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

Wentian Li

SFI WORKING PAPER: 1991-01-002

SFI Working Papers contain accounts of scientific work of the author(s) and do not necessarily represent the views of the Santa Fe Institute. We accept papers intended for publication in peer-reviewed journals or proceedings volumes, but not papers that have already appeared in print. Except for papers by our external faculty, papers must be based on work done at SFI, inspired by an invited visit to or collaboration at SFI, or funded by an SFI grant.

©NOTICE: This working paper is included by permission of the contributing author(s) as a means to ensure timely distribution of the scholarly and technical work on a non-commercial basis. Copyright and all rights therein are maintained by the author(s). It is understood that all persons copying this information will adhere to the terms and constraints invoked by each author's copyright. These works may be reposted only with the explicit permission of the copyright holder.

www.santafe.edu
Generating Non-trivial Long-range Correlations and 1/f Spectra by Replication and Mutation

Wentian Li
Santa Fe Institute, 1120 Canyon Road, Santa Fe, NM 87501
January 31, 1991

Abstract. This paper aims at the goal of understanding the statistical features of nucleic acids sequences from the knowledge of the dynamical process that produces them. On one hand, the 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 or replication can easily produce long range correlations. These long range correlations could be destroyed by the mutation with various degree in different sequence manipulation models. On the other hand, the mutual information functions for several human nucleic acids sequences are determined. It is observed that intron sequences (non-coding sequences) tend to have longer correlation lengths than the exon sequences (protein coding sequences).

1. Introduction

Ever since the terrestrial genesis, the molecules which are capable of replication have been playing the essential role in the life on earth. The replicators are basically nucleic acids sequences — 1-dimensional strands consist of nucleotide bases. The arrangement of nucleotide bases on a nucleic acids sequence is transformed to the arrangement of amino acids 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 arrangements of the nucleotide bases on nucleic acids sequences result from more than three billion years’ evolution (see, for example, Refs.[1, 2]).

Now, we ask the following question: can we understand why the nucleic acids sequences have the arrangement of the bases as is observed today? Or, can we understand the statistical features of these nucleic acids sequences from some models of the 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), the 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. First, we do not have the complete knowledge of the arrangement of the nucleotide bases in a large portion of nucleic acids sequences. There are, however, great efforts towards improving the situation, notably the human genome project [3]. 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 the nucleic acids sequences. Thirdly, we still know very little about how life started, and do not have a good guess of the initial condition.

This paper attempts to make a small contribution towards the final answer of the question. At one end of matching models with the 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 occasional, the initial condition. At another end of the matching, I will calculate the mutual information function [4, 5], one of the most important statistical quantities of the sequence, of several nucleic acids sequences. Not completely surprising, the mutual information function of a nucleic acids sequence depends 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 the intron segments differs from that of the 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 acids sequences. As the title suggested, the goal is quite limited: I will examine the long range correlations in nucleic acids sequences as produced by elongation (or condensation followed by the replication). The presence, or the absence for that purpose, of the long range correlation 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 the 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 the structure of the sequence manipulation rules and the 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 acids sequences; Section 6 studies the 1/f spectrum in one of the intron sequences; and finally, the section 7 contains discussions and possible future research directions.
2. Mutual information function: measure of correlation in symbolic sequences

It is not clear which statistical property is most appropriate for characterizing a nucleic acids sequence, and, by comparing that of the nucleic acids sequence and that derived from the model, for checking the validity of a 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 [4]. If the block entropies are determined for all block sizes, the statistical feature of the sequence is determined too. Unfortunately, the block entropy cannot be calculated accurately for large block lengths when the sequence length itself is limited. In the previous studies of the entropies of natural language texts (ranging from English [6, 7] to Arabic [8]) and the nucleic acids sequences [9, 10, 11], 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 the function of the distance between them. There are several ways to measure the correlation of two variables, for example, the average of the product of the two variables subtracting the product of the average of each variable. If this definition of correlation is used, we have the conventional autocorrelation function:

\[ \Gamma(d) = \sum_{\alpha\beta} x_\alpha x_\beta P_{\alpha\beta}(d) - \left( \sum_{\alpha} x_\alpha P_\alpha \right)^2, \]  

(2.1)

where \( x_\alpha \)'s are all possible values, \( P_\alpha \) is the density of the symbol with value \( x_\alpha \), and \( P_{\alpha\beta}(d) \) is the probability of having a symbol with the value \( x_\alpha \) followed \( d \) sites away by a symbol with the value \( x_\beta \). 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 [4], and the correlation function becomes the mutual information function [5]:

\[ M(d) = \sum_{\alpha\beta} P_{\alpha\beta}(d) \log \frac{P_{\alpha\beta}(d)}{P_\alpha P_\beta}. \]  

(2.2)

Note that because each term inside the summation is weighted by a \( P_{\alpha\beta}(d) \), in certain approximation, the \( M(d) \) behaves like \( P_{\alpha\beta}(d)^2 \), in comparison with \( \Gamma(d) \sim P_{\alpha\beta}(d) \). For more discussions on the relation between the mutual information function and the autocorrelation function, see Ref.[5].

There might be other definitions of the correlation function such as the one based on “Chi-square” (e.g., Chapter 13 of [12]). To avoid confusion, I
will use $C(d)$ to represent any one of them, i.e.,

$$C(d) = \{ \Gamma(d), M(d), \ldots \}. \tag{2.3}$$

With correlation functions such as the mutual information function being defined, we can quantitatively define the 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. The periodic structure need not to be exact for having this long range correlation, they can only approximately be so. We call the long range correlations resulted from the periodic structures in the sequence trivial long range correlations.

Sometimes, the $C(d)$ is non-zero at almost all larger $d$'s, with no dominant peaks in $C(d)$. Usually, such slow decay of the $C(d)$ can be approximated by power law functions (algebraic decay) — $1/d^\alpha$. In particular, if the decay is so slow that the exponent $\alpha$ is close to zero (when $\alpha$ is exactly equal to zero, $C(d)$ decays slower than any power law functions; they are, e.g., logarithmic functions), the sequence can be called 1/f noise because its power spectrum behaves like $1/f$, where $f$ is the frequency (to 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: $e^{-d/d_0}$, the $d_0$ 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 the $C(d)$ is almost zero beyond certain distance $d'_0$, and this distance is a good 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 the statistical features of the galaxy distribution such as the power law two-point correlation function are the results of the expansion of space, the long-range gravitational interaction, and the initial stages of the big bang which determine the starting configuration. Varying either one of the conditions, one may not reproduce the statistics in the real data. 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 the sequence length is changed, in order to determine the statistical properties of the sequence.
In the following, I will review three types of the "sequence manipulation rules": (1) those generating sequences from one end to another with short memories; (2) those updating sequences parallelly according to local dynamics with the sequence length fixed; (3) those updated 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 some general conclusions concerning the statistical properties of the 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 some brief discussions 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 will be added one by one at the end of the sequence. This class includes the well known examples of Markov chains [13] and regular languages [14]. In 1-step Markov chains, the probability of having a 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 $\lambda$ is the largest eigenvalue (excluding the trivial value of 1) of the Markov transition matrix [13].

Regular languages, studied in the framework of the formal language [14], 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 bigger, then degenerate these symbols again; see Refs.[15] for details), but again $C(d)$ behaves like $\lambda^d$, with $\lambda$ as the largest eigenvalue (excluding the value of 1) of the more complicated 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. Also note that when the largest non-trivial eigenvalue is negative (largest in magnitude), $C(d)$ oscillates.

(2) The second type of the sequence manipulation rules could be considered
as one of the spatially-extended dynamical systems, with the latter includes 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 [16, 17]. 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 Ref.[15]. 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 [18], and by our previous discussion, the correlation function is exponential (either monotonic or oscillating).

Generally speaking, if a cellular automaton rule is capable of generating correlation length much longer than the local coupling range, 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. This rule can then be said to be on the “edge of chaos.” In fact, the existence of a large correlation value at long distances is used to locate the region of the cellular automata rule space where the transition from periodic to chaotic dynamics occurs [19].

(3) The third type of the sequence manipulations contains rules that update symbols according to local dynamics and increase the sequence length at the same time. One might call them context-sensitive Lindenmayer systems [20] or context-sensitive development systems [21], or perhaps “expanding cellular automata”. These systems are rarely discussed from the perspective of the statistical properties of the limiting sequence. Even the 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 the 2-symbol 3-input cellular automata which is $2^8 = 256$.

A direct consequence of the 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 $\lambda$ 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
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 is studied, and it has already shown an enormous amount of complexity [22, 23, 24].

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 copy of itself, it will mutate. In other words, the probability of having replication $P_{\text{replication}}$ is $1 - P_{\text{mutation}}$, with $P_{\text{mutation}}$ 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 only that the replication is complementary; (3) the whole sequence replicates and the copy is linked to the original sequence; and (4) similar to the third case only 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 the 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:

\[
\begin{align*}
    a & \rightarrow \begin{cases} 
        aa : 1 - p \\
        b : p 
    \end{cases} \\
    b & \rightarrow \begin{cases} 
        bb : 1 - p \\
        a : p 
    \end{cases}
\end{align*}
\]

(4.1)

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

This model is first proposed in Ref.[25] as a model for spatial $1/f$ spectra in open dynamical systems. More properties of the model are discussed in
Replication and mutation

Ref.[26]. I will not repeat all the details here, only to outline the basic features which are essential to the main theme of this paper.

Eq.(4.1) 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 $\alpha$ and type $\beta$ 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 the elongation, and $\alpha'$ and $\beta'$ can be any two types that could be different from the type $\alpha$ and type $\beta$. The most general expression for this relation is a multi-distance matrix equation:

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

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.(4.2) $\{P_{\alpha\beta}(d)\}$, 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), assuming that on average, the distance $d$ is elongated to the distance $kd$, where $k$ is the average elongation ratio. For Eq.(4.1), $k = 2 - p$. Furthermore, assuming that distances $d'$s around the distance $(2 - p)d$ also contribute to the scaling function. With all these approximations, it can be shown [26] that the joint probability behaves like

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

and for Eq.(4.1)

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

The autocorrelation function is proportional to the joint probability (the mutual information function is roughly proportional to the square of the joint probability), which also decays as a power law function. 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.(4.1) 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 $c \approx 0$, $1 - c \approx 1$, and the power spectrum is called $1/f$ spectrum, or, $1/f$ noise. The curious thing about $1/f$ noise is that
Replication and mutation

it appears almost everywhere [30]. Our model suggests that it is possible to find spatial $1/f$ spectra in sequences produced by elongation and mutation, which can perhaps provide an insight to 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 to the sequence. It is also a probabilistic context-free Lindenmayer system, represented by the following:

$$
\begin{align*}
 a &\rightarrow \begin{cases} 
 ab & : 1 - p \\
 b & : p 
\end{cases} \\
 b &\rightarrow \begin{cases} 
 ba & : 1 - p \\
 a & : p 
\end{cases}
\end{align*}
$$

(4.5)

Fig.3 illustrates the sequence generation process.

The statistical properties of the sequences generated by Eq.(4.5) is quite different from those generated by Eq.(4.1). First of all, when $p = 0$ and if the initial seed is a single symbol, Eq.(4.1) generates a homogeneous sequence containing a string of the same symbols, whereas Eq.(4.5) generates an almost periodic sequence called Thue-Morse sequence [27, 28, 29]. Secondly, related to the first difference, the largest non-trivial eigenvalue of the transition matrix (largest in magnitude) for Eq.(4.5) is negative, compared with the positive value for Eq.(4.1). It can easily argued that this negative eigenvalue will introduce an oscillation term whose wavelength is varying with the distance. Third, in some sense, the order present in the Thue-Morse sequence is more easily destroyed by mutation than for the homogeneous sequence. The reason is that the order in the Thue-Morse sequence is the 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.(4.5) 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 condenses 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:

$$
\begin{align*}
 \ldots a \ldots &\rightarrow \begin{cases} 
 \ldots a \ldots a & : 1 - p \\
 \ldots b \ldots & : p 
\end{cases} \\
 \ldots b \ldots &\rightarrow \begin{cases} 
 \ldots b \ldots b & : 1 - p \\
 \ldots a \ldots & : p
\end{cases}
\end{align*}
$$

(4.6)

Fig.5 illustrates the sequence generation process.
Replication and mutation

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 be used with care.

Multiple copies of the same segment or the same gene in one single nucleic acid sequence is quite common [31, 32, 33]. It is also suggested that oligomeric repeat could be an early mechanism for the nucleic acid sequences to explore possible coding schemes [34]. Considering these facts, this type of models needs more attentions and theoretical investigations.

The sequence generated by Eq.(4.6) 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 functions of the limiting sequences are shown in Fig.6 with two different mutation rates. As mentioned above, this plot does not characterize all the structures in the sequence, since there exist correlations at much long distances. From the plot (at the mutation rate \(p = 0.01\)) one can see that the peaks supposedly at the multiples of 3 suffer a shift after \(d = 18\). The subtle structures are quickly destroyed with larger mutation probabilities.

The lack of the scaling in the limiting sequence is due to the lack of the scaling in the equation describing transition of the joint probabilities. Roughly speaking, the transition equation is like

\[
P_{\alpha\beta}(d')^{t+1} = 2 \sum_{d=d'} \lambda(d) P_{\alpha\beta}(d)^t + \sum_{d=0-d'} \lambda(d) P_{\beta\alpha}(d)^t,
\]

(4.7)

where \(\lambda(d)\) is the largest non-trivial eigenvalue of the corresponding transition matrix. Note that the index of 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.,

\[
\ldots a \ldots \rightarrow \begin{cases} \ldots a \ldots b : 1 - p \\ \ldots b \ldots : p \end{cases} \\
\ldots b \ldots \rightarrow \begin{cases} \ldots b \ldots a : 1 - p \\ \ldots a \ldots : p \end{cases}
\]

(4.8)

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 there is a tendency for the \( M(d) \) to decrease, but also the local peaks beyond \( d = 24 \) are shifted.

5. Mutual information functions of several human nucleotide sequences

As promised in the first section, I will present the mutual function of nucleic acids sequences. I would like to discuss two facts observed in the human nucleotide sequences: (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. For mutual information function of other nucleic acids sequences, especially those of the complete genomes, will be included in the forthcoming paper [35].

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 [36], and as an algorithm for speeding up the calculation of autocorrelation functions [37].

The autocorrelation function is defined only for numerical sequences. So the question of how to get a numerical series from the nucleic acids sequences has been handled in different ways. There are following approaches: (1) Using other physical quantities instead of the symbolic sequence [38, 39]. It is assumed 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 on a site, the numerical value on that site is 1, otherwise, the value is 0. So far, this approach is the most popular one [40, 41, 42, 36, 43]; (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. The autocorrelation function or the power spectrum for the three components of the vector sequence can be calculated, and the overall autocorrelation function takes contributions from that of each component [44]. This idea is very neat, but has not been applied to nucleic acids 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 Ref. [45]. Nevertheless, when 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 should be useful for the nucleic acids 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 [46]. I choose these sequences because they have relatively longer sequence lengths,
Replication and mutation

which makes the calculation of the joint probability as well as the mutual information more reliable. These 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 pair. 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 for all five exon sequences. Due to the finite statistics, even two uncorrelated variables can have residue mutual information [5]. 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 densities for all four symbols A (Adenine), C (Cytosine), G (Guanine), and T (Thymine).

The region of the crossing between two mutual information functions (one for the original nucleic acids sequence and another for the corresponding random sequence) gives the distance at which the correlation starts to become negligible. In other words, it is a good estimation of the correlation length. Roughly speaking, the correlation lengths for the five exon sequences are on the order of 10, except the 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 for all five intron sequences, as well as those for the corresponding random sequences. The correlation lengths seem to be around (or more than) 20, except the sequence HUMCFVII whose correlation length is substantially longer. In order to see how long the correlation length is, I plot the mutual information function for sequence HUMCFVII again (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).

By the way, Fig.9 and Fig.10 confirm the previous findings that correlation in nucleic acids sequences oscillates [40, 41], and the periodicity of the oscillation tends to be 3 for exon sequences and 2 for intron sequences [43] (see, in particular, the exon HUMHSP90 and the intron 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 acids sequence analysis (see, for example, [47]). It is known that intron and exon sequences do have different statistical properties, and it will be interesting to know if 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. This 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 actually good approximations for codon sequences. As a result, the correlation length is at most a few codon lengths, i.e., a few multiples of 3. A value of 10 for the correlation length is consistent with this picture.

In fact, the mutual information function for the letter sequences (all alphabets as well as the punctuations and the blank space) or letter-type sequences (with a smaller number of symbols, including only the vowel, consonant, punctuation and the blank space) of the English texts exhibit the similar behavior. Fig.12 shows the mutual information function of the JFK's speech (sequence length is equal to 7391, the text is taken from [48]). The correlation length is around 10 — also a few multiples of the average length of English words. More results of the mutual information functions of the letter sequences in English is in Ref.[49].

To understand 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. If this is true, the value of the correlation length should depend on how frequent this repetition occurs, and how long the repeated segment is. Another understanding of the long range correlation in introns might be that the 2- or 3-dimensional structures of RNA's require certain correlation in the primary sequence. The effects of the secondary structure of RNA sequences on the grammar of the formal language that describes the sequence is mentioned in [50]. It should be interesting to be able 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.

6. Partial 1/f spectrum in the nucleic acids sequence HUMCFVII

A 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 greatly. 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 (Fig.11(a)) at the multiples of 17. But other than this repetition, there seem to have many other repetitions in the sequence as well.

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

In order to calculate the power spectrum, I convert the 4-symbol sequence into 2-symbol sequences 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 $2^{12} = 4096$ (the original sequence length is 5640, which means 1544 bases are deleted, including the complete fourth segment and part

---

1 I thank C. Burk for discussions on this point.
Replication and mutation

of the third segment). For the program of calculating the power spectrum, see, e.g., Chapter 12 of Ref.[12]).

The two spectra are very similar, but the fitting of the 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 higher frequency components are more related to the periodic structures mentioned above, for example, the peak at $\log_{10}(f) = \log_{10}(4096/17) = 2.38$. Nevertheless, these higher frequencies spectral components are basically flat. The boundary separating the lower frequency $1/f$ spectrum and the higher frequency white spectrum is arbitrary, and is chosen by a personal judgement. The scaling of the $1/f$ spectrum spans a 1.5 decades, out of the total 3.3 decades ($\log_{10}(4096/2) = 3.31$). To emphasize that this $1/f$ spectrum is not perfect, 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 the protein coding sequences, because their correlation length is too short. Besides the intron segments, the "junk genes", the satellite sequences that are between two genes are also potential candidates. Unfortunately (maybe fortunate 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 acids sequences. These rules are of 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 the high level control and regulation, driven by evolutionary pressures, and involves other macromolecules. There are attempts to model the nucleic acid sequence manipulation in more complicated environment, see, for example, the approach of typogenetics [51, 52]. In this model, the sequence is being 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 [14]. It is not known 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 prebiotic environment since the high level instructions resulted in simple terms from the low level interaction of a population of sequences. There are several attempts to simulate the reaction networks, one example is the hypercycle [53], and another
is the autocatalytic networks [22, 23, 24]. Again, the statistical properties of the population of sequences are not the focal point of these studies (see, however, a recent study mentioned in Ref.[54]).

Even with the single-sequence, fixed-dynamical-rule models, there are many other variations which are potentially relevant to the evolution of nucleic acid sequences. In particular, adding insertions and deletions to the models should be desirable. Shepherd has studied the effects of introducing insertion to the periodic sequences [40]. 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 of the periodic sequence with defects decays linearly. In other words, the effect of the mismatch propagates linearly. On the other hand, the insertions will have very little effects if the original sequence is random.

In conclusion, this paper discusses the long range correlations generated by local replication followed by an insertion (elongation) or sequence replication followed by a condensation. 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 it is still a long way to go before we can comprehend all the statistical features of the 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 interests and studies on this subject.

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 the period the main ideas presented here were discussed, and I appreciate the financial support from the CNLS. I also thank Rick Bagley, Christian Burks, Chris Fields and Walter Fontana for discussions or recommending references, and Alan Lapede for providing the nucleic acids 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). The paper is dedicated to the memory of Li, Chunqing (1908–1990).

References


Replication and mutation


Replication and mutation


[34] 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.


Figure 1: Illustration of the sequence manipulation rule (4.1), 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 the sequences generated by rule (4.1) 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 (4.5), 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 the sequences generated by rule (4.5) 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 (4.6), in which a symbol can either make an extra copy of the same symbol (solid arrows), or do not copy but mutate itself (shaded arrows), then the copies sequence is linked to the original sequence (encircled by the rectangle).

Figure 6: Mutual information function $M(d)$ of the sequences generated by rule (4.6) 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 (4.8), 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 copies sequence is linked to the original sequence (encircled by the rectangle).

Figure 8: Mutual information function $M(d)$ of the sequences generated by rule (4.8) 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 $N = 100,000$. 

$M(d)$ (in log scale) 

$d$ (in log scale)
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 sequence.
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 sequence.
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 the $M(d)$ of the corresponding random sequence; (b) the same HUMCFVII intron sequence with the period 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 the vowel, consonant, punctuation and the blank space. The sequence length is $N = 7391$. Also shown is the $M(d)$ of the corresponding random sequence.
Figure 13: The location of the four intron segments of the HUMCFVII and the sequence itself.
Figure 14: The power spectrum $P(f)$ of the binary sequence derived from 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 the total 5640 bases. Half of the Fourier components are redundant, and only the first half of the spectrum is plotted (the maximum value on the x-axis is $\log_{10}(4096/2) = 3.31$). Four neighboring spectrum components are averaged into one value (which leaves $2^{9} = 512$ on the plot). The best fitting line for the first 20 points using power law function $P(f) \sim 1/f^{\alpha}$ gives $\alpha \approx 0.93$.

Figure 15: The power spectrum $P(f)$ of the binary sequence derived from 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 fitting line for the first 20 points (out of 512 spectrum components) using power law function $P(f) \sim 1/f^{\alpha}$ gives $\alpha \approx 0.76$. 