ELSEVIER

Physica D 75 (1994) 392–416

PHYSICA D

Understanding long-range correlations in DNA sequences

Wentian Li [a], Thomas G. Marr [a], Kunihiko Kaneko [b]

[a] *Cold Spring Harbor Laboratory, P.O. Box 100, Cold Spring Harbor, NY 11724, USA*

[b] *Department of Pure and Applied Sciences, University of Tokyo, Komaba, Meguro, Tokyo 153, Japan*

Abstract

In this paper, we review the literature on statistical long-range correlation in DNA sequences. We examine the current evidence for these correlations, and conclude that a mixture of many length scales (including some relatively long ones) in DNA sequences is responsible for the observed $1/f$-like spectral component. We note the complexity of the correlation structure in DNA sequences. The observed complexity often makes it hard, or impossible, to decompose the sequence into a few statistically stationary regions. We suggest that, based on the complexity of DNA sequences, a fruitful approach to understand long-range correlation is to model duplication, and other rearrangement processes, in DNA sequences. One model, called "expansion-modification system", contains only point duplication and point mutation. Though simplistic, this model is able to generate sequences with $1/f$ spectra. We emphasize the importance of DNA duplication in its contribution to the observed long-range correlation in DNA sequences.

1. Introduction

Some time ago, two of the authors (WL and KK) noticed the existence of statistical long-range base–base correlation in DNA sequences [31,30], which was later observed independently by two other groups [41,56]. Statistically, such long-range correlation can be detected by measuring the two-point autocorrelation function and examining if it decays slower than exponentially. What is surprising in these findings is not just the existence of long-range correlation, but the particular form of the correlation structure – the $1/f$-like spectral component [31,56]. If the power spectrum, which is the Fourier transform of the autocorrelation function, obeys a power-law behavior (e.g. $1/f$), it indicates that the two-point correlation function also decays as a power-law function, in contrast to an exponential decay. Thus, the $1/f$-like spectra imply that there is no single finite length scale for the base–base correlation: there should be many different length scales in DNA sequences.

This point we want to emphasize, that what is interesting about the study of DNA sequences is more a particular non-trivial form of the correlation structure than the mere fact that statistical correlation at longer distances exists, reveals certain complexity of the statistical correlation structure

SSDI 0167-2789(94)00112-4

in DNA sequences. There are cases where correlation does exist at long distances, but its existence is easily explainable, and those sequences lack a multi-length-scaled correlation structure and thus simple. For example, a sequence with an exponential correlation function but nevertheless a relatively long correlation length has a broad Lorenzian spectrum. Another example is a random packing of two types of blocks: its power spectrum has a $1/f^2$ component. None of these cases are as complex as sequences with $1/f$-like spectra.

One might, then, imagine that the statistical character of DNA sequences is very simple and universal – all have a scale-invariance $1/f$ power spectrum. This is not the case. The $1/f$-like spectral component observed in DNA sequences is not the sole spectral component; the white noise component also contributes to the spectrum. Besides, not all DNA sequences exhibit long-range correlations. Another complication in the study of DNA sequences lies in the finiteness of the sequence: the longest length scale of correlation is always bounded. When a long-range correlation is claimed, it is a claim that the longest length scale is not an order of magnitude smaller than the sequence length, not a claim that a power-law correlation function extends to infinity. The current studies only show clearly the existence of statistical correlations up to the length scale of about 1–10 kb (1 kb = 1000 bases). The correlation structure of DNA sequences at even larger scales has not yet been thoroughly studied.

What is the biological significance of these long-range correlations? First we have to note that the statistical correlation is not equivalent to a "cause-and-effect" correlation. A long-range statistical correlation in DNA sequences means the base appearance or base density tends to co-vary at regions separated by a long distance. It is a completely different concept when one says that one region of the DNA sequence has a biological influence on another that is far apart. It is not easy to make a straightforward connection between the long-range correlation from the viewpoint of biological function.

On the other hand, we know very well that complexity stored in an object may be created by a relatively simple process. For example, imagine the dynamical process that doubles the whole sequence (plus a random mutation afterward). Although the doubling process is extremely simple, it actually generates sequences that are multi-length-scaled. One of the authors (WL) has studied a similar model system, called expansion-modification systems [27,29], which can conceptually be related to the mutation and duplication processes in the evolution of DNA sequences. In fact, it was the discovery of this model being able to generate sequences with $1/f$ spectra that led two of the authors (WL and KK) to expect the existence of $1/f$ spectra in real DNA sequences [31,32], and motivated our analysis of DNA sequences.

There are three major purposes of the present paper. The first one is to provide an introduction and a survey of the long-range correlations in DNA sequences. Second, we try to clarify some confusions appeared after the publication of the first round of papers. This discussion is needed because many misunderstandings of the finding still exist. Third, we discuss the dynamical origin of such long-range correlations, in connection with the study of the expansion-modification system.

The organization of the paper is as follows. We start in Section 2 by reviewing some of the measures used to study the correlation structure of DNA sequences, then we graphically illustrate our main theme that there are fluctuations of base densities at many different length scales (Section 2.1). Later, we review the current knowledge of the correlation structure of DNA sequences (Section 2.2). A set of new calculation of power spectra of base density fluctuations is included in Section 2.2. These calculations confirm some previous results.

We review four "controversial" topics in Section 2.3. The first is on how good the scale-invariant

property of DNA sequences is. The second is on whether the DNA sequences with long-range correlation can be decomposed into subregions of different base densities, each with subregions of white noise with no correlations. The third is on whether the $1/f$-like spectra are actually Lorenzian spectra. And the fourth is on whether coding sequences have a common correlation structure that is different from that derived from non-coding sequences. Both the second and the third topics are related to the question on whether DNA sequences are simple or "complex". If a sequence is complex, it cannot be decomposed into spatially separable simple sub-sequences, and it cannot have only a single length scale. Complex sequences have many different length scales and each of the length scale acquires its contribution from throughout the sequence. It will be demonstrated that many DNA sequences are complex.

In Section 3, we study theoretically the "dynamical origin" of the correlation structure of DNA sequence. Section 3.1 reviews the expansion-modification system which is seemingly simple, but is capable of generating sequences with a fair amount of complexity: in this case, the $1/f$ spectrum. In this model, the only processes involved are a point duplication and a point mutation, each occurring with a fixed probability. Although this model is simplistic in describing changes in DNA sequences, both duplication and point mutation are common events for DNA sequences. Since both processes can create new patterns in the sequence, these are potentially important for evolution. This topic is discussed in Section 3.2.

In summary, we use this opportunity to review what we currently know about the statistical long-range correlation in DNA sequences. There are many issues one can question: whether long-range correlation is a common feature of DNA sequences; whether the correlation structure of DNA sequences at 1 Mb (1 Mb= 10^6 bases) is similar to that at 1 kb; whether the correlation structure has any relevance to our understanding of biological function; whether the statistical structure of 1-dimensional DNA sequences can reveal features of the 3-dimensional structure of the chromosome, and so on. We do not have definitive answers to these questions yet. With more and more long stretchs of DNA sequences become available, it is an exciting time for those who are interested in exploring and understanding the large-scale statistical structure of DNA sequences.

2. Statistical correlation structures of DNA sequences

Correlation structure of a length-N DNA sequence is the correlations between two nucleotides of any distance d ($< N$). The intention here is to characterize the correlation structure of a particular DNA sequence with a finite length, rather than to generalize the result to infinitely long sequences, nor to a subregion of the sequence. So we are not particularly concerned about the concept of "statistical stationarity". It is similar to the case when one says that a sample density of G of a particular DNA sequence is 0.21: it does not necessarily imply that the density of G of other sequences is also 0.21, nor does it claim that any subregion of length $n < N$ has the same density of 0.21.

Correlation structure of a finite sequence can be studied by several sample measures of correlation. We list these in three categories:

Direct measures of correlation. We first define a sample "mutual information function": if the indices α and β run through the four nucleotides (G,C,T,A), we define (the cyclic boundary is

used) [1]

$$n_{\alpha\beta}(d) \equiv \sum_{i=1}^{N} 1[x_i = \alpha \text{ and } x_{i+d} = \beta] \tag{2.1}$$

as the total number of counts of a particular joint nucleotide-pair type (α, β) with the two separated by a fixed distance d, and

$$n_\alpha \equiv \sum_{i=1}^{N} 1[x_i = \alpha] \tag{2.2}$$

as the number of counts of the nucleotide type α. Note that $n_{\alpha\alpha}(0) = n_\alpha$.

The sample mutual information function can be defined as (the notation $\hat{}$ is used to indicate an estimate) [2]:

$$\begin{aligned}\widehat{M(d)} &= \frac{1}{N} \sum_{\alpha=(G,C,T,A)} \sum_{\beta=(G,C,T,A)} n_{\alpha\beta}(d) \log\left(\frac{N n_{\alpha\beta}(d)}{n_\alpha n_\beta}\right) \\ &= \log N + \frac{1}{N} \sum_{\alpha=(G,C,T,A)} \sum_{\beta=(G,C,T,A)} n_{\alpha\beta}(d) \log\left(\frac{n_{\alpha\beta}(d)}{n_\alpha n_\beta}\right).\end{aligned} \tag{2.3}$$

This sample mutual information function is an estimator of the mutual information function:

$$M(d) \equiv \sum_\alpha \sum_\beta P_{\alpha\beta}(d) \log\left(\frac{P_{\alpha\beta}(d)}{P_\alpha P_\beta}\right), \tag{2.4}$$

where $P_{\alpha\beta}(d)$ is the joint probability of (α, β) symbol pair, and P_α is the density of symbol α. This quantity was first introduced in information theory [49], and recently is applied to, for example, the study of chaotic nonlinear dynamics [51,16,21,19], symbolic sequence analysis [20,28], learning features from experiments [46], nonlinear prediction [34], improving neural network performance [12], identifying active sites in AIDS virus sequence [23]. Mutual information is now considered a standard measure of correlation (see page 634 of [45]). The famous "PAM" (Point Accepted Mutation) matrix element used to measure the point mutation rate from one amino acid to another is of the form of log-likelihood [9], and the average of all these matrix elements is exactly a mutual information [2]. See also a textbook introduction in [47].

One may also use the traditional covariance or autocorrelation function to calculate the same-symbol correlation (note that cross-correlations between different types of symbols need to be defined separately). Here is a sample covariance function:

$$\widehat{cov(d)} = \sum_{\alpha=(G,C,T,A)} \left(\frac{n_{\alpha\alpha}(d)}{N} - \frac{n_\alpha^2}{N^2}\right), \tag{2.5}$$

[1] An alternative sampling method is to not use the cyclic boundary condition, so $n_{\alpha\beta}(d) \equiv \sum_{i=1}^{N-d} 1[x_i = \alpha \text{ and } x_{i+d} = \beta]$. Because the number of point sampled is less than N, one might want to introduce some correction factor to compensate this fact.

[2] Other alternatives include: (1) to use $\log_{10}$ instead of $\log_e$; and (2) to normalize the function by its value at $d = 0$, i.e., $\hat{m}(d) = \widehat{M(d)}/\widehat{M(0)}$.

which is an estimator of the covariance function:

$$cov(d) \equiv \sum_{\alpha=(G,C,T,A)} \left(P_{\alpha\alpha}(d) - P_\alpha^2\right). \tag{2.6}$$

And one can define a sample autocorrelation function – the sample covariance function normalized by its value at $d = 0$:

$$\widehat{\Gamma(d)} = \sum_{\alpha=(G,C,T,A)} \left(\frac{n_{\alpha\alpha}(d) - n_\alpha^2/N}{n_\alpha - n_\alpha^2/N}\right), \tag{2.7}$$

which is an estimator of the autocorrelation function:

$$\Gamma(d) \equiv \frac{\sum_{\alpha=(G,C,T,A)} \left(P_{\alpha\alpha}(d) - P_\alpha^2\right)}{\sum_{\alpha=(G,C,T,A)} \left(P_\alpha - P_\alpha^2\right)}. \tag{2.8}$$

Again, there are alternative definitions, either concerning the sampling method (e.g., whether or not to use the cyclic boundary condition) or concerning the definition of covariance function for DNA sequences (e.g., whether or not to take the absolute value for each term in Eq. (2.6) before adding these together).

Frequency domain characterization. Due to various reasons (see, e.g., pages 7,8 of [43], including being easier to interpret the result), power spectrum, instead of autocorrelation function, is often used to characterize the statistical structure of a sequence. Here is a sample measure of power spectrum of DNA sequence[3]:

$$\widehat{S(f)} = \frac{1}{N^2} \sum_{\alpha=(G,C,T,A)} \left| \sum_{j=1}^{N} 1[x_j = \alpha] e^{-i2\pi f j} \right|^2. \tag{2.9}$$

The frequency is $f = k/N$ $(k = 1,2,\ldots N/2)$. By the convolution theorem (also known as "Wiener–Khinchin theorem" in this context), a power spectrum is the Fourier transform of the corresponding covariance function or autocorrelation function. We call a sequence "$1/f^\alpha$ spectrum" if its power spectrum behaves as an inverse power-law with the exponent α: $S(f) \sim 1/f^\alpha$. Two extreme cases are trivial: $\alpha \approx 0$ corresponds to white noise, while random walk sequences, step functions, or any sequences with a linear autocorrelation function lead to $\alpha \approx 2$.

Cumulative variables. The approach of using cumulative variables to study correlation structure of DNA sequences can suppress fluctuations in $cov(d)$ or $\Gamma(d)$ [41], but it measures essentially the same thing. Following Ref. [41], we define the purine–pyrimidine sequence as (purine = (A,G), pyrimidine = (C,T)): b_i is 1 if $x_i = (A, G)$, and 0 if $x_i = (C, T)$. The cumulative variable $y_k(l)$ is a sum of b_i within a window of size l: $y_k(l) \equiv \sum_{i=k}^{k+l-1} b_i$. The variance of $y_k(l)$ series $(k = 1,2,\cdots,N-l+1)$ – $var_y(l)$ (subscript y indicates the y series)– can be estimated by

$$\widehat{var_y}(l) = \frac{1}{N} \sum_{k=1}^{N-l+1} y_k(l)^2 - \frac{1}{N^2} \left(\sum_{k=1}^{N-l+1} y_k(l) \right)^2. \tag{2.10}$$

[3] There are also other definitions of power spectrum for multi-symbol sequences (e.g., [52]).

It can be shown that $var_y(l)$ is a sum of covariance functions from the original purine–pyrimidine sequence [41,22] (subscripts b indicates the original binary series):

$$var_y(l) = l \cdot cov_b(0) + 2(l-1) \cdot cov_b(1) + 2(l-2) \cdot cov_b(2) \cdots + 2 \cdot cov_b(l-1). \quad (2.11)$$

From Eq. (2.11), if $\{b_i\}$ is a white noise, then all $cov_b(d)$'s are zero except $cov_b(0)$, so $var_y(l)$ is a linear function of l. Similarly, if $\{b_i\}$ is generated by a first-order Markov chain with only 1-step correlations, $cov_b(d)$ decreases with d exponentially, so $var_y(l)$ is almost linearly proportional with l, plus an exponential correction. On the other hand, if $cov_b(l)$ decays slower than exponential functions, $var_y(l)$ may not change with l linearly. In particular, a power-law function in $cov_b(d)$ leads to a power-law function in $var_y(l)$.

Any one of the three approaches (direct calculation, frequency domain characterization, and cumulative variables) can be used to analyze the correlation structure of DNA sequences. In the following, we will first visualize the base composition fluctuation in some long DNA sequences, then review the study of correlation structure that has appeared in recent publications, and discuss some issues currently under debate.

2.1. Large scale fluctuations in base composition

One of the easiest ways to obtain information from a DNA sequence is to "view" it. Good visualization can help us develop intuition about the sequence. Since the study of correlation structure is essentially the study of fluctuation of base densities at different length scales, we view the base densities within a moving window of different sizes (W). Besides the window size, another parameter is the moving distance from one window to another (D).

To simplify the work, we plot two types of densities instead of four base densities. The first is the density for G or C ("$G + C$ content") within the jth moving window:

$$\rho_{G+C}(j) = \frac{1}{W} \sum_{i=(j-1)D+1}^{(j-1)D+W} 1[\text{ if } x_i = (G,C)], \quad j = 1,2,\ldots. \quad (2.12)$$

If the moving distance D is equal to the window size W, windows are non-overlapping; if $D < W$, windows are overlapping. Similarly we can calculate the purine density:

$$\rho_{\text{purine}}(j) = \frac{1}{W} \sum_{i=(j-1)D+1}^{(j-1)D+W} 1[\text{ if } x_i = (A,G)], \quad j = 1,2,\ldots. \quad (2.13)$$

Fig. 1 shows ρ_{G+C} and ρ_{purine} for the complete DNA sequence of *Saccharomyces cerevisiae* (also called budding yeast) chromosome 3 (GenBank locus name: SCCHRIII; accession number: X59720) [38] with window sizes of $W = 1000$, 10000, and 100000 bases, respectively (the moving distances D is fixed at 1000 bases). The sequence length is $N = 315{,}338$ bases. The average density of $G + C$, 0.3855, and the average density of purine, 0.4998, are drawn by the horizontal lines.

The first impression of Fig. 1 is that the smoother fluctuations with larger window sizes look quite differently from that with the smaller window size. In other words, the small-scaled fluctuation behaves independently from those of large-scaled fluctuation, and both can be very complicated. Also note that the fluctuation of $G + C$ content and that of purine density are completely different, giving a caution to those studies that reduce a DNA sequence to one of the binary sequences. The

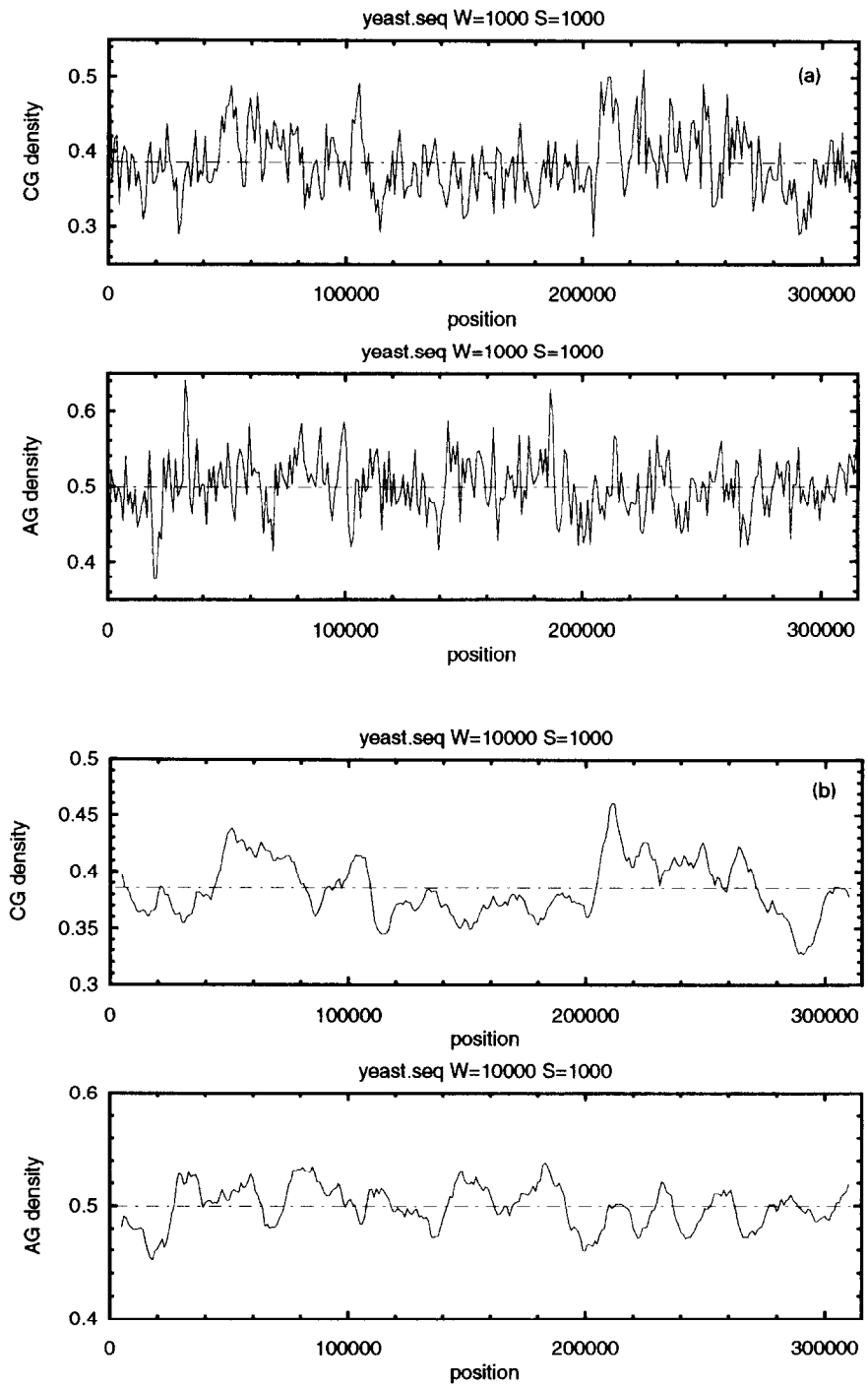

Fig. 1. $G + C$ content and purine density of budding yeast chromosome 3 with various window sizes ($W = 1000$, 10000, 100000). The moving distance is fixed at $D = 1000$. The sequence length is $N = 315,338$ bases. The moving windows in the second and the third case are overlapping.

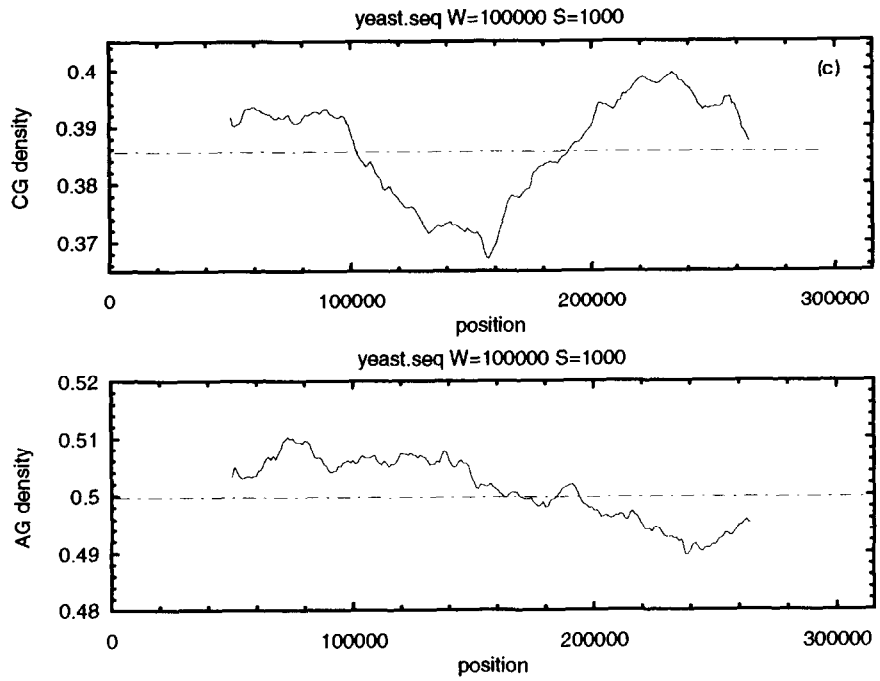

Fig. 1 — continued.

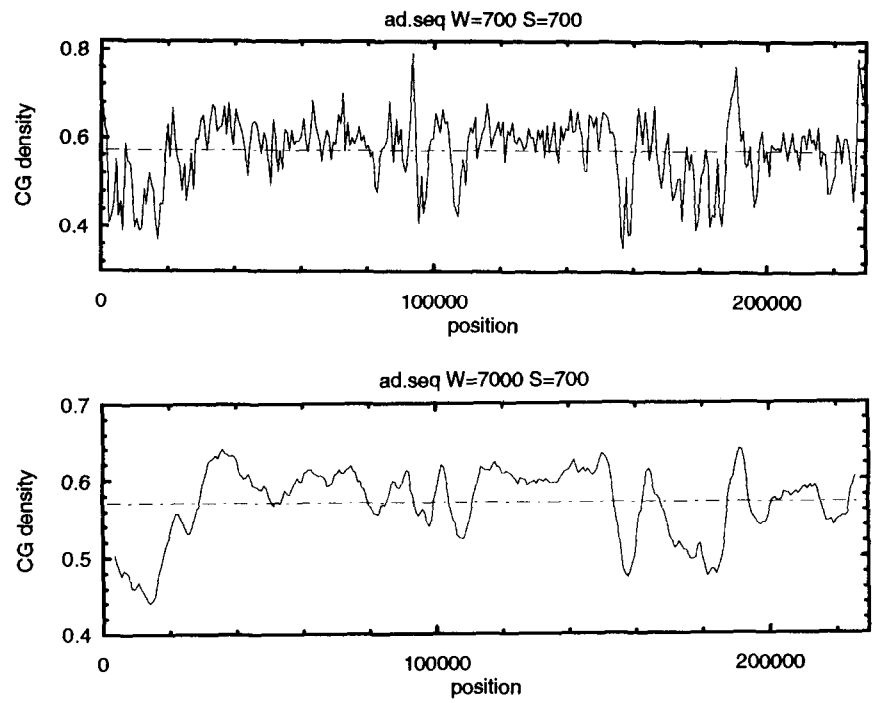

Fig. 2. Fluctuation of $G + C$ content in human cytomegalovirus (strain AD169) sequence with window sizes $W = 700$, 7000 The moving distance is fixed: $D = 700$. The sequence length is $N = 229,354$ bases.

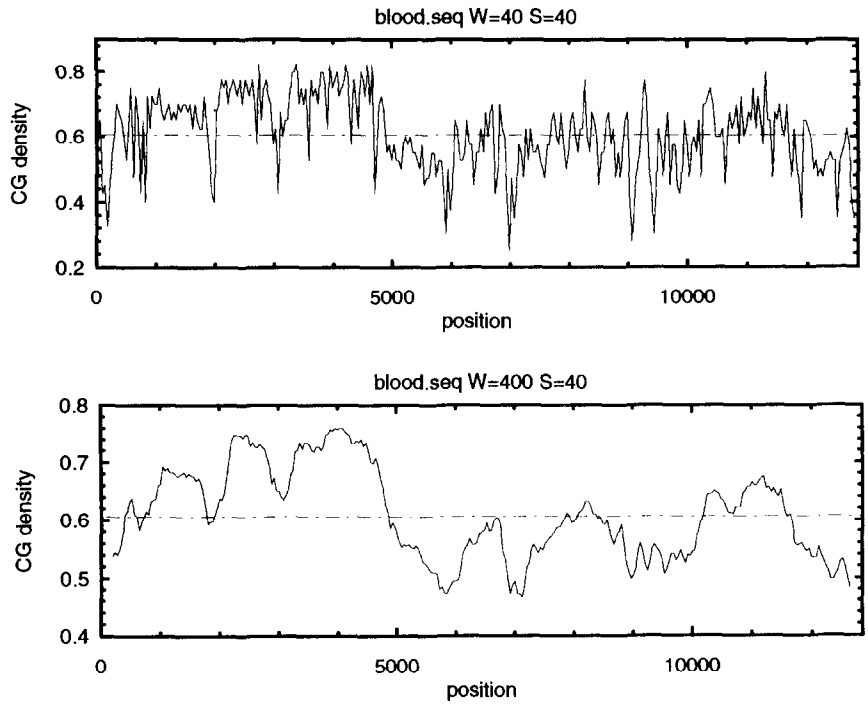

Fig. 3. Fluctuation of $G + C$ content in human blood coagulation factor VII gene sequence with window sizes $W = 40, 400$. The moving distance is fixed: $D = 40$. The sequence length is $N = 12,850$ bases.

fluctuation of $G + C$ content with the window size $W = 100000$ is reminiscent of the "isochore" in frog, chicken, mouse, and human genomes studied by Bernardi's group [3] [4].

The density fluctuation in budding yeast chromosome 3 sequence illustrates well the possible existence of many different length scales. It is not uncommon in this type of situation that a $G + C$ rich region can contain many subregions that are $G + C$ poor, or vice versa. Here we plot a few more density fluctuations of other DNA sequences whose statistical structure has been calculated in one way or another elsewhere. To save space, we only plot the ρ_{G+C} fluctuations. These sequences are:

- The complete sequence of a human cytomegalovirus (strain AD169, with the GenBank locus name: HEHCMVCG, and accession number X17403) [8]. The sequence length is $N = 229,354$ (Fig. 2). The power spectrum of its base sequence was calculated in [56].
- The complete sequence of human blood coagulation factor VII gene (GenBank locus name: HUMCFVII, and accession number: J02933) [37]. The sequence length is $N = 12,850$ (Fig. 3). Both the mutual information function and power spectrum of this sequence is calculated in [31,30].
- The complete sequence of bacteriophage lambda (GenBank locus name: LAMCG; accession number: J02459) [48]. The sequence length is $N = 48,502$. All studies of this sequence were by using the cumulative variable method, see Refs. [41,22,42,25].

[4] An interesting observation of the budding yeast chromosome 3 sequence is that only coding region is responsible for the $G + C$ rich/poor variations [50]. The large-scale variation of $G + C$ content along the sequence is not observed in non-coding segments [50].

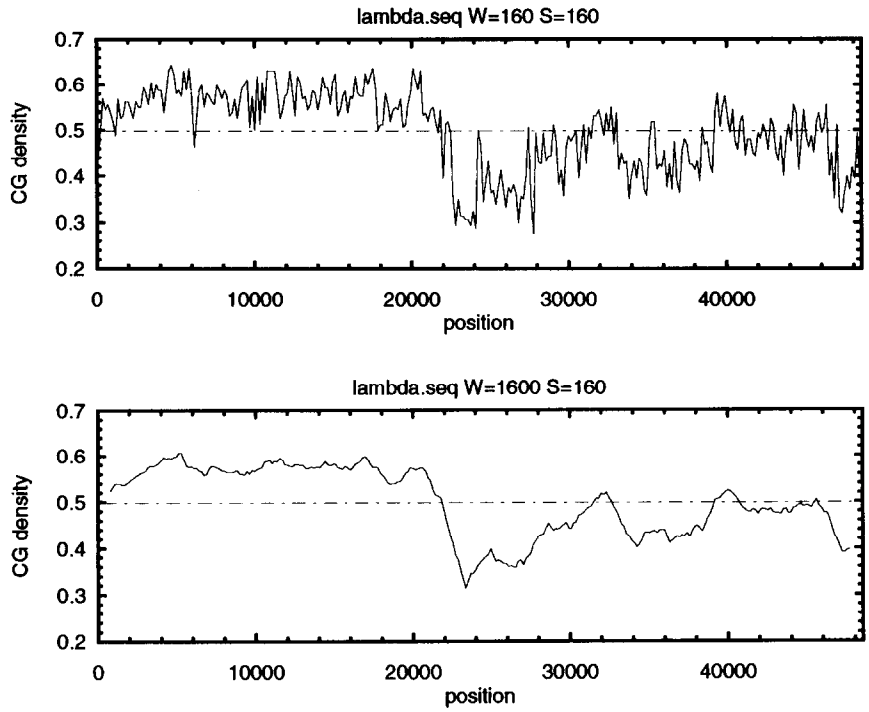

Fig. 4. Fluctuation of $G + C$ content in bacteriophage lambda sequence with window sizes $W = 160$, 1600. The moving distance is fixed at $D = 160$. The sequence length is $N = 48,502$ bases.

To make the plots in Figs. 1–4 somehow comparable, we choose the moving distances to be about 1/300 of the sequence length ($D = 700$, 40 and 160 for Fig. 2, 3 and 4). The first plot in each figure has $W = D$, and the second plot has $W = 10D$. Among the four sets of plots, it seems that bacteriophage lambda sequence shows the least change as the window size is increased by ten-fold. We will have more comments on these four sequences in the next section.

2.2. *Published results of statistical correlation structure of DNA sequences*

In Ref. [30], the mutual information function of five coding sequences and five non-coding sequences from human DNA is calculated. The sequence length ranges from around 2 kb to 16 kb. It is observed that correlation structure as measured by the mutual information function can differ from one sequence to another. For these sequences (ten in all), the correlation decays to a negligible value at less than 10 bases for coding sequences, and at slightly more than 10 bases for non-coding sequences, with the exception of one coding sequence and one non-coding sequence.

That exception in non-coding sequence has the longest correlation length among the 10 cases, which is the blood coagulation factor VII gene whose base composition fluctuation is plotted in Fig. 3. The correlation structure of the sequence is examined in more detail in Ref. [31], using the complete sequence instead of only the non-coding region (though 76% of the sequence is non-coding). In particular, it is observed that the power spectrum of the sequence has a $1/f^{\alpha}$ spectral component, with $\alpha \approx 0.84$, 0.57 and 0.53 when the first, the middle, and the last $2^{13} = 8192$ bases are taken for the spectral analysis (the sequence length is $N = 12850$). This was the first example that the $1/f^{\alpha}$ spectral component in DNA sequence was shown.

The similar $1/f^{\alpha}$ spectral component is also observed in a DNA sequence with much longer length: the human cytomegalovirus [56], whose base composition fluctuation is plotted in Fig. 2. The sequence is cut into 3 or 4 pieces, and $2^{16} = 65536$ bases from each piece are used for a spectrum calculation. It should be noted that correlation structure at length scales longer than 65 kb is not studied in [56] even though the sequence length N is approximately 229 kb.

Another interesting study in [56] is the average spectrum of *all* sequences in GenBank with $2^{11} = 2048$ base in each spectrum calculation. As such, this study reflects only the average correlation structure up to the length scale of 2 kb. The exponent α in the $1/f^{\alpha}$ spectral component is amazingly close to 1, highly reminiscent of the similar spectral analysis of music time series [58]. A reason why a much better quality $1/f$ spectral component is observed in this case is that there are many more different length scales mixed into one statistic, and a good $1/f$ spectrum requires many different length scales.

The base–base correlations in a 300 kb long sequence – the complete sequence of chromosome 3 of budding yeast, whose base composition fluctuation is plotted in Fig. 1 – are calculated in [60] for all possible types of base-pair: $\widehat{I_{\alpha\beta}(d)} = N\frac{n_{\alpha\beta}(d)}{n_{\alpha}n_{\beta}}$. A difference from 1 for $\widehat{I_{\alpha\beta}(d)}$ indicates a correlation. It is observed that $\widehat{I_{\alpha\alpha}(d)}$'s approach 1 from above, while $\widehat{I_{\alpha\beta}(d)}$'s ($\beta \neq \alpha$) approach 1 usually from below, but sometimes alternatingly from above and below [60]. Also, $\widehat{I_{\alpha\alpha}(d)}$'s reaches 1 at longer distances than $\widehat{I_{\alpha\beta}(d)}$'s, i.e., the longest length scale for yeast chromosome 3 is decided by the base-pairs of the same type, which is about 1 kb [60].

All these published studies of correlation structure examine base–base correlation. To give a more complete picture, we include here (Figs. 5–8) power spectra of density fluctuations of the four DNA sequences illustrated in Figs. 1–4. In this calculation, a sequence is partitioned into $n = 2^m$ non-overlapping windows, and the densities of four types of nucleotide are calculated ($\rho_A(j)$, $\rho_C(j)$, $\rho_G(j)$ and $\rho_T(j)$, for $j = 1, 2, ..., n$). The overall power spectrum of the density fluctuation is the sum of four power spectra of each nucleotide type sequence. If the sequence length is long, the spectral analysis for density fluctuation is computationally easier than that of the base sequence because of its smaller number of data points.

The window sizes used in Figs. 5–8 are around a few dozen bases (76, 55, 25 and 23 bases, respectively). The ranges of x and y axes are chosen in such a way that the cross-diagonal line corresponds to exactly $1/f$ (Figs. 5, 6 and 7) or $1/f^2$ (Fig. 8). Roughly speaking, the low frequency power spectra are close to $1/f^{0.5}$ in Fig. 5, and close to $1/f$ in Figs. 6 and 7, but close to $1/f^2$ in Fig. 8. Figs. 6 and 7 are consistent with the previous results that human cytomegalovirus (strain AD169) sequence [56] and human blood coagulation factor VII gene sequence [31] have $1/f$-like spectra. Fig. 8 helps us to understand both sides of the argument on whether there is a long-range correlation in the bacteriophage lambda sequence [41,42,22,25]. We can see that the low frequency spectrum is $1/f^2$, which is essentially trivial and easily explainable.

2.3. Review of some controversial topics

Another line of approach to the study of correlation structure in DNA sequences first appeared in [41]. Although it claims less concerning the correlation structure than [31] and [56] (the latter emphasizes the particular correlation structure: the $1/f^{\alpha}$ spectral component), this paper has generated much more controversy. In this subsection, we will focus on reviewing four controversial topics, with three generated by that paper and one by Ref. [31]). These are: (1) how good is

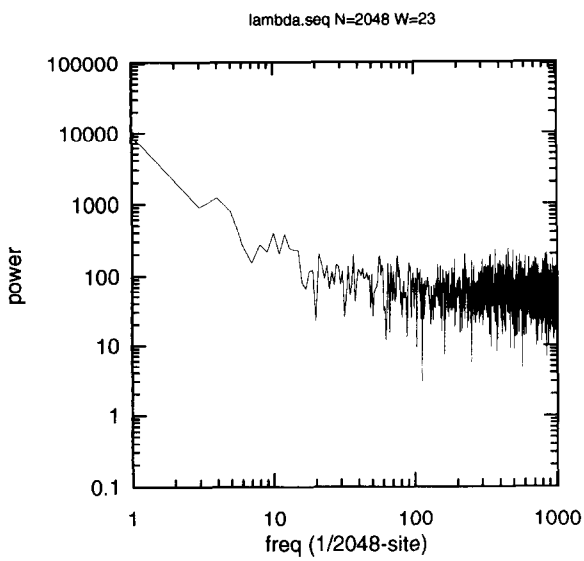

Fig. 5.

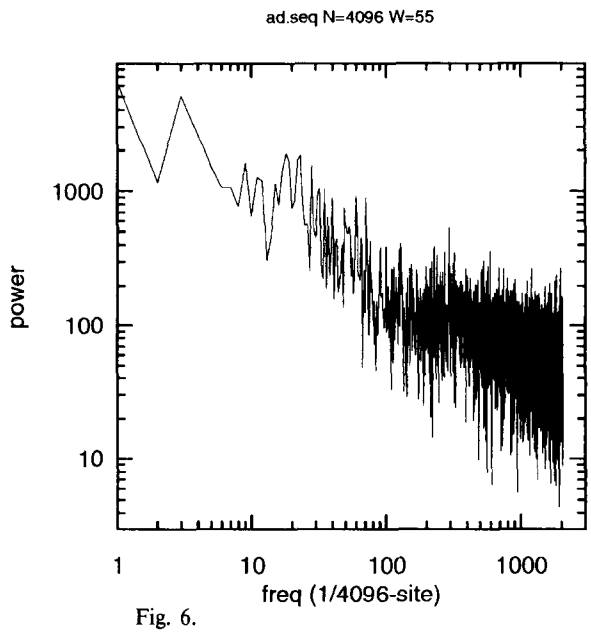

Fig. 6.

Fig. 5. Power spectrum of the density fluctuation of the budding yeast chromosome 3 sequence. The window size is 76, and the number of non-overlapping windows is $2^{12} = 4096$. The total number of bases included in the calculation is 311,296 ($= 4096 \times 76$) with the last 4042 ($= 315,338 - 311,296$) bases unused.

Fig. 6. Power spectrum of the density fluctuation of the human cytomegalovirus (strain AD169) sequence. The window size is 55, and the number of non-overlapping windows is $2^{12} = 4096$. The total number of bases included in the calculation is 225,280 ($= 4096 \times 55$) with the last 4074 ($= 229,354 - 225,280$) bases unused.

the scale invariance in the correlation structure in DNA sequences; (2) whether the observed non-white-noise feature in DNA sequences is purely a consequence of the sequence being composed of subregions with different base densities, whereas each subregion is a white noise; (3) whether the $1/f$-like spectral components in DNA sequences are actually Lorenzian spectra; and (4) to what degree coding and non-coding sequences have different statistical correlation structures (this topic was actually suggested earlier in [30]).

On scale invariance. As we have seen in Eq. (2.11), function $var_y(l)$ changes with l much more smoothly than function $cov_b(d)$ changes with distance d. It seems hard to observe $cov_b(d)$ as a power-law function of d in DNA sequences, but much easier to observe $var_y(l)$ as a power-law function of l, at least up to some upper limit.

Several publications show that the power-law behavior of $var_y(l)$ breaks down at larger values of l's [44,7,22]. It is quite natural to assume a mixture of a few length scales instead of a continuous distribution of length scales as usually the case for scale-invariant systems. Most important, if there is an upper limit for the largest length scale, the assumption of scale invariance is no longer valid beyond that length scale, and one may not be able to reliably calculate the scaling exponent. See also [35] for a $var_y(l)$ versus l plot from the budding yeast chromosome 3 sequence.

On whether a complex sequence is decomposable to subregions with simple correlation structures. Based on the observation of mainly one sequence (bacteriophage lambda, whose base composition

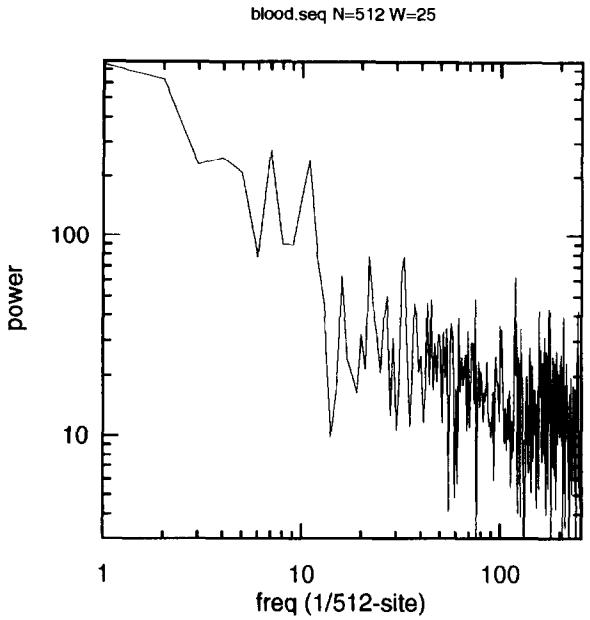

Fig. 7.

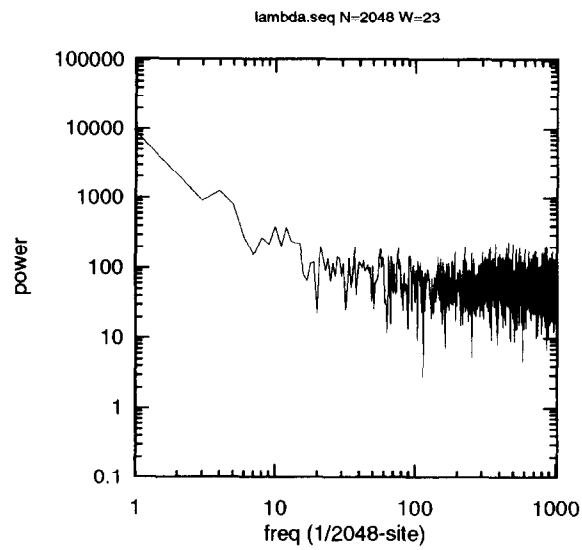

Fig. 8.

Fig. 7. Power spectrum of the density fluctuation of the human blood coagulation factor VII gene sequence. The window size is 25, and the number of non-overlapping windows is $2^9 = 512$. The total number of bases included in the calculation is 12800 ($= 512 \times 25$) with the last 50 ($= 12850 - 12800$) bases unused.

Fig. 8. Power spectrum of the density fluctuation of the bacteriophage lambda sequence. The window size is 23, and the number of non-overlapping windows is $2^{11} = 2048$. The total number of bases included in the calculation is 47104 ($= 2048 \times 23$) with the last 1398 ($= 48502 - 47104$) bases unused. Note that the scale of the y-axis is expanded so that the cross-diagonal line corresponds exactly to 1/frequency^2.

fluctuation is plotted in Fig. 4), it is claimed that the long-range correlation observed in [41] can be fully accounted for by the difference of base compositions in subregions (might also be called "patches" or "domains"), such as the difference of $G + C$ content [22]. Besides the problem that this sequence was not even listed as an example of long-range correlation in [41], there are other problems in the argument in [22].

First, the effect of different $G + C$ content or purine density at different regions can be removed by "detrending", as in the traditional approach for studying non-stationary time series such as the gross national product time series [17]. The detrending is carried out for the bacteriophage lambda sequence as well as other controlled sequences [42]. It is shown that in those detrended sequences, both a trivial linear function and a non-trivial scaling function are possible for $var_y(l)$ [42]. Not surprisingly, $G + C$ content is not responsible for *all* features of correlation structure of a DNA sequence. And often it is completely irrelevant to the observed long-range correlation.

Second, even if the effect of different base composition in different subregions is not removed, it can be easily recognized as a trivial source of correlation. The power spectra of density fluctuation presented at the end of last section illustrates this point: the bacteriophage lambda sequence has a $1/f^2$ spectral component while other three sequences have a $1/f$-like spectrum. Although both types of spectra are not white noise, it is the $1/f$-like spectra that reveals the more interesting multi-length-scaled long-range correlation.

To emphasize the trivial contribution of subregions with different base compositions to the correlation structure, a simple calculation of the covariance function is carried out in the Appendix for both the case with two subregions and the case with many subregions. If the correlation structure within any subregion is a white noise, the overall covariance function is a constant term contributed from squares of base density differences, assuming the distance is smaller than the subregion sizes (see the Appendix). When the distance is larger, more samplings are obtained from crossing two neighboring subregions. Even if we add this term, thus a constant covariance function becoming distance-dependent, the distance-dependent form is still simple (linear). Such simple functional form will not lead to a $1/f$-like spectrum.

The central issue in this discussion is whether DNA sequences are simple or complex. The white noise is the simplest type, and patching fixed-length subregions may be the next simplest type. Only when we patch subregions with many lengths that satisfy a power-law distribution, it is possible to have a $1/f$-like spectrum. Nevertheless, it is not the patching of white noise subregions that is responsible for the complexity of the sequence. Rather, it is the extra condition introduced to the length distribution (i.e., the power law distribution) that makes the sequence multi-length-scaled and complex. As we examine more and more DNA sequences, we will realize that most DNA sequences are not simple.

On whether the spectra of DNA sequences are Lorenzian spectra. In a recent study [4], it is questioned whether a Lorenzian spectrum is "mistaken" as a $1/f^{\alpha}$ spectral component. Following the notation in [4], suppose the autocorrelation function is $\Gamma(d) = \sum_{d_0} a(d_0)e^{-d/d_0}$ when $d = 1, 2, \cdots$, while $\Gamma(0) = 1$ ($a(d_0)$ is the weight of contribution from a particular length scale d_0), then the power spectrum via the discrete Fourier transform is:

$$S(f) = 1 + \sum_{d_0} a(d_0) \frac{\cos(2\pi f) - e^{-1/d_0}}{(e^{1/d_0} + e^{-1/d_0})/2 - \cos(2\pi f)}. \tag{2.14}$$

Each term with a fixed d_0 value is a Lorenzian-type spectrum, which approaches $2d_0/(1 + (2\pi d_0 f)^2)$ in small f and large d_0 limit. The latter standard expression of the Lorenzian spectrum is obtained by the continuous Fourier transform of the exponential function.

The analysis carried out in [4] is the average spectrum for flanking sequences (those located before or after a gene), with the DNA sequence being converted to a binary sequence, and it seems that $2^{12} = 4096$ bases are taken for each spectrum calculation (see the black dots in Fig. 2 of [4]).

First we note that the power spectra presented in [31] and [56] are not Lorenzian spectra: there is no tendency at low frequencies for the spectrum to flatten out which is a signature of the Lorenzian spectrum. So whether the spectrum in the Fig. 2 of [4] is a Lorenzian spectrum or not does not affect the claim of $1/f$-like spectra in [31] and [56] – as different segments of DNA sequences may have completely different statistical structures. On the other hand, if the flanking sequence studied in [4] is indeed a Lorenzian spectrum, it will not be mistaken as a $1/f$-like spectrum because the low frequency components will be off the $1/f$ line (e.g., see the first black dot in Fig. 2 of [4]) [5].

[5] The second spectrum in Fig. 2 of [4], as represented by open circles, is for an artificially joined long sequence of flanking segments. As it is hard to imagine that unrelated flanking segments can have any correlation between them, it is not surprising that the resulting long sequence will lack a long-range correlation.

Second, the connection between $1/f$ spectra and Lorenzian spectra has long before been suggested [55]. A Lorenzian spectrum represents a specific length scale, whereas a $1/f$ spectrum indicates a mixture of many different length scales. Naturally, a superposition of many Lorenzian spectra may lead to a $1/f$-like spectrum. One can easily perform some tests by adding more and more Lorenzian spectra with larger and larger length scales, and it will be observed that the overall spectrum can look very much $1/f$-like. In particular, the exact $1/f$ spectrum corresponds to the case of $a(d_0) \propto 1/d_0$. Nevertheless, the decomposition of a $1/f$-like spectrum to many Lorenzian spectra can be arbitrary, so is the number of length scales as well as their values.

The third point is on the flat plateau at high frequencies. The existence of this plateau depends on the fact that $\sum_{d_0} a(d_0) < 1$, i.e., there is a "discontinuous" drop in $\Gamma(d)$ from $d = 0$ to $d = 1$. If $\sum_{d_0} a(d_0) = 1$, a Lorenzian spectrum extends its $1/f^2$ tail to high frequencies and there will be no identity confusion with $1/f$ spectrum. This drop in autocorrelation seems to be the source of the "white noise component" observed in [31] and [56].

On the difference of correlation structure between coding and non-coding regions. In higher organisms, the mapping of a stretch of DNA to the corresponding amino acid sequence is frequently not continuous, but is interrupted by the regions called "introns". Introns are non-coding regions, since its DNA contents do not translate to part of the amino acid sequence. The parts that are translated to the amino acid sequence are called "exons". Much effort is spent on automatic recognition of intron/exon regions, and currently the success rate for a computer recognition is about 60–70 % [14].

If we know which sequence is coding and which is non-coding, we can determine the correlation structure for each sequence, and then see whether there is a common property among all coding sequences or among all non-coding sequences. There does not seem to be *a priori* reason that coding sequences should have a different correlation structure from that of non-coding sequences.

In [13], the correlation coefficient between the base composition in two co-moving windows is studied as a function of the distance between the two windows. This correlation coefficient decays much slower in human DNA than *E. coli* DNA [13]. Non-coding regions are common, of course, in human DNA, and rare in *E. coli*.

In [30], it is observed that non-coding sequences consistently have longer correlation length than coding sequences. Nevertheless, it is not clear whether the result can be generalized to other cases because there are few sequences being examined and all sequences are human DNA.

In [41], it is shown that all complementary DNA (cDNA) (they are made by reverse transcription from the mature mRNA and thus do not contain introns) and other intron-less DNAs such as prokaryote DNAs being studied do not exhibit long-range correlation (judged by whether $var_y(l)$ is a linear function of l). However, many intron-containing DNAs do have long-range correlation (judged by whether $var_y(l)$ is a non-linear function of l). Two more examples are given in [5] showing again that intron-less sequences do not exhibit long-range correlation.

One suggestion is that intron-containing sequences have long-range correlation because intron and exons may have totally different statistical properties and when they are mixed in the sample statistics, a spatial structure might be detected [36], as shown for example by the calculation in Appendix (e.g., Eq. (A.7)). It implies that if the similar calculation is carried out for intron only, no long-range correlation will be detected. Such calculation has already been done in [30] and clearly intron alone can exhibit long-range correlation.

Instead of separating exon and intron, in Ref. [56], DNA sequences were grouped according to

their Genbank categories[6]. Besides the fact that the exponent α in the $1/f^{\alpha}$ spectral component is slightly different from one category to another [56], there is a striking qualitative difference between the spectra of bacteria and phage sequences and those of others, Note that bacteria and phage sequences contain mostly coding sequences – these sequences are more "compact" or "efficient" in the context of protein-making.

In another study [57], it is shown that in a certain grouping among a few GenBank categories (group 1 contains primate, rodent, mammal, vertebrate DNAs, group 2 contains invertebrate, plant DNAs, and group 3 contains virus, organelle, phage DNAs), intron sequences and exon sequences do not exhibit dramatic difference in their $1/f^{\alpha}$ spectral component. Some of the grouping is questionable, such as whether organelle DNAs share anything in common with virus DNAs. Also, one important question not addressed is about non-coding sequences other than introns, i.e., the intergenic sequences: whether intergenic sequences have a different correlation structure from that of intron sequences.

Despite the inconclusiveness of our knowledge concerning the extent the correlation structure differs between coding and non-coding sequences, we make the following speculations:

- As mentioned before, there is no *a priori* reason to believe that the correlation structure should be different between coding and non-coding sequences. However, there are major physical characteristics of coding sequences, most notably the constraint of triplet (codon) usage in translation (for a review of this topic see several chapters in [10]). One reason to expect that correlation structure might be different in the two is that coding regions are likely to be under different evolutionary constraints than non-coding regions.
- If this is correct, the way correlation structure diverges between coding and non-coding sequences would depend on what have been the main driving forces in the DNA changes in non-coding regions, and how often each mechanism of DNA change has been in effect. For example, if point mutation is the major driving force to change non-coding regions, non-coding regions would look more like a random sequence than coding regions. On the other hand, if duplication of larger segments is the major driving force, non-coding regions would become less random and more regular than coding regions.
- Will the difference of correlation structure between coding and non-coding regions ever be useful in an algorithm for automatic recognition of coding regions? It would be less useful if other methods are more accurate and if the DNA sequences are known accurately. But if the sequence data is error-prone (it might be the case for a brute-force sequencing in the early stage of human genome project), one needs to re-evaluate each method [53]. Since the calculation of correlation structure is rather insensitive to a frame shift due to point deletion or point insertion, the result will be the same whether such error exists or not in the sequence.

Finally, since it has been shown that long-range correlation exists in budding yeast chromosome 3 sequence [60,35], it is interesting to note that 30% of the sequence is estimated to be non-coding and most of them are intergenic sequences instead of introns [15].

[6] As pointed out in [7], these categories are not of the equal taxonomic rank.

3. Dynamical origins of long-range correlation in DNA sequences

If we can recreate the dynamical process that led to the current DNA sequences, we should be able to understand why these sequences have the correlation structure they have today. For example, we can take the point of view that the Markov chain model implicitly assumes that the sequence is created from one end to another: at each moment, a new site is considered, and a new base is emitted from the source of the Markov chain to become the last base of the sequence. The probability of picking each base-type depends on the previous base (or the previous few bases for higher-order Markov chains) in the sequence. Such process clearly is not a natural one for modeling DNA evolution, and it is thus not surprising that the Markov chain fails to characterize many features of the correlation structure of DNA sequences.

A different class of models carries out sequence manipulation, i.e., updating an existing sequence by certain rules. For example, cellular automata are a type of sequence manipulation that synchronously updates bases according only to local environments (see [59,54] for a general discussion and [18] for recent work). The idea to use cellular automata to model DNA sequence change is considered in Ref. [6]. We might ask the following question: what is the resulting correlation structure if one applies a cellular automaton rule repeatedly on a sequence that is initially white noise?

This question is studied in [26]. The answer is provided only for a special class of situations: if the dynamics of a cellular automaton is periodic, then the sequence can be characterized by regular languages (almost the same as Markov chains), and the autocorrelation function of the limiting sequence is exponential – long-range correlation occurs (if it does) only within the framework of exponential functions [26]. For more general situations, we make a few comments:

- Since cellular automata updates bases locally, it is usually hard to propagate local effects cooperatively to generate a long-range correlation.
- For some cellular automata, their temporal dynamics are irregular and their spatial configuration is unstable – these are termed "chaotic"[7]. Since a local effect propagates very fast in chaotic cellular automata, we would expect some long-range correlation exists in the limiting sequence. Nevertheless, such long-range correlation tends to be a weak signal as compared with the high degree of randomness created by the chaotic dynamics. In this case, the long-range correlation is expected to be statistically insignificant.
- The best candidate in cellular automata for generating detectable long-range correlation has a combination of the ability to propagate local effects and a low randomness level. Such rules are the "edge-of-chaos" cellular automata [24,33]. In fact, because the correlation at intermediate ranges is small for cellular automata with both periodic and chaotic dynamics, the existence of such correlation has been used to locate the edge-of-chaos region in a cellular automata rule space [33].
- In a special class of cellular automata that maintain propagating gliders (solitons), the spatial pattern can be very complicated [1]. In several cases, the spatial spectra after averaged over

[7] To avoid confusion, this definition for a cellular automaton being chaotic [40,33] is different from the definition in continuous-variable non-linear dynamical systems: here, we have *linear* propagation of perturbation in *position space*, there, it is the *exponential* propagation of perturbation in *variable space*. One way to reconcile the two is to map a (e.g. binary) sequence to a real value so that the binary representation of the real value is that binary sequence with a particular choice of the decimal point. By doing so, a linear divergence in the position space corresponds to an exponential divergence in the variable space.

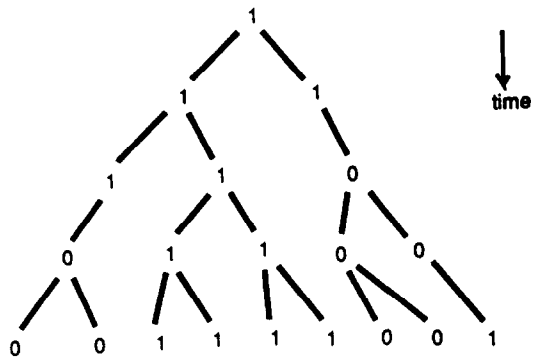

Fig. 9. An illustration of the dynamical process as defined by the expansion-modification system (reproduced from Fig. 1 of [29]).

500 time steps are $1/f^{\alpha}$ ($\alpha > 0.5$) at low frequencies (see Fig. 4.2b, 4.7, and 4.9 of [1]). But note that such low-frequency $1/f^{\alpha}$ spectra often appear during the transient, also note that different length scales may be picked up at different time steps (because a time average has been performed in [1]).

Since cellular automata are again not clearly the most "natural" model for DNA evolution, we turn our attention to sequence manipulations that increase the sequence length. In the following two subsections, we will first review the "expansion-modification" system, and then argue about the relevance of models like expansion-modification systems to DNA evolution. Some of our points were already presented in [32].

3.1. Expansion-modification systems

The prototype of the expansion-modification systems is defined as follows (for two symbols) [27,29]:

$$\begin{array}{ll} t & t+1 \\ 1 \rightarrow & \begin{cases} 11 : 1-p \\ 0 \ : p \end{cases} \\ 0 \rightarrow & \begin{cases} 00 : 1-p \\ 1 \ : p. \end{cases} \end{array} \tag{3.1}$$

To describe in words, the symbol 1 at time t is updated to either two symbols 11 (with a probability $1-p$), or the symbol 0 (with a probability p). Similar action is applied to the symbol 0. An illustration of this dynamical process is in Fig. 9. An example of the spatial-temporal pattern is presented in Fig. 10.

When the probability for the "switch" operation ($= p$) is small, the limiting sequence exhibit a perfect $1/f^{\alpha}$ ($\alpha \approx 1$) spectrum (unlike DNA sequences with $1/f^{\alpha}$ spectral component, spectra of sequences generated by Eq. (3.1) do not contain a white noise component). The exponent α is a function of p. One such spatial $1/f$ spectrum generated by this system is presented in Fig. 11.

Unlike cellular automata, the expansion-modification system uses a different way to propagate local effects to global scale: it forces the old neighbors farther away by creating new neighbors. This

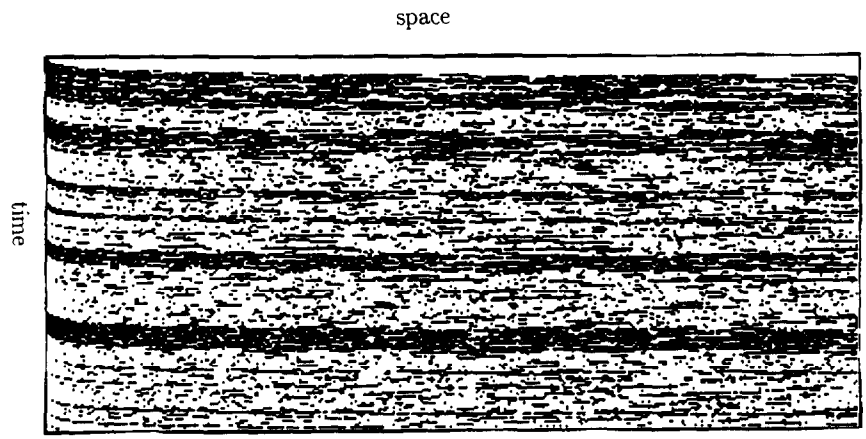

Fig. 10. An example of the spatial-temporal pattern of the expansion-modification system at $p = 0.1$ (reproduced from Fig. 2b of [29]).

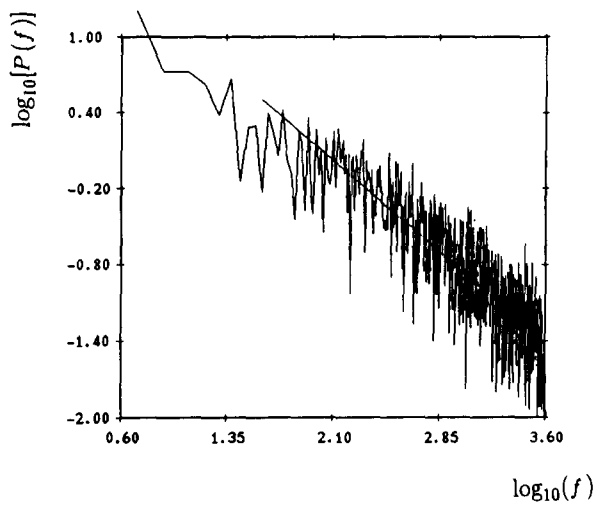

Fig. 11. Spatial spectrum of a sequence generated by the expansion-modification system at $p = 0.1$ (reproduced from Fig. 3b of [29]).

mechanism avoids the difficulty of having both elements of chaos and non-chaos at the same time as in the case of a cellular automaton.

A simple argument shows that power-law covariance or autocorrelation functions can be *maintained* by the expansion-modification system. The argument goes as follows (we use subscript to represent the time step): first, there is a linear expansion of the distance after each time step:

$$d_{t+1} \approx d_t(2-p). \tag{3.2}$$

Then, the covariance at time step t, distance d_t is essentially the same with that at time step $t+1$, distance d_{t+1}, with perhaps a proportionality factor λ. Using Eq. (3.2), we have

$$cov(d_{t+1})_{t+1} \approx \lambda cov(d_t)_t \approx \lambda cov\left(\frac{d_{t+1}}{2-p}\right)_t. \tag{3.3}$$

Assuming that the covariance function is invariant by the updating (another way to say this is that the function reaches the asymptotic limit), so we can drop all subscripts:

$$cov(d) \approx \lambda cov\left(\frac{d}{2-p}\right). \tag{3.4}$$

A power-law function $cov(d) \sim 1/d^{\beta}$ is a solution of Eq. (3.4), as long as

$$\beta = -\frac{\log(\lambda)}{\log(2-p)}. \tag{3.5}$$

When $\beta \approx 0$, the power spectrum $1/f^{1-\beta}$ becomes $1/f$. For β being 0 requires λ being close to 1, but p away from 1.

This argument is used in [27,29]. Actually, the condition for $cov(d)$ being 1-step invariant can be relaxed to τ-step ($\tau > 1$) invariant:

$$cov(d) = \lambda^{\tau} cov\left(\frac{d}{(2-p)^{\tau}}\right). \tag{3.6}$$

Since we only deal with positive, real values here, as the function $cov(d) \sim 1/d^{\beta}$ is inserted to the equation, the expression for β is the same. As τ goes to infinity, the same result holds.

Note that our argument about power-law covariance function being maintained by some rule with a linear expansion of distance does not really show *how* the power-law covariance function is *generated, before being maintained.* The latter might be called an "operational" proof.

When the expansion-modification system was first studied, we had thought about many potential applications of the model, ranging from particle showers in high-energy experiments, the expansion of the universe, grammar of natural languages, to composition of musical notes. But it occurred to one of us (KK) that no other examples were more compelling than the evolution of DNA sequences. The expectation that expansion-modification system could be relevant to DNA sequences was thus conceived before we actually analyzed and eventually observed cases of long-range correlation in DNA sequences.

Interestingly, on one hand, we have the observation that the exponent α in $1/f^{\alpha}$ spectral component differs from one Genbank category to another [56], and on the other hand, the exponent α in the $1/f^{\alpha}$ power spectrum generated by the expansion-modification system depends on the mutation rate p [27,29]. If we believe that DNA sequences are updated according to some rules similar to the expansion-modification system, then the variation of the exponent α in DNA sequences might reflect a difference of degree of point mutation relative to point duplications.

3.2. Duplication events in DNA evolution

We use the phrase "DNA duplication" in a general way here to describe the general situation when segments of DNA get copied and inserted elsewhere in the genome. This can include situations when DNA is duplicated to adjacent segments, or otherwise. The result in either situation is the development of tandem (adjacent) or dispersed repeating segments. Because of the underlying mechanisms involved, the duplicated segments can appear as not exact copies.

The expansion-modification system described in Eq. (3.1) represents one special case of duplication, that involves single point duplication that leads to a single tandem repeat per duplication

event. Similar tandem repeats are though to be common during DNA replication where the replication machinery causes segments to be duplicated through unequal pairing during mitosis. Clearly, our model is simpler than the real situation and we make no claim about the biological reality of this model.

Besides the duplication processes discussed above, there are many other sources of variation generation in DNA sequences. They fall into two broad categories, first there are the point mutations, such as point deletions and point insertions, and secondly there are the chromosomal level variations such as recombination, translocations, and deletions, for example. Chromosomal level (meaning essentially any relatively large segment) variation processes are thought to be major sources for generation of raw material for evolution to work on, such as in the shuffling of modular elements described in [11].

The duplication of relatively large segments of DNA provides interesting sources of variation for evolution. For example, entire genes are known to duplicate and insert elsewhere in the genome, leaving the original copy intact and operational, while the duplicated copy can be subject to further modification and evolutionary selection. We will not address the level of organization at which evolution operates here, but rather assume that a mechanism exists that can ultimately influence the evolutionary potential of duplicated segments. Perhaps no one more advocates the role of DNA duplication in evolution than S. Ohno [39]. The principle argument here is that when entire genes or otherwise functional units are duplicated, the working original is left alone, while the duplicated copy can be subject to a different set of evolutionary dynamics.

While genome length per se does not scale with complexity of the organism, it is true that longer genomes on an absolute scale can contain a larger collection of distinct elements than smaller ones. Increased genome length puts additional constraints on the existence of certain functional elements, such as origins of replication, since replication during the cell cycle is time-constrained, so that more origins are needed to replicate the longer segments during a fixed period of time. There may be other similar constraints that exist that are necessary for higher-order genome structure that are related to linear dimension. The type of analysis we have described here may provide insight into what the dynamics and constraints are in DNA organization. Clearly, there is much room for this type of analysis as the international human genome program provides biology the opportunity to ask such fundamental questions regarding DNA structure, organization, and dynamics.

Acknowledgements

WL would like to thank Chung-Kang Peng, Michael Zhang, Roderic Guigó for communicating results before publication, and Jim Fickett for providing valuable comments to the first draft of the paper. We also thank the referee who made numerous suggestions on improvement of the manuscript. The participation of WL to the Oji International Symposium on Complex Systems was made possible by a financial support from the Fujihara Foundation and Japanese Society for Promotion of Science. The work of WL and TGM at Cold Spring Harbor Laboratory is supported by the grant from DOE (DE-FG02-91ER61190). TGM is also supported by NIH grant 2-R01-HE0020304.

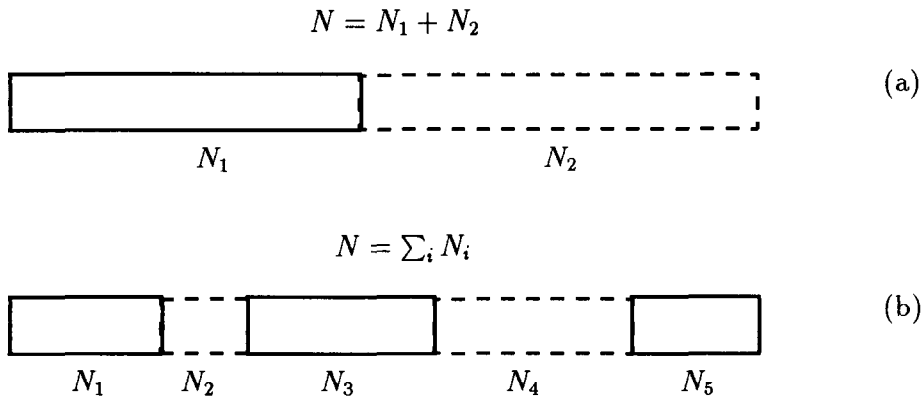

Fig. A.1. Illustration for the calculation in the Appendix: (a) two-subregion case; (b) multi-subregion case.

Appendix A. Calculation of covariance functions for sequences decomposable to many white-noise subregions

Two subregions case (same-symbol correlation). Fig. A.1a illustrates this situation. The overall sequence length is N, and the length for two subregions are N_1 and N_2. The ratio of the length of two subregions relative to the whole is $f_1 \equiv N_1/N$ and $f_2 \equiv N_2/N$.

Suppose the count of nucleotide type α in two subregions are $n_{\alpha(1)}$ and $n_{\alpha(2)}$, respectively, we have

$$\widehat{P_\alpha} = \frac{n_\alpha}{N} = \frac{n_{\alpha(1)} + n_{\alpha(2)}}{N} = \frac{N_1}{N}\widehat{P_{\alpha(1)}} + \frac{N_2}{N}\widehat{P_{\alpha(2)}} = f_1\widehat{P_{\alpha(1)}} + f_2\widehat{P_{\alpha(2)}}. \tag{A.1}$$

A similar decomposition of the joint probability is

$$\begin{aligned}\widehat{P_{\alpha\beta}(d)} &= \frac{n_{\alpha\beta(1)}(d) + n_{\alpha\beta(2)}(d) + n_{\alpha\beta(12)}(d)}{N} \\ &= f_1\widehat{P_{\alpha\beta(1)}(d)} + f_2\widehat{P_{\alpha\beta(2)}(d)} + \frac{n_{\alpha\beta(12)}(d)}{N},\end{aligned} \tag{A.2}$$

where $n_{\alpha\beta(12)}(d)$ is the number of cross-subregion counts of the (α, β) base-pair out of total d counts. It is usually negligible as compared with other two terms if $d << N_1, N_2, N$.

The differences of base composition between two subregions are:

$$\Delta P_\alpha \equiv P_{\alpha(1)} - P_{\alpha(2)} \neq 0, \quad \Delta P_\beta \equiv P_{\beta(1)} - P_{\beta(2)} \neq 0. \tag{A.3}$$

It can be shown that the covariance function of the whole sequence as defined by Eq. (2.5) is related to the covariance functions of the two subregions in the following way:

$$\begin{aligned}\widehat{cov(d)} &= \sum_{\alpha=(G,C,T,A)} \left(\widehat{P_{\alpha\alpha}(d)} - \widehat{P_\alpha}^2\right) \\ &= f_1\widehat{cov_1(d)} + f_2\widehat{cov_2(d)} + f_1f_2 \sum_{\alpha=(G,C,T,A)} (\widehat{\Delta P_\alpha})^2 + \sum_{\alpha=(G,C,T,A)} \frac{n_{\alpha\alpha(12)}(d)}{N} \\ &\approx f_1\widehat{cov_1(d)} + f_2\widehat{cov_2(d)} + f_1f_2 \sum_{\alpha=(G,C,T,A)} (\widehat{\Delta P_\alpha})^2 \qquad (\text{if } d << N_1, N_2, N)\end{aligned}$$

$$\approx f_1 f_2 \sum_{\alpha=(G,C,T,A)} (\widehat{\Delta P_\alpha})^2 \text{ (if both subregions are white noise and } d > 0). \tag{A.4}$$

The above formula shows that the difference of base composition alone contributes a *constant* term to the overall covariance or autocorrelation function. The d-dependent term of $cov(d)$ is of the order of $o(d/N)$ if d is smaller than N_1, N_2:

$$\sum_{\alpha=(G,C,T,A)} \frac{n_{\alpha\alpha(12)}(d)}{N} \approx \sum_{\alpha=(G,C,T,A)} \frac{d}{N} \widehat{P_{\alpha(1)}} \widehat{P_{\alpha(2)}}. \tag{A.5}$$

To summarize, the overall correlation structure of the sequence, if it is decomposable to two subregions with white noise but different base compositions, is very simple: it is a constant plus a small linear term. This correlation function does not share similar features with those of complex, multi-length-scaled sequences with $1/f^\alpha$ spectral component.

Two subregions case (different-symbol correlation). One can also easily derive the correlation between bases of different type ($\alpha \neq \beta$):

$$\begin{aligned} \widehat{cov_{\alpha\beta}}(d) &= \widehat{P_{\alpha\beta}}(d) - \widehat{P_\alpha}\widehat{P_\beta} \\ &\approx f_1 \widehat{cov_{\alpha\beta(1)}}(d) + f_2 \widehat{cov_{\alpha\beta(2)}}(d) + f_1 f_2 \widehat{\Delta P_\alpha}\widehat{\Delta P_\beta} \qquad (\text{if } d << N_1, N_2, N) \\ &\approx f_1 f_2 \widehat{\Delta P_\alpha}\widehat{\Delta P_\beta} \qquad (\text{if both subregions are white noise and } d > 0) \\ &= -f_1 f_2 (\widehat{\Delta P_\alpha})^2 \qquad (\text{if the sequence is binary}) \end{aligned} \tag{A.6}$$

Interestingly, the observation in the budding yeast chromosome 3 sequence that $\widehat{I_{\alpha\alpha}(d)}$'s tend to be larger than 1, while $\widehat{I_{\alpha\beta}(d)}$'s ($\beta \neq \alpha$) tend to be smaller than 1 [60] can be compared with our result Eq. (A.4) and Eq. (A.6) that covariance functions between same base type acquire a positive contribution from the base composition difference, whereas those between different base types acquire a negative contribution.

Many subregions. All the above results can be generalized to $K > 2$ different subregions, each of them is a white noise. Fig. A.1(b) illustrates the situation, with the overall sequence length as N, the length of i'th subregion as N_i, and $f_i \equiv N_i/N$. For example, we can show that

$$\begin{aligned} \widehat{cov}(d) &= \sum_{\alpha=(G,C,T,A)} \left[\sum_{i=1}^{K} f_i \widehat{P_{\alpha\alpha(i)}}(d) + \sum_{i=1}^{K-1} \frac{n_{\alpha\alpha(i,i+1)}(d)}{N} - \left(\sum_{i=1}^{K} f_i \widehat{P_{\alpha(i)}} \right) \left(\sum_{j=1}^{K} f_j \widehat{P_{\alpha(j)}} \right) \right] \\ &= \sum_{i=1}^{K} f_i \widehat{cov_i}(d) + \sum_{\alpha=(G,C,T,A)} \sum_{i=1}^{K} \sum_{j>i} f_i f_j (\widehat{\Delta P_{\alpha(ij)}})^2 + \sum_{\alpha=(G,C,T,A)} \sum_{i=1}^{K-1} \frac{n_{\alpha\alpha(i,i+1)}(d)}{N} \\ &\approx \sum_{i=1}^{K} f_i \widehat{cov_i}(d) + \sum_{\alpha=(G,C,T,A)} \sum_{i=1}^{K} \sum_{j>i} f_i f_j (\widehat{\Delta P_{\alpha(ij)}})^2 \qquad (\text{if } d << N_i, N) \\ &\approx \sum_{\alpha=(G,C,T,A)} \sum_{i=1}^{K} \sum_{j>i} f_i f_j (\widehat{\Delta P_{\alpha(ij)}})^2 \quad (\text{if all subregions are white noise and } d > 0) \end{aligned} \tag{A.7}$$

where $n_{\alpha\alpha(i,i+1)}$ is the count of (α,α) base-pairs crossing the i'th and the $(i+1)$'th subregions.

References

[1] Y. Aizawa, I. Nishikawa and K. Kaneko, Soliton turbulence in one-dimensional cellular automata, Physica D 45 (1990) 307–327.
[2] S.F. Altschul, Amino acid substitution matrices from an information theoretic perspective, J. Mol. Biol. 219 (1991) 555–565.
[3] G. Bernardi, The isochore organization of the human genome, Ann. Rev. Gen. 23 (1989) 637–661.
[4] B. Borštnik, D. Pumpernik and D. Lukman, Analysis of apparent $1/f^{\alpha}$ spectrum in DNA sequences, Europhys. Lett. 23 (1993) 389–394.
[5] S.V. Buldyrev, A.L. Goldberger, S. Havlin, C-K. Peng, M. Simon, F. Sciortino and H.E. Stanley (comment), Phys. Rev. Lett. 71 (1993) 1776.
[6] C. Burks and D. Farmer, Towards modeling DNA sequences as automata, Physica D 10 (1984) 157–167.
[7] C.A. Chatzidimitriou-Dreismann and D. Larhammar, (Scientific Correspondence), Nature 361 (1993) 212–213.
[8] M. Chee et al., Analysis of the protein coding content of human cytomegalovirus strain AD169, Current Top. Microbiol. Immunol. 154 (1990) 125–169.
[9] M.O. Dayhoff, R.M. Schwartz and B.C. Oreutt, A model of evolutionary change in proteins, in: Atlas of Protein Sequence and Structure, M.O. Dayhoff, ed., vol. 5, suppl. 3 (National Biomedical Research Foundation, 1978).
[10] Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences, R.F. Doolittle, ed. Methods in Enzymology 183 (Academic Press, New York, 1990).
[11] R.F. Doolittle and P. Bork, Evolutionarily mobile modules in proteins, Scientific American (Oct 1993) 50–56.
[12] R. Farber, A. Lapedes and K. Sirotkin, Determination of eukaryotic protein coding region using neural networks and information theory, J. Mol. Biol. 226 (1992) 471–479.
[13] J.W. Fickett, D.C. Torney and D.R. Wolf, Base compositional structure of genomes, Genomics 13 (1992) 1056–1064.
[14] J.W. Fickett and C-S. Tung, Assessment of protein coding measures, Nucleic Acids Research 20 (1992) 6441–6450.
[15] J.W. Fickett and R. Guigó, Estimation of protein coding density in a corpus of DNA sequence data, Nucleic Acids Research 21 (1993) 2837–2844.
[16] A.M. Fraser, Information and entropy in strange attractors, Ph.D Thesis (Univ. of Texas, Austin, 1988);
A.M. Fraser and H.L. Swinney, Independent coordinates for strange attractors from mutual information, Phys. Rev. A 33 (1986) 1134–1140;
J.A. Vastano and H.L. Swinney, Information transport in spatio-temporal systems, Phys. Rev. Lett. 60 (1988) 1773–1776.
[17] C.W.J. Granger and M. Hatanaka, Spectral Analysis of Economic Time Series (Princeton Univ. Press, Princeton, NJ, 1964).
[18] H. Gutowitz, ed., Cellular Automata: Theory and Experiment, Physica D 45 (1990); (North-Holland, Amsterdam, 1990); (MIT Press, Cambridge, MA, 1990).
[19] H.-P. Herzel and W. Ebeling, The decay of correlations in chaotic maps, Phys. Lett. 111 (1985) 1–4.
[20] H. Herzel, Complexity of symbol sequences, Syst. Anal. Model. Simul. 5 (1988) 435–444.
[21] K. Kaneko, Lyapunov analysis and information flow in coupled map lattices, Physica D 23 (1986) 436–447.
[22] S. Karlin and V. Brendel, Patchiness and correlations in DNA sequences, Science 259 (1993) 677–680.
[23] B.T.M. Korber, R.M. Farber, D.H. Wolpert and A.S. Lapedes, Covariation of mutation in the V3 loop of HIV-1: an information theoretic analysis, Proc. National Academy of Science (USA) 90 (1993) 7176–7180.
[24] C.G. Langton, Computation at the edge of chaos, Physica D 42 (1990) 12–37;
C.G. Langton, Life at the edge of chaos, in: Artificial Life II, C.G. Langton, C. Taylor, J.D. Farmer and S. Rasmussen, eds. (Addison-Wesley, Reading, MA, 1992).
[25] D. Larhammar and C.A. Chatzidimitriou-Dreismann, Biological origins of long-range correlations and compositional variations in DNA, Nucleic Acids Research 21 (1993) 5167–5170.
[26] W. Li, Power spectra of regular languages and cellular automata, Complex Systems 1 (1987) 107–130; (errata) 2 (1989) 725.
[27] W. Li, Spatial 1/f spectra in open dynamical systems, Europhys. Lett. 10 (1989) 395–400.
[28] W. Li, Mutual information function versus correlation functions, J. Statist. Phys. 60 (1990) 823–837.
[29] W. Li, Expansion-modification systems: a model for spatial 1/f spectra, Phys. Rev. A 43 (1991) 5240–5260.
[30] W. Li, Generating nontrivial long-range correlations and 1/f spectra by replication and mutation, Int. J. Bifurcation and Chaos 2 (1992) 137–154.
[31] W. Li and K. Kaneko, Long-range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence, Europhys. Lett. 17 (1992) 655–660.

[32] W. Li and K. Kaneko (Scientific Correspondence), Nature 360 (1992) 635–636.
[33] W. Li, N.H. Packard, C.G. Langton, Transition phenomena in cellular automata rule space, Physica D 45 (1990) 77–94.
[34] T.P. Meyer, Long Range Predictability of High Dimensional Chaotic Dynamics, Ph.D Thesis (Univ. of Illinois, Urbana-Champaign, 1992);
T.P. Meyer and N.H. Packard, Local forecasting of high-dimensional chaotic dynamics, in: Nonlinear Modeling and Forecasting, M. Casdagli and S. Eubank, eds. (Addison-Wesley, Reading, MA, 1991).
[35] P.J. Munson, R.C. Taylor and G.S. Michaels (Scientific Correspondence), Nature 360 (1992) 636.
[36] S. Nee (Scientific Correspondence), Nature 357 (1992) 450.
[37] P. O'Hara et al., Nucleotide sequence of the gene coding for human factor VII, a vitamin K-dependent protein participating in blood coagulation, Proc. of National Academy of Sciences 84 (1987) 5158–5162.
[38] S.G. Oliver et al. The complete DNA sequence of yeast chromosome III, Nature 357 (1992) 38–46.
[39] S. Ohno, Evolution by Gene Duplication (Springer, Berlin, 1970).
[40] N.H. Packard, Complexity of growing patterns in cellular automata, in: Dynamical Systems and Cellular Automata, J. Demongeot, E. Goles and M. Techuente, eds. (Academic Press, New York, 1985).
[41] C-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simon and H.E. Stanley, Long-range correlations in nucleotide sequences, Nature 356 (1992) 168–170.
[42] C-K. Peng, S.V. Buldyrev, S. Havlin, M. Simon, H.E. Stanley and A.L. Goldberger, Mosaic organization of DNA nucleotides, Phys. Rev. E 49 (1994) 1685–1689.
[43] D.B. Percival and A.T. Walden, Spectral Analysis for Physical Applications (Cambridge Univ. Press, Cambridge, 1993).
[44] V.V. Prabhu and J.-M. Claverie (Scientific Correspondence), Nature 359 (1992) 782.
[45] W.H. Press, S.A. Teukolsky, W.T. Vetterling and B.R. Flannery, Numerical Recipes in C: the Art of Scientific Computing , 2nd edition (Cambridge Univ. Press, Cambridge, 1992).
[46] F.C. Richards, Learning Two-Dimensional Spatial Dynamics from Experimental Data, Ph.D Thesis (Univ. of Illinois, Urbana-Champaign, 1991);
T.P. Meyer, F.C. Richards and N.H. Packard, Learning algorithm for modeling complex spatial dynamics, Phys. Rev. Lett. 63 (1989) 1735–1738;
F.C. Richards, T.P. Meyer and N.H. Packard, Extract cellular automaton rules directly from experimental data, Physica D 45 (1990) 189–202.
[47] S. Sakamoto, M. Ishiguro and G. Kitagawa, Akaike Information Criterion Statistics (Reidel, Dordrecht, Holland).
[48] F. Sanger et al., Nucleotide sequence of bacteriophage λ DNA, J. Mol. Biol. 162 (1982) 729–773.
[49] C.E. Shannon and W. Weaver, The Mathematical Theory of Communication (Univ. of Illinois Press, Champaign, IL, 1949).
[50] P.M. Sharp and A.T. Lloyd, Regional base composition variation along yeast chromosome III: evolution of chromosome primary structure, Nucleic Acids Research 21 (1993) 179–183.
[51] R. Shaw, The Dripping Faucet as a Model Chaotic System (Aerial Press, 1984).
[52] B. Silverman and R. Linsker, A measure of DNA periodicity, J. Theor. Biol. 118 (1986) 295–300.
[53] D.J. States and D. Botstein, Molecular sequence accuracy and the analysis of protein coding regions, Proc. National Academy of Science (USA) 88 (1991) 5518–5522.
[54] T. Toffoli and N. Margolus, Cellular Automata Machine – A New Environment for Modeling (MIT Press, Cambridge, MA, 1987).
[55] A. Van de Ziel, On the noise spectra of semi-conductor noise and of flicker effect, Physica 16 (1950) 359–372.
[56] R. Voss, Evolution of long-range fractal correlations and 1/f noise in DNA base sequences, Phys. Rev. Lett. 68 (1992) 3805–3808.
[57] R. Voss (reply to the comment), Phys. Rev. Lett. 71 (1993) 1777.
[58] R.F. Voss and J. Clarke, 1/f noise in music and speech, Nature 258 (1975) 317–318.
[59] S. Wolfram, ed., Theory and Applications of Cellular Automata (World Scientific, Singapore, 1986).
[60] M. Zhang and T. Marr, Large-scale structure of yeast chromosome III, J. Comput. Biol. (1994), to appear.