed lines. The correlations can spread along solid lines. Thus nucleotides that are far away along the chain are still closely correlated since they descend from the same ancestor. The above values of mutation probabilities correspond to the long range power-law correlations with $\gamma = 0.25$.

ing a nucleotide to a single ancestor can be regarded as a one-step Markovian process with a matrix of transition probabilities

$$\mathbf{P} = \begin{pmatrix} 1 - p_Y & p_X \\ p_Y & 1 - p_X \end{pmatrix}. \tag{38}$$

Simple calculations of the eigenvector corresponding to the largest eigenvalue $\lambda = 1$ described in section "Markovian Processes" gives the frequencies of nucleotides X and Y after many steps: $f_X = p_X/(p_X + p_Y)$ and $f_Y = p_Y/(p_X + p_Y)$. In addition, Markovian analysis predicts that all dependence coefficients along any branch of the tree decay as λ_2^k, where k is number of generations, and $\lambda_2 = 1 - p_X - p_Y$ is the second largest eigenvalue.

Let us compute the dependence coefficients between two nucleotides which are at distance r from each other in the resulting sequence. The reason of why the correlations are now long-range is obvious. Indeed, the nucleotides which are $r = 2^{k'}$ apart from each other in space are only $2k'$ apart from each other in time, since they are both descendants of one common ancestor $k' = \log_2 r$ generations before. As we see above, the correlations decay exponentially with k' and hence as a power law with r. After some elementary algebra, we get that all dependence coefficients D_{XX}, D_{XY}, D_{YX}, and D_{YY} decay as power law

$$D(r) \sim r^{-\gamma} \tag{39}$$

where

$$\gamma = -\frac{2\ln|p_X + p_Y - 1|}{\ln 2}. \tag{40}$$

If the deletions may occur with some probability $P_d < 1/2$, the number of descendants of one common ancestor grows as $z^{k'}$ where $z = 2(1 - P_d)$ and k' is the number of generations. Thus, replacing ln2 by ln$z$ in the denominator of the expression for (40), we get

$$\gamma = -\frac{2\ln|p_X + p_Y - 1|}{\ln 2(1 - p_d)}. \tag{41}$$

The true long range correlations correspond to the case $\gamma < 1$, or $(p_X + p_Y - 1)^2(1 - p_d) > 1/2$, which means that the mutation rates must be very small: $p_X + p_Y \approx 0$ or alternatively very large: $p_X + p_Y \approx 2$, while the deletion rate must be small. This simple example shows that the exponent of the power law crucially depends on the parameters of the model. In real DNA sequences, the duplication unit is rather a gene or a part of a gene coding for a protein domain. One can generalize this model assuming that coding sequences X and Y can duplicate, and with some probability jump from place to place effectively mimicking mutations X to Y and Y to X in the above scheme. One can also introduce various point mutation rates for nucleotides in the sequences X and Y. These alternations may change the formula for γ, but the model will still produce power law decaying correlations $D(r) \sim (r/\langle n \rangle)^{-\gamma}$, where $\langle n \rangle$ is the average length of sequences X and Y. The problem with the application of this model to a real situation is that the model has many parameters, describing point mutations, duplications and deletions, while resulting in a single observable parameter γ.

Alternation of Nucleotide Frequencies

Let us assume that a nucleotide sequence consists of two types of patches,[57] in one of which the frequency of nucleotide X is f_{X1} while in the other it is f_{X2}. The patches can alternate at random, so that after a patch of type 1 a patch of type 2 can follow with probability 1/2 and

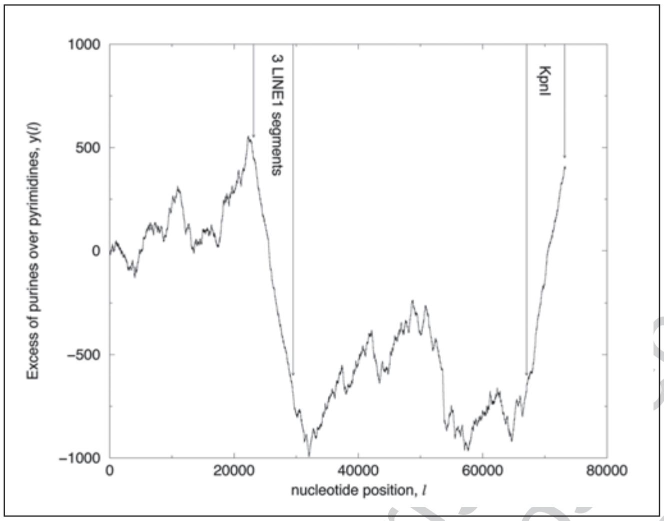

Figure 13. Purine-pyrimidine landscape representation (excess of purines over pyrimidines) of the human beta globin chromosomal region (GenBank accession HUMHBB) of the total length $L = 73{,}308$. The overall frequency of purines (50.27%) is almost equal to the frequency of pyrimidines (49.73%). HUMHBB contains a KpnI repeat from position 67071 to position 73195. This region is very purine rich with 58.57% of purines. KpnI repeat belongs to the LINE1 family of repetitive elements. A region from 23,137 to 29,515 is very purine poor (41.43%). It contains 3 segments of LINE1 repetitive elements inserted into the opposite DNA strand, so that all purines are exchanged with pyrimidines.

vice versa. Let us assume that the lengths of these patches l are distributed according to the same probability distribution $P(l)$.

The motivation for this model could be the insertion of transposable elements,[53,66] e.g., LINEs and SINEs into the opposite strands of the DNA molecule. It is known that LINE-1 sequence has 59% purines (A,G) and 41% pyrimidines (T,C).[66-68] Obviously, due to A-T and C-G complementarity, if LINE-1 is inserted into the opposite strand, it will have 41% purines and 59% pyrimidines (see Fig. 13).

Of course, much more complex models with many parameters can be introduced. These types of models are similar to hidden Markov processes.[16,69] However we will study only the above simple case in order to understand under which conditions this model can lead to power-law correlations.

Let us compute the correlation function for this model. Obviously, the average frequency of a nucleotide in the entire sequence is $f_X = (f_{X1} + f_{X2})/2$, so if both nucleotides k and $k + r$ belong to the same patch their correlation $D_{XX}(r)$ will be $f_{X1}^2 - f_X^2$ if it is a patch of type 1 or $f_{X2}^2 - f_X^2$ otherwise. Since both events have the same probability 1/2, the overall correlation $D_{xx}(r) = \left(f_{X1}^2/2 + f_{X2}^2/2\right)\prod(r) = \prod(r)\left(f_{X1} - f_{X2}\right)^2/4$, where

$$\prod(r) = \sum_{l=1}^{\infty} P(r+l)l / \sum_{l=1}^{\infty} P(l)l \qquad (42)$$

is the probability that a randomly chosen pair of nucleotides at distance r belongs to the same patch.

If the distribution of patch sizes is exponential $P(l) = \lambda^{l-1}(1-\lambda)$, the overall correlation is easy to compute using summation of geometric series $D_{XX}(r) = (f_{X1} - f_{X2})^2\lambda^r/4$, which decays exponentially with r. However, this correlation can be extremely small comparatively to the

"white noise level" $D_{XX}(0) = f_X(1 - f_X)$, and thus can be very difficult to detect. For example, if $f_{X1} = 0.3$ and $f_{X1} = 0.2$ $D_{XX}(0) = 3/16$, while $D_{XX}(1) = \lambda/400$, which is almost 100 times smaller even for very large $\lambda \to 1$.

If we have the distribution of patch sizes decaying for $l \to \infty$ as a power law

$$P(l) \sim l^{-\mu}, \tag{43}$$

where $\mu > 2$, one can show that $\prod(r) \sim r^{2-\mu}$. This can be easily seen if one approximates summation by integration in the expression (42) for $\prod(r)$. Thus, in this case the correlations are indeed power law with

$$\gamma = \mu - 2. \tag{44}$$

For $\mu > 3$, we have $\gamma > 1$ and the power spectrum of the model is finite for $f \to 0$, which means $\beta = 0$, $\alpha = 1/2$. The value $\lim_{f \to 0} S(f) \sim \sum_{l=1}^{\infty} l^2 P(l) / \sum_{l=1}^{\infty} lP(l)$ has the meaning of the weighted average patch length, i.e., the average length of the patch containing a randomly selected base pair.

The case $2 < \mu < 3$ is equivalent to the behavior of the displacement in the so called Lévy walks,[70,71] i.e., walks in which distribution of step lengths are taken from a power law with exponent μ. In this case, $\beta = 3 - \mu$, $\alpha = 2 - \mu/2$.

If $\mu \leq 2$, the sums in (42) do not converge, this means that summation in Eq. (42) must be taken up to the largest $l \approx L$, where L is the total sequence length. Thus $\prod(r) \sim (L - r)/L = 1 - r/L$ and we can assume $\gamma = 0$, $\beta = 1$, $\alpha = 1$.

Figure 14A shows the behavior of the correlation function of a sequence for which $p_{X1} = 0.3$, $p_{X2} = 0.2$ and $P(l) = l^{-3/2} - (l + 1)^{-3/2}$, corresponding to $\mu = 2.5$. In this case $\prod(r) = \sum_{l=r+1}^{\infty} l^{-3/2} / \sum_{l=1}^{\infty} l^{-3/2} \sim r^{-0.5}$. We present the results of correlation analysis for a very long sequence of $L = 223 \approx 8 \cdot 10^6$. One can see good agreement with Eq. (44). For a short sequence, $L = 2^{13} = 8192$, there is no agreement: the correlations sink below random fluctuations, whose amplitude is equal to $C(0)/\sqrt{L}$. This means that the sequence must be very long so that the long range correlations can be seen on top of random noise.

Figure 14B shows the power spectrum for the case of $N = L = 2^{23}$ obtained by averaging power spectra for 2048 non-overlapping windows of size $N = 4096$. The power spectrum is almost flat corresponding to the white noise level $C(0) = 3/16$. If the white noise level is subtracted, the long-range correlations become apparent (Fig. 14C). Indeed the graph of $|S(f) - C(0)|$ on a log-log scale is a perfect straight line with slope –0.57 in a good agreement with the theoretical prediction. The DFA method gives exponent $\alpha(r)$ monotonically increasing from an uncorrelated value 0.5 for small r to $\alpha = 1 - \gamma/2 = 0.75$ for large r. Similar situation is observed in coding DNA, in which the long range correlations may exist but are weak comparatively to the white noise level. These correlations are limited to the third nucleotide in each codon[72] and can be detected if the white noise level is subtracted.

If the length of the largest patch is comparable with the length of the entire sequence as in case $\mu \leq 2$, $\beta = 1$, the global average frequency f_X of a nucleotide cannot be accurately determined no matter how large is the entire sequence length. The average frequency we obtain will be always the frequency of the largest patch. This behavior known as non-stationarity is observed in many natural systems in which different parts are formed under different conditions. Non-stationarity makes the correct subtraction of the white noise level problematic, since its calculation involves estimation of $C(0) \sim f_X(1 - f_X)$, which depends on f_X.

Applying subtraction of the white noise level procedure, Richard Voss[34,35] found that both coding and noncoding DNA sequences from any organism, have exponent $\beta \approx 1$, corresponding to the $1/f$ noise. Note that $\beta = 1$ is exactly the case when this procedure is not quite reliable. Earlier[73] he applied the same type of analysis to the music of different composers from J.-S. Bach to the Beatles and showed that all their music is just $1/f$ noise! No matter how

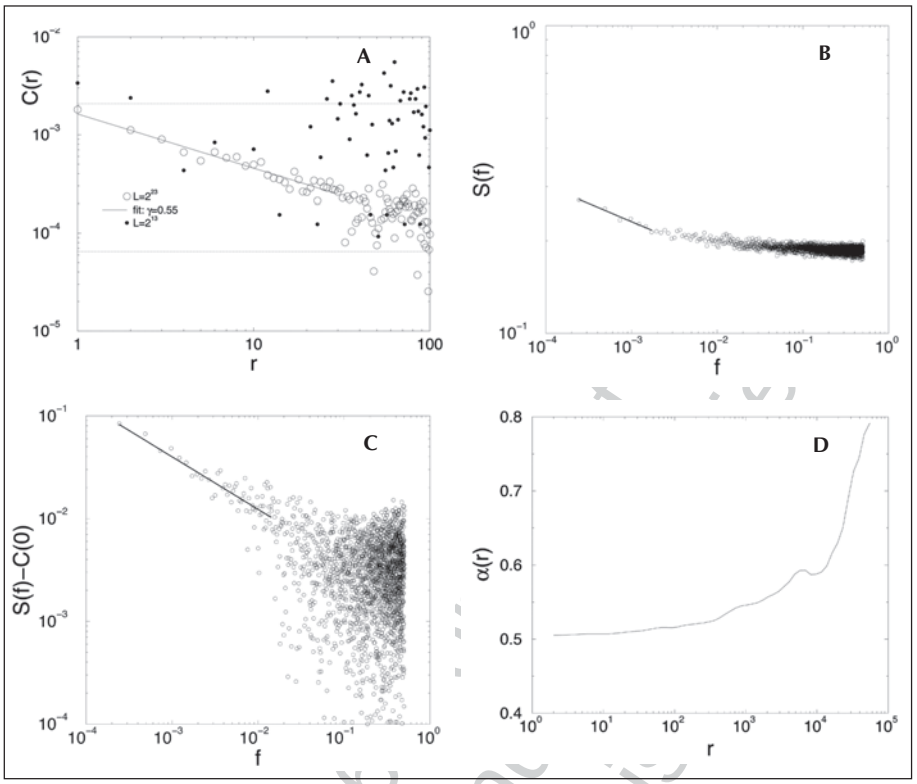

Figure 14. A) Correlation functions for the Lévy walk model for a long $L = 2^{23}$ and a short $L = 2^{13}$ sequences. The lower and upper horizontal lines show random noise levels for the long and short sequences, respectively. B) The power spectrum for the long sequence. The spectrum is almost flat indicating that long-range correlations are small comparatively to the white noise level. Effective exponent $\beta = 0.12$ is very small. C) After the white noise level is subtracted, the power spectrum shows long-range correlations with exponent $\beta = 0.57$. D) The effective exponent $\alpha(r)$ obtained by the DFA method.

intriguing this observation might seem, the explanation is somewhat trivial. The case $\mu < 2$, ($\beta = 1$) i.e., the case when the length of the largest patch is comparable with the entire sequence length is indeed likely to be true for music as well as for DNA. In music, fast pieces follow slow pieces, while in DNA, CG rich isochores follow CG poor ones.

It is interesting to note that similar long range correlations with exponent $\alpha = 0.57$ have been found in human writings.[74,75] These correlations can be explained by the changes in local frequenies of letters caused by changes in the narrative which excessively uses the names of currently active characters.

In DNA, these patches may represent different structural elements of 3D chromosome organization, e.g., the DNA double helix with period 10.5 bp,[76] nucleosomes about 200 bp long,[76] 30 nm fiber, looped domains of about 10^5 bp, and chromatin bands or isochores[72,77] that may consist of several million nucleotides. Such hierarchical structure of several length-scales may produce effective long-range power law correlations. In fact,[78,79] it is enough to have three discrete sizes $r = 100$, $r = 1000$ and $r = 10000$ of these patches in the distribution $\prod(r)$ in order to get a sufficiently straight double logarithmic plot of the power spectrum over three decades in the frequency range.

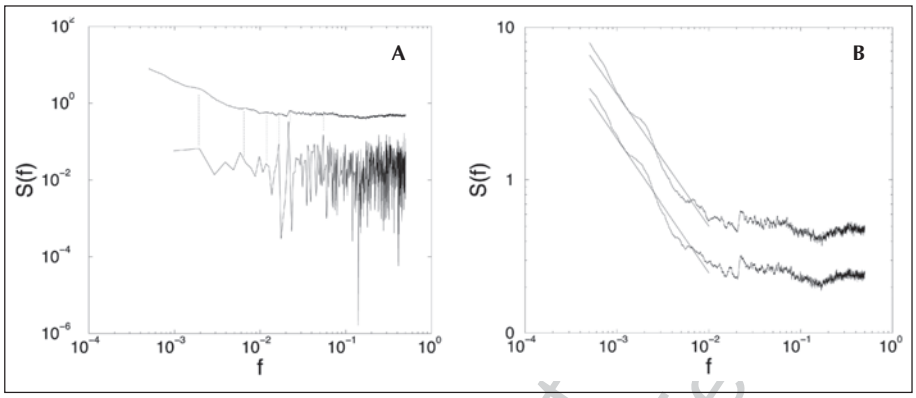

Figure 15. A) Power spectrum of the "chromosome" of length $L = 2^{20}$ (upper curve) in comparison with the power spectrum of the inserted "transposon" $\ell = 2^{10}$ (lower curve) in the insertion-deletion model. Dotted lines indicate peaks present in both sequences. B) Power spectra of the "chromosome" after 1024 iterations (lower curve) and after 16384 iterations (upper curve) showing that the model reaches steady state after $L/\ell = 2^{10}$ iterations.

An interesting model can reproduce some feature of the human genome, namely the abundance of interspersed repeats or retroposons,[68] virus-like sequences that can insert themselves into different places of the chromosomes by reverse transcriptase. An example of such a sequence is LINE-1, which we discussed earlier in this section.

Suppose, we have an initially uncorrelated "chromosome" consisting of L base-pairs with equal concentration of purines and pyrimidines and a "transposon" of length $\ell << L$ with strong strand bias (60% purines) and no correlations (Fig. 15A). Let us assume that at every simulation step our "transposon" can be inserted at random places into one of the two opposite strands of the "chromosome" with equal probabilities. In order to keep the length of the chromosome constant, let us delete exactly ℓ nucleotides selected at random after each insertion. After approximately L/ℓ insertions, the power spectrum of the "chromosome" reaches a steady state shown in Figure 15B. In this example, we use $L = 2^{20}$, $\ell = 2^{10}$. Note the presence of strong peaks in the flat spectral part for $f > 0.01$. And a steep slope with average slope $\beta \approx 0.8$ for $0.0005 < f < 0.01$. One can easily see (Fig. 15A) that the power spectrum of the "transposon" which is kept unchanged during the entire simulation has strong peaks coinciding with the peaks of the resulting "chromosome". This example shows that the presence of many copies of interspersed repeats (some of which have partially degraded) can lead to the characteristic peaks at high frequencies larger than the inverse length of the retroposons and strong power-law like correlations at low frequencies comparable with the inverse length of the retroposons.

Models of Long Range Anti-Correlations

Another interesting situation may exist in coding DNA which preserves the reading frame. The reading frame is a non-interrupted sequence of codons each consisting of three nucleotides. One of the most fundamental discoveries of all time, is the discovery of the universal genetic code, i.e., that in all leaving organisms, with very few exceptions, each of the twenty amino acids is encoded by the same combinations of three nucleotides or codons. Since there are $4^3 = 64$ different codons and only 20 amino acids, some amino acids are encoded by several codons. In the different codons used for coding the same amino acid, the first letter is usually preserved. Since the amino acid usage is non-uniform, the same is true for the codon usage, particularly for the frequency of the first letter in the codon. It is known[80] that for

all coding sequences in the GenBank, there is a preference for purine in the first position in the codon (32% G and 28% A) and for weakly bonded pair in the second position (31% A and 28% T). This preference exists for any organism in the entire phylogenetic spectrum and is the basis for the species independence of mutual information.[44]

Accordingly, let us generate many patches of different length l in which the frequencies of a certain nucleotide at positions $3k + 1 + c$, $3k + 2 + c$, and $3k + c$ are f_1, f_2 and f_3. Here c is a random offset which is constant within each patch and can take values 0,1,2 with equal probabilities. Following Herzel and Grosse,[43] we will call this construction a random exon model.

All the correlation properties of the random exon model can be computed analytically. But even without lengthy algebra, it is clear that the correlation function will oscillate with period three being positive at positions $r = 3k$ and negative at positions $r = 3k + 1$ and $r = 3k + 2$. The envelope of these oscillations will decay, either exponentially if the patch length is distributed exponentially or as a power law if the distribution of patches is a power law $P(l) \sim l^{-\mu}$. Accordingly, in the power spectrum, there will be either a finite strong peak at frequency $f = 1/3$ with intensity proportional to the weighted average patch length or a power law singularity $|f - 1/3|^{\mu-3}$ if $2 < \mu < 3$. If $\mu \leq 2$, it will be $1/f$-singularity $|f - 1/3|^{-1}$.

Figure 16A,B shows the correlation function for the random exon model with $f_1 = 0.29$, $f_2 = f_3 = 0.2$ and a power law distribution of reading frame lengths with $\mu = 2.5$. Figure 16C

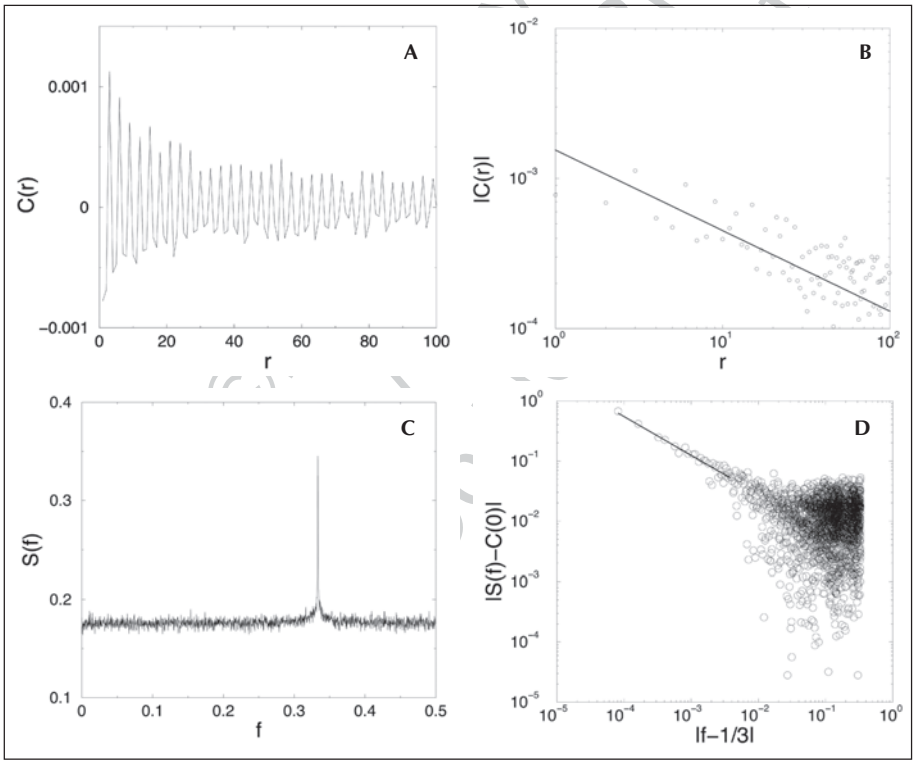

Figure 16. A) Correlation function for the random exon model with power-law distribution of reading frame lengths $P(l) \sim l^{-2.5}$ oscillates with period three. B) The log-log plot of the absolute value of the correlation function for the same sequence. The power-law correlations with exponent $\gamma = 0.57$ are clearly seen. C) The power spectrum of the same sequence. It is almost flat with a strong peak at $f = 1/3$. D) The log-log plot of the power spectrum with the subtracted white noise level. The power law correlations with $\beta = 0.64$ are clearly seen.

shows the power spectrum of this sequence and finally Figure 16D shows the log-log plot of $|P(f) - C(0)|$ versus $|f - 1/3|$. One can see approximate straight line behavior with the slope 0.6. The DFA analysis fails to show the presence of any power law correlations except for a small bump at $r = 3$ (not shown).

Analysis of DNA Sequences

In this section, we finally present the analysis of real DNA sequences. The examples of the previous sections show us that among different methods of analysis, the power spectrum usually gives the best results. In contrast to $C(r)$ method, it provides natural averaging of the long range correlations from a broad interval of large distances $[r,r + \Delta r]$ adding them up into a narrow range of low frequencies $[1/r - \Delta r/r^2, 1/r]$. Thus, the power spectrum restores useful information which cannot be seen from $C(r)$ quickly sinking below the white noise level for large r. On the other hand, the power spectrum does not smooth out the details on the short length scales corresponding to high frequencies as DFA does. Also it is much less computationally intensive than the two other methods. Once the intuition on how to use the power spectrum analysis is developed, it can be applied to DNA sequences with the same success as in X-ray crystallography, especially, today when the length of the available DNA sequences becomes comparable with the number of atoms in the nano-scale experimental systems. Not surprisingly, power spectra of the DNA from different organisms have distinct characteristic peaks,[81] similarly to the X-ray diffraction patterns of different substances. Accordingly, in this section, we will use only the power spectrum analysis.

In the beginning of 1990, when the first long DNA sequences became available, an important practical question was to find coding regions in the "sea" of noncoding DNA which constitutes 97% of human genome. The problem was not only to determine genes, i.e., the regions which are transcribed in the process of RNA transcription, but also the exons, the smaller segments of genes which remain in the messenger RNA after the noncoding introns are spliced out. Only the information from exons is translated into proteins by the ribosomes.[51,52] That is why, the claim of reference 31 that the non-coding DNA sequences have stronger power law correlations than the coding ones attracted much attention and caused a lot of controversy.[34] The results of reference 31 were based on the studies of a small subset of sequences using DNA landscape technique (see Fig. 13). Later these results were confirmed by the DFA method, the wavelet,[55,72,82] the power spectrum[80] and modified standard deviation analyses.[83] However, the difference between coding and noncoding DNA appeared to be not as dramatic as it was originally proposed. In Figure 17 we present the results[80] of the analysis of 33301 coding and 29453 noncoding sequences of the eukaryotic organisms. These were all the genomic DNA sequences published in the GenBank release of August 15th, 1994 whose length was at least 512 nucleotides. The power spectrum is obtained by averaging power spectra calculated by FFT of all non-overlapping intervals of length $N = 2^9 = 512$ contained in the analyzed sequences. The conclusions hold not only for the average power spectrum of all eukaryotes but also for the average power spectra of each organism analyzed separately.

Unlike the graphs for Ising model, the log-log graphs for coding and non-coding DNA are not straight but have three distinct regimes for high (H) ($f > 0.09$), medium (M) $0.012 < f < 0.09$ and low (L) $f < 0.012$ frequencies. The slopes β_M in the region of medium frequencies can be obtained by the least square linear fit. For RY mapping rule (**see Section VI**) presented in Figure 17 for coding DNA, we see $\beta_M = 0.03$ which corresponds to the white noise, while for non-coding DNA we see weak power-law correlations with $\beta_M = 0.21$. Reference 80 contains the tables of the exponents β_M obtained for various eukaryotic organisms for seven different mapping rules (RY, SW, KM, A, C, G, T). For all the rules and all the organisms, the exponent β_M for the averaged power spectra of non-coding regions is always larger than β_M for coding regions. For some rules, such as SW, the exponent β_M is negative for coding DNA and

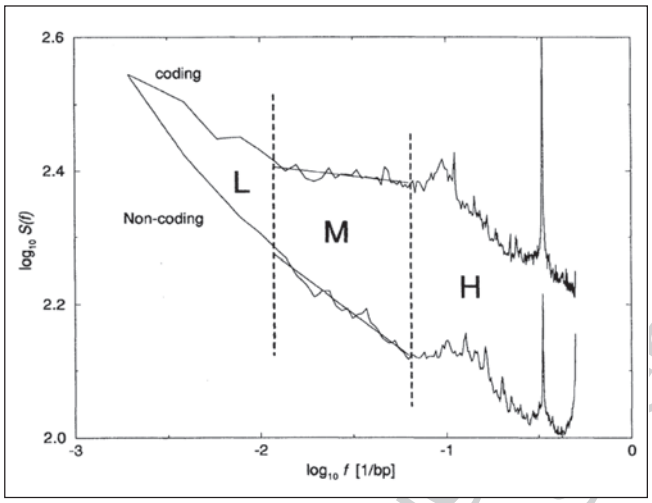

Figure 17. The RY power spectrum obtained by averaging power spectra of all eukaryotic sequences longer than 512 bp, obtained by FFT with window size 512. Upper curve is average over 29,453 coding sequences; lower curve is average over 33,301 noncoding sequences. For clarity, the power spectra are shifted vertically by arbitrary quantities. The straight lines are least squares fits for second decade (Region M). The values of β_M for coding and noncoding DNA obtained from the slopes of the fits are 0.03 and 0.21, respectively. (From ref. 80.)

is close to zero for non-coding DNA. But the algebraic values of the exponents for non-coding DNA is always larger than for coding DNA. The histogram of values of β_M computed for individual 512-bp sequences has a roughly Gaussian shape with standard deviation $\sigma = 0.3$ which is several times larger than the difference between mean values of β_M for coding and non-coding DNA. This makes the use of fractal exponent β_M impractical for finding coding regions.[84]

A much more important characteristic of the power spectrum is the height of the peak at the codon frequency $f = 1/3$, which was included in the standard gene finding tool boxes.[85,86] Figure 17 shows that the peak for coding regions is several times higher than for non-coding ones. The presence of the weak peak in the noncoding regions can be attributed to the non-identified genes or to pseudo-genes which have recently (on the evolutionary time scale) become inactive (like olfactory genes for humans). The presence of the peak can be explained by the non-uniform codon usage, (see section "Models of Long Range Anti-Correlations", Fig. 16).

Another interesting and distinctive feature of non-coding DNA is the presence of the peak at $f = 1/2$ as in the anti-ferromagnetic Ising model. This peak can be attributed to the presence of long tandem repeats ...CACACA... and …TGTGTG… which are prolific in non-coding DNA but very rare in the coding (see next section).

Presently, when several complete or almost complete genomes are just a mouse-click away, it is easy to test if the true power-law long-range correlations do exist in the chromosomes of different species. Figure 18A,B shows power spectra of the 88 million base-pair contig of the human chromosome XIV computed according to the seven mapping rules described in section "Correlation Analysis of DNA Sequences". A very interesting feature of the human genome is the presence of the strong peaks at high frequencies. These peaks are much stronger than the peak at $f = 1/3$ for coding DNA. It is plausible that these peaks are due to the hundreds of thousands almost identical copies of the SINE and LINE repeats,[87] which constitute a major portion of human genome.[68] If one compares the peaks in the power spectrum of the chromosome, with

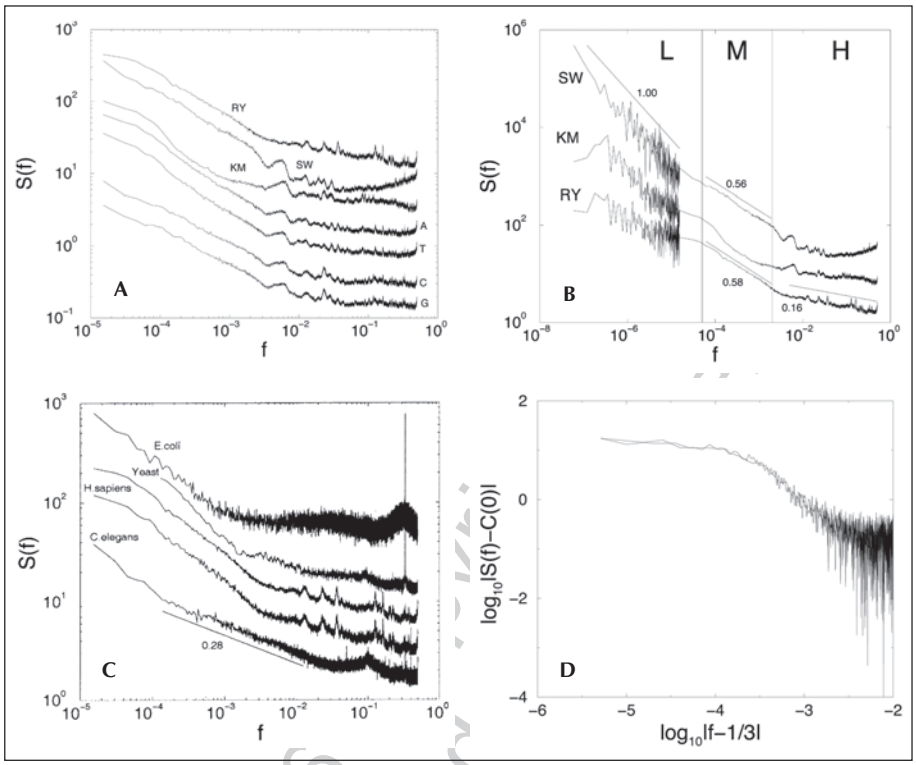

Figure 18. A) Power spectra for seven different mapping rules computed for the Homo sapiens chromosome XIV, genomic contig NT_026437. The result is obtained by averaging 1330 power spectra computed by FFT for non-overlapping segments of length $N = 2^{16} = 65536$. B) Power spectra for SW, RY, and KM mapping rules for the same contig extended to the low frequency region characterizing extremely long range correlations. The extension is obtained by extracting low frequencies from the power spectra computed by FFT with $N = 2^{24} = 16M$ base pairs. Three distinct correlation regimes can be identified. High frequency regime ($f < 0.003$) is characterized by small sharp peaks. Medium frequency regime ($0.5 \cdot 10^{-5} < f < 0.003$) is characterized by approximate power-law behavior for RY and SW mapping rules with exponent $\beta_M = 0.57$. Low frequency regime ($f < 0.5 \cdot 10^{-5}$) is characterized by $\beta = 1.00$ for SW rule. The high frequency regime for RY rule can be approximated by $\beta_H = 0.16$ in agreement with the data of Figure 17. C) RY Power spectra for the entire genome of *E. coli* (bacteria), *S. cerevisae* (yeast) chromosome IV, *H. sapiens* (human) chromosome XIV and the largest contig (NT_032977.6) on the chromosome I; and *C.elegans* (worm) chromosome X. It can be clearly seen that the high frequency peaks for the two different human chromosomes are exactly the same, while they are totally different from the high frequency peaks for other organisms. One can also notice the presence of enormous peaks for $f = 1/3$ in *E. coli* and yeast, indicating that their genomes do not have introns, so that the lengths of coding segments are very large. The *C. elegans* data can be very well approximated by power law correlations $S(f) \sim f^{-0.28}$ for $10^{-4} < f < 10^{-2}$. D) Log-log plot of the RY power spectrum for *E. coli* with subtracted white noise level versus $|f - 1/3|$. It shows a typical behavior for a signal with finite correlation length, indicating that the distribution of the coding segments in *E. coli* has finite average square length.

the peaks in the power spectra of various SINE and LINE sequences, one can find that some of these peaks coincide as in the model of insertion-deletion discussed in section "Alternation of Nucleotide Frequencies". The absence of these peaks in the genomes of primitive organisms (see Fig. 18C) is in agreement with the fact that these organisms lack interspersed repeats.

It is clear that the long-range correlations lack universality, i.e., they are different for different species and strongly depend on the mapping rule. The slopes of the power spectra change with frequency and undergo sharp crossovers which do not coincide for different organisms. The strongest correlations with the spectral exponent $\beta = 1$ are present for *SW* rule at low frequencies, indicating the presence of the isochores. The middle frequency regime which can be particularly well approximated by power law correlations in *C. elegans* can be explained by the generalized duplication-mutation model of W. Li in which duplications and mutations occur on the level genes, consisting of several hundred base pairs. The high frequency correlations, sometimes characterized by small positive slopes of the power spectra can be attributed to the presence of simple sequence repeats (see next section). In contrast, the high frequency spectrum of the bacterium *E. coli* is almost flat with the exception of the huge peak at $f = 1/3$. Bacterial DNA practically does not have noncoding regions, thus (in agreement with refs. 31,72,80,82) it does not have long range correlations on the length scales smaller than the length of a typical gene. Large peaks at $|f - 1/3|$ in the power spectra of *E. coli* and yeast are consistent with fact that these primitive organisms do not have introns and therefore their open reading frames are very long. The spectrum of *E. coli* printed versus $|f - 1/3|$ shows a horizontal line for $f \to 1/3$ on a double logarithmic plot indicating that the length distribution of the open reading frames has finite second moment.

Distribution of Simple Repeats

The origin, evolution, and biological role of tandem repeats in DNA, also known as microsatellites or simple sequence repeats (SSR), are presently one of the most intriguing puzzles of molecular biology. The expansion of such SSR has recently become of great interest due to their role in genome organization and evolutionary processes.[88-100] It is known that SSR constitute a large fraction of noncoding DNA and are relatively rare in protein coding sequences.

SSR are of considerable practical and theoretical interest due to their high polymorphism.[97] The formation of a hairpin structure during replication is believed to be the cause of the *CAG* and *CTG* repeat expansions, which are associated with a broad variety of genetic diseases. Among such diseases are fragile X syndrome,[101] myotonic dystrophy, and Huntington's disease[94,102] SSR of the type $(CA)_\ell$ are also known to expand due to slippage in the replication process. These errors are usually eliminated by the mismatch-repair enzyme MSH2. However, a mutation in the MSH2 gene leads to an uncontrolled expansion of repeats—a common cause of ovarian cancers.[103] Similar mechanisms are attributable for other types of cancer.[85,92,93] Telomeric SSR, which control DNA sequence size during replication, illustrate another crucial role of tandem repeats.[51]

Specifically, let us consider the distribution of the most simple case of SSR—repeats of identical dimers *XYXY...XY* ("dimeric tandem repeats"). Here *X* and *Y* denotes one of the four nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T). Dimeric tandem repeats are so abundant in noncoding DNA that their presence can even be observed by global statistical methods such as the power spectrum. For example, Figures 17 and 18A-C show presence of a peak at $(1/2)\text{bp}^{-1}$ in the power spectrum of **noncoding** DNA (corresponding to repetition of dimers) and the absence of this peak in **coding** DNA. The abundance of dimeric tandem repeats in noncoding DNA suggests that these repeats may play a special role in the organization and evolution of noncoding DNA.

First, let us compute the number of repeats in an uncorrelated sequence. Suppose that we have a random uncorrelated sequence of length $2L$ which is a mixture of all 16 possible types of dimers *XY*, each with a given frequency f_{XY}, The probability that a randomly selected dimer belongs to a dimeric tandem repeat $(XY)_\ell$ of length ℓ can be written as

$$P_{XY}(\ell) = f_{XY}^{\ell} \cdot (1 - f_{XY})^2 \cdot \ell, \tag{45}$$

where $(1 - f_{XY})$ is the terminating factor responsible for not producing an additional unit XY at the beginning (or end) of the repeating sequences and the factor ℓ takes into account ℓ possible positions of a dimer XY in a repeat $(XY)_\ell$. Since the total number of dimers in our sequence is L, the number of dimers in the repeats $(XY)_\ell$ is $LP_{XY}(\ell) = \ell N_{XY}(\ell)$, where $N_{XY}(\ell)$ is the total number of repeats $(XY)_\ell$ in a sequence of length $2L$. Finally,

$$N_{XY}(\ell) = f_{XY}^{\ell} \cdot (1 - f_{XY})^2 \cdot L \sim e^{-\ell \cdot |\ln f_{XY}|}, \tag{46}$$

which decreases exponentially with the length of the tandem repeat. Thus, a semi-logarithmic plot of $N_{XY}(\ell)$ versus ℓ must be a straight line with the slope

$$-k_{\text{unc}} = \ln(f_{XY}). \tag{47}$$

In order to compare the prediction of this simple model with real DNA data, we estimate f_{XY} for the real DNA as follows: (i) divide the DNA sequence into L non-overlapping dimers, (ii) count n_{XY}, the total number of occurrences of a dimer XY in this sequence, and calculate

$$f_{XY} \equiv \frac{n_{XY}}{L}. \tag{48}$$

Indeed, most dimeric tandem repeats in **coding** DNA produce linear semilogarithmic plots, (Fig. 19A) but with slopes significantly different from those predicted by (47). The deviation of the slopes from prediction (47) can be explained by the short order Markov correlations.[106,107]

On the other hand, semilogarithmic plots of the length distributions of dimeric repeats for noncoding parts (Fig. 19C) are usually not straight, but display negative slope with constantly decreasing absolute value which indicates that their probability decays less rapidly than exponentially. Indeed, these distributions can be better fit by straight lines on a double logarithmic plot (Fig. 19D)

$$N_{XY}(\ell) \sim \ell^{-\mu}. \tag{49}$$

A simple model to explain the power law behavior (49) was presented in references 106 and 108. The mechanism proposed in references 106 and 108 is based on random multiplicative processes, which can reproduce the observed non-exponential distribution of repeats. The increase or decrease of repeat length can occur due to unequal crossover[51,109] or slippage during replication.[92,100,110,111] It is reasonable to assume (see ref. 110 and refs. therein) that in these types of mutations, the new length ℓ' of the repeat is not a stepwise increase or decrease of the old length ℓ, but is defined as a product $\ell' = \ell r$, where r is some random variable.

For simplicity we neglect point mutations and assume that with conditional probability $\pi(r, \ell)$ in a single mutation, a repeat of length ℓ can expand or shrink to a repeat of length $r\ell$, where the function $\pi(r, \ell)$ is normalized:

$$\int_0^\infty \pi(r, \ell) dr = 1. \tag{50}$$

After t steps of evolution the length of the repeat is given by

$$\ell_t = \prod_{i=1}^{t} r_i \ell_0, \tag{51}$$

where r_i is a random variable taken from a distribution with probability density $\pi(r, \ell)$. Such a process is called a random multiplicative process and, in many cases, leads in the long time limit ($t \to \infty$) to a stable distribution of repeat length $P(\ell)$. According to Eq. (51), repeats may fluctuate in length and even disappear. Thus, to prevent the extinction of repeats, one can either set a non-zero probability for a repeat to reappear, or set $\pi(r, \ell) = 0$ when $r\ell \leq 1$. Both ways are mathematically equivalent and might be biologically controlled by point mutations.

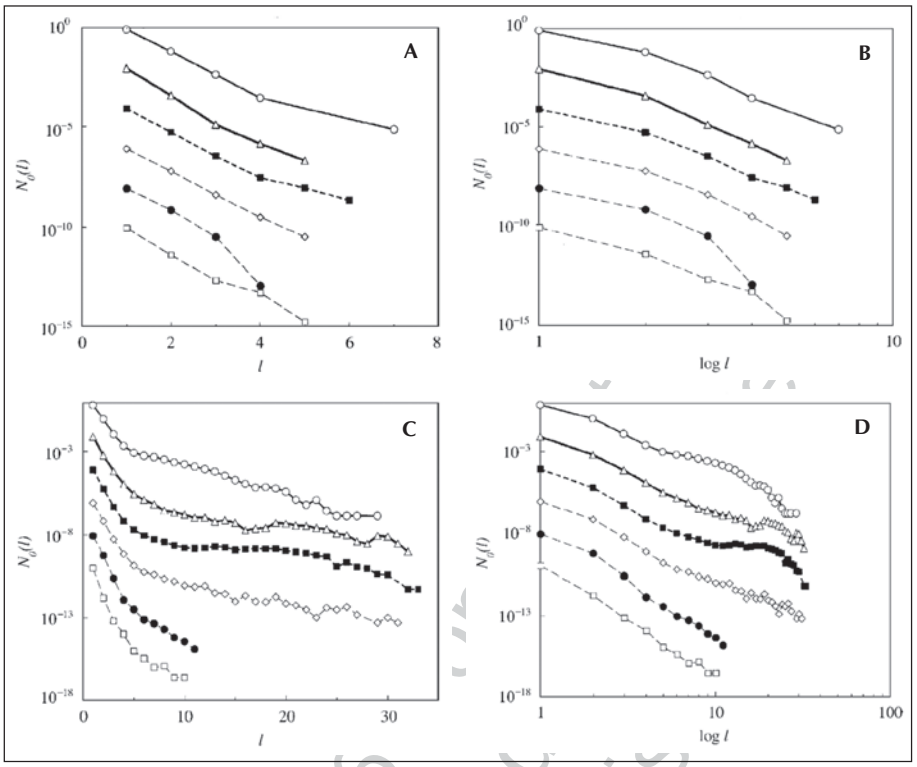

Figure 19. The combined plot of the normalized number $N_0(\ell) \equiv N_{XY}(\ell)/N_{XY}(1)$ of repeats for six groups of dimeric tandem repeats in human genome averaged over analogous repeats in each group: $(AA)_\ell$ and $(TT)_\ell$ (°); $(TA)_\ell$ and $(AT)_\ell$ (Δ); $(CA)_\ell$,$(AC)_\ell$,$(TG)_\ell$ and $(GT)_\ell$ (■); $(GA)_\ell$,$(AG)_\ell$, $(TC)_\ell$ and $(CT)_\ell$ (◊); $(CC)_\ell$ and $(GG)_\ell$ (●); $(GC)_\ell$ and $(CG)_\ell$ (□). Semi-logarithmic plot for coding DNA (A), double-logarithmic plot for coding DNA (B), semi-logarithmic plot for noncoding DNA (C), and double-logarithmic plot for noncoding DNA (D). For clarity, we separate plots for these six groups by shifting them by a factor of 100 on the ordinate. The values of μ for six groups of repeats in (D) are 3.6, 3.3, 3.2, 4.1, 6.7, and 5.4 from top to bottom, fitting range is $\ell > 5$. The values of μ for strongly bonded repeats *GC,CG* and *CC,GG* are significantly larger than for other repeats. (From ref. 107.)

If we take the logarithm of both parts of Eq. (51) and change variables to $z \equiv \ln \ell$, the process becomes a random diffusion process in semi-infinite space $z > 0$ in which a particle makes steps $v_i = \ln r_i$. The distribution of steps $\tilde{\pi}(v,z)$ can be related to the original distribution of growth-rates, $\pi(r, \ell)$. Indeed, in the continuum limit $\tilde{\pi}(v,z)dv = \pi(r, \ell)dr$, or $\tilde{\pi}(v,z) = \pi(e^v, e^z)e^v$.

A classical example of such a process is Brownian motion in a potential field $U(z)$, which leads for $t \to \infty$ to a Boltzmann probability distribution (3). The strength and the direction of the potential force $f(z) = -dU/dz$ depends on the probability distribution $\tilde{\pi}(v,z)$. If probability to go up is larger than the probability to go down, the force acts upward, so the particle travels upward indefinitely and no stable probability distribution is observed. (This situation corresponds to the uncontrollable expansion of repeats as in some types of cancers.) If the distribution $\tilde{\pi}(v,z)$ does not depend on z, the force is constant. If the force is constant and acting down as the gravitational force on the Earth, the final probability distribution decays exponentially with z as the density of Earth's atmosphere

$$P(z) \sim e^{-kz}, \tag{52}$$

where is a positive constant which depends on the distribution $\tilde{\pi}(v,z) = \tilde{\pi}(v)$.

Using the theory of Markovian processes (Section III), one can show that the final probability distribution $P(z)$ must satisfy an equation analogous to (9) in which $P(z)$ plays the role of eigenvector $\mathbf{a}_1$ and $\tilde{\pi}(v,z)$ plays the role of transition matrix **P**. In the continuum limit we have $P(z) = \int_{-\infty}^{\infty} P(z-v)\,\tilde{\pi}(v,z-v)dv$, which in case $\tilde{\pi}(v,z) = \tilde{\pi}(v)$ has solution (52) and $k > 0$ must satisfy equation

$$\int_{-\infty}^{\infty} \exp(kv)\tilde{\pi}(v)dv = 1. \tag{53}$$

After transforming back to our original variables, the solution (52) can be rewritten in the form of a power law,

$$P(\ell) = \ell^{-\mu} \tag{54}$$

where $\mu = k + 1 > 1$. Accordingly (53) must be rewritten in the form

$$\int_0^{\infty} r^{\mu-1} \cdot \pi(r)dr \tag{55}$$

Equation (55) always has a trivial solution $\mu = 1$ (due to the normalization (50)). However, Eq. (55) may also have additional roots, $\mu > 1$. If it does not have such roots then the final distribution does not exist. This case corresponds to the uncontrollable expansion of repeats.

Let us discuss two examples, in which Eq. (55) has simple solutions. For example, if $\pi(r)$ is a step-function

$$\pi(r) = \begin{cases} 1/2, & 0 \le r \le 2 \\ 0, & r < 1, r > 2, \end{cases} \tag{56}$$

equation (55) becomes

$$\frac{1}{2} \cdot \frac{2^{\mu}}{\mu} = 1. \tag{57}$$

Eq. (57) has a solution $\mu = 2$. The above case can serve as the simplest model of unequal crossover,[108] after which a repeat of length ℓ becomes of length $\ell \cdot (1 + r)$ in the first allele and of length $\ell \cdot (1 - r)$ in the second allele. If both alleles have equal probability of becoming fixed in the population, we arrive to Eq. (56).

In another simple example we take

$$\pi(r) = \pi_1 \cdot \delta(r - 1/2) + \pi_2 \cdot \delta(r - 2), \tag{58}$$

where $\pi_1 + \pi_2 = 1$ and $\delta(r)$ is the Dirac delta-function, i.e., with probability π_1 the repeat can shrink by factor of two and with probability π_2 it can grow by factor of two. In this case, Eq. (55) can be written as

$$\pi_1 \cdot \left(\frac{1}{2}\right)^{\mu-1} + \pi_2 \cdot 2^{\mu-1} = 1, \tag{59}$$

which has a root $\mu = 1 + \log_2(\pi_1/\pi_2)$. If probability to grow is larger than probability to shrink, $\pi_2 > \pi_1$, we have $\mu < 1$, which, as we see above, leads to an uncontrollable expansion of repeats as in some diseases. These simple examples show that our multiplicative model is capable to explain the power law distribution of simple repeats with any exponent $\mu > 1$.

In the general case of discrete multiplicative processes, one cannot obtain analytical solutions. However, numerical simulations[106] show that Eq. (54) still provides a good approximation for large ℓ. The deviation of the actual distributions (Fig. 19) from an exact power law can

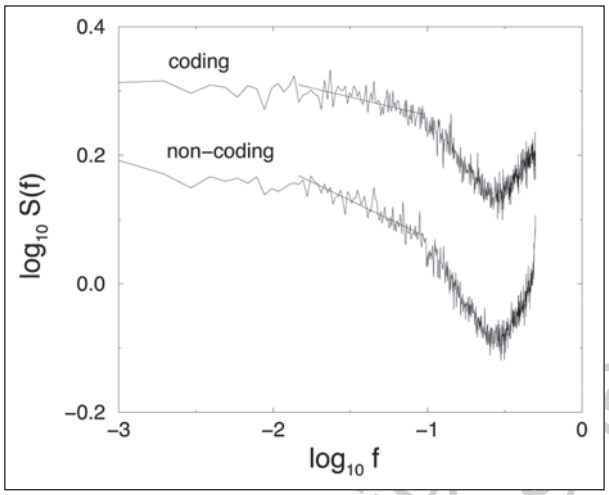

Figure 20. RY power spectra of the random dimeric tandem repeat model for coding and noncoding DNA for the mammalian sequences.

be explained if one takes into account that the distribution of growth rates $\pi(r, \ell)$ may depend on the length of the repeat ℓ.[107,112] This is especially plausible for the slippage during replication mechanism, since the ability for a DNA molecule to form hairpins clearly depends on the length of a segment involved in the slippage and on its biophysical properties and thus must depend on the type of repeat. Therefore, it is not surprising that different types of repeats have different length distribution. For example, the distribution of $(AC)_\ell$ and $(TG)_\ell$ repeats in vertebrates have plateaus in the range $10 < \ell < 30$. In contrast, the distributions of $(CC)_\ell$, $(CG)_\ell$ and $(GG)_\ell$, and repeats decay much faster than other repeats which include weakly bonded base pairs.

A different model proposed by reference 113 can also reproduce long tails in the repeat length distribution. This model assumes the stepwise change in repeat length with the mutation rate proportional to the repeat length. It is possible to map this model to a random multiplicative process with a specific form of distribution $\pi(r, \ell)$, where $r = \ell'/\ell$, ℓ is the original length and ℓ' is a repeat length after a time interval during which several stepwise mutations can occur.

From the analysis in section "Alternation of Nucleotide Frequencies", it follows that simple tandem repeats randomly distributed along the sequence can produce long-range power-law correlations if, and only if, $\mu < 3$. However, in almost all real DNA sequences $\mu > 3$, which means that simple tandem repeats alone cannot explain long-range correlations. On the other hand, simple tandem repeats may be the primary source of the difference in correlation properties of coding and noncoding sequences at relatively short length scales of $\ell \approx 100$ bp.[78,79] In order to test such a possibility, we construct a random dimeric repeat model by randomly selecting all possible repeats $(XY)_\ell$ from the empirically observed distribution $N_{XY}(\ell)$ and concatenating them into an artificially constructed sequence of nucleotides. Figure 20 shows the power spectra of two sequences produced by random concatenations of various dimeric repeats taken from the noncoding and coding mammalian DNA. These power spectra show a slight difference in the spectral exponent β_M in the region of medium frequencies analogous to Figure 17. Note that the difference in the spectral exponents in the random repeat model is smaller than in real sequences. However, here we consider only dimeric tandem repeats, thus

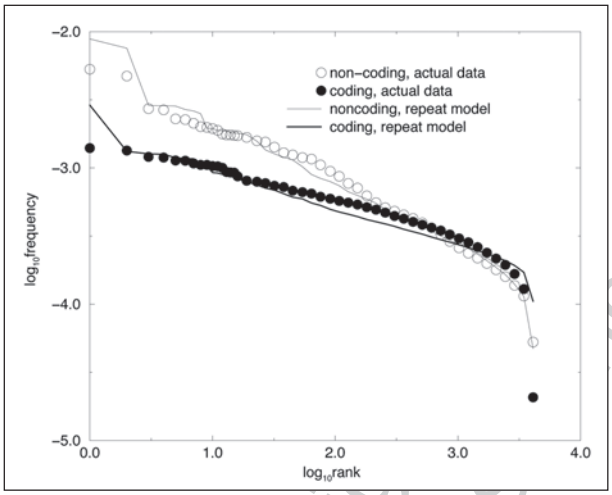

Figure 21. Log-log plot of the frequency of 6-letter words (hexamers) versus their rank for invertebrate coding and non-coding sequences in comparison with the same graphs produced by the random dimeric repeat model.

neglecting the repeats of other types. We also neglect the possibility of imperfect repeats interrupted by several point mutations.

Finally, dimeric tandem repeats can explain the difference observed in the distribution of *n*-letter words in coding and non-coding DNA (see Fig. 21). As an example, we show the rank-frequency of the 6-letter words (hexamers) for invertebrate coding and noncoding sequences in the form of the so called Zipf plots.[114] For natural languages, Zipf graphs show that the frequency of a word in a text is inverse proportional to its rank. For example, in an English text, the most frequent word is "the" (rank 1), the second most frequent word is "of" (rank 2), the third most frequent word is "a" (rank 3) and so on. Accordingly, the frequency of word "of" is roughly two times smaller than the frequency of word "the" and the frequency of word "a" is roughly three times smaller than the frequency of "the". Thus on the log-log scale, the Zipf graph is a straight line with the slope -1. In a DNA sequence, there is no precise definition of the "word", so one can define "word" as any string of the fixed number of consecutive nucleotides that can be found in the sequence. One can notice that the Zipf graph for non-coding DNA is approximately straight but with a slope smaller than 1, while for coding DNA, the graph is more curvy and is less steep. This observation led Mantegna et al [115,116] to conclude that noncoding DNA have some properties of natural languages, namely redundancy. Accordingly, noncoding DNA may contain some "hidden language". However, this conjecture was strongly opposed by the bioinformatics community.[117] Indeed, Zipf graphs of coding and non-coding DNA can be trivially explained by the presence of dimeric tandem repeats (Fig. 21).

To conclude, noncoding DNA may not contain any hidden "language" but it definitely has lot of hidden biological information. For example, it contains transcription regulatory information which is very difficult to extract. Application of correlation analysis may help to solve this problem.[118]

Conclusion

Long range correlations of different length scales may develop due to different mutational mechanisms. The longest correlations, on the length scales of isochores may originate due to

base-substitution mutations during replication (see ref. 77). Indeed, it is known that different parts of chromosomes replicate at different stages of cell division. The regions rich in C+G replicate earlier than those rich in A+T. On the other hand, the concentration of C+G precursors in the cell depletes during replication. Thus the probability of substituting A/T for C/G is higher in those parts of the chromosome that replicate earlier. These unequal mutation rates may lead to the formation of isochors.[77] Correlations on the intermediate length scale of thousands of nucleotides may originate due to DNA shuffling by insertion or deletion[57,58] of transposable elements such as LINES and SINES[66,68,119] or due to a mutation-duplication process proposed by W. Li[56] (see also ref. 120).

Finally, the correlations on the length scale of several hundreds of nucleotides may evolve due to simple repeat expansion[106,108] As we have seen in the previous section, the distributions of simple repeats are dramatically different in coding and noncoding DNA. In coding DNA they have an exponential distribution; in noncoding DNA they have long tails that in many cases may be fit by a power law function. The power law distribution of simple repeats can be explained if one assumes a random multiplicative process for the mutation of the repeat length, i.e., each mutation leads to a change of repeat length by a random factor with a certain distribution (see ref. 106). Such a process may take place due to errors in replication[110] or unequal crossing over (see ref. 108 and refs. therein). Simple repeat expansion in the coding regions would lead to a loss of protein functionality (as, e.g., in Huntington's disease[110]) and to the extinction of the organism.

Thus the weakness of long-range correlations in coding DNA is probably related to the coding DNA's conservation during biological evolution. Indeed, the proteins of bacteria and humans have many common templates, while the noncoding regions can be totally different even for closely related species. The conservation of protein coding sequences and the weakness of correlations in the amino acid sequences[121] are probably related to the problem of protein folding. Monte-Carlo simulations of protein folding on the cubic lattice suggest that the statistical properties of the sequences that fold into a native state resemble those of random sequences.[122]

The higher tolerance of noncoding regions to various mutations, especially to mutations involving the growth of DNA length—e.g., duplication, insertion of transposable elements, and simple repeat expansion—lead to strong long-range correlations in the noncoding DNA. Such tolerance is a necessary condition for biological evolution, since its main pathway is believed to be gene duplication by chromosomal rearrangements, which does not affect coding regions.[123] However, the payoff for this tolerance is the growth of highly correlated junk DNA.

Acknowledgments

I am grateful to many individuals, including H.E. Stanley, S. Havlin, C.-K. Peng, A.L. Goldberger, R. Mantegna, M.E. Matsa, S.M. Ossadnik, F. Sciortino, G.M. Viswanathan, N.V. Dokholyan, I. Grosse, H. Herzel, D. Holste, and M. Simons for major contributions to those results reviewed here that represent collaborative research efforts. Financial support was provided by the National Science Foundation and National Institutes of Health (Human Genome Project).

References

1. Stauffer D, Stanley HE. From Newton to Mandelbrot: A Primer in Theoretical Physics. Heidelberg, New York: Springer-Verlag, 1990.
2. Stanley HE. Introduction to Phase Transitions and Critical Phenomena. London: Oxford University Press, 1971.
3. Stauffer D, Aharony A. Introduction to Percolation Theory. Philadelphia: Taylor & Francis, 1992.
4. de Gennes PG. Scaling Concepts in Polymer Physics. Ithaca: Cornell University Press, 1979.

5. Barabási AL, Stanley HE. Fractal Concepts in Surface Growth, Cambridge: Cambridge University Press, 1995.
6. Mandelbrot BB. The Fractal Geometry of Nature. San Francisco: WH Freeman, 1982.
7. Feder J. Fractals. New York: Plenum, 1988.
8. Bunde A, Havlin S, eds. Fractals and Disordered Systems. Berlin: Springer-Verlag, 1991.
9. Bunde A, Havlin S, eds. Fractals in Science. Berlin: Springer-Verlag, 1994.
10. Garcia-Ruiz JM, Louis E, Meakin P et al, eds. Growth Patterns in Physical Sciences and Biology. New York: Plenum, 1993.
11. Grosberg AY, Khokhlov AR. Statistical Physics of Macromolecules, New York: AIP Press, 1994.
12. Bassingthwaighte JB, Liebovitch LS, West BJ. Fractal Physiology. New York: Oxford University Press, 1994.
13. Vicsek T. Fractal Growth Phenomena. Singapore: World Scientific, 1992.
14. Vicsek T, Shlesinger M, Matsushita M, eds. Fractals in Natural Sciences. Singapore: World Scientific, 1994.
15. Guyon E, Stanley HE. Fractal Formes. Amsterdam: Elsevier, 1991.
16. Li W. The study of correlation structures of DNA sequences: a critical review. Computers Chem 1997; 21:257-271.
17. Baxter RJ. Exactly Solvable Models in Statistical Mechanics. London: Academic Press, 1982.
18. Azbel MY. Random two-component, one-dimensional Ising model for heteropolymer melting. Phys Rev Lett 1973; 31:589-593.
19. Azbel MY, Kantor Y, Verkh L et al. Statistical Analysis of DNA Sequences. Biopolymers 1982: 21:1687-1690.
20. Azbel MY. Universality in a DNA statistical structure. Phys Rev Lett 1995; 75:168-171.
21. Feller W. An introduction to ptobability theory and its applications. Vols. 1-2. New York: Jhon Wiley & Sons, 1970.
22. Binder K, ed. Monte Carlo Methods in Statistical Physics. Berlin: Springer-Verlag, 1979.
23. Karlin S, Brendel V. Patchiness and correlations in DNA sequences. Science 1993; 259:677-680.
24. Grosberg AY, Rabin Y, Havlin S et al. Crumpled globule model of the 3-dimensional structure of DNA. Europhys Lett 1993; 23:373-378.
25. des Cloizeaux, J. Short range correlation between elements of a long polymer in a good solvent. J Physique 1980; 41:223-238.
26. Bak P. How Nature Works. New York: Springer 1996.
27. Bak P, Tang C, Wiesenfeld K. Self-organised criticality: an explanation of 1/f noise. Phys Rev Lett 1987; 59:381-384.
28. Bak P, Sneppen, K. Punctuated equilibrium and criticality in a simple model of evolution. Phys Rev Lett 1993; 71:4083-4086.
29. Paczuski M, Maslov S, Bak, P. Avalanche dynamics in evolution, growth and depinning models. Phys Rev E 1996; 53:414-443.
30. Jovanovic B, Buldyrev SV, Havlin S et al. Punctuated equilibrium and history-dependent percolation. Phys Rev E 1994; 50, R2403-2406.
31. Peng C-K, Buldyrev SV, Goldberger AL et al. Nature 1992; 356:168.
32. Li W, Kaneko K. Long-range correlations and partial 1/f a spectrum in a noncoding DNA sequence. Europhys Lett 1992; 17:655.
33. Nee S. Uncorrelated DNA walks. Nature 1992; 357:450-450.
34. Voss R. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Phys Rev Lett 1992; 68:3805-3808.
35. Voss R. Long-Range Fractal Correlations in DNA Introns and Exons. Fractals 1994; 2:1-6.
36. Maddox J. Long-range correlations within DNA. Nature 1992; 358:103-103.
37. Munson PJ, Taylor RC, Michaels GS. DNA correlations. Nature 1992; 360:636-636.
38. Amato I. Mathematical biology-DNA shows unexplained patterns writ large. Science 1992; 257:747-747.
39. Prabhu VV, Claverie J-M.Correlations in intronless DNA. Nature 1992; 359:782-782.
40. Chatzidimitriou-Dreismann CA, Larhammar D. Long-range correlations in DNA. Nature 1993; 361:212-213.
41. Li W, Kaneko K. DNA correlations, Nature 1992; 360:635-636.

42. Karlin S, Cardon LR. Computational DNA sequence analysis. Annu Rev Microbiol 1994; 48:619-54.
43. Herzel H, Grosse I. Correlations in DNA sequences: The role of protein coding segments. Phys Rev E 1997; 55:800-810.
44. Grosse I, Herzel H, Buldyrev SV et al. Species independence of mutual information in coding and noncoding DNA. Phys Rev E 2000; 61:5624-5629.
45. Holste D, Grosse I, Herzel H et al. Optimization of coding potentials using positional dependence of nucleotide frequencies. J Theor Biol 206:525-537.
46. Berthelsen CL, Glazier JA, Skolnick MH. Global fractal dimension of human DNA sequences treated as pseudorandom walks. Phys Rev A 1992; 45:8902-8913.
47. Borovik AS, Grosberg AY, Frank-Kamenetski MD. Fractality of DNA texts. J Biomolec Struct Dyn 1994; 12:655-669.
48. Li WT. Are isochore sequences homogeneous? Gene 2002; 300:129-139.
49. Bernaola-Galvan P, Carpena P, Roman-Roldan R et al. Study of statistical correlations in DNA sequences. Gene 2002; 300:105-115.
50. Oliver JL, Carpena P, Roman-Roldan R et al. Isochore chromosome maps of the human genome. Gene 2002; 300:117-127.
51. Alberts B, Bray D, Lewis J et al. Molecular Biology of the Cell. New York: Garland Publishing, 1994.
52. Watson JD, Gilman M, Witkowski J et al. Recombinant DNA. New York: Scientific American Books, 1992.
53. Chen CF, Gentles AJ, Jurka J et al. Genes, pseudogenes, and Alu sequence organization across human chromosomes 21 and 22. Proc Natl Acad Sci USA 2002; 99:2930-2935.
54. Altschul SF, Madden TL, Schaffer AA et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucl Acids Res 1997; 25:3389-3402.
55. Audit B, Vaillant C, Arneodo A et al. Wavelet analysis of DNA bending profiles reveals structural constraints on the evolution of genomic sequences. J Biol Phys 2004; 30:33-81.
56. Li WH. Expansion-modification systems: A model for spatial 1/f spectra. Phys Rev A 1991; 43:5240-5260.
57. Buldyrev SV, Goldberger AL, Havlin S et al. Generalized Levy Walk Model for DNA Nucleotide Sequences. Phys Rev E 1993; 47:4514-4523.
58. Buldyrev SV, Goldberger AL, Havlin S et al. Fractal Landscapes and Molecular Evolution: Modeling the Myosin Heavy Chain Gene Family. Biophys J 1993; 65:2673-2681.
59. Vieira MD, Herrmann HJ. A growth model for DNA evolution. Europhys Lett 1996; 33:409-414.
60. Hansen JP, McDonald IR. Theory of Simple Liquids. London: Academic Press, 1976.
61. Abramowitz M, Stegun IA, eds. Handbook of Mathematical Functions. New York: Dover, 1965
62. Press WH, Flannery BP, Teukolsky SA et al. Numerical Recipes. Cambridge: Cambridge Univ Press, 1989.
63. Burrus CS, Parks TW. DFT/FFT and Convolution Algorithms. New York: John Wiley and Sons, Inc. 1985.
64. Peng CK, Buldyrev SV, Havlin S et al. Mosaic Organization of DNA Sequences. Phys Rev E 1994; 49:1685-1689.
65. Chen Z, Ivanov PC, Hu K et al. Effect of nonstationarities on detrended fluctuation analysis. Phys Rev E 2002; 65:041107.
66. Jurka J, Walichiewicz T, Milosavljevic A. Prototypic sequences for human repetitive DNA. J Mol Evol 1992; 35:286-291.
67. Hattori M, Hidaka S, Sakaki Y. Sequence analysis of a KpnI family member near the 3' end of human beta-globin gene. Nucleic Acids Res 1985; 13:7813-7827.
68. Hwu RH, Roberts JW, Davidson EH et al. Insertion and/or deletion of many repeated DNA sequences in human and higher apes evolution. Proc Natl Acad Sci USA 1986; 83:3875-3879.
69. Churchill GA. Hidden Markov chains and the analysis of genome structure. Computers Chem 1992; 16:107-116.
70. Zolotarev VM, Uchaikin VM. Chance and Stability: Stable Distributions and their Applications. Utrecht: VSP BV, 1999.
71. Shlesinger MF, Zaslavsky GM, Frisch U, eds. Lévy Flights and Related Topics in Physics. Berlin: Springer-Verlag, 1995.

72. Arneodo A, D'Aubenton-Carafa Y, Audit B et al. What can we learn with wavelets about DNA sequences? Physica A 1998; 249:439-448.
73. Voss RF, Clarke J. 1/f noise in music: music from 1/f noise. J Acoust Soc Amer 1978; 63:258-263.
74. Schenkel A, Zhang J, Zhang, YC. Long Range Correlation in Human Writings. Fractals 1993; 1:47-57.
75. Amit M, Shmerler Y, Eisenberg E et al. Language and codification dependence of long-range correlations in texts. Fractals 1994; 2:7-13.
76. Trifonov EN. 3-, 10.5-, 200- and 400-base periodicities in genome sequences. Physica A 1998; 249:511-516.
77. Gu X. Li WH. A model for the correlation of mutation-rate with gc content and the origin of gc-rich isochores. J Mol Evol 1994; 38:468-475.
78. Viswanathan GM, Buldyrev SV, Havlin S et al. Quantification of DNA patchiness using correlation measures. Biophys J 1997; 72:866-875.
79. Viswanathan GM, Buldyrev SV, Havlin S et al. Long-range correlation measures for quantifying patchiness: Deviations from uniform power-law scaling in genomic DNA. Physica A 1998; 249:581-586.
80. Buldyrev SV, Goldberger AL, Havlin S et al. Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. Phys Rev E 1995; 51:5084-5091.
81. Nyeo SL, Yang IC, and Wu CH. Spectral classification of archaeal and bacterial genomes. J Biol Syst 2002; 10:233-241.
82. Arneodo A, Bacry E, Graves PV et al. Characterizing long-range correlations in dna-sequences from wavelet analysis. Phys Rev Lett 1995; 74:3293-3296.
83. Nikolaou C, Almirantis Y. A study of the middle-scale nucleotide clustering in DNA sequences of various origin and functionality, by means of a method based on a modified standard deviation. J Theor Biol 2002; 217:479-492.
84. Ossadnik SM, Buldyrev SV, Goldberger AL et al. Correlation approach to identify coding regions in DNA sequences. Biophys J 1994; 67:64-70.
85. Uberbacher EC, Mural RJLocating protein-coding regions in human dna-sequences by a multiple sensor neural network approach. Proc Natl Acad Sci USA 1991; 88:11261-11265.
86. Fickett JW, Tung CS. Assessment of protein coding measures. Nucleic Acids Research 1992; 20:6441-6450.
87. Holste D, Grosse I, Beirer S et al. Repeats and correlations in human DNA sequences. Phys Rev E 2003; 67:061913.
88. Bell GI. Roles of repetitive sequences. Comput Chem, 1992; 16:135-143
89. Bell GI. Repetitive DNA sequences: some considerations for simple sequence repeats. Comput Chem 1993; 17:185-190.
90. Bell GI. Evolution of simple sequence repeats. Comput Chem 1996; 20:41-48.
91. Bell GI. and Jurka J. The length distribution of perfect dimer repetitive DNA is consistent with its evolution by an unbiased single step mutation process. J Mol Evol 1997; 44:414-421.
92. Richards RI, Sutherland GR. Simple repeat DNA is not replicated simply. Nature Genetic 1994; 6:114-116.
93. Richards RI, Sutherland GR. Simple tandem DNA repeats and human genetic disease. Proc Natl Acad Sci USA 1995; 92:3636-3641.
94. Chen X, Mariappan SV, Catasti P et al. Hairpins are formed by the single DNA strands of the fragile X triplet repeats: structure and biological implications. Proc Natl Acad Sci USA 1995; 92:5199-5203.
95. Gacy AM, Goellner G, Juramic N et al. Trinucleotide repeats that expand in human disease form hairpin structures in vitro. Cell 1995; 81:533-540.
96. Orth K, Hung J, Gazdar A et al. Genetic instability in human ovarian cancer cell lines. Proc Natl Acad Sci USA 1994; 91:9495-9499.
97. Bowcock AM, Ruiz-Linares A, Tomfohrde J et al. High resolution of human evolutionary trees with polymorphic microsatellites. Nature 1994; 368:455-457.
98. Olaisen B, Bekkemoen M, Hoff-Olsen P et al. VNTR mutation and sex. In: Pena SDJ, Chakraborty, R, Epplen JT et al, eds. DNA Fingerprinting: State of the Science. Basel: Springer-Verlag, 1993.

99. Jurka J. Pethiyagoda G. Simple repetitive DNA sequences from primates: compilation and analysis. J Mol Evol 1995; 40:120-126.
100. Li YC, Korol AB, Fahima T et al. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol Ecol 2002; 11:2453-2465.
101. Kremer E, Pritchard M, Lynch M et al. Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. Science 1991; 252:1711-1714.
102. Huntington's Disease Collaborative Research Group. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 1993; 72:971-983.
103. Ionov Y, Peinado MA, Malkhosyan S et al. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for clonic carcinogenesis. Nature 1993; 363:558-561.
104. Kunkel TA. Slippery DNA and diseases. Nature 1993; 365:207-208.
105. Wooster R, Cleton-Jansen AM, Collins N et al. Instability of short tandem repeats (microsatellites) in human cancers. Nat Genet 1994; 6:152-156.
106. Dokholyan NV, Buldyrev SV, Havlin S et al. Distribution of base pair repeats in coding and noncoding DNA sequences. Phys Rev Lett 1997; 79:5182-5185.
107. Dokholyan NV, Buldyrev SV, Havlin S et al. Distributions of dimeric tandem repeats in noncoding and coding DNA sequences. J Theor Biol 2000; 202:273-282.
108. Dokholyan NV, Buldyrev SV, Havlin S et al. Model of unequal chromosomal crossing over in DNA sequences. Physica A 1998; 249:594-599.
109. Charlesworth B, Sniegowski P, Stephan W. The evolutionary dynamics of repetative DNA in eukaryotes. Nature 1994; 371:215-220.
110. Wells RD. Molecular basis of genetic instability of triplet repeats. J Biol Chem 1996; 271:2875-2878.
111. Levinson G, Gutman GA. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 1987; 4:203-221.
112. Buldyrev SV, Dokholyan NV, Havlin S et al. Expansion of tandem repeats and oligomer clustering in coding and noncoding DNA sequences. Physica A 1999; 273:19-32.
113. Kruglyak S, Durrett RT, Schug MD et al. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci USA 1998; 95:10774-10778.
114. Zipf KG. Human Behavior and the Principle of Least Effort. Redwood City: Addison-Wesley 1949.
115. Mantegna RN, Buldyrev SV, Goldberger AL et al. Linguistic features of noncoding DNA sequences. Phys Rev Lett 1994; 73:3169-3172.
116. Mantegna RN, Buldyrev SV, Goldberger AL et al. Phys Rev E 1995; 2939.
117. Bonhoeffer S, Herz AVM, Boerlijst MC et al. Explaining ''linguistic features'' of noncoding DNA. Science 1996; 271:14-15.
118. Makeev VJ, Lifanov AP, Nazina AG et al. Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. Nucl Acids Res 2003; 31:6016-6026.
119. Jurka J, Kohany O, Pavlicek A et al. Duplication, coclustering, and selection of human Alu retrotransposons. Proc Natl Acad Sci USA 2004; 101:1268-1272.
120. Stanley HE, Afanasyev V, Amaral L AN et al. Anomalous fluctuations in the dynamics of complex systems: From DNA and physiology to econophysics. Physica A 1996; 224 302-321.
121. Pande V, Gosberg A Ya, Tanaka T. Nonrandomness in protein sequences - evidence for a physically driven stage of evolution, Proc Natl Acad Sci USA 1994; 91:12972-12975.
122. Shakhnovich EI, Gutin AM. Implications of thermodynamics of protein folding for evolution of primary sequences. Nature 1990; 346:773-775.
123. Li W-H, Marr TG, Kaneko K. Understanding long-range correlations in DNA sequences. Physica D 1994; 7:392-416.