Generating Non-trivial Long-Range Correlations and 1/f Spectra by Replication and Mutation

Wentian Li

Santa Fe Institute, 1660 Old Pecos Trail, Suite A, Santa Fe, NM 87501 [*]

Received May 29, 1991
published in International Journal of Bifurcation and Chaos, 2, 137-154 (1992)

Abstract

This paper aims at understanding the statistical features of nucleic acid sequences from the knowledge of the dynamical process that produces them. Two studies are carried out: first, mutual information function of the limiting sequences generated by simple sequence manipulation dynamics with replications and mutations are calculated numerically (sometimes analytically). It is shown that elongation and replication can easily produce long-range correlations. These long range correlations could be destroyed in various degrees by mutation in different sequence manipulation models. Second, mutual information functions for several human nucleic acid sequences are determined. It is observed that intron sequences (non-coding sequences) tend to have longer correlation lengths than exon sequences (protein-coding sequences).

[*] Present address: Box 167, Rockefeller University, 1230 York Avenue, New York, NY 10021.

1 Introduction

Ever since terrestrial genesis, the molecules which are capable of replication have been playing an essential role in life on earth. The replicators are basically nucleic acid sequences — 1-dimensional strands consist of nucleotide bases. An arrangement of nucleotide bases on a nucleic acid sequence is transformed into an arrangement of amino acid in the protein, which in turn determines the 3-dimensional structure of the protein, and consequently, many aspects of the biochemical reactions in biological systems. The arrangement of the nucleotide bases on nucleic acid sequences results from more than three billion years of evolution (see, for example, [Watson *et. al.*, 1987] [Horgan, 1991]).

Now, we ask the following question: can we understand why the nucleic acid sequences have the arrangement of bases observed today? Or, can we understand the statistical features of these nucleic acid sequences from some models of evolution? The question is similar to what has been asked in cosmology on whether one can explain the galaxy distribution from the known physics laws (e.g., gravitational interaction), evolutionary scenarios (e.g., expansion of the universe from the big bang), and a set of simple assumptions (e.g., the initial condition of the universe). In cosmology, it is a simple matter of setting up the model and the initial condition, running the simulation on computer, and comparing the results with the observation data.

The research on the evolution of life is far behind that on the evolution of the universe. There are several reasons for this. First, we do not have complete knowledge of the arrangement of the nucleotide bases of nucleic acid sequences. There are, however, great efforts towards improving the situation, notably the human genome project [Watson, 1990]. Secondly, there is no simple universal force, such as the gravitational interaction in the evolution of the universe, that controls all aspects of the evolution of nucleic acid sequences. Thirdly, we still know very little about how life started; that is, we do not have a good guess of the initial condition.

This paper attempts to make a very small contribution towards an ultimate answer of the question. At one end of the matching between models and reality, I will study a few simple sequence manipulation rules with only replications and mutations. Similar to the dynamical systems with spatial degrees of freedom, where the randomness of the spatial configuration can sometimes be related to how chaotic the dynamical rule is, the statistical properties of the sequences generated by these simple rules are also crucially determined by the structure of the rule, the parameter setting, and occasionally, the initial condition.

At another end of the matching, I will calculate the mutual information function [Shannon, 1948] [Shannon & Weaver, 1949] [Li, 1990], one of the most important statistical quantities of the sequence, of several nucleic acid sequences. Not completely surprising, the mutual information function of a nucleic acid sequence has been found to depend on whether the sequence is a protein-coding (exon) or a non-coding (intron) segment. It certainly hints that the dynamical process which controls the updating of intron segments differs from that of exon segments.

The intention of the paper is not to claim that the models studied here can reproduce the statistical properties of the current nucleic acid sequences. As the title suggests, the goal is quite limited: I will examine the long-range correlations in nucleic acid sequences as produced by elongation, or replication followed by a ligation. The presence or the absence, for that matter, of long-range correlations then teaches us something about the dynamical process itself.

This paper is organized as follows: Section 2 will review the main statistical quantity to be used in the paper — the mutual information function. The related definitions such as power spectrum, 1/f spectrum, long-range correlation, and non-trivial long-range correlation will also be given for easy reference; Section 3 will review some known results on the relation between structure of the sequence manipulation rules and statistical properties of the generated sequences; Section 4 will discuss four sequence manipulation rules with only replications (or elongations) and mutations; Section 5 will present the results on mutual information functions of several human nucleic acid sequences; Section 6 studies the $1/f$ spectrum in one of the intron sequences; and finally, section 7 contains discussions and possible future research directions.

2 Mutual information function: measure of correlation in symbolic sequences

It is not clear what statistical property is most appropriate for characterizing a nucleic acid sequence, and, by comparing this property of a nucleic acid sequence with the one derived from the theory, for checking the validity of the theory. Some statistical quantities are too specialized for our purpose, for example, the CG dimer (cytosine and guanine) density. One can imagine many different ways to modify the model to make a CG rich sequence, and we simply cannot discriminate among these models by knowing this density only.

The single-site entropy can give much information on whether all symbols are equally used in a sequence, but it does not say how symbols are arranged in the sequence. A better quantity is the block entropy, which measures the degree of equal distribution of all blocks with a fixed length [Shannon, 1948]. If block entropies are determined for all block sizes, the statistical feature of the sequence is rather completely determined. Unfortunately, the block entropy cannot be calculated accurately for very large block lengths when the sequence length itself is limited. In previous studies of the entropy of natural language texts (ranging from English [Shannon, 1951] [Cover & King, 1978] to Arabic [Wanes *et. al.*, 1976]) and nucleic acid sequences [Gatlin, 1966, 1968, 1972] [Smith, 1969], the block length has never gone up to a very large value.

A better quantity to statistically characterize the arrangement of the nucleotide bases in the sequence is the *correlation function*, defined as the *correlation* between two bases as a function of the distance between them. There are several ways to measure the correlation of two variables, for example, the average value of the product of the two variables subtracting the product of the average value of each variable. If this definition of correlation is used, we have the conventional *autocorrelation function*:

$$, (d) = \sum_{\alpha\beta} x_\alpha x_\beta P_{\alpha\beta}(d) - (\sum_{\alpha} x_\alpha P_\alpha)^2, \tag{1}$$

where x_α's and x_β's are all possible values of the variable, P_α is the density of the symbol α with value x_α, and $P_{\alpha\beta}(d)$ is the probability of having a symbol with the value x_α followed d sites away by a symbol with the value x_β. The autocorrelation function is widely used in the correlation analysis in numerical sequences.

If the sequence is purely symbolic, there is no value attached to each symbol, and we measure the correlation by *mutual information* [Shannon, 1948], and the correlation function becomes the *mutual information function* [Li, 1990]:

$$M(d) = \sum_{\alpha\beta} P_{\alpha\beta}(d) \log \frac{P_{\alpha\beta}(d)}{P_\alpha P_\beta}. \tag{2}$$

Zero $M(d)$ at some distance d implies zero $, (d)$ at that distance, but the reverse may not be true. As a consequence, mutual information function is a more sensitive measure of correlation than the autocorrelation function. Note that because each of the $\log[P_{\alpha\beta}(d)/P_\alpha P_\beta]$ term in the summation is weighted by a $P_{\alpha\beta}(d)$, in certain approximations, the $M(d)$ behaves like $P_{\alpha\beta}(d)^2 - \mathit{const.}$, whereas $, (d) \sim P_{\alpha\beta}(d) - \mathit{const.}$. For more discussions on the relation between the mutual information function and the autocorrelation function, see [Li, 1990].

There might be other definitions of the correlation function such as the one based on "Chi-square" (e.g., Chapter 13 of [Press *et. al.*, 1988]). To avoid confusion, I will use $C(d)$ to represent any one of them, i.e.,

$$C(d) = \{, (d), M(d), \cdots\}. \tag{3}$$

With correlation functions such as the mutual information function being defined, we can quantitatively define terms such as long-range correlation, non-trivial long-range correlation, correlation length, 1/f spectrum (1/f noise), etc.

First of all, the sequences with long-range correlations are those whose $C(d)$'s decay very slowly and remain at non-zero values at some large distances. It includes the case of the periodic sequences, whose $C(d)$'s have peaks when the distances are equal to the multiples of the periodicity. We do not need exact periodicity for having this long-range correlation; an approximate periodicity is good enough. Because there is nothing profound about the periodic structures, we consider these long-range correlations to be trivial.

Sometimes, the $C(d)$ is non-zero at almost all larger d's, with no dominant peaks in $C(d)$. Such slow decay of the $C(d)$ can usually be approximated by power law functions (algebraic decay) — $1/d^{\alpha}$. In particular, if the decay is so slow that the exponent α is close to zero (when α is exactly equal to zero, $C(d)$ decays slower than any power law functions; they are, for example, logarithmic functions), the sequence can be called $1/f$ noise because its power spectrum behaves like $1/f$, where f is the frequency. $1/f$ spectrum will be mentioned again in section 4 and section 6.

If the decay of $C(d)$ is fast, it can be approximated by an exponential function: e^{-d/d_0}, where d_0 is called the *correlation length* because the correlation value becomes very small as $d > d_0$. In a numerical calculation of $C(d)$, one might observe that $C(d)$ is almost zero beyond certain distance d_0', and this distance is sometimes used as an estimate of the correlation length d_0.

In the next section, I will review the results concerning which dynamical systems typically produce sequences with exponential decay correlation functions, and which produce sequences with algebraic correlation functions. These results provide a potential method for inferring the underlying dynamical process from the observed data sequences.

3 Different dynamical models generate sequences with different correlation functions

In cosmology, it is known that statistical features of galaxy distribution such as the power law two-point correlation function are the results of the expansion of space, the long-ranged gravitational interaction, and the initial stages of the big bang which determined the starting configuration. Varying either one of the conditions, one may not reproduce the statistics in the observational data. Similarly, for dynamical systems applied to 1-dimensional sequences, it is important to know the dynamical rule (how the symbols in the sequence are updated), the initial condition, and whether or not the sequence length is changed, in order to determine the statistical properties of the sequence.

In the following, I will review three types of sequence manipulation rules: (1) those putting symbols sequentially with short memories; (2) those updating sequences parallelly according to local dynamics with the sequence length fixed; and (3) those updating parallelly according to local dynamics with the sequence length increased. Obviously, these three types represent only a small portion of all possible sequence manipulation rules. Other sequence manipulation rules will not be discussed in detail in this paper, since I cannot provide general conclusions concerning the statistical properties of the limiting sequences, except for simulating the rule on case-by-case bases. There are, however, some discussions in the next section on rules which link the copied sequence with the original sequence (then the dynamics is not local), and a passing mentioning in the last section on sequence manipulation rules with high-level control, and the dynamics of a population of sequences.

(1) The first type of sequence manipulation rules is actually "sequence-producing rules." The symbols are added one by one at the end of the sequence with the rule having a short memory of what has already been in the sequence. This class includes the well-known Markov chains [Karlin, 1968] [Karlin & Taylor, 1981] and regular languages [Hopcroft & Ullman, 1979]. In the 1-step Markov chains, the probability of having a particular new symbol in the end of the sequence depends only on the last symbol already in the sequence. All such probabilities are included in the Markov transition matrix, and the correlation function $C(d)$ behaves like $\lambda^d = e^{-\log(1/\lambda)d}$, where λ is the largest eigenvalue (excluding the trivial eigenvalue equal to 1) of the Markov transition matrix [Karlin & Taylor, 1981].

Regular languages, studied in the framework of the formal language theory [Hopcroft & Ullman, 1979], are very similar to Markov chains. The difference between the two is that in regular languages, the probability of a symbol to be followed by another symbol depends on the "history", which can be determined by checking the grammar of the regular language, usually represented by a directed graph. A regular language can become a Markov chain by increasing the number of symbols — so that the same symbol with different histories is considered as different symbols, or by increasing the memory — so that a finite block rather a symbol determines the probability of having a new symbol. The calculation of $C(d)$ for sequences generated by regular language grammars is more complicated (one has to increase the number of the symbols and make the transition matrix larger, then degenerate these symbols again; see [Li, 1987] for details), but again $C(d)$ behaves like λ^d, with λ as the largest eigenvalue (excluding the value of 1) of the expanded transition matrix.

Formally speaking, Markov chains and regular languages always produce sequences with exponentially decayed $C(d)$. Nevertheless, if the largest non-trivial eigenvalue of the transition matrix is very close to 1, the correlation length $d_0 \approx 1/\log(1/\lambda)$ can be extremely long, and by the Taylor expansion of the exponential function, $C(d)$ can decay as a linear function. The power spectrum corresponding to this linear $C(d)$ behaves like $1/f^2$ (see, e.g., appendix of [Li, 1991b]). Also note that when the largest non-trivial eigenvalue is negative (largest in magnitude), $C(d)$ oscillatory.

(2) The second type of the sequence manipulation rules can be considered as one of the spatially extended dynamical systems, which include coupled maps, coupled oscillators, and for an example of the real system, the turbulence flow. One starts from an initial sequence whose length is fixed during the dynamics, and updates the sequence by local rules. The best example of this type of rules is the cellular automata [von Neumann, 1966; Wolfram, 1983; Toffoli & Margolus, 1987]. For each symbol in the sequence, by examining the local configuration around that site and checking the rule table which tells what new symbol will replace the old one according to the local configuration, one can update all symbols in the sequence one by one.

The statistical properties of the limiting sequence depend on what the initial sequence is, and which cellular automaton rule is applied. Suppose the initial sequence is random with no correlations, the only thing that determines the statistical properties of the limiting sequence is the rule table. The connection between the rule and the correlation function of the limiting sequence is studied in [Li, 1987]. In particular, it is known that if the dynamics is periodic (i.e., the sequence repeats itself, with or without a spatial shift, after a finite number of time steps), the limiting sequence can be characterized by some regular language grammar [Wolfram, 1984], and by our previous discussion, the correlation function is exponential (either monotonic or oscillatory).

Generally speaking, if a cellular automaton rule is capable of generating correlation length much longer than the range of local coupling, that rule will have other interesting properties such as long transient times, marginal instability with respect of perturbations, and poor convergence of most of the statistical quantities. The rule can then be said to be on the "edge of chaos." In fact, the existence of a large value of correlation at long distances is used to locate the region of the cellular automata rule space where the transition from periodic to chaotic dynamics occurs [Li *et. al.*, 1990].

(3) The third type of the sequence manipulation rules contains rules that update symbols according to local dynamics and the sequence length is increased at the same time. One might

call them context-sensitive Lindenmayer systems [Lindenmayer, 1968] or context-sensitive "development systems" [Węgzyn *et. al.* 1990], or perhaps "expanding cellular automata". These systems are rarely discussed from the perspective of the statistical properties of limiting sequence. Even simple context-sensitive Lindenmayer systems contain huge number of possible rules. For example, in 2-symbol 3-input context-sensitive Lindenmayer systems, suppose each symbol will expand to a block with two symbols, the total number of the possible rules is $4^8 = 65536$ (8 possible input configurations and 4 possible expanded blocks). This number is much larger than the number of rules for 2-symbol 3-input cellular automata which is $2^8 = 256$.

A direct consequence of elongation of sequences is that it is quite easy to generate long-range correlations, even if there is no local interaction (context-free)! In the examples to be discussed in the next section, the correlation function of the limiting sequence can be a power law function $1/d^\alpha$, $\alpha \approx \log(\lambda)/\log(k)$, where λ is the largest non-trivial eigenvalue of the transition matrix (to be defined later) and k is the average elongation ratio. This result seems to be applicable to a large class of context-free Lindenmayer systems.

4 Four sequence manipulation rules with replication/elongation and mutation

One plausible picture of the prebiotic evolution is that first, mononucleotides were condensed into short polymers (oligonucleotides), and some of them happened to be able to replicate, making more copies of themselves. Then, the polymerizations, ligations, cleavages and other reactions occur constantly in a population of mononucleotides, oligonucleotides and polynucleotides, and the average sequence length becomes longer and longer. Some much simplified model based on the above picture has been studied, and it has already shown an enormous amount of complexity [Kauffman, 1986] [Farmer *et. al.*, 1986] [Bagley, 1991].

Here I will not attempt to propose a realistic model for the prebiotic evolution for a population of sequences. Instead, I will concentrate on models with only replications and mutations, and assuming that if a symbol does not make a copy of itself, it will mutate. In other words, the probability of having replication $p_{replication}$ is $1 - p_{mutation}$, with $p_{mutation}$ as the probability for mutation. Certainly it is not the best assumption because there should be a probability for neither replication nor mutation, i.e., preservation. The advantage for assuming only two operations is that there is only one parameter to tune.

The four sequence manipulation rules with replication/elongation and mutation are: (1) the monomer replicates and the extra copy is inserted back to the sequence causing local elongation; (2) similar to the first case but specifying that the replication is complementary; (3) the whole sequence replicates and the copy is ligated to the original sequence; and (4) similar to the third case but specifying that the replication is complementary. All the replications are not perfect with a chance of having mutations.

(1) The first model is the following: suppose there are two symbols in the sequence, a and b; at each time step, each symbol can either expand to two same symbols (with probability $1-p$), or mutate to another symbol (with probability p). The expansion part can also be pictured as a symbol replicating an extra copy of itself and then that copy is inserted near its parental symbol. Perhaps elongation is the better word than replication to describe the process. In formula, the model is:

$$a \rightarrow \left\{ \begin{array}{lcl} aa & : & 1-p \\ b & : & p \end{array} \right. ;$$

$$b \rightarrow \left\{ \begin{array}{lcl} bb & : & 1-p \\ a & : & p \end{array} \right. . \tag{4}$$

Fig.1 illustrates a particular realization of the above sequence generation process.

This model is first proposed by the author as a model for spatial $1/f$ spectra in open dynamical systems [Li, 1989a]. More details of the model are discussed in [Li, 1991a]. I will not repeat all the details here, only enough to outline the basic features which are essential to the main theme of this paper.

Eq.(4.4) is a probabilistic context-free Lindenmayer system. Even though there is no interaction among the symbols, i.e., context-free, the rule can still generate long-range correlations purely by elongation. To be more specific, suppose the joint probability for two symbols of type α and type β separated by a distance d is $P_{\alpha\beta}(d)$; $P_{\alpha\beta}(d)^t$ at time t leads to $P_{\alpha'\beta'}(d')^{t+1}$ at time $t+1$ by the updating. We have $d' > d$ because of elongation, and α' and β' can be any two symbols that are different from the type α and type β. The most general expression for the updating of $P_{\alpha\beta}(d)$ is a multi-distance matrix equation:

$$P_{\alpha'\beta'}(d')^{t+1} = \sum_{d} \sum_{\alpha\beta} T(\alpha\beta d \rightarrow \alpha'\beta' d') P_{\alpha\beta}(d)^t, \tag{5}$$

where $T(\alpha\beta d \rightarrow \alpha'\beta' d')$'s comprise the transition matrix (note: the transition matrix in Markov chains characterizes the transition from one symbol to another; here, the transition is from one symbol pair to another symbol pair).

The invariant solution of Eq.(5) $\{P_{\alpha\beta}(d)\}$, or simply $P(d)$, is a self-consistent, multi-scaling function, and each scaling exponent is related to the largest non-trivial eigenvalue for the transition matrix $T(\alpha\beta d \rightarrow \alpha'\beta' d')$ bridging the distances d and d'.

To approximate the multi-scaling function with a single scaling function (or almost single scaling function), assume that on average, the distance d is elongated to the distance kd, where k is the average elongation ratio. For Eq.(4.4), $k = 2 - p$. Furthermore, assume that distances d''s around the distance $(2-p)d$ also contribute to the scaling function. With all these approximations, it can be shown [Li, 1991a] that the joint probability behaves like

$$P(d) \sim \frac{1}{d^c} \quad \text{with} \quad c = \frac{\sum_{d' \approx kd} \log(\lambda(d'))}{\log(k)}, \tag{6}$$

and for Eq.(4.4)

$$c \approx 1 - \frac{\log(2-3p)}{\log(2-p)}. \tag{7}$$

The autocorrelation function is proportional to the joint probability and the mutual information function is roughly proportional to the square of the joint probability, so they all decay as power law functions. When the mutation probability p is very small, $c \approx 0$. It means the correlation function decays extremely slowly. To check this, I plot the mutual information function in Fig.2 (in log-log scale) for sequences generated by Eq.(4.4) at two different mutation rates. The power law decay of $M(d)$ with small exponent is indeed observed.

It is known that if the correlation function is $1/d^c$ $(0 < c < 1)$, the power spectrum which is the Fourier transformation of the correlation function is $1/f^{1-c}$ (f is the frequency). If $c \approx 0$, then $1 - c \approx 1$, and the power spectrum is called $1/f$ spectrum, or $1/f$ noise if the phase spectrum is random. The curious thing about $1/f$ noise is that it appears almost everywhere [Press, 1978; Musha, *et. al.*, 1991]. Our model suggests that it is possible to find spatial $1/f$ spectra in sequences produced by elongation and mutation, which perhaps provides an insight into the result to be presented in section 6.

(2) The second model is similar to the first, except that each symbol replicates a symbol that is complementary to itself (e.g., symbol a makes a copy of symbol b) and then inserts that copy into the sequence. It is also a probabilistic context-free Lindenmayer system, represented by the following:

$$a \quad \rightarrow \quad \left\{ \begin{array}{lcl} ab & : & 1-p \\ b & : & p \end{array} \right. ;$$

$$b \;\rightarrow\; \begin{cases} ba & : \; 1-p \\ a & : \; p \end{cases} . \tag{8}$$

Fig.3 illustrates the sequence generation process.

The statistical properties of the sequences generated by Eq.(4.8) is quite different from those generated by Eq.(4.4). First of all, when $p = 0$ and if the initial seed is a single symbol, Eq.(4.4) generates a homogeneous sequence containing a string of the same symbols, whereas Eq.(4.8) generates an "almost periodic" sequence called Thue-Morse sequence [Thue, 1906] [Morse, 1921] [Cheng, *et. al.*, 1988] [Cheng & Savit, 1990]. Secondly, related to the first difference, the largest non-trivial eigenvalue of the transition matrix (largest in magnitude) for Eq.(4.8) is negative, compared with the positive value for Eq.(4.4). It can be easily argued that this negative eigenvalue will introduce an oscillation term whose wavelength is varying with the distance. Thirdly, in some sense, the order present in the Thue-Morse sequence is more easily destroyed by mutation than that in the homogeneous sequence. The reason is that the order in the Thue-Morse sequence is an almost periodic structure; and once the mutation is introduced, the distance between two almost repeating segments shifts. Fig.4 shows the mutual information function for the sequences generated by Eq.(4.8) at several parameters. Notice that some of the peaks in the mutual information function for the original Thue-Morse sequence ($d = 6, 8, 12, 16, 20, 22, 24, 26, 34, 36, \ldots$) remain when the mutation rate is $p = 0.05$ (i.e., $d = 6, 8, 12, 16, 22, 24$), but not the peaks at longer distances (i.e., $d = 26, 34, 36, \ldots$).

(3) The third model considers the case when the sequence replicates an imperfect copy of itself, then ligates the copy sequence with the original one. The replication is direct (e.g., a copies another a), and there is a probability for mutation. The rule is:

$$\cdots a \cdots \;\rightarrow\; \begin{cases} \cdots a \cdots a & : \; 1-p \\ \cdots b \cdots & : \; p \end{cases} ;$$

$$\cdots b \cdots \;\rightarrow\; \begin{cases} \cdots b \cdots b & : \; 1-p \\ \cdots a \cdots & : \; p. \end{cases} . \tag{9}$$

Fig.5 illustrates the sequence generation process.

This type of sequence manipulation rules can easily create long-range correlation, and the range of the correlation becomes longer and longer as the sequence length becomes longer. This feature makes the rule not fit to be described by Lindenmayer systems, either context-free or context-sensitive, because the rule is highly non-local. The longest range of correlation is always comparable with the sequence length. In fact, the sequence is not stationary by the standard definition, and the concept of correlation function should only be used with care.

Multiple copies of the same segment or the same gene in one single nucleic acid sequence is quite common [Britten & Kohne, 1968, 1970] [Long & Dawid, 1980] [Jelinek & Schmid, 1982]. It is also suggested that oligomeric repeat could be an early mechanism for the nucleic acid sequences to explore possible coding schemes [Oono, 1987]. Considering these facts, this type of models needs more attention and theoretical investigations.

The sequence generated by Eq.(4.9) starting from a single seed is very boring, with almost no structure. Instead, I will simulate a case when the starting segment is *abb* with length three. The mutual information function of the limiting sequences is shown in Fig.6 with two different mutation rates. As mentioned above, there are correlations at lengths comparable to the sequence length itself, whereas the maximum distance shown in Fig.6 is 100, so not all structures in the sequence are shown in the figure. From the plot (at the mutation rate $p = 0.01$), one can see that the peaks supposedly at the multiples of three suffer a shift after $d = 18$. The subtle structure in the sequence produced with zero mutation rate are quickly destroyed by the larger mutation probabilities.

The lack of the scaling in the limiting sequence is due to the lack of the scaling in the equation describing the updating of $P_{\alpha\beta}(d)$. Roughly speaking, the equation updating $P_{\alpha\beta}(d)$ is like

$$P_{\alpha\beta}(d')^{t+1} = 2\sum_{d\approx d'} \lambda(d) P_{\alpha\beta}(d)^t + \sum_{d\approx N-d'} \lambda(d) P_{\beta\alpha}(d)^t, \tag{10}$$

where $\lambda(d)$ is the largest non-trivial eigenvalue of the corresponding transition matrix. Note that the order of the index on the joint probability is reversed into $\beta\alpha$ in the second summation. It is not clear how to derive an approximate invariant solution from this equation.

(4) The last model is revised from the previous model by replacing the direct replication with the complementary replication, i.e.,

$$\begin{aligned} \cdots a \cdots \quad &\rightarrow \quad \begin{cases} \cdots a \cdots b & : \quad 1-p \\ \cdots b \cdots & : \quad p, \end{cases} \\ \cdots b \cdots \quad &\rightarrow \quad \begin{cases} \cdots b \cdots a & : \quad 1-p \\ \cdots a \cdots & : \quad p, \end{cases} \end{aligned} \tag{11}$$

illustrated in Fig.7.

Again, there is no interesting structure in the limiting sequence if the initial seed is a single symbol. If we start from a segment *abb* with length three, the mutual information function for the limiting sequences is shown in Fig.8. Without mutation, the mutual information function of the limiting sequence reaches maximum at $d = 2, 3, 6, 9, 12, 18, 24, 36, 48, \ldots$, whereas in Fig.8

(for mutation rate $p = 0.01$), not only is there a tendency for the $M(d)$ to decrease, but the local peaks beyond $d = 24$ are also shifted.

5 Mutual information functions of several human nucleotide sequences

As promised in the first section, I will present the result of mutual information function of nucleic acid sequences. I would like to discuss two facts observed in human nucleotide sequences which I analyzed: (1) intron (non-coding) segments tend to have longer correlation lengths than exon (protein-coding) segments; (2) the correlation length for some intron sequences can be so long that part of the power spectrum is close to a $1/f$ spectrum. The mutual information function of other nucleic acid sequences, especially those of complete genomes, will be included in the forthcoming paper [Li, in preparation].

There have been several correlation analysis for nucleotide sequences and amino acid sequences, using basically the autocorrelation function. Occasionally, power spectra are also used for detecting periodicity in protein sequences [Liquori, *et. al.*, 1986], and as an algorithm for speeding up the calculation of autocorrelation functions [Felsenstein, *et. al.*, 1982].

The autocorrelation function is defined only for numerical sequences. The question of how to get a numerical series from the nucleic acid sequences has been handled in different ways. There are the following approaches: (1) using other physical quantities instead of the base sequence [Trifonov & Sussman, 1980] [Kubota, *et. al.*, 1981], assuming that these physical quantities are closely related to the underlying primary sequence; (2) calculating the correlation of sites with a particular property: if this property is present at a site, the numerical value on that site is one, otherwise, the value is zero. So far, this is the most popular approach [Shepherd, 1981] [Fickett, 1982] [MaLachlax & Karn, 1983] [Arquést & Michel, 1987, 1990a, 1990b]; (3) considering each of the 4 symbols as a vertex of the 3-simplex (i.e., the tetrahedron). Then a 4-symbol sequence becomes a vector sequence with three component sequences. The autocorrelation function or the power spectrum for the three component sequences can be calculated, and the overall autocorrelation function or the power spectrum takes contributions from each component sequence [Silverman & Linsker, 1986]. This idea is very neat, but has not been applied to nucleic acid sequence analysis very often.

For sequences with only short-range correlations, Markov chain approximation should be good enough, and one only needs to determine all the elements in the transition matrix. For sequences with median range correlations, Markov chains with higher orders can be applied, see [Tavaré & Giddings, 1989]. Nevertheless, if the correlation length is much longer as a result of tandem or interspersed repeat, one has to calculate the correlation function up to very large distances. It is this fact that the discussion presented in this section could be useful for the nucleic acid sequence analysis.

To start the calculation, I take five exon segments and five intron segments from human DNA sequences. All the data are from GenBank [Burks, *et. al.*, 1989]. I choose these sequences because they have relatively longer sequence lengths, which makes the calculation of the joint probability as well as the mutual information more reliable. The five exon sequences are:

- Human coagulation factor VIII:C (anti-hemophilic factor) mRNA
  (name: HUMFVIII, length: 7056);
- Human alpha-2-macroglobulin mRNA
  (name: HUMA2M, length: 4425);
- Human ceruloplasmin (ferroxidase) mRNA
  (name: HUMCERP, length: 3198);
- Human 90-kDa heat-shock protein gene
  (name: HUMHSP90, length: 2175);
- Human factor I (C3b/C4b inactivator) mRNA
  (name: HUMFISP, length: 1752).

The unit of length is the nucleotide base (or base-pair due to the double-strand structure of DNA molecules). The five intron sequences are:

- Human serum albumin gene
  (name: HUMALBGC, length: 16349);
- Human proopiomelanocortin (POMC) gene
  (name: HUMPOMC, length: 6594);
- Human blood coagulation factor VII gene
  (name: HUMCFVII, length: 5640);
- Human haptoglobin gene (alpha-2 allele)
  (name: HUMHPARS1, length: 5017);

- Human alpha-tubulin gene (b-alpha-1)
  (name: HUMTUBAG, length: 1980).

Fig.9 (a)–(e) show the mutual information functions of all five exon sequences. Due to the finite statistics, even two uncorrelated variables can have a non-zero residue mutual information [Li, 1990]. In order to subtract the finite size effect, I include the mutual information function for the corresponding random sequences in these plots (two for each). By "corresponding", I mean that the sequence has the same sequence length and the same composition of the four symbols: A (Adenine), C (Cytosine), G (Guanine), and T (Thymine).

The crossing region between the two mutual information functions, one for the original nucleic acid sequence and another for the corresponding random sequence, gives the distance at which the correlation becomes negligible. In other words, it is a good estimation of the correlation length. Roughly speaking, the correlation lengths for the five exon sequences are of the order of 10 or less, except one sequence HUMHSP90 whose correlation length seems to be much longer. Curiously, for this sequence, the two mutual information functions cross at around $d \sim 5$, but then they are separated again at longer distances.

Fig.10 (a)–(e) show the mutual information functions of all five intron sequences, as well as those of the corresponding random sequences. The correlation lengths seem to be around 20 or more, except one sequence HUMCFVII whose correlation length is substantially longer. In order to see how long the correlation length is, I plot the mutual information function of the sequence HUMCFVII again in Fig.11(a) up to much longer distances. The two $M(d)$'s intersect around $d \sim 600 - 1000$ (the sequence length itself is 5640).

Note that Fig.9 and 10 confirm the previous findings that correlation in nucleic acid sequences oscillates [Shepherd, 1981] [Fickett, 1982], and the periodicity of the oscillation tends to be three for exon sequences and two for intron sequences [Arquést, 1987, 1990a, 1990b] (see, in particular, the exon sequence HUMHSP90 and the intron sequence HUMALBGC). In addition to these known results, our mutual information functions show a new feature which has not been discussed before, that intron sequences tend to have more slowly decaying mutual information functions than exons, or *intron sequences tend to have longer correlation lengths*.

It is not clear of whether this observation holds for other exon or intron sequences, and whether it can be turned into some practical tool for distinguishing introns and exons. Identifying protein-coding regions in DNA sequences is a classical problem in nucleic acid sequence analysis (see, for example, [Stormo, 1987, 1990]). It is known that intron and exon sequences do have different statistical properties, and it will be interesting to establish that the correlation length is one of them.

In some hand-waving arguments, one could understand why the exon sequences tend to have correlation length around 10. Exon sequences consist of codons, which can be considered as "words" in the "sentence" which is the exon sequence itself. Typically, there is a distinct structure within a codon and not all possible three-base configurations appear in the sequence with equal probability. The structure of codons and their uneven distribution impose a strong correlation at short distances. On the other hand, the correlation between codons is weak, and Markov chains are in fact good approximation for codon sequences. As a result, the correlation length is at most a few codon lengths, i.e., a few multiples of three. A value of 10 for the correlation length is consistent with this picture.

In fact, the mutual information function for the letter sequences (alphabets as well as punctuations, and blank spaces) or letter-type sequences (with a smaller number of symbols, including only vowel, consonant, punctuation and blank space) of the English texts exhibit the similar behavior. Fig.12 shows the mutual information function of the JFK's speech (the text is taken from [Graham, 1970], with the sequence length equal to 7391). The correlation length is around 10 — also a few multiples of the average length of English words. More results of the mutual information function of letter sequences in English is in [Li, 1989b].

Estimating the correlation length of introns seems to be more difficult. One understanding of the long-range correlation in intron sequences is perhaps that there exist highly repeated segments.[1] If this is true, the value of the correlation length should depend on how frequent this repetition occurs, how long the repeated segment is, and how far apart the repeated segments are. Another understanding of the long-range correlation in introns might be that the secondary structures of RNA molecules require certain correlation in the primary sequence. Similar discussion on the effects of the secondary structure of RNA sequences on the formal language grammar that describes the sequence can be found in [Searls, 1990]. It should be interesting to understand the intron/exon difference, including the difference of the correlation length, from the evolutionary point of view. It will bring us closer to the theme discussed throughout this paper, that the statistical properties of the sequences should be strongly related to the dynamics which generate them.

[1] I thank C. Burks for discussions on this point.

6 Partial $1/f$ spectrum in the nucleic acid sequence HUMCFVII

The persistence of large correlation values at longer distances indicates there are structures with length scales comparable to the sequence length itself, and it causes an increase of the power spectrum at lower frequencies. One case of this situation is the $1/f$ noise, or sequences whose power spectra are $P(f) \sim 1/f^{\alpha}$, with $\alpha \approx 1$.

The extremely slow decay of the mutual information function in intron sequence HUMCFVII fits the above description. The sequence HUMCFVII is composed of four intron segments, from position 586–1653, 1720–4293, 4455–6382, and 6408–6477. Both the location of four segments of HUMCFVII and the actual sequence are shown in Fig.13. Because the sequence is almost all introns, the deletion of a small fraction of the exons is not expected to effect our conclusion. From Fig.13, it can be seen that the second segment contains a highly repetitive structure, with the periodicity equal to 17. Indeed, there are peaks in the mutual information function (see Fig.11(a)) at the multiples of 17. Besides this repetition, there seem to be other repetitions in the sequence as well.

To check that this period of 17 repetitions are not the only source of the long-range correlation, Fig.11(b) shows the mutual information function of the same sequence with the period of 17 segments being deleted (the sequence length is now 4808 as compared with the original length of 5640). Although all peaks at multiples of 17 disappear, the correlation length is still as high as 500.

In order to calculate the power spectrum, I convert the 4-symbol sequence to 2-symbol sequences either by grouping A and G (both of them are purines, R), T and C (both of them are pyrimidines, Y); or, by grouping T and A (they are complementary to each other), C and G (they are also complementary to each other). The power spectra (in log-log scale) for the two converted binary sequences are shown in Fig.14 and Fig.15 respectively with the sequence length being cut at $2^{12} = 4096$ (1544 bases are deleted from the 5640 bases, including the complete fourth segment and part of the third segment). For the program of calculating power spectrum, see, for example, Chapter 12 of [Press *et. al.*, 1988]).

The two spectra are very similar, but the fitting of lower frequency components of the spectrum gives $P(f) \sim 1/f^{0.93}$ for the first plot, and $P(f) \sim 1/f^{0.76}$ for the second plot. The high frequency spectral components are basically flat. The peaks correspond to the period 17 patterns ($\log_{10}(f) = \log_{10}(4096/17) = 2.38$). The separation between the low frequency $1/f^{\alpha}$ spectrum and the high frequency white spectrum is arbitrary, and has been chosen by a personal judgement. The scaling of the $1/f$ spectrum spans 1.5 decades, out of a total 3.3 decades ($\log_{10}(4096/2) = 3.31$). To emphasize that this $1/f$ spectrum can only characterize a small portion of the spectrum, I call it *partial* $1/f$ spectrum.

A question raised is how widespread partial $1/f$ spectra like this are in nucleic acid sequences? We have already excluded all protein-coding sequences, because their correlation length is typically very short. Besides intron segments, "junk genes" — the segments in between two genes — are also potential candidates for sequences with long-range correlation. Unfortunately (perhaps fortunately for biologists?), there have been so far no junk genes sequences available in the GenBank.

7 Discussions and conclusions

The models described in section 4 have several simplistic aspects which make them hardly realistic for the evolution of nucleic acid sequences. These rules act at a very low level. They are more like models for physical systems instead of biological systems. In contemporary biological organisms, the change of nucleic acid sequences is under high-level control and regulation, driven by evolutionary pressures, and involves other macromolecules. There have been attempts to increase the degree of complexity of the sequence manipulation rules; see, for example, an approach called typogenetics [Hofstadter, 1979] [Morris, 1988, 1989]. In typogenetics, the sequence is examined by a moving head. The moving head imposes operations such as cutting the sequence or inserting new symbols by looking at the local symbol configurations and consulting a high-level code. When the moving head finally stops, a new sequence, or a new set of sequences, is produced. The moving head approach is reminiscent of the Turing machine [Hopcroft & Ullman, 1979]. It is not known in typogenetics what the connection is between the high-level code (as well as the initial sequence) and the statistical features of the final sequence(s).

In typogenetics, one has to provide a high-level code, which is presumably based on the knowledge of chemistry. Identifying the high-level code directly from chemistry can be very difficult in contemporary biological systems. Nevertheless, it might be relatively easier for a prebiotic environment since the high-level instructions were rare. Even if some high-level rules exist, they resulted in simple terms from the low-level interaction of a population of sequences. There are several attempts to study this "reaction networks", one example is the hypercycle [Eigen, 1971] [Eigen & Schuster, 1979] [Eigen *et. al.*, 1988], and another is the autocatalytic networks [Kauffman, 1986] [Farmer *et. al.*, 1986] [Bagley, 1991]. Again, the statistical properties of a population of sequences are not the focal point of these studies (see, however, a recent study mentioned in [Kauffman, 1990]).

Even with a single-sequence, fixed-dynamics-rule models, there are many variations beyond the scope of this paper which are potentially relevant to the evolution of nucleic acid sequences. In particular, the addtion of insertions and deletions to our models should be desirable. Shepherd has studied the effects of introducing insertion to the periodic sequences [Shepherd, 1981]. By examining simple examples (for example, the sequence $\ldots abababab \ldots$ before insertion, and $\ldots ababbabab \ldots$ after), it can be shown that the correlation function of periodic sequences with defects decays linearly. In other words, the effect of the mismatch propagates linearly. On the other hand, if the original sequence is random, the insertions will have very little effect.

In conclusion, this paper discusses the long-range correlations generated by local replication followed by an insertion (elongation) or sequence replication followed by a ligation. The existence or the absence of the long-range correlation is used to infer, to some extents, the dynamical process which produces the sequence. Indeed, it is observed in this paper that protein-coding (exons) and non-coding (intron) segments have different correlation lengths — those in introns are typically longer than those in exons. Although there is still a long way to go before we can comprehend all the statistical features of contemporary nucleic acid sequences from the evolution process — like what has been partially achieved in cosmology on explaining the statistical features of the galaxy distribution — it is hoped that this paper will stimulate more interest and studies on this subject.

Note Added in Press

Two highly relevant papers have been published since the completion of this paper. The first one [Li & Kaneko, 1992] carries out a symbolic spectral analysis of the sequence HUMCFVII and suggests a parallel between the repetitive segments in intron sequences and those in music notes. The second one [Peng, *et. al.*, 1992] converts purine-pyrimidine binary sequences into random walks, and these “DNA walks” are graphically displayed. The main conclusion of the current paper that intron sequences have longer correlation lengths than exon sequences is confirmed in [Peng, *et. al.*, 1992].

Acknowledgement

I would like to thank Kuni Kaneko who invited me to the Center for Nonlinear Studies (CNLS) at Los Alamos in the summer of 1989; during that period the main ideas presented here were discussed. I appreciate the financial support from the CNLS. I also thank Rick Bagley, Christian Burks, Chris Fields and Walter Fontana for discussions or for recommending references; and Alan Lapedes for providing the nucleic acid sequences. The second draft of this paper is based partly on the talk I gave at the "recognizing gene" workshop held at the Aspen Center for Physics in the summer of 1990. I acknowledge the support from the workshop. The work at Santa Fe Institute is funded by the grants from the MacArthur Foundation, from the National Science Foundation (PHY-87-14918), and from the Department of Energy (DE-FG05-88ER25054). This paper is dedicated to the memory of Li, Chunqing (1908–1990).

References

[1] D.G. Arquést and C.J. Michel [1987] "A purine-pyrimidine motif verifying an identical presence in almost all gene taxonomic groups," *Journal of Theoretical Biology*, **128**, 457–461.

[2] D.G. Arquést and C.J. Michel [1990a] "Periodicities in coding and noncoding regions of the genes," *Journal of Theoretical Biology*, **143**, 307–318.

[3] D.G. Arquést and C.J. Michel [1990b] "A model of DNA sequence evolution," *Bulletin of Mathematical Biology*, **52**(6), 741–772.

[4] R.J. Bagley [1991] *The functional self-organization of autocatalytic networks in a model of evolution of biogenesis*, Ph.D Thesis, University of California at San Diego; and private communications.

[5] R. Britten and D. Kohne [1968] "Repeated sequences in DNA," *Science*, **161** (3841), 529–540;

[6] R. Britten and D. Kohne [1970] "Repeated segments of DNA," *Scientific American*, **222**(4), 24–31.

[7] C. Burks *et. al.* [1989] "GenBank: Current status and future direction," *Methods in Enzymology*, **183**, 1–22.

[8] Z. Cheng, R. Savit and R. Merlin [1988] "Structure and electronic properties of Thue-Morse lattices," *Physical Review B*, **37**(9), 4375–4382;

[9] Z. Cheng and R. Savit [1990] "Structure factor of substitutional sequences," *Journal of Statistical Physics*, **60**, 383–393.

[10] Thomas M. Cover and Roger C. King [1978] "A convergent gambling estimate of the entropy of English," *IEEE Transactions on Information Theory*, **24**(4), 413–421.

[11] M. Eigen [1971] "Self-organization of matter and the evolution of biological macromolecules," *Die Naturwissenschaftern*, **58**, 465–523;

[12] M. Eigen and P. Schuster [1979] *The hypercycle: a principle of natural self-organization* (Springer-Verlag).

[13] M. Eigen, J. McCaskill and P. Schuster [1988] "The molecular quasi-species," *Journal of Chemical Physics*, **92**, 149–263.

[14] J.D. Farmer, S.A. Kauffman, and N.H. Packard [1986] "Autocatalytic replication of polymers," *Physica D*, **22**, 50–67.

[15] J. Felsenstein, S. Sawyer and R. Kochin [1982] "An efficient method for matching nucleic acid sequences," *Nucleic Acids Research*, **10**(1), 133–139.

[16] J.W. Fickett [1982] "Recognition of protein coding regions in DNA sequences," *Nucleic Acids Research*, **10**(17), 5303–5318.

[17] L. Gatlin [1966] "The information content of DNA, " *Journal of Theoretical Biology*, **10**, 281–300;

[18] L. Gatlin [1968] "The information content of DNA. II, " *Journal of Theoretical Biology*, **18**, 181–194;

[19] L. Gatlin [1972] *Information Theory and the Living Systems* (Columbia University Press).

[20] J. Graham [1970] *Great American Speeches: 1898–1963* (ACC).

[21] D. Hofstadter [1979] *Gödel, Escher, Bach* (Basic Books).

[22] J. E. Hopcroft and J. D. Ullman [1979] *Introduction to Automata Theory, Language and Computation* (Addison-Wesley).

[23] J. Horgan [1991] "In the beginning ...," *Scientific American*, **264**(2), 116–125.

[24] W. Jelinek and C. Schmid [1982] "Repetitive sequences in eukaryotic DNA and their expression," *Ann. Reviews of Biochemistry*, **51**, 813–844.

[25] S. Karlin [1968] *A First Course in Stochastic Process* (Academic Press).

[26] S. Karlin and H. M. Taylor [1981] *A Second Course in Stochastic Processes* (Academic Press).

[27] S.A. Kauffman [1986] "Autocatalytic sets of proteins," *Journal of Theoretical Biology*, **119**, 1–24.

[28] S. Kauffman [1990] "Random grammars: a new class of models for functional integration and transformation in the biological, neural and social sciences," Santa Fe Institute preprint 90-020.

[29] Y. Kubota, S. Takahashi, K. Nishikawa, and T. Ooi [1981] "Homology in protein sequences expressed by correlation coefficients," *Journal of Theoretical Biology*, **91**, 347–361.

[30] W. Li [1987] "Power spectra of regular languages and cellular automata," *Complex Systems*, **1**(1), 107–130.

[31] W. Li [1989a] "Spatial 1/f spectra in open dynamical systems," *Europhysics Letters*, **10**(5), 395–400.

[32] W. Li [1989b] "Mutual information functions of natural language texts," Santa Fe Institute preprint 89-009.

[33] W. Li [1990] "Mutual information functions versus correlation functions," *Journal of Statistical Physics*, **60**(5/6), 823–837.

[34] W. Li [1991a] "Expansion-modification systems: a model for spatial 1/f spectra," *Physical Review A*, **43**(10), 5240–5260.

[35] W. Li [1991b] "Absence of $1/f$ spectra in Dow Jones daily average," *International Journal of Bifurcation and Chaos*, **1**(3), 583–597.

[36] W. Li and K. Kaneko [1992] "Long-range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence," *Europhysics Letters*, **17**(7), 655–660.

[37] W. Li, N. Packard and C. Langton [1990] "Transition phenomena in cellular automata rule space," *Physica D*, **45**(1-3), 77–94.

[38] A. Lindenmayer [1968] "Mathematical models for cellular interactions in development (I. filaments with one-sided inputs)," *Journal of Theoretical Biology*, **18**, 280–299.

[39] A.M. Liquori, A. Ripamonti, C. Sadun, S. Ottani, and D. Braga [1986] "Pattern recognition of sequence similarities in globular proteins by Fourier analysis: a novel approach to molecular evolution," *Journal of Molecular Evolution*, **23**, 80–87.

[40] E. Long and I. Dawid [1980] "Repeated genes in eukaryotes," *Ann. Reviews of Biochemistry*, **49**, 727–764.

[41] A.D. McLachlax and J. Karn [1983] "Periodic features in the amino acid sequence of nematode myosin rod," *Journal of Molecular Biology*, **164**, 605–626.

[42] H. Morris [1988] *Typogenetics: a logic of artificial propagating entities*, Ph.D Thesis, University of British Columbia.

[43] H. Morris [1989] "Typogenetics: a logic for artificial life," in *Artificial Life*, ed. C. Langton (Addison-Wesley).

[44] H.M. Morse [1921] "Recurrent geodesics on a surface of negative curvature," *Transactions of American Mathematical Society*, **22**, 84–100.

[45] T. Musha, S. Sato, and M. Yamamoto, eds. [1991] *Proceedings of the International Conference on Noise in Physical Systems and* $1/f$ *Fluctuations* (Ohmsha, Ltd.).

[46] S. Ohno [1987] "Early genes that were oligomeric repeats generated a number of divergent domains on their own," *Proceedings of National Academy of Science*, **84**, 6486–6490.

[47] C-K. Peng, *et. al.* [1992] "Long-range correlations in nucleotide sequences," *Nature*, **356**, 168–170.

[48] W.H. Press [1978] "Flicker Noise in Astronomy and Elsewhere," *Comments On Astronomy*, **7**(4), 103–119.

[49] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling [1988] *Numerical Recipes in C* (Cambridge University Press).

[50] M. Queffelec [1987] *Substitution Dynamical Systems: Spectral Analysis.*

[51] D.B. Searls, "The computational linguistics of biological sequences," UNISYS Center for Advanced Information Technology preprint CAIT-KSA-9010 (1990).

[52] C.E. Shannon [1948] "The Mathematical Theory of Communication," *Bell Syst. Techn. Journal* **27**, 379-423;

[53] C. E. Shannon [1951] "Prediction and entropy of printed English," *Bell Syst. Tech. Journal* 50-64.

[54] C.E. Shannon and W. Weaver [1949] *The Mathematical Theory of Communication* (University of Illinois Press).

[55] J.C.W. Shepherd [1981] "Periodic correlations in DNA sequences and evidence suggesting their origin in a comma-less genetic code," *Journal of Molecular Evolution*, **17**, 94–102.

[56] B. Silverman and R. Linsker [1986] "A measure of DNA periodicity," *Journal of Theoretical Biology*, **118**, 295–300.

[57] T.F. Smith [1969] "The genetic code, information density, and evolution," *Mathematical Biosciences*, **4**, 179–187.

[58] G.D. Stormo [1987] "Identifying coding sequences," in *Nucleic Acid and Protein Sequence Analysis: A Practical Approach*, eds. M.J. Bishop and C.J. Rawlings (IRL Press);

[59] G.D. Stormo [1990] "Identifying regulatory sites from DNA sequence data," in *Structure and Methods. Vol I: Human Initiative and DNA Recombination*, eds. R.H. Sarma and M.H. Sarma (Adenine Press).

[60] S. Tavarè and B.W. Giddings [1989] "Some statistical aspects of the primary structure of nucleotide sequences," in *Mathematical Methods for DNA Sequences*, ed. M.S. Waterman (CRC Press).

[61] A. Thue [1906] *Norske Vid. Selsk. Skr.*, **7**, 1.

[62] T. Toffoli and N. Margolus [1987] *Cellular Automata Machine — A New Environment for Modeling* (MIT Press).

[63] E.N. Trifonov and A.J. Sussman [1980] "The pitch of chromatin DNA is reflected in its nucleotide sequence," *Proceedings of National Academy of Science*, **77**, 3816–3820.

[64] J. von Neumann [1966] *Theory of Self-Reproducing Automata*, ed. A.W. Burks (University of Illinois Press).

[65] M.A. Wanes, A.I. Zayed, M.M. Shaker, and E.H. Taha [1976] "First-, second-, and third-order entropies of Arabic text," *IEEE Transactions on Information Theory*, **22**(1), 123.

[66] J.D. Watson [1990] "The human genome project: past, present, and future," *Science*, **248**, 44–49.

[67] J.D. Watson, N.H. Hopkins, J.W. Roberts, J.A. Steitz, and A.M. Weiner [1987] *Molecular Biology of the Gene*, Chapter 21 (Benjamin/Cummings).

[68] Stefan Węgzyn, Jean-Charles Gille, and Pierre Vidal [1990] *Development Systems* (Springer-Verlag).

[69] S. Wolfram [1983] "Statistical mechanics of cellular automata," *Review of Modern Physics*, **55**, 601–644.

[70] S. Wolfram [1984] "Computation theory of cellular automata," *Communications in Mathematical Physics*, **96**, 15–57.

Figure 1: Illustration of the sequence manipulation rule (4.4), in which a symbol can either be elongated to two same symbols (solid arrows) or mutate to a different symbol (shaded arrows).

Figure 2: Mutual information function $M(d)$ of sequences generated by rule (4.4) at mutation probabilities $p = 0.0492 \approx 0.05$ and $p = 0.299 \approx 0.3$. The initial condition is a single symbol a, and the sequence length $N = 100,000$.

Figure 3: Illustration of the sequence manipulation rule (4.8), in which a symbol can either be elongated to one same symbol followed by a different symbol (solid arrows), or mutate to a different symbol (shaded arrows).

Figure 4: Mutual information function $M(d)$ of sequences generated by rule (4.8) at mutation probabilities $p = 0.0496 \approx 0.05$ and $p = 0.298 \approx 0.3$. The initial condition is a single symbol a, and the sequence length $N = 100,000$.

Figure 5: Illustration of the sequence manipulation rule (4.9), in which a symbol either makes an extra copy of the same symbol (solid arrows), or does not copy but mutates itself (shaded arrows), then the copied sequence is ligated to the original sequence (encircled by the rectangle).

Figure 6: Mutual information function $M(d)$ of sequences generated by rule (4.9) at mutation probabilities $p = 0.00993 \approx 0.01$ and $p = 0.0493 \approx 0.05$. The initial condition is a symbol string abb, and the sequence length $N = 100,000$.

Figure 7: Illustration of the sequence manipulation rule (4.11), in which a symbol can either make an extra copy of a symbol different from itself (solid arrows), or do not copy but mutate itself (shaded arrows), then the copied sequence is ligated to the original sequence (encircled by the rectangle).

Figure 8: Mutual information function $M(d)$ of sequences generated by rule (4.11) at mutation probabilities $p = 0.0101 \approx 0.01$ and $p = 0.0496 \approx 0.05$. The initial condition is a symbol string abb, and the sequence length $N = 100,000$.

Figure 9: Mutual information function $M(d)$ of five exon sequences from human genome. The sequences are (a) HUMFVIII, with length $N = 7056$; (b) HUMA2M, $N = 4425$; (c) HUMCERP, $N = 3198$; (d) HUMHSP90, $N = 2175$; and (e) HUMFISP, $N = 1752$. The mutual information functions of two corresponding random sequences are also included for each case. The correlation length can be estimated by the distance at which $M(d)$ of the exon sequence intersects with the $M(d)$ of the random sequences.

Figure 10: Mutual information function $M(d)$ of five intron sequences from human genome. The sequences are (a) HUMALBGC, with length $N = 16349$; (b) HUMPOMC, $N = 6594$; (c) HUMCFVII, $N = 5640$; (d) HUMHPARS1, $N = 5017$; and (e) HUMTUBAG, $N = 1980$. The mutual information functions of two corresponding random sequences are also included for each case. The correlation length can be estimated by the distance at which $M(d)$ of the intron sequence intersects with the $M(d)$ of the random sequences.

Figure 11: The mutual information function $M(d)$ of (a) the intron sequence HUMCFVII (up to distance $d = 1000$, as compared with the maximum distance $d = 100$ in Fig.10(c)). Also shown is $M(d)$ of a corresponding random sequence; (b) the same HUMCFVII intron sequence with the period of 17 segments (832 bases) being deleted.

Figure 12: $M(d)$ of the letter-type sequence derived from the letter sequence of the JFK's inaugural speech. The four letter types are vowel, consonant, punctuation and blank space. The sequence length is $N = 7391$. Also shown is the $M(d)$ of a corresponding random sequence.

Figure 13: The location of the four intron segments of HUMCFVII and the sequence itself.

Figure 14: The power spectrum $P(f)$ of a binary sequence derived from the intron sequence HUMCFVII. The first symbol includes nucleotides A and G (purines), and the second symbol includes T and C (pyrimidines). The number of bases included in the calculation is $2^{12} = 4096$ out of total 5640 bases. Half of the Fourier components are redundant, and only the first half of the spectrum is plotted (the maximum value on x-axis is $\log_{10}(4096/2) = 3.31$). Four neighboring spectrum components are averaged into one value (which leaves $2^9 = 512$ points on the plot). The best-fit line for the first 20 points using the power law function $P(f) \sim 1/f^{\alpha}$ gives $\alpha \approx 0.93$.

Figure 15: The power spectrum $P(f)$ of a binary sequence derived from the intron sequence HUMCFVII. The first symbol includes nucleotides T and A, and the second symbol includes C and G (see Fig.14 for a comparison). The best-fit line for the first 20 points (out of 512 spectrum components) using the power law function $P(f) \sim 1/f^{\alpha}$ gives $\alpha \approx 0.76$.