Multiscaling in Expansion-modification Systems: An Explanation for Long Range Correlation in DNA

Ricardo Mansilla[*]
Faculty of Mathematics and Computer Science,
University of Havana, Cuba
and
Institute of Physics, UNAM, Mexico

Germinal Cocho
Institute of Physics, UNAM, Mexico

Inspired by a model first studied by W. Li in [7], scaling properties of the correlation function for expansion-modification systems are developed. The existence of several characteristic exponents is proved and the relationship of this fact with long-range correlations in DNA is established. Comparison between theoretical exponents and those obtained from numerical experiments and real biological sequences is also included.

1. Introduction

The discovery of the DNA molecule has revolutionized our way of thinking about biological evolution [1]. One of the most important challenges of our times is the understanding of its dynamics, because, notwithstanding the huge amount of sequenced base pairs, the rate of understanding the function of this data is far behind its rate of acquisition.

The studies of long-range correlation in DNA [2–5, 8–17] have been a fruitful line of research in recent years. A symbolic sequence is said to have long-range correlations if its power spectrum scales as $1/d^c$, where d is the distance between symbols in the sequence and $c \approx 1$. It is well known that the power-law correlation function is equivalent to the power-law power spectrum [19]. Hence, sometimes long-range correlations are defined in terms of a power-law spectrum.

Most of the papers about this subject report experimental evidence on long-range correlations [3–5, 8, 9, 14, 16]. Among them, accurate numerical studies have shown that the spectrum of real sequences [3], as well as those simulated by computers [8, 9], do not fit very well to

[*] Electronic mail address: mansilla@fenix.ifisicacu.unam.mx.

a power-law function. In [3] the existence of three spectral regimes corresponding to relatively high-frequency, middle-frequency, and low-frequency scaling regions are reported. Li's simulations [7, 19] also suggest that power spectrums possess more than one exponent and the analytical approximation of the real exponent [19] does not fit very well to data. In general, for low frequencies as well as for the high ones, the power spectrum "breaks" [3] and a better fit is obtained with a function of the type $\sum_i 1/f^{c_i}$ where f stands for frequency.

Only a few papers have developed theoretical models designed to explain long-range correlations and, as far as we know, nothing has been done to understand the existence of several regimes in power spectrums and hence in correlation functions.

The aim of this paper is to prove that for the model proposed in [19], the correlation function has behavior of the form $\sum_i K_i/d^{\phi_i(d)}$. We obtain constants K_i and asymptotic upper and lower bounds for the functions $\phi_i(d)$. We also prove that one of these exponents has a more important contribution to correlation functions than the others and show its fitness with respect to those obtained in simulations and in real sequences.

The structure of this paper is as follows. In section 2 we present the model and develop our main results. We obtain a closed form for the correlation function and also asymptotic expressions for its exponent. Section 3 is for discussion and section 4 for conclusions. All the heavy calculations can be found in the appendices.

2. Fundamental results

The evolution of nucleotide sequences is an instance in which two competing processes play an important role in determining the statistical properties of the sequences. Among the group of modifications that DNA sequences experience, replications and point mutations are, in some sense, antagonistic mechanisms. Replications insert substrings in several sites creating long-range correlations, while point mutations tend to destroy them. If the evolution of DNA consisted just of replications, the limiting sequence would be periodic; if on the other hand the point mutation rate was too high, the limiting sequence would be random. Only when the two processes are in an appropiate balance do the nucleotide sequences show nontrivial long-range correlations as observed in nature.

Hence, a model which describes that feature of real sequences should take into account replication and point mutation. We will consider in our work binary sequences. This fact does not reduce the scope of our conclusions because there are several ways of coding DNA sequences using binary digits [3]. Moreover, that kind of coding has been used to study long-range correlations in DNA.

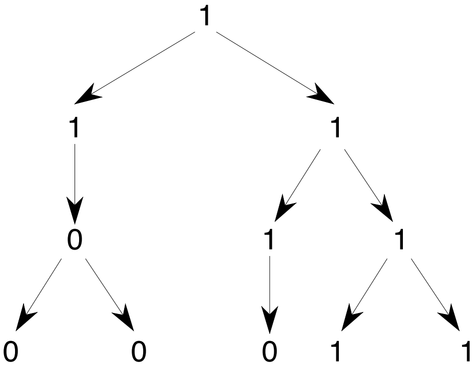

Figure 1. A realization of the expansion-modification process.

The simplest expansion-modification system [17] is the two-symbol system in which the expansion process rewrites one symbol into two identical symbols and the modification process switches one symbol to another different symbol. This process captures the mechanism behind point mutations and insertions and is in good agreement with some experimental facts [11] about the length of inserted strings and some features of the point mutation dynamics.

We first introduce some notations. Let $x^t = \cdots\alpha_0^t\alpha_1^t\cdots$ be a binary sequence. It is mapped in $x^{t+1} = \cdots\alpha_0^{t+1}\alpha_1^{t+1}\cdots$ by means of the following rules: each symbol α_i^t is substituted by two identical symbols with probability $1-p$ or switches to the other symbol with probability p, that is,

$$0 \to \begin{cases} 00 & 1-p \\ 1 & p \end{cases}; \qquad 1 \to \begin{cases} 11 & 1-p \\ 0 & p. \end{cases}$$

A particular realization of this process is shown in Figure 1. If α' and β' represent two symbols of x^t separated by distance d', upon applying the rewriting rule, this symbol pair would lead to another symbol pair α,β in x^{t+1} separated by a longer distance d. Let $P_{\alpha,\beta}^t(d)$ be the joint probability of having the symbol pair α,β separated by the distance d.

Assuming that the transition probability from an α',β' pair initially at distance d' to an α,β pair at distance d is $T(\alpha',\beta',d' \to \alpha,\beta,d)$, the joint probability satisfies the dynamical equation:

$$\begin{bmatrix} P_{0,0}^{t+1}(d) \\ P_{0,1}^{t+1}(d) \\ P_{1,0}^{t+1}(d) \\ P_{1,1}^{t+1}(d) \end{bmatrix} = \sum_{d'=[d/2]}^{d} \begin{bmatrix} T(00 \to 00) & . & . & T(11 \to 00) \\ . & . & . & . \\ . & . & . & . \\ T(00 \to 11) & . & . & T(11 \to 11) \end{bmatrix} \begin{bmatrix} P_{0,0}^{t}(d') \\ P_{0,1}^{t}(d') \\ P_{1,0}^{t}(d') \\ P_{1,1}^{t}(d') \end{bmatrix}, \qquad (2.1)$$

where we have written $T(\alpha',\beta' \to \alpha,\beta)$ instead of $T(\alpha',\beta',d' \to \alpha,\beta,d)$ for simplicity. From now on, the square brackets [.] stand for the integer part of a real number. The summation index in equation (2.1) begins at $d' = [d/2]$ because symbols α', β' initially at a distance smaller than $[d/2]$ cannot transform into α, β at distance d.

The transition probabilities $T(\alpha', \beta', d' \to \alpha, \beta, d)$ can be grouped into three types.

- $T_0(d', d, p)$: Keep both symbols unchanged, for instance: $T(0, 0 \to 0, 0)$.
- $T_1(d', d, p)$: Change one symbol, for instance: $T(0, 1 \to 1, 1)$.
- $T_2(d', d, p)$: Change both symbols, for instance: $T(1, 1 \to 0, 0)$.

Hence, equation (2.1) can be written as:

$$P^{t+1}(d) = \sum_{d'=[d/2]}^{d} T(d', d, p) P^t(d') \tag{2.2}$$

where

$$P^t(d) = \begin{bmatrix} P^t_{0,0}(d) \\ P^t_{0,1}(d) \\ P^t_{1,0}(d) \\ P^t_{1,1}(d) \end{bmatrix}$$

$$T(d', d, p) = \begin{bmatrix} T_0 & T_1 & T_1 & T_2 \\ T_1 & T_0 & T_2 & T_1 \\ T_1 & T_2 & T_0 & T_1 \\ T_2 & T_1 & T_1 & T_0 \end{bmatrix}.$$

In this matrix we have written T_S instead of $T_S(d', d, p)$ in order to simplify notation.

Suppose there is a time invariant condition in the $t \to +\infty$ limit and the superscript can be dropped [19]. Then equation (2.2) takes the following form:

$$P(d) = \sum_{d'=[d/2]}^{d} T(d', d, p) P(d') \tag{2.3}$$

or

$$P(d) = (I - T(d, d))^{-1} \sum_{d'=[d/2]}^{d-1} T(d', d, p) P(d'). \tag{2.4}$$

Equation (2.4) is a recursive definition for vector $P(d)$ that is difficult to handle. We now obtain a closed form expression for $P(d)$. Let $G(d, n)$ be the multiple-index set $(i_1, \ldots, i_n)$ which holds the following conditions.

1. $i_1 = 1,\ i_n = d$.
2. $i_1 < i_2 < \cdots < i_n$.
3. For every $l = 1, \ldots, n-1$: $[i_{l+1}/2] \le i_l$.

Then it can be proved (see appendix A1 for the details) that:

$$P(d) = \left\{ \sum_{n=2}^{d} \sum_{(i_1,\ldots,i_n)\in G(d,n)} M(i_{n-1}, i_n)\ldots M(i_1, i_2) \right\} P(1) \tag{2.5}$$

where $I(d) = (I - T(d,d,p))^{-1}$ and $M(d',d) = I(d)T(d',d,p)$. We ignore the dependence of $M(d',d)$ on p just for simplicity. As we also show in appendix A1, the summation in equation (2.5) could begin at $n = 1 + [\log_2 d]$. See equation (A1.9) and the discussion above.

We would like to stress the meaning of the $G(d,n)$ sets. Matrix $M(d',d)$ is in some sense a transition matrix. Therefore the product $M(i_{n-1}, i_n)\ldots M(i_1, i_2)$ is the transition probability matrix from an α',β' pair initially side by side to reach the distance d in n time steps (recall that $i_1 = 1$ and $i_n = n$ always). Then $G(d,n)$ represents the set of ways to reach a distance d starting from distance 1 in n time steps.

Note that $P(d)$ depends on $P(1)$. In sequences of four symbols (such as DNA sequences), $P(1)$ is related to the dimers structure. As we have shown [11, 20] $P(1)$ distinguishes the noncoding regions of the DNA molecule. That is why we use it in [11] as a fitness function for an evolutionary model studied there. Because the matrices in equation (2.5) depend on $M(d',d)$ our next step is to study the structure of matrix $T(d',d,p)$. It can be proved that:

$$T(d',d,p) = \begin{bmatrix} T_0 & T_1 & T_1 & T_2 \\ T_1 & T_0 & T_2 & T_1 \\ T_1 & T_2 & T_0 & T_1 \\ T_2 & T_1 & T_1 & T_0 \end{bmatrix} = Q \begin{bmatrix} \pi_1 & 0 & 0 & 0 \\ 0 & \pi_2 & 0 & 0 \\ 0 & 0 & \pi_2 & 0 \\ 0 & 0 & 0 & \pi_3 \end{bmatrix} Q^{-1} = QDQ^{-1}$$

where:

$$\pi_1 = T_0 + 2T_1 + T_2; \quad \pi_2 = T_0 - T_2; \quad \pi_3 = T_0 - 2T_1 + T_2 \tag{2.6}$$

$$Q = \begin{bmatrix} 1 & -1 & 0 & 1 \\ 1 & 0 & -1 & -1 \\ 1 & 0 & 1 & -1 \\ 1 & 1 & 0 & 1 \end{bmatrix}; \qquad Q^{-1} = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ -1/2 & 0 & 0 & 1/2 \\ 0 & -1/2 & 1/2 & 0 \\ 1/4 & -1/4 & -1/4 & 1/4 \end{bmatrix}.$$

In the same way:

$$I(d) = Q \begin{bmatrix} (1-v_1)^{-1} & 0 & 0 & 0 \\ 0 & (1-v_2)^{-1} & 0 & 0 \\ 0 & 0 & (1-v_2)^{-1} & 0 \\ 0 & 0 & 0 & (1-v_3)^{-1} \end{bmatrix} Q^{-1} = Q\Xi Q^{-1}$$

where v_1, v_2, and v_3 are the eigenvalues of the matrix $T(d,d,p)$. Using the above expression, equation (2.5) can be written as:

$$P(d) = Q \left\{ \sum_{n=[\log_2 d]+1}^{d} \sum_{(i_1,..i_n)\in G(d,n)} \Lambda(i_{n-1}, i_n)\ldots\Lambda(i_1, i_2) \right\} Q^{-1}P(1) \tag{2.7}$$

where $\Lambda(i_{k-1}, i_k) = \Xi(i_k)D(i_{k-1}, i_k)$.

Denote by:

$$H(d) = \sum_{n=[\log_2 d]+1}^{d} \sum_{(i_1,\ldots,i_n)\in G(d,n)} \Lambda(i_{n-1}, i_n)\ldots\Lambda(i_1, i_2)$$

$$H(d) = \begin{bmatrix} H_1(d) & 0 & 0 & 0 \\ 0 & H_2(d) & 0 & 0 \\ 0 & 0 & H_2(d) & 0 \\ 0 & 0 & 0 & H_3(d) \end{bmatrix}$$

where

$$H_k(d) = \sum_{n=[\log_2 d]+1}^{d} \sum_{(i_1,\ldots,i_n)\in G(d,n)} \delta_k(i_1, \ldots, i_n) \tag{2.8}$$

$$\delta_k(i_1, \ldots, i_n) = \frac{\pi_k(i_{n-1}, i_n)\ldots\pi_k(i_1, i_2)}{(1 - v(i_n))\ldots(1 - v(i_2))}. \tag{2.9}$$

Equation (2.5) allows us to have closed expressions for $P_{\alpha,\beta}(d)$ and hence for the correlation function $\Gamma(d)$ [21]. As proven in the following result, $\Gamma(d)$ only depends on $H_1(d)$ and $H_3(d)$ for almost all sequences.

Proposition 1. Let $\Theta = \cdots\alpha_{-1}\alpha_0\alpha_1\alpha_2\cdots$ be an infinite string of binary symbols. Let us denote by $P_0(\Theta), P_1(\Theta)$ the densities of zeros and ones respectively in string Θ. Then if $P_1(\Theta) \neq 1/2$ we have $H_2(d) \equiv 1$.

Proof. We will use some properties of probabilities $P_{\alpha,\beta}(d)$ in strings of binary symbols [21]. For every $d \geq 1$:

$$P_{0,1}(d) = P_{1,0}(d)$$
$$P_{0,0}(d) = 1 - 2P_1(\Theta) + P_{1,1}(d)$$
$$P_{1,0}(d) = P_1(\Theta) - P_{1,1}(d).$$

It is not difficult to see that:

$$\begin{bmatrix} P_{0,0}(d) \\ P_{0,1}(d) \\ P_{1,0}(d) \\ P_{1,1}(d) \end{bmatrix} = \frac{1}{4} M(P_1(\Theta), P_{1,1}(1)) \begin{bmatrix} H_1(d) \\ H_2(d) \\ H_3(d) \end{bmatrix} \tag{2.10}$$

where

$$M(P_1(\Theta), P_{1,1}(1)) = \begin{bmatrix} 1 & 2(1 - 2P_1(\Theta)) & 1 - 4P_1(\Theta) + 4P_{1,1}(1) \\ 1 & 0 & -(1 - 4P_1(\Theta) + 4P_{1,1}(1)) \\ 1 & 0 & -(1 - 4P_1(\Theta) + 4P_{1,1}(1)) \\ 1 & -2(1 - 2P_1(\Theta)) & 1 - 4P_1(\Theta) + 4P_{1,1}(1) \end{bmatrix}.$$

Now because of $P_{00}(d) = 1 - 2P_1(\Theta) + P_{1,1}(d)$, from equation (2.10) we have:

$$(1 - 2P_1(\Theta))H_2(d) = 1 - 2P_1(\Theta).$$

This completes the proof. ■

Therefore we could write the correlation function as

$$\Gamma(d)=H_1(d)+(1-4P_1(\Theta)+P_{1,1}(d))H_3(d)-(P_1^2(\Theta)-4P_1(\Theta)+2). \qquad (2.11)$$

Our next step is to obtain the upper and lower bounds for $H_1(d)$ and $H_3(d)$ as a function of $\delta_k(i_1,\ldots,i_n)$. Hence upper and lower bounds for $\delta_k(i_1,\ldots,i_n)$ would become the upper and lower bounds for $H_1(d)$ and $H_3(d)$. We briefly introduce some notations.

Let $(i_1,\ldots,i_n)\in G(d,n)$ and $d_0\in \mathrm{N}$. Let us denote by $l(i_1,\ldots,i_n)$ the set of indices which are smaller than d_0 and by $u(i_1,\ldots,i_n)$ those which are bigger:

$$l(i_1,\ldots,i_n)=\{i_r\in(i_1,\ldots,i_n)\colon i_r<d_0\}$$
$$u(i_1,\ldots,i_n)=\{i_r\in(i_1,\ldots,i_n)\colon i_r\geq d_0\}.$$

It can be proved (see appendix A2 for details) that:

$$L_j(p,d_0,d,n)\geq\delta_j(i_1,\ldots,i_n)\leq U_j(p,d_0,d,n)\qquad j=1,3$$

where

$$L_j(p,d_0,d,n)=c_j(i_1,\ldots,i_n)\Phi_l^j(p,d_0,d,n)\frac{e^{-\left(\frac{1-p}{2p}S(d,n)\right)}}{M(d,n)} \qquad (2.12)$$

$$U_j(p,d_0,d,n)=c_j(i_1,\ldots,i_n)\Phi_u^j(p,d_0,d,n)\frac{e^{-\left(\frac{p}{2(1-p)}S(d,n)\right)}}{M(d,n)} \qquad (2.13)$$

$$\Phi_l^j(p,d_0,d,n)=\prod_{l(i-1,\ldots,i_n)}\frac{\phi_l^j(p)}{1-v_j(i_k)}$$

$$\Phi_u^j(p,d_0,d,n)=\prod_{l(i_1,\ldots,i_n)}\frac{\phi_u^j(p)}{1-v_j(i_k)}$$

$$S(d,n)=\sum_{u(i_1,\ldots,i_n)}(i_k-1)$$

$$M(d,n)=\prod_{u(i_1,\ldots,i_n)}(i_k-1)$$

$$\phi_l^1(p)=\frac{e^{\left(\frac{1}{2(1-p)}\right)}+(1-p)^2e^{-\left(\frac{3}{2p}\right)}+2(1-p)}{\sqrt{2\pi p(1-p)}}$$

$$\phi_u^1(p)=\frac{e^{\left(\frac{1}{p}\right)}+(1-p)^2e^{\left(\frac{3}{1-p}\right)}+2(1-p)}{\sqrt{2\pi p(1-p)}}$$

$$\phi_l^3(p)=\frac{e^{\left(\frac{1}{2(1-p)}\right)}+(1-2p)^2e^{-\left(\frac{3}{2p}\right)}-2(1-2p)}{\sqrt{2\pi p(1-p)}}$$

$$\phi_u^3(p)=\frac{e^{\left(\frac{1}{p}\right)}+(1-2p)^2e^{\left(\frac{3}{1-p}\right)}+2(1-2p)}{\sqrt{2\pi p(1-p)}}.$$

Remark. It is easy to see that $\phi_l^3(p) \geq 0$ if $0.0716 \leq p$. In all that follows we suppose that p satisfies the above mentioned condition. Notice that the allowed values of p covers 92.84 percent of the interval $[0, 1]$.

We would also like to note that, as shown in equations (2.12) and (2.13), upper and lower bounds for $\delta_j(i_1, \ldots, i_n)$ depend on $S(d, n)$ and $M(d, n)$. We will obtain upper and lower bounds for $S^*(d, n) = \sum_{(i_1,\ldots,i_n)\in G(d,n)}(i_k - 1)$ and $M^*(d, n) = \prod_{(i_1,\ldots,i_n)\in G(d,n)}(i_k - 1)$, which together with equations (2.12) and (2.13), will be used to obtain uniform bounds for $\delta_j(i_1, \ldots, i_n)$ on $G(d, n)$.

The expression $S^*(d, n)$, as well as $M^*(d, n)$, reach their maximum values in those members of the set $G(d, n)$ in which the last elements are consecutive, for example, $i_n = d$, $i_{n-1} = d - 1$, $i_{n-2} = d - 2$, and so on. This condition constrains the first ones to be as sparse as possible, but fulfilling the condition $[i_{k+1}/2] = i_k$, for example,

$$(i_1, i_2, \ldots, i_{n-r-1}, i_{n-r}, \ldots, i_n) = (1, 3, \ldots, 2^{n-r-1} - 1, d - r, \ldots, d).$$

Therefore we should find an index r such that:

$$\begin{cases} \left[\frac{d-r}{r}\right] \leq 2^{n-r-1} - 1 \\ 2^{n-r-2} - 1 < \left[\frac{d-r-1}{2}\right]. \end{cases} \tag{2.14}$$

If $d - r$ is even, the above conditions are equivalent to:

$$\begin{cases} d + 2 \leq 2^{n-r} + r \\ 2^{n-(r+1)} + r + 1 \leq d. \end{cases} \tag{2.15}$$

If $d - r$ is odd, the conditions of equation (2.14) are equivalent to:

$$\begin{cases} d + 1 \leq 2^{n-r} + r \\ 2^{n-(r+1)} + r + 1 \leq d + 1. \end{cases} \tag{2.16}$$

In order to obtain such an index r we will solve the equation

$$2^{n-r} + r = d + 1. \tag{2.17}$$

Figure 2 shows the graph of the difference between solutions of the equations:

$$2^{n-r} + r = d$$
$$2^{n-r} + r = d + 2$$

as a function of d. This justifies the search of index r using equation (2.17).

It can be proved that the only feasible solution of equation (2.17) can be expressed by means of the following series expansion:

$$r = \sum_{k=1}^{\infty} \frac{1}{k!} \frac{(\ln 2)^k P_k(d)}{(d \ln 2 - 1)^{2k-1}} \left(n - \frac{\ln d}{\ln 2}\right)^k \tag{2.18}$$

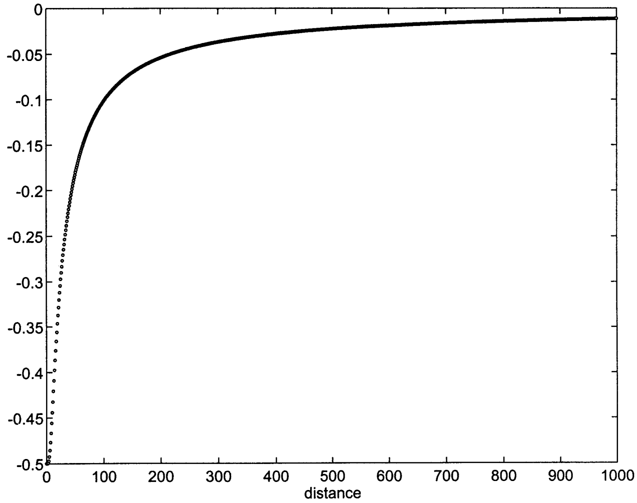

Figure 2. Difference between solutions of equations $2^{n-x}+x=d$ and $2^{n-x}+x=d+2$ for $n=15$. As can be seen, this difference tends to zero as d tends to infinity. This justifies the use of equation (2.17).

where $P_k(d)$ is a polynomial in d of degree $k-1$. The expression $n-(\ln d/\ln 2)$ is always positive because $n\geq[\log_2 d]+1$.

The above series could be written as:

$$r=\sum_{k=1}^{\infty}\frac{1}{k!}\frac{P_k(d)}{(d\ln 2-1)^{k-1}}\left(\frac{n\ln 2-\ln d}{d\ln 2-1}\right).$$

Let us note that

$$0\leq\frac{n\ln 2-\ln d}{d\ln 2-1}\leq\frac{d\ln 2-\ln d}{d\ln 2-1}$$

and it is easy to see that

$$\lim_{d\to\infty}\frac{P_k(d)}{(d\ln 2-1)^k}=\frac{1}{\ln 2}.$$

Therefore, as a solution of equation (2.17) we will use the approximation done by the first term of the series in equation (2.18):

$$r_u=\frac{d\ln 2}{d\ln 2-1}\left(n-\frac{\ln d}{\ln 2}\right). \tag{2.19}$$

We want to remark on the accuracy of this approximation. In Table 1 are shown some couples of (d,n), the exact values of $n-r$, and the approximation obtained from equation (2.19) (labeled $n-r_u$).

d	n	$n-r$	$n-r_u$
30	10	5	4.64
30	20	4	4.14
33	11	5	4.77
60	20	6	5.55
60	40	5	5.06
62	7	6	5.92
80	13	7	6.19
81	18	7	6.12
120	40	7	6.50
120	80	6	6.01
240	80	8	7.47
240	160	7	6.98

Table 1. Exact values of $n-r$ and their estimates $n-r_u$.

From the results obtained we could have upper bounds for $S^*(d,n)$ and $M^*(d,n)$:

$$S^*(d,n) \leq S_u(d,n) \text{ and } M^*(d,n) \leq M_u(d,n) \tag{2.20}$$

where:

$$S_u(d,n) = d^{\left(1-\frac{n}{d\ln d}\right)} + 2d\left(n+2-\frac{\ln d}{\ln 2}\right) + \frac{\ln d}{\ln 2}\left(2n-1+\frac{\ln d}{\ln 2}\right) \tag{2.21}$$

$$M_u(d,n) = 2^{\left(\frac{(n-r_u)(n-r_u-1)}{2}\right)}\frac{d!}{(d-r_u-1)!}. \tag{2.22}$$

We proceed in the same way for the lower bounds. The expression $S^*(d,n)$, as well as $M^*(d,n)$, reach their minimum values in those members of the set $G(d,n)$ in which the last elements are as sparse as possible. This condition constrains the first ones to be consecutive, because the condition $i_k < i_{k+1}$, $k = 1,\ldots,n-1$ must be fulfilled:

$$(i_1,\ldots,i_r,i_{r+1},\ldots,i_n) = (1,\ldots,r,i_{r+1},\ldots,i_n)$$

where:

$$\left[\frac{i_{l+1}}{2}\right] \leq i_l; \quad l = r+1,\ldots,n-1.$$

Therefore, we should find an index r such that:

$$\begin{cases} 2^{n-(r+1)}(r+1) < d+1 \\ d+1 < 2^{n-r}r. \end{cases} \tag{2.23}$$

To obtain such an index, we will solve the equation

$$2^r(n-r) = d+1. \tag{2.24}$$

It can be proved that the only feasible solution of equation (2.24) is expressed by means of the following series expansion:

$$r = \sum_{k=1}^{\infty} \frac{1}{k!} \frac{(\ln 2)^k P_k(d)}{(d \ln 2 - 1)^{2k-1}} (d-n)^k \tag{2.25}$$

where $P_k(d)$ is a polynomial of degree $k-1$. More precisely:

$$P_k(d) = (k-1)!(\ln 2)^{k-2} d^{k-1} + \cdots + d.$$

Besides, it is not difficult to see that

$$0 \le \frac{d \ln 2 - n \ln 2}{d \ln 2 - 1} \le \frac{d \ln 2 - \ln d}{d \ln 2 - 1} \le 1$$

and

$$\lim_{d \to \infty} \frac{P_k(d)}{(d \ln 2 - 1)^k} = \frac{1}{\ln 2}.$$

Hence, for d large enough:

$$r \approx \frac{1}{\ln 2} \sum_{k=1}^{\infty} \frac{1}{k} \left(\frac{d \ln 2 - n \ln 2}{d \ln 2 - 1} \right)^k = \frac{1}{\ln 2} \ln \left[\frac{d \ln 2 - 1}{n \ln 2 - 1} \right]. \tag{2.26}$$

We will take this as an approximation to the solution of equation (2.24). In Figure 3 we present the graphs of the whole series (upper plot) and the approximation by equation (2.26) (lower plot) for $d = 1024$ and $11 \le n \le 1000$. In Table 2 we present some couples of (d, r), the corresponding values of r, and the approximation obtained from equation (2.26) (labeled r_l).

From all of the above:

$$S^*(d,n) \ge S_l(d,n); \qquad M^*(d,n) \ge M_l(d,n) \tag{2.27}$$

where

$$S_l(d,n) = \frac{(n - r_l)(n - r_l - 1) - 2n}{2} + (d+1) \left\{ \frac{(d-n) \ln 2 - 1}{2d \ln 2 - 1} \right\}$$
$$- \frac{1}{\ln 2} \ln \left[\frac{d \ln 2 - 1}{n \ln 2 - 1} \right] + 1 \tag{2.28}$$

$$M_l(d,n) = d(n - r_l - 1)!(d+1)^{r_l} 2^{-\left(\frac{r_l(r_l+1)}{2}\right)} e^{\left(\frac{2}{d+1}(2^{r_l}-1)\right)}. \tag{2.29}$$

We finally have all the elements to construct upper and lower bounds for $H_j(d)$. Let us denote the following:

$$c_j^u(d_0, n) = \max_{(i_1, \ldots, i_n) \in G(d,n)} c_j(i_1, \ldots, i_n)$$
$$c_j^l(d_0, n) = \min_{(i_1, \ldots, i_n) \in G(d,n)} c_j(i_1, \ldots, i_n)$$

$$\begin{aligned}
Q^l(d_0, n) &= \min_{(i_1,\ldots,i_n)\in G(d,n)} W^l(i_1,\ldots,i_n)\\
Q^u(d_0, n) &= \max_{(i_1,\ldots,i_n)\in G(d,n)} W^u(i_1,\ldots,i_n)\\
W^l(i_1,\ldots,i_n) &= \prod_{l(i_1,\ldots,i_n),\ k\neq 1} \sqrt{i_k - 1}\, e^{\left(\frac{1-p}{2p}\sum_{l(i_1,\ldots,i_n)}(i_k-1)\right)}\\
W^u(i_1,\ldots,i_n) &= \prod_{l(i_l,\ldots,i_n),\ k\neq 1} \sqrt{i_k - 1}\, e^{\left(\frac{p}{2(1-p)}\sum_{l(i_1,\ldots,i_n)}(i_k-1)\right)}.
\end{aligned}$$

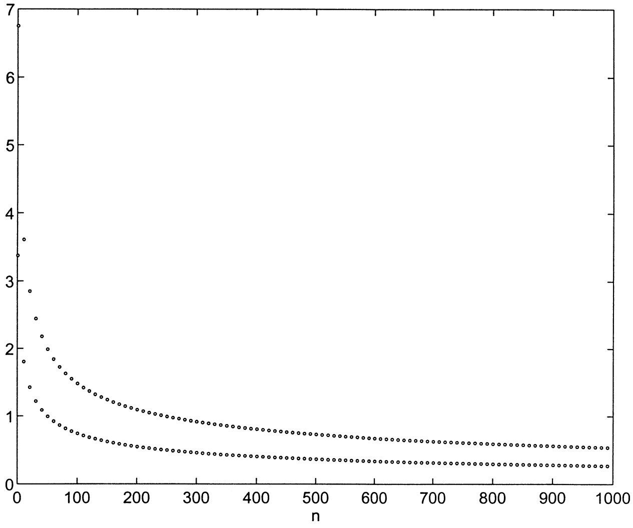

Figure 3. Graph of the whole series (upper plot) and the approximation (lower plot) for $d = 1024$ and $11 \leq n \leq 1000$.

d	n	$n-r$	$n-r_l$
30	10	9	8.26
30	20	20	19.37
33	11	11	9.27
60	20	19	18.34
60	40	40	39.39
62	7	4	3.55
80	13	11	10.23
81	18	16	15.73
120	40	39	38.77
120	80	80	79.40
240	80	79	78.34
240	160	160	159.41

Table 2. Exact values of $n-r$ and their estimates $n-r_l$.

Then it can be proved (see appendix A3 for the details) that:

$$H_j(d) \le \sum_{n=[\log_2 d]+1}^{d} c_j^u(d_0,n) Q^u(d_0,n) e^{-\epsilon_u(d,n,p)} \tag{2.30}$$

$$H_j(d) \ge \sum_{n=[\log_2 d]+1}^{d} c_j^l(d_0,n) Q^l(d_0,n) e \tag{2.31}$$

where

$$\epsilon_u(d,n,p) = \epsilon_u^a(d,n,p) + o\left(\frac{1}{d},\frac{1}{n}\right) \tag{2.32}$$

$$\epsilon_l(d,n,p) = \epsilon_l^a(d,n,p) + o\left(\frac{1}{d},\frac{1}{n}\right) \tag{2.33}$$

$$\begin{aligned}\epsilon_u^a(d,n,p) = {} & \frac{pd}{4(1-p)} + \frac{n^2}{2} + n\left[2\ln n + \left(1+\frac{1}{\ln 2}\right)(\ln n - \ln d)\right. \\ & \left. -\frac{5-2p}{2(1-p)}\right] + \frac{1}{2}\left(1+\frac{1}{\ln 2}\right)\left(\frac{\ln d}{\ln 2}\right)^2 \\ \epsilon_l^a(d,n,p) = {} & \frac{d}{2}\left[1+(2n+5)\left(\frac{1-p}{p}\right)\right] \\ & +\frac{1}{2}\frac{\ln d}{\ln 2}\left[n + \frac{1-p}{p}\left(1-\frac{d}{2}\right) - \frac{\ln 2}{2}\right] + \frac{1}{2}(2n-1)\frac{1-p}{p}.\end{aligned}$$

Note that each term of the sums in equations (2.30) and (2.31) correspond to different $G(d,n)$ sets. Each of them has associated exponents $\epsilon_u(d,n,p)$ and $\epsilon_l(d,n,p)$. As we prove below, certain sets of $G(d,n)$ make a major contribution to the function $H_j(d)$. We also prove the uniqueness of such a set $G(d,n)$.

If the symbols α,β are at distance d', then between them there are $d'-1$ other symbols. Taking into account the six possible cases mentioned in appendix A2, it is easy to prove that the expected values for the distance is $\overline{d} = (d'-1)(2-p)+2$. Therefore, for a certain value of n there exists an element $(i_1,\ldots,i_n) \in G(d,n)$, such that, for every $k = 1,\ldots,n-1$: $i_{k+1} = [(2-p)(i_k-1)+2]$.

Let $d \in \mathrm{N}$ and $0 < p < 1$. Let $n(d,p)$ be the positive integer such that the set $G(d,n(d,p))$ contains the element $(i_1,\ldots,i_{n(d,p)})$ which holds the following condition: for every $k = 1,\ldots,n(d,p)-1$: $i_{k+1} = [(2-p)(i_k-1)+2]$. The set $G(d,n(d,p))$ represents, given d and p, the most probable ways that two symbols reach the distance d along the sequence in $n(d,p)$ steps.

Proposition 2. For $p \to 0$ and $d \to \infty$ we have:

$$n(d,p) = 1 + \left[\frac{\ln(d(1-p)+p)}{\ln(2-p)}\right]. \tag{2.34}$$

Proof. Although $i_{k+1} = [(i_k - 1)(2 - p) + 2]$, we have:

$$i_{k+1} = (2 - p)i_k + p + \epsilon_k,$$

where $0 \leq \epsilon_k < 1$. It can be proved by induction that:

$$i_k = (2 - p)^{n-1} + p\left\{\frac{(2 - p)^{n-1} - 1}{1 - p}\right\} + \epsilon_1(2 - p)^{n-2} + \cdots + \epsilon_{k-1}.$$

In particular, for $k = n(d, p)$,

$$d = (2 - p)^{n(d,p)-1} + p\left\{\frac{(2 - p)^{n(d,p)-1} - 1}{1 - p}\right\}$$
$$+\epsilon_1(2 - p)^{n(d,p)-2} + \cdots + \epsilon_{n(d,p)-1}.$$

From this equation and the condition imposed on ϵ_i we have:

$$n(d, p) \leq 1 + \frac{\ln(d(1 - p) + p)}{\ln(2 - p)} < n(d, p)$$
$$+\frac{1}{\ln(2 - p)}\left\{\ln 2 + \ln\left(1 - \frac{1}{2(2 - p)^{n(d,p)-1}}\right)\right\}.$$

From this expression and for $d \to \infty, p \to 0$ we obtain equation (2.34). ■

Remark. Let us note that:

$$\lim_{p \to 0} n(d, p) = 1 + \left[\frac{\ln d}{\ln 2}\right] = 1 + [\log_2 d]$$
$$\lim_{p \to 1} n(d, p) = d$$

in agreement with the fact: $1 + [\log_2 d] \leq n \leq d$. It also agrees with the interpretation of p. If $p \to 0$ then $1 - p \to 1$ and expansion happens more often than modification, hence few steps are needed for a pair of symbols α', β' initially side by side on the sequence to reach the distance d. If $p \to 1$ then modification happens more often than expansion and many steps are needed to reach the distance d.

3. Discussion

We have obtained a good agreement with numerical simulations. In Figure 4 we show the graph of $\epsilon_u^a(d, n(d, p), p)$ for $p = 0.1$ (lower plot) and that obtained from averaging 10 simulations with the same values of p (upper plot) assuming that only the term corresponding to $G(d, n(d, p))$ exists in the simulation, for example, $\Gamma(d) \approx K/d^{\epsilon_u^a(d,n,p)}$ and $n = n(d, p)$. We want to remark on the coincidence in the shape of both plots. In fact, the difference between them is related with the terms that were not taken into account, for example, those with $n < n(d, p)$ and for

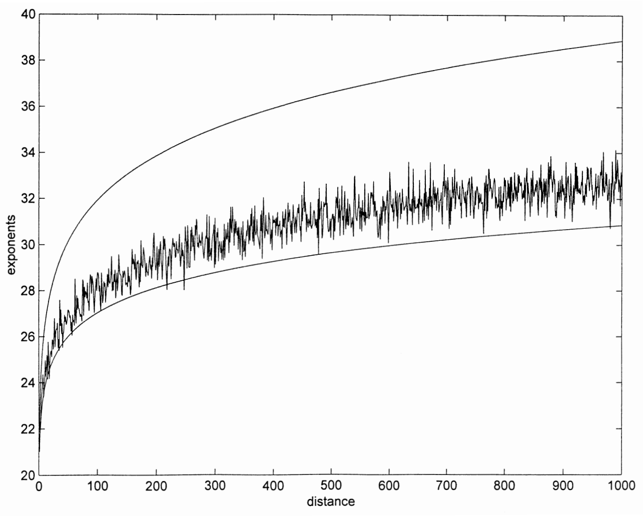

Figure 4. Graph of $\epsilon_u^a(d,n(d,p),p)$ (lower plot), $\epsilon_l^a(d,n(d,p),p)$ (upper plot) for $p = 0.1$, and that obtained from averaging 10 simulations with the same values of p (middle plot).

$n > n(d,p)$. In [4] a magnitude inversely proportional to the correlation function is studied. For that magnitude a fitness of the form $F(d) = d^{\phi(d)}$ is obtained. In Figure 1 of that paper a plot of the exponent $\phi(d)$ is shown. We also want to emphasize the coincidence in shape of that exponent and the lower plot in Figure 4 of our present work.

In Figure 1 of [3] the graph of the averaged power spectrums for all 33301 coding and for all 29453 noncoding sequences of the GENBANK larger than 512 bp is shown. As the authors remark, there are three spectral regimes. In our opinion, the existence of several exponents is the best explanation for that behavior.

Furthermore, the middle-frequency scaling region reported by Buldyrev *et al.* [3] corresponds to the terms in equation (2.5) related with $G(d,n(d,p))$ and some close neighbors ($n \approx n(d,p)$).

4. Conclusion

We have studied the correlation function of an expansion-modification system, which grasps the main features of the mutational process occurring in the evolution of the DNA molecule. We obtain bounds for the exponent in the correlation function and show the resemblance between the theoretical exponent and those obtained from simulation. We also give an explanation for the existence of several regions in the power spectrums of real sequences.

We have studied more complicated models with an alphabet of three symbols. A generalization of the rule studied in this paper might be:

$$0 \to \begin{cases} 00 & 1-p_1-p_2 \\ 1 & p_1 \\ 2 & p_2 \end{cases}$$
$$1 \to \begin{cases} 11 & 1-p_1-p_2 \\ 2 & p_1 \\ 0 & p_2 \end{cases}$$
$$2 \to \begin{cases} 22 & 1-p_1-p_2 \\ 0 & p_1 \\ 1 & p_2 \end{cases} .$$

The difficulties to obtain analytical results are enormous. In that case, transition matrices $T(d', d, p_1, p_2)$ are of order 9×9. However, we have made computer simulations and very interesting results arise [23]. At some distance $d(p_1, p_2)$, which depends on the selection of the probabilities p_1 and p_2, the correlation function has a positive maximum. Interestingly, this type of behavior is observed in real intergenic sequences [23].

Acknowledgments

The authors are thankful to G. Martinez-Meckler, R. Bujalich, and P. Miramontes for their helpful comments as well as to the anonymous referee. This work was partially (R. Mansilla) supported by CONACYT, Mexico.

Appendix A1

Here we prove equation (2.5). First we give some definitions.

Definition A1.1. Let $d \geq n > 1$ be integers. Let $G(d, n)$ be the set of elements of the form $(i_1, \ldots, i_n) \in \{1, \ldots, d\}^n$ which hold the following conditions.

1. $i_1 = 1, i_n = d$.
2. $i_1 < i_2 < \cdots < i_n$.
3. For every $l = 1, \ldots, n-1$: $[i_{l+1}/2] \leq i_l$.

Examples of such sets are:

$$G(3,2) = \{(1,3)\}$$
$$G(4,3) = \{(1,2,4), (1,3,4)\}$$
$$G(d,d) = \{(1,2,\ldots,d)\}$$
$$G(4,2) = \emptyset.$$

In general, $G(d, 2) = \emptyset$ for every $d \geq 4$. To prove this, let us note that $G(d, 2) = \{(1, d)\}$ and as condition 3 of Definition A1.1 must hold, then $[d/2] \leq 1$, which is only possible for $d \leq 3$.

Definition A1.2. Let $d > k \geq n > 1$ be integers which satisfy the following conditions.

1. $[d/2] \leq k$.

2. $n \geq [\log_2 k] + 1$.

Let us denote by $(k, d) \wedge G(k, n)$ the set of elements of the form $(i_1, \ldots, i_n, d)$ such that $(i_1, \ldots, i_n) \in G(k, n)$.

Lemma A1.1. Let $d > n > 1$. Denote:

$$u(d, n) = \min\{d - 1, 2^n - 1\}$$
$$l(d, n) = \max\{[d/2], n\}.$$

Then:

$$G(d, n + 1) = \bigcup_{l(d,n)}^{u(d,n)} (k, d) \wedge G(k, n). \tag{A1.1}$$

Proof. Consider $(i_1, \ldots, i_n, d) \in (k, d) \wedge G(k, n)$. This implies that $k \geq n$ from Definition A1.1; $[d/2] \leq i_n = k$ from condition 1 of Definition A1.2; $k \leq d - 1$ from Definition A1.2, and $k \leq 2^n - 1$ from condition 2 of Definition A1.2. Hence we have $l(d, n) \leq k \leq u(d, n)$. Besides $(i_1, \ldots, i_n) \in G(k, n)$. Let us show that $(i_1, \ldots, i_n, d) \in G(d, n + 1)$. Obviously $(i_1, \ldots, i_n, d) \in \{1, \ldots, d\}^{n+1}$; $i_1 = 1$ because $(i_1, \ldots, i_n) \in G(k, n)$ and also $i_{n+1} = d$. The above guarantees condition 1 of Definition A1.1. Besides, $i_1 < \cdots < i_n$ because $(i_1, \ldots, i_n) \in G(k, n)$ and $k = i_n < d$. Hence, condition 2 of Definition A1.1 also holds. Lastly, condition 3 of Definition A1.1 holds for $l = 1, \ldots, n - 1$ because $(i_1, \ldots, i_n) \in G(k, n)$. From condition 1 of Definition A1.2, condition 3 of Definition A1.1 holds for $l = n$. Hence $(i_1, \ldots, i_n, d) \in G(d, n + 1)$. Let us see the opposite inclusion.

Let $(i_1, \ldots, i_{n+1}) \in G(d, n + 1)$. This implies that $i_{n+1} = d$. Let k be the element i_n. From condition 3 of Definition A1.1 we have $[d/2] \leq k$, therefore, condition 1 of Definition A1.2 holds. Obviously $k \leq d - 1$. On the other hand, $k \geq n$ because $1 = i_1 < \cdots < i_n = k$. Let us prove that $(i_1, \ldots, i_n) \in G(k, n)$. First, $(i_1, \ldots, i_n) \in \{1, \ldots, k\}^n$ because $(i_1, \ldots, i_{n+1}) \in G(d, n + 1)$. Consequently condition 1 of Definition A1.1 holds. As already stated, condition 2 also holds. Condition 3 of Definition A1.1 is true for $l = 1, \ldots, n - 1$ because once again $(i_1, \ldots, i_{n+1}) \in G(d, n+1)$. All the above implies that $(i_1, \ldots, i_n) \in G(k, n)$ and therefore $n \geq [\log_2 k] + 1$. Hence $l(d, n) \leq k \leq u(d, n)$.

As $(i_1, \ldots, i_n) \in G(k, n)$ then $(i_1, \ldots, i_{n+1}) \in (k, d) \wedge G(k, n)$ for certain k. But this implies that

$$G(d, n+1) \subset \bigcup_{l(d,n)}^{u(d,n)} (k, d) \wedge G(k, n)$$

and completes the proof. ■

Remarks

1. If $u(d, n) = 2^n - 1$, then for $2^n - 1 < k \leq d - 1$ we have $G(k, n) = \emptyset$. It is not difficult to prove that $G(k, n) \neq \emptyset$ if and only if $n \geq [\log_2 k] + 1$. Hence, if $2^n - 1 < k$ we have $G(k, n) = \emptyset$.

2. Let us note that it is not possible for $u(d, n) = 2^n - 1$ and $l(d, n) = n$. If $d - 1 > 2^n - 1$ then $[d/2] > 2^{n-1}$, but $2^{n-1} \geq n$ for $n \geq 2$, therefore it is impossible that $[d/2] < n$.

From the above remarks, we obtain that the equation

$$G(d, n+1) = \bigcup_{k=n}^{d-1} (k, d) \wedge G(k, n) \tag{A1.2}$$

is also true.

Corollary A1.1. Let $\theta(d, n) = \operatorname{card} G(d, n)$. Then we have:

$$\theta(d, n+1) = \sum_{k=l(d,n)}^{u(d,n)} \theta(k, n). \tag{A1.3}$$

Proof. It is straightforward from equation (A1.1) and the fact that $G(d, n)$ sets do not intersect each other. ■

Remark. From the remarks following Lemma A1.1, the expression

$$\theta(d, n+1) = \sum_{k=n}^{d-1} \theta(k, n) \tag{A1.4}$$

is also true.

Table 3 shows the values of $\theta(d, n)$ for $2 \leq d \leq 13$ and $2 \leq n \leq 13$. It can be proved (see appendix B for the details) that:

$$\theta(d, k) \leq \frac{(2d - k - 3)}{k - 3} \binom{d - 4}{d - k}. \tag{A1.5}$$

Let us denote $I(d) = (I - T(d', d))^{-1}$ and $M(d', d) = I(d) T(d', d)$. Then we have Theorem A1.1.

d/*n*	2	3	4	5	6	7	8	9	10	11	12	13	14
2	1												
3	1	1											
4	0	2	1										
5	0	2	3	1									
6	0	1	5	4	1								
7	0	1	6	9	5	1							
8	0	0	6	15	14	6	1						
9	0	0	6	21	29	20	7	1					
10	0	0	4	26	50	49	27	8	1				
11	0	0	4	30	76	99	76	35	9	1			
12	0	0	2	31	105	175	175	111	44	10	1		
13	0	0	2	33	136	280	350	286	155	54	11	1	
14	0	0	1	30	165	415	630	636	441	209	65	12	1

Table 3. Values of cardinals of $G(d, n)$ sets.

Theorem A1.1. For every $d \in \mathrm{N}$, $d \geq 2$:

$$P(d) = \left\{ \sum_{n=2}^{d} \sum_{(i_1,\ldots,i_n)\in G(d,n)} M(i_{n-1}, i_n) \ldots M(i_1, i_2) \right\} P(1). \tag{A1.6}$$

Proof. The proof will be by induction on d. The property is true for $d = 3$:

$$P(3) = \left\{ \sum_{(i_1,i_2)\in G(3,2)} M(i_1, i_2) + \sum_{(i_1,i_2,i_3)\in G(3,3)} M(i_2, i_3) M(i_1, i_2) \right\} P(1).$$

Let suppose that it is true for $k = 2, \ldots, d-1$ and prove that it is also true for $k = d$:

$$\begin{aligned} P(d) &= I(d) \sum_{k=2}^{d-1} T(k, d) P(k) \\ &= I(d) \sum_{k=2}^{d-1} T(k, d) \left\{ \sum_{n=2}^{k} \sum_{(i_1,\ldots,i_n)\in G(k,n)} M(i_{n-1}, i_n) \ldots M(i_1, i_2) \right\} P(1) \\ &= \left\{ \sum_{k=2}^{d-1} \sum_{n=2}^{k} \sum_{(i_1,\ldots,i_n)\in G(k,n)} \right. \\ &\qquad \left. M(k, d) M(i_{n-1}, i_n) \ldots M(i_1, i_2) \right\} P(1). \end{aligned} \tag{A1.7}$$

Let us denote by:

$$\sigma(k, n) = \sum_{(i_1,\ldots,i_n)\in G(k,n)} M(k, d) M(i_{n-1}, i_n) \ldots M(i_1, i_2).$$

Then it is not difficult to see that:

$$\sum_{k=2}^{d-1}\sum_{n=2}^{k}\sigma(k,n)=\sum_{n=2}^{d-1}\sum_{k=n}^{d-1}\sigma(k,n).$$

Hence equation (A1.7) can be written as:

$$P(d)=\left\{\sum_{n=2}^{d-1}\sum_{k=n}^{d-1}\sum_{(i_1,\dots,i_n)\in G(k,n)}M(k,d)M(i_{n-1},i_n)\dots M(i_1,i_2)\right\}P(1).\ \text{(A1.8)}$$

Now the multiple-index of the product $M(k,d)M(i_{n-1},i_n)\dots M(i_1,i_2)$ is $(i_1,\dots i_n,d)$. From Lemma A1.1 we have: $(i_1,\dots i_n,d)\in(k,d)\wedge G(k,n)\subset G(d,n+1)$. From equation (A1.2) we can write:

$$\sum_{(i_1,\dots i_n+1)\in G(d,n+1)}M(i_n,i_{n+1})\dots M(i_1,i_2)$$
$$=\sum_{k=n}^{d-1}\sum_{(i_1,\dots i_n)\in G(k,n)}M(k,n)M(i_{n-1},i_n)\dots M(i_1,i_2).$$

Therefore, equation (A1.8) can be written:

$$P(d)=\left\{\sum_{n=2}^{d-1}\sum_{(i_1,\dots i_{n+1})\in G(d,n+1)}M(i_n,i_{n+1})\dots M(i_1,i_2)\right\}P(1).$$

Now making the change of variable $n=m-1$ the above expression can be written as:

$$P(d)=\left\{\sum_{m=3}^{d}\sum_{(i_1,\dots i_m)\in G(d,m)}M(i_{m-1},i_m)\dots M(i_1,i_2)\right\}P(1).$$

We could add the term for $m=2$ because $G(d,2)=\emptyset$ for every $d\geq 4$ and the inner sum would be zero. Finally:

$$P(d)=\left\{\sum_{n=2}^{d}\sum_{(i_1,\dots i_n)\in G(d,n)}M(i_{n-1},i_n)\dots M(i_1,i_2)\right\}P(1)$$

where we have changed m by n. This completes the proof. ■

Remark. Some terms of equation (A1.6) are equal to zero. As we have pointed out, $G(d,n)\neq\emptyset$ if and only if $n\geq[\log_2 d]+1$. Besides:

$$\sum_{(i_1,\dots i_n)\in G(d,n)}M(i_{n-1},i_n)\dots M(i_1,i_2)=0$$

if $n<[\log_2 d]+1$ because in each product $M(i_{n-1},i_n)\dots M(i_1,i_2)$ there exists at least a couple i_{r-1},i_r which does not satisfy condition 3 of

Definition A1.1 and therefore $M(i_{r-1}, i_r) = 0$. This explains why we could add the term for $n = 2$ at the end of the proof of Theorem A1.1. Hence the following expression remains true:

$$P(d) = \left\{ \sum_{n=[\log_2 d]+1}^{d} \sum_{(i_1,\ldots,i_n)\in G(d,n)} M(i_{n-1}, i_n) \ldots M(i_1, i_2) \right\} P(1). \tag{A1.9}$$

Appendix A2

In this appendix we will obtain upper and lower bounds for $\delta_j(i_1, \ldots, i_n)$. The first step will be to obtain upper and lower bounds for $v_k(d)$ and $\pi_k(d', d)$ as functions of d, d', and p. First, we give some definitions which can be found in [19]. There, in Figure 10, all the possible cases in which two binary symbols previously separated at distance d' would be at distance d in the next time step are shown. These cases are labeled as A_1, A_2, A_3, B_1, B_2, and C. Their probabilities are (see Equation B5 of Appendix B in [19]):

$$P(A_1) = (1-p)^2 \binom{d'-1}{2d'-d+1} p^{2d'-d+1}(1-p)^{d-d'-2} \tag{A2.1a}$$

$$P(A_2) = (1-p)^2 \binom{d'-1}{2d'-d} p^{2d'-d}(1-p)^{d-d'-1} \tag{A2.1b}$$

$$P(A_3) = (1-p)^2 \binom{d'-1}{2d'-d-1} p^{2d'-d-1}(1-p)^{d-d'} \tag{A2.1c}$$

$$P(B_1) = p(1-p) \binom{d'-1}{2d'-d} p^{2d'-d}(1-p)^{d-d'-1} \tag{A2.1d}$$

$$P(B_2) = p(1-p) \binom{d'-1}{2d'-d-1} p^{2d'-d-1}(1-p)^{d-d'} \tag{A2.1e}$$

$$P(C) = p^2 \binom{d'-1}{2d'-d-1} p^{2d'-d-1}(1-p)^{d-d'}. \tag{A2.1f}$$

The coefficients of matrix $T(d', d, p)$ can be built in terms of their probabilities (see also appendix B):

$$T_0(d', d, p) = P(A_1) + 2P(A_2) + P(A_3) \tag{A2.2a}$$

$$T_1(d', d, p) = P(B_1) + P(B_2) \tag{A2.2b}$$

$$T_2(d', d, p) = P(C). \tag{A2.2c}$$

Now from equations (2.6), (A2.2a), (A2.2b), and (A2.2c) we have:

$$\pi_1(d', d, p) = p^{2d'-d-1}(1-p)^{d-d'} \{A(d', d)p^2 + 2B(d', d)p + C(d', d)\} \tag{A2.3a}$$

$$\pi_2(d', d, p) = p^{2d'-d-1}(1-p)^{d-d'} \{A(d', d)p^2 + 2B(d', d)p(1-p) + C(d', d)(1-2p)\} \tag{A2.3b}$$

$$\pi_3(d',d,p) = p^{2d'-d-1}(1-p)^{d-d'}\left\{A(d',d)p^2 \right.$$
$$\left. +2B(d',d)p(1-2p) + C(d',d)(1-2p)^2\right\} \qquad \text{(A2.3c)}$$

where:

$$A(d',d) = \binom{d'-1}{2d'-d+1}$$
$$B(d',d) = \binom{d'-1}{2d'-d}$$
$$C(d',d) = \binom{d'-1}{2d'-d-1}.$$

It is not difficult to see that if $d' = d$, then the only possible cases are A_3, B_2, and C. Therefore the eigenvalues of matrix $T(d',d',p)$ are:

$$v_1(d') = p^{d'-1} \qquad \text{(A2.4a)}$$
$$v_2(d') = p^{d'-1}(1-2p) \qquad \text{(A2.4b)}$$
$$v_3(d') = p^{d'-1}(1-2p)^2. \qquad \text{(A2.4c)}$$

Theorem A2.1. Let $\pi_1(d',d,p)$, $\pi_3(d',d,p)$ be as defined by equations (A2.3a) and (A2.3c). Then if d' is big enough and $0 \le p < 1/2$ we have:

$$\phi_l^1(p)\frac{e^{-\left(\frac{(d'-1)(1-p)}{2p}\right)}}{\sqrt{d'-1}} \le \pi_1(d',d,p) \le \phi_u^1(p)\frac{e^{-\left(\frac{(d'-1)p}{2(1-p)}\right)}}{\sqrt{d'-1}} \qquad \text{(A2.5a)}$$
$$\phi_l^3(p)\frac{e^{-\left(\frac{(d'-1)(1-p)}{2p}\right)}}{\sqrt{d'-1}} \le \pi_3(d',d,p) \le \phi_u^3(p)\frac{e^{-\left(\frac{(d'-1)p}{2(1-p)}\right)}}{\sqrt{d'-1}} \qquad \text{(A2.5b)}$$

where:

$$\phi_l^1(p) = \frac{e^{\left(\frac{1}{2(1-p)}\right)} + 1(1-p)^2 e^{-\left(\frac{3}{2p}\right)} + 2(1-p)}{\sqrt{2\pi p(1-p)}}$$
$$\phi_u^1(p) = \frac{e^{\left(\frac{1}{p}\right)} + (1-p)^2 e^{\left(\frac{3}{1-p}\right)} + 2(1-p)}{\sqrt{2\pi p(1-p)}}$$
$$\phi_l^3(p) = \frac{e^{\left(\frac{1}{2(1-p)}\right)} + (1-2p)^2 e^{-\left(\frac{3}{2p}\right)} - 2(1-2p)}{\sqrt{2\pi p(1-p)}}$$
$$\phi_u^3(p) = \frac{e^{\left(\frac{1}{p}\right)} + (1-2p)^2 e^{\left(\frac{3}{1-p}\right)} + 2(1-2p)}{\sqrt{2\pi p(1-p)}}.$$

Proof. Let $d_0 \in \mathrm{N}$ such that for every $d \ge d_0$, the approximation of the local limit theorem remains valid:

$$\binom{d}{r}p^r(1-p)^{d-r} \approx \frac{1}{\sqrt{2\pi p(1-p)d}}e^{-\left(\frac{1}{2}\left(\frac{r-dp}{\sqrt{dp(1-p)}}\right)^2\right)}$$

(d_0 could be 25, see page 84 of [22]). Then from equations (A2.3a), (A2.3c), and under the supposition that $d' \geq d_0$, we have:

$$\pi_1 = \frac{\sqrt{w}}{\sqrt{2\pi p(1-p)}} e^{-\left(\frac{u^2}{2wp(1-p)}\right)} \left\{ e^{-\left(\frac{w+2u}{2(1-p)}\right)} \right. \\ \left. +(1-p)^2 e^{-\left(\frac{w-2u}{2(1-p)}\right)} + 2(1-p) \right\} \tag{A2.6a}$$

$$\pi_3 = \frac{\sqrt{w}}{\sqrt{2\pi p(1-p)}} e^{-\left(\frac{u^2}{2w(1-p)}\right)} \left\{ e^{-\left(\frac{w-2u}{2(1-p)}\right)} \right. \\ \left. +(1-2p)^2 e^{-\left(\frac{w+2u}{2(1-p)}\right)} - 2(1-2p) \right\} \tag{A2.6b}$$

where:

$$u = 1 - \frac{2d' - d}{p(d'-1)}; \qquad w = \frac{1}{p(d'-1)}.$$

From condition 3 of Definition A1.1 we have $d' + 1 \leq d \leq 2d' + 1$, hence:

$$-\frac{1}{p(d'-1)} \leq \frac{2d'-d}{p(d'-1)} \leq \frac{1}{p}$$

and then:

$$e^{-\left(\frac{3}{2p}\right)} \leq e^{-\left(\frac{w-2u}{2(1-p)}\right)} \leq e^{\left(\frac{3}{1-p}\right)} e^{\left(\frac{1}{2(1-p)}\right)} \leq e^{-\left(\frac{w+2u}{2(1-p)}\right)} \leq e^{\left(\frac{1}{p}\right)}.$$

From the above inequalities and equations (A2.6a) and (A2.6b) we could obtain:

$$\phi_l^1(p) \frac{e^{-\left(\frac{u^2}{2w(1-p)}\right)}}{\sqrt{d'-1}} \leq \pi_1(d', d, p) \leq \phi_u^1(p) \frac{e^{-\left(\frac{u^2}{2w(1-p)}\right)}}{\sqrt{d'-1}} \tag{A2.7a}$$

$$\phi_l^3(p) \frac{e^{-\left(\frac{u^2}{2w(1-p)}\right)}}{\sqrt{d'-1}} \leq \pi_3(d', d, p) \leq \phi_u^3(p) \frac{e^{-\left(\frac{u^2}{2w(1-p)}\right)}}{\sqrt{d'-1}}. \tag{A2.7b}$$

Let us note that:

$$\frac{u^2}{2w(1-p)} = \frac{p(d'-1)}{2(1-p)} \left(1 - \frac{2d'-d}{p(d'-1)}\right)^2.$$

It is not difficult to prove that if $d' + 1 \leq d \leq 2d' + 1$, then:

$$1 \leq \left| 1 - \frac{2d'-d}{p(d'-1)} \right| \leq \frac{1-p}{p}.$$

The above condition, along with equations (A2.7a) and (A2.7b) complete the proof. ■

Definition A2.1. Let $(i_1, \ldots, i_n) \in G(d, n)$ and $d_0 \in \mathrm{N}$. Let us denote by $l(i_1, \ldots, i_n)$ the set of indices which are smaller than d_0 and by $u(i_1, \ldots, i_n)$ those which are bigger:

$$l(i_1, \ldots, i_n) = \{i_r \in (i_1, \ldots, i_n) \colon i_r < d_0\}$$
$$u(i_1, \ldots, i_n) = \{i_r \in (i_1, \ldots, i_n) \colon i_r \geq d_0\}.$$

Definition A2.2. Let $(i_1, \ldots, i_n) \in G(d, n)$. Denote by:

$$c_j(i_1, \ldots, i_n) = \prod_{l(i_1,\ldots,i_n)} \frac{\pi_j(i_k, i_{k+1}, p)}{1 - v(i_{k+1})} \qquad j = 1, 3.$$

Corollary A2.1. Under the conditions of Theorem A2.1 we have:

$$L_j(p, d_0, d, n) \leq \delta_j(i_1, \ldots, i_n) \leq U_j(p, d_0, d, n) \tag{A2.8}$$

where:

$$L_j(p, d_0, d, n) = c_j(i_1, \ldots, i_n)\Phi_l^j(p, d_0, d, n)\frac{e^{-\left(\frac{1-p}{2p}S(d,n)\right)}}{M(d, n)} \tag{A2.9a}$$

$$U_j(p, d_0, d, n) = c_j(i_1, \ldots, i_n)\Phi_u^j(p, d_0, d, n)\frac{e^{-\left(\frac{p}{2(1-p)}S(d,n)\right)}}{M(d, n)} \tag{A2.9b}$$

$$\Phi_l^j(p, d_0, d, n) = \prod_{l(i_1,\ldots,i_n)} \frac{\phi_l^j(p)}{1 - v_j(i_k)}$$

$$\Phi_u^j(p, d_0, d, n) = \prod_{l(i_1,\ldots,i_n)} \frac{\phi_u^j(p)}{1 - v_j(i_k)}$$

$$S(d, n) = \sum_{u(i_1,\ldots,i_n)} (i_k - 1)$$

$$M(d, n) = \prod_{u(i_1,\ldots,i_n)} (i_k - 1).$$

Proof. The expression $\delta_j(i_1, \ldots, i_n)$ can be written:

$$\delta_j(i_1, \ldots, i_n) = \prod_{l(i_1,\ldots,i_n)} \frac{\pi_j(i_k, i_{k+1})}{1 - v_j(i_{k+1})} \prod_{u(i_1,\ldots,i_n)} \frac{\pi_j(i_k, i_{k+1})}{1 - v_j(i_{k+1})},$$

therefore:

$$\delta_j(i_1, \ldots, i_n) = c_j(i_1, \ldots, i_n) \prod_{u(i_1,\ldots,i_n)} \frac{\pi_j(i_k, i_{k+1})}{1 - v_j(i_{k+1})}.$$

Now from equations (A2.5a) and (A2.5b) we have:

$$\Phi_l^j(p, d_0, d, n)\frac{e^{-\left(\frac{1-p}{2p}S(d,n)\right)}}{M(d,n)} \leq \prod_{u(i_1,\ldots,i_n)} \frac{\pi_j(i_k, i_{k+1})}{1 - v_j(i_{k+1})}$$

$$\Phi_u^j(p, d_0, d, n)\frac{e^{-\left(\frac{p}{2(1-p)}S(d,n)\right)}}{M(d,n)} \geq \prod_{u(i_1,\ldots,i_n)} \frac{\pi_j(i_k, i_{k+1})}{1 - v_j(i_{k+1})}$$

which completes the proof. ■

Appendix A3

Here we obtain upper and lower bounds for $H_j(d)$. Let us first give some definitions.

Definition A3.1. Let us denote for:

$$c_j^u(d_0, n) = \max_{(i_1,\ldots,i_n)\in G(d,n)} c_j(i_1, \ldots, i_n)$$

$$c_j^l(d_0, n) = \min_{(i_1,\ldots,i_n)\in G(d,n)} c_j(i_1, \ldots, i_n).$$

Proposition A3.1. The following inequalities hold:

$$L_j(p, d_0, d, n) \geq c_j^l(d_0, n)\Phi_l^j(p, d_0, d, n)Q^l(d_0, n)\frac{e^{-\left(\frac{1-p}{2p}S_u(d,n)\right)}}{M_u(d,n)} \tag{A3.1}$$

$$U_j(p, d_0, d, n) \leq c_j^u(d_0, n)\Phi_u^j(p, d_0, d, n)Q^u(d_0, n)\frac{e^{-\left(\frac{p}{2(1-p)}S_l(d,n)\right)}}{M_l(d,n)} \tag{A3.2}$$

where:

$$Q^l(d_0, n) = \min_{(i_1,\ldots,i_n)\in G(d,n)} W^l(i_1, \ldots, i_n)$$

$$Q^u(d_0, n) = \max_{(i_1,\ldots,i_n)\in G(d,n)} W^u(i_1, \ldots, i_n)$$

$$W^l(i_1, \ldots, i_n) = \prod_{l(i_1,\ldots,i_n),\ k\neq 1} \sqrt{i_k - 1}e^{\left(\frac{1-p}{2p}\sum_{l(i_1,\ldots,i_n)}(i_k-1)\right)}$$

$$W^u(i_1, \ldots, i_n) = \prod_{l(i_1,\ldots,i_n),\ k\neq 1} \sqrt{1_k - 1}e^{\left(\frac{p}{2(1-p)}\sum_{l(i_1,\ldots,i_n)}(i_k-1)\right)}.$$

Proof. From equations (A2.9a) and (A2.9b) we have:

$$L_j(p, d_0, d, n) \geq c_j^l(d_0, n)\Phi_l^j(p, d_0, d, n)W^l(i_1, \ldots, i_n)\frac{e^{-\left(\frac{1-p}{2p}S^*(d,n)\right)}}{M^*(d,n)}$$

$$U_j(p, d_0, d, n) \leq c_j^u(d_0, n)\Phi_u^j(p, d_0, d, n)W^u(i_1, \ldots, i_n)\frac{e^{-\left(\frac{p}{2(1-p)}S^*(d,n)\right)}}{M^*(d,n)}.$$

From these inequalities and equations (2.20) and (2.27) we have equations (A3.1) and (A3.2). ■

Proposition A3.2. If $0.08 \le p \le 0.25$, then for d large enough and $j = 1, 3$:

$$\Phi_u^j(p, d_0, d, n) \le e^{n\left(2\ln 2+\frac{1}{p}+\frac{p^{d_0}}{1-p}\right)}(2\pi(1-p)p)^{-\left(\frac{n}{2}\right)} = B_u(p, d_0, n) \tag{A3.3}$$

$$\Phi_l^j(p, d_0, d, n) \ge (2\pi(1-p)p)^{-\left(\frac{1}{2}\right)} e^{\left(\frac{1}{2(1-p)}-\ln 5\right)} = B_l(p, d_0, n). \tag{A3.4}$$

Proof. If $p < 0.25$, we could prove that

$$\phi_u^j(p) \le 4e^{\left(\frac{1}{4}\right)}(2\pi(1-p)p)^{-\left(\frac{1}{2}\right)}.$$

Besides, from:

$$1 - p^k = e^{\ln(1-p^k)} = e^{-p^k+o(p^k)}$$

we have, for d large enough,

$$\prod_{u(i_1,\ldots,i_n)} (1 - p^{i_k}) \ge e^{-(p^{d_0}} + \cdots + p^d) + o(p^{d_0}) = e^{-\left(\frac{p^{d_0}}{1-p}+o(p^{d_0})\right)}.$$

Therefore:

$$\Phi_u^j(p, d_0, d, n) \le e^{n\left(2\ln 2+\frac{1}{p}+\frac{p^{d_0}}{1-p}\right)}(2\pi(1-p)p)^{-\left(\frac{n}{2}\right)}.$$

If $0.08 < p$, then:

$$\frac{1}{5}(2\pi(1-p)p)^{\left(\frac{1}{2}\right)} e^{\left(\frac{1}{2(1-p)}\right)} \le \phi_l^j(p)$$

and from the above:

$$(2\pi(1-p)p)^{-\left(\frac{1}{2}\right)} e^{\left(\frac{1}{2(1-p)}-\ln 5\right)} \le \Phi_l^j(p, d_0, d, n).$$

This completes the proof. ■

Proposition A3.3. Under the same conditions of Proposition A3.2 we have:

$$\sum_{(i_1,\ldots,i_n)\in G(d,n)} \delta_j(i_1, \ldots, i_n) \le K^u(p, d_0, n) \frac{e^{-\left(\frac{p}{2(1-p)} S_l(d,n)\right)}}{M_l(d, n)} \frac{(2d-n-3)}{n-3}\binom{d-4}{n-4} \tag{A3.5}$$

$$\sum_{(i_1,\ldots,i_n)\in G(d,n)} \delta_j(i_1, \ldots, i_n) \ge K^l(p, d_0, n) \frac{e^{-\left(\frac{1-p}{2p} S_u(d,n)\right)}}{M_u(d, n)} \tag{A3.6}$$

where:

$$K^u(p, d_0, n) = c_j^u(d_0, n) B_u(p, d_0, n) Q^u(d_0, n)$$
$$K^l(p, d_0, n) = c_j^l(d_0, n) B_l(p, d_0, n) Q^l(d_0, n).$$

Proof. Let $(i_1, \ldots, i_n) \in G(d, n)$. Then from Corollary A2.1 and Propositions A3.2 and A3.3 we have:

$$\delta_j(i_1, \ldots, i_n) \le K^u(p, d_0, n) \frac{e^{-\left(\frac{p}{2(1-p)} S_l(d,n)\right)}}{M_l(d, n)} \tag{A3.7}$$

$$\delta_j(i_1, \ldots, i_n) \ge K^l(p, d_0, n) \frac{e^{-\left(\frac{1-p}{2p} S_u(d,n)\right)}}{M_u(d, n)}. \tag{A3.8}$$

From equation (A3.8) we obtain equation (A3.6) straightforwardly. From equations (A3.7) and (B.9) we obtain equation (A3.5) to complete the proof. ■

Equations (A3.5) and (A3.6) can be written as:

$$\sum_{(i_1,\ldots,i_n)\in G(d,n)} \delta_j(i_1, \ldots, i_n) \ge c_j^l(d_0, n) Q^l(d_0, n) e^{-\epsilon_l(d,n,p)} \tag{A3.9}$$

$$\sum_{(i_1,\ldots,i_n)\in G(d,n)} \delta_j(i_1, \ldots, i_n) \le c_j^u(d_0, n) Q^u(d_0, n) e^{-\epsilon_u(d,n,p)} \tag{A3.10}$$

where:

$$\epsilon_u(d, n, p) = \epsilon_u^1(d, n, p) + \epsilon_u^2(d, n, p) + \epsilon_u^3(d, n, p) + \epsilon_u^4(d, n, p) \tag{A3.11}$$

$$\begin{aligned}
\epsilon_u^1(d, n, p) = {} & \frac{p}{2(1-p)}(d+1)\left[\frac{(d-n)\ln 2 - 1}{d \ln 2 - 1}\right] \\
& + \frac{1}{\ln 2} \ln\left[\frac{d \ln 2 - 1}{n \ln 2 - 1}\right] \Bigg\{ \ln(d+1) \\
& - \frac{1}{2}\left[\ln\left(\frac{d \ln 2 - 1}{n \ln 2 - 1}\right) + \ln 2\right] - \frac{p}{1-p} \Bigg\} \\
\epsilon_u^2(d, n, p) = {} & (n - r_l - 1)\left\{\frac{n - r_l}{2} + \ln(n - r_l - 1)\right\} + \frac{p}{1-p}(1-n) \\
& + (d-n)\left\{\ln(d-n) + \frac{2}{d+1}\left(\frac{\ln 2}{n \ln 2 - 1}\right)\right\} \\
\epsilon_u^3(d, n, p) = {} & \ln d + \frac{1}{2}\ln(n - r_l - 1) + \frac{\ln(d-n)}{2} \\
& + \ln(n-4)\left(n - \frac{7}{2}\right) + \ln(n-3)
\end{aligned}$$

$$\epsilon_u^4(d,n,p) = -\left\{\frac{\ln(d-4)}{2} + (d-4)\ln(d-4) + \ln(2d-n-3)\right.$$
$$\left. +n\left(2\ln 2 + \frac{1}{p} + \frac{p^{d_0}}{1-p}\right)\right\}$$
$$\epsilon_l(d,n,p) = \epsilon_l^1(d,n,p) + \epsilon_l^2(d,n,p) + \epsilon_l^3(d,n,p) \qquad \text{(A3.12)}$$

$$\epsilon_l^1(d,n,p) = \frac{1-p}{p}\left\{d^{(1-\frac{n}{d\ln d})} + 2d\left(n+2-\frac{\ln d}{\ln 2}\right)\right.$$
$$\left. +\frac{\ln d}{\ln 2}\left(2n-1+\frac{\ln d}{\ln 2}\right)\right\}$$
$$\epsilon_l^2(d,n,p) = \frac{\ln 2}{4}(n-r_u)(n-r_u-1) + \frac{\ln d}{2} + \frac{d\ln d}{2} + (d-r_u-1)$$
$$\epsilon_l^3(d,n,p) = -\frac{1}{2}\{d + \ln(d-r_u-1) + (d-r_u-1)\ln(d-r_u-1)\}.$$

It can be proved that when $d \to \infty$ the following asymptotic expansions hold:

$$\epsilon_u(d,n,p) = \epsilon_u^a(d,n,p) + o\left(\frac{1}{d},\frac{1}{n}\right) \qquad \text{(A3.13)}$$
$$\epsilon_l(d,n,p) = \epsilon_l^a(d,n,p) + o\left(\frac{1}{d},\frac{1}{n}\right) \qquad \text{(A3.14)}$$

where:

$$\epsilon_u^a(d,n,p) = \frac{pd}{4(1-p)} + \frac{n^2}{2} + n\left[2\ln n + \left(1+\frac{1}{\ln 2}\right)(\ln n - \ln d)\right.$$
$$\left. -\frac{5-2p}{2(1-p)}\right] + \frac{1}{2}\left(1+\frac{1}{\ln 2}\right)\left(\frac{\ln d}{\ln 2}\right)^2$$
$$\epsilon_l^a(d,n,p) = \frac{d}{2}\left[1 + (2n+5)\left(\frac{1-p}{p}\right)\right] + \frac{1}{2}\frac{\ln d}{\ln 2}\left[n + \frac{1-p}{p}\left(1-\frac{d}{2}\right)\right.$$
$$\left. -\frac{\ln 2}{2}\right] + \frac{1}{2}(2n-1)\frac{1-p}{p}.$$

Now from equation (2.8) and equations (A3.9) and (A3.10) we have:

$$H_j(d) \le \sum_{n=[\log_2 d]+1}^{d} c_j^u(d_0,n)Q^u(d_0,n)e^{-\epsilon_u(d,n,p)} \qquad \text{(A3.15)}$$

$$H_j(d) \ge \sum_{n=[\log_2 d]+1}^{d} c_j^l(d_0,n)Q^l(d_0,n)e^{-\epsilon_l(d,n,p)}. \qquad \text{(A3.16)}$$

From the preceding inequalities and equation (2.11) the upper and lower bounds for the correlation function can be obtained.

Appendix B: The proof of equation (A1.5)

Lemma B.1. Under the same conditions of Lemma A1.1, let S be the set:

$$S = \{(n, d) \in \mathrm{N}^2 : d > n > 1, [d/2] \geq n, d \text{ even}\}$$

then:

$$\theta(d, n+1) = \begin{cases} \theta(d-1, n) + \theta(d-1, n+1) - \theta(d/2-1, n) & (n, d) \in S \\ \theta(d-1, n) + \theta(d-1, n+1) & (n, d) \notin S. \end{cases}$$

Proof. We analyze several cases.

I. If $u(d, n) = d - 1$, then:

$$\theta(d, n+1) = \sum_{k=l(d,n)}^{d-2} \theta(k, n) + \theta(d-1, n). \tag{B.1}$$

I.1. If $l(d, n) = n$; then from $u(d, n) = d - 1$, it follows straightforwardly that $u(d-1, n) = d - 2$ and because $l(d-1, n) = l(d, n)$ we have:

$$\theta(d-1, n+1) = \sum_{k=l(d,n)}^{d-2} \theta(k, n). \tag{B.2}$$

Now from equation (A.1) we have:

$$\theta(d, n+1) = \theta(d-1, n+1) + \theta(d-1, n). \tag{B.3}$$

I.2. If $l(d, n) = [d/2]$; then it is not difficult to see that:

$$[(d-1)/2] = \begin{cases} d/2 - 1 & d \text{ is even} \\ [d/2] & d \text{ is odd.} \end{cases} \tag{B.4}$$

From the above it follows that if d is odd: $l(d, n) = l(d-1, n)$ and therefore we obtain equations (B.2) and (B.3).

If d is even, we analyze two cases.

I.2.1. If $d/2 = n$; then $d/2 - 1 < n$ and:

$$l(d-1, n) = n = d/2 = l(d, n)$$

from this equation we obtain equations (B.2) and (B.3).

I.2.2. If $d/2 - 1 \geq n$; then $l(d-1, n) = d/2 - 1$ therefore $l(d-1, n) = l(d, n) - 1$. From the previous equation we have:

$$\theta(d-1, n+1) - \theta(d/2-1, n) = \sum_{k=l(d,n)}^{d-2} \theta(k, n).$$

Now, from equation (B.1):

$$\theta(d, n+1) = \theta(d-1, n) + \theta(d-1, n+1) - \theta(d/2-1, n).$$

II. Let us suppose that $u(d,n) = 2^n - 1$. Hence $2^n \leq d$. Besides, from Remark 2 following Lemma A1.1 we have: $l(d,n) = [d/2]$. We analyze two cases.

II.1. $2^n = d$; then $u(d,n) = d - 1$ and we are in case I.

II.2. $2^n < d$; then: $n \leq [\log_2(d-1)]$ and therefore: $\theta(d-1,n) = 0$.

Besides, because $2^n < d$, $l(d-1,n) = [(d-1)/2]$ we have:

$$\theta(d-1,n+1) = \sum_{k=[(d-1)/2]}^{2^n-1} \theta(k,n) = \begin{cases} \theta(d-1,n+1) & d \text{ is odd} \\ \theta(d-1,n+1) - \theta(d/2-1,n) & d \text{ is even.} \end{cases}$$

We obtain the above condition from equation (B.4). This completes the proof. ■

From Lemma B.1 we have: $\theta(d,n+1) \leq \theta(d-1,n) + \theta(d-1,n+1)$. In order to obtain an upper bound for $\theta(d,n)$ we will study the following sequence defined recursively:

$$\omega(d,n) = \omega(d-1,n-1) + \omega(d-1,n) \tag{B.5}$$

with the following boundary conditions:

C1. $\omega(d,2) = 0 \quad \forall d \geq 4$;

C2. $\omega(d,d-1) = d-2 \quad \forall d \geq 3$.

It is not difficult to see that $\theta(d,n) \leq \omega(d,n)$. Now we obtain some properties of $\omega(d,n)$.

Let us consider the vector $[\omega(d,d-1), \omega(d,d-2), \ldots, \omega(d,3)]^t$. From equation (B.5) and conditions C1 and C2 we have:

$$\begin{bmatrix} \omega(d,d-1) \\ \vdots \\ \omega(d,3) \end{bmatrix} = \begin{bmatrix} d-2 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} 0 \\ \omega(d-1,d-2) \\ \vdots \\ \omega(d-1,3) \end{bmatrix}$$

where:

$$B_{11} = [0]_{1\times 1}$$
$$B_{12} = [\,0 \quad \ldots \quad 0\,]_{1\times(d-4)}$$
$$B_{21} = [\,0 \quad \ldots \quad 0\,]^t_{1\times(d-4)}$$
$$B_{22} = \begin{bmatrix} 1 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 1 \end{bmatrix}_{(d-4)\times(d-4).}$$

Lemma B.2. Let $d \geq 5$. Then we have:

$$\begin{bmatrix} \omega(d,d-1) \\ \omega(d,d-2) \\ \vdots \\ \omega(d,3) \end{bmatrix} = b_{d-4} + \sum_{k=0}^{d-5} A_{d-4}A_{d-5}\dots A_{d-k-4}b_{d-k-5} \tag{B.6}$$

where:

$$A_{d-r} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

$$B_{11} = [0]_{(r-3)\times(r-3)};\ B_{21}^t = B_{12} = [\,0 \quad \cdots \quad 0\,]_{(r-3)\times(d-r)}$$

$$B_{22} = \begin{bmatrix} 1 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 1 \end{bmatrix}_{(d-r)\times(d-r)}$$

$$b_{d-r} = [\,0 \quad \cdots \quad d-r+2 \quad \cdots \quad 0\,]^t_{1\times(d-r)}$$

where the nonzero element of b_{d-r} is in position $r-3$.

Proof. The proof will be by induction on d.

For $d = 5$ we have:

$$\begin{bmatrix} \omega(5,4) \\ \omega(5,3) \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ \omega(4,3) \end{bmatrix}.$$

Let us suppose that it is true for d and prove that it is also true for $d+1$:

$$\begin{bmatrix} \omega(d+1,d) \\ \omega(d+1,d-1) \\ \vdots \\ \vdots \\ \omega(d+1,3) \end{bmatrix} = \begin{bmatrix} d-1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 1 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 0 \\ \omega(d,d-1) \\ \omega(d,d-2) \\ \vdots \\ \omega(d,3) \end{bmatrix}.$$

From the hypothesis of induction, the last $d-3$ components of the vector $[\,0 \quad \omega(d,d-1) \quad \cdots \quad \omega(d,3)\,]^t$ can be written by using equation (B.6) as:

$$\begin{bmatrix} \omega(d+1,d) \\ \vdots \\ \omega(d+1,3) \end{bmatrix} = \begin{bmatrix} d-1 \\ \vdots \\ 0 \end{bmatrix} + A_{d-3}\left\{ b_{d-4} + \sum_{k=0}^{d-5} A_{d-4}\dots A_{d-k-4}b_{d-k-5} \right\}.$$

We have added to the element of $b_{d-4} + \sum_{k=0}^{d-5} A_{d-4}\dots A_{d-k-4}b_{d-k-5}$ a row and/or a column of zeros to obtain the same dimension. This completes the proof. ■

Lemma B.3. Under the same conditions of Lemma B.2 let a_{ij}, $1 \leq i$, and $j \leq d-3$ be the coefficients of matrix $A_{d-4}\dots A_{d-r}$, then:

$$a_{ij} = \begin{cases} \binom{r-4}{i-2} & 2 \leq i \leq r-2 \text{ and } j = r-2 \\ \binom{r-3}{r-3-j+i} & r-1 \leq j \leq d-3 \text{ and } j-r+3 \leq i \leq j \\ 0 & \text{in the other cases.} \end{cases} \tag{B.7}$$

Proof. The proof will be by induction on r. The property is obviously true for $r = 4$. Let us suppose that it is true for r and prove that it is also true for $r+1$. Let us denote by b_{ij} the coefficient of matrix $A_{d-(r+1)}$, then:

$$b_{ij} = \begin{cases} 1 & i = j; \quad j = r-1, \ldots, d-3 \\ 1 & i = j-1; \quad j = r-1, \ldots, d-3 \\ 0 & \text{in the other cases.} \end{cases}$$

Let c_{ij} be the coefficients of matrix $A_{d-4} \ldots A_{d-r} A_{d-(r+1)}$. Then, if $j = r-1$:

$$c_{i(r-1)} = \sum_{k=1}^{d-3} a_{ik} b_{k(r-1)} = a_{i(r-1)} b_{(r-1)(r-1)} = \binom{r-3}{r-3-(r-1)+i} = \binom{r-3}{i-2}$$

for $2 \leq i \leq r-1$. In the other cases $c_{i(r-1)} = 0$. Besides, if $r \leq j \leq d-3$ then:

$$\begin{aligned} c_{ij} &= \sum_{k=1}^{d-3} a_{ik} b_{kj} = a_{i(j-1)} b_{(j-1)j} + a_{ij} b_{jj} \\ &= \binom{r-3}{r-2-j+i} + \binom{r-3}{r-3-j+i} = \binom{r-2}{r-2-j+i} \end{aligned}$$

for $j-r+2 \leq i \leq j$. In the other cases $c_{ij} = 0$. This completes the proof. ■

Corollary B.1. Under the same assumptions as Lemma B.3:

$$A_{d-4} \ldots A_{d-r} b_{d-(r+1)} = \begin{bmatrix} 0 \\ (d-r+1)\binom{r-4}{0} \\ (d-r+1)\binom{r-4}{1} \\ \vdots \\ (d-r+1)\binom{r-4}{r-4} \\ \vdots \\ 0 \end{bmatrix}.$$

Proof. The only nonzero element of vector $b_{d-(r+1)}$ is in position $r-2$ and has the value $d-r+1$. From Lemma B.3, the column $r-2$ of matrix $A_{d-4} \ldots A_{d-r}$ is:

$$\begin{bmatrix} 0 & \binom{r-4}{0} & \binom{r-4}{1} & \cdots & \binom{r-4}{r-4} & \cdots & 0 \end{bmatrix}^t.$$

This completes the proof. ■

Proposition B.1. Let $d \geq 5$. Then:

$$\omega(d, d-1) = \sum_{n=i-2}^{d-5} (d-n-3) \binom{n}{i-2} \qquad \text{for } i \geq 2. \tag{B.8}$$

Proof. From Lemma B.2 and Corollary B.1 we have:

$$\begin{bmatrix} \omega(d,d-1) \\ \omega(d,d-2) \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \omega(d,3) \end{bmatrix} = \begin{bmatrix} d-2 \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{bmatrix} + \sum_{k=0}^{d-5} (d-k-3) \begin{bmatrix} 0 \\ \binom{k}{0} \\ \vdots \\ \vdots \\ \binom{k}{k} \\ \vdots \\ 0 \end{bmatrix}.$$

From this expression we immediately obtain the result.

From equation (B.8) it can be proved that for $k \geq 3$:

$$\omega(d,k) = \frac{(2d-k-3)}{k-3} \binom{d-4}{d-k}.$$

And from the above result and the definition of $\omega(d,k)$ we have:

$$\theta(d,k) \leq \frac{(2d-k-3)}{(k-3)} \binom{d-4}{d-k}. \blacksquare \tag{B.9}$$

References

[1] C. Burks and D. Farmer, "Towards Modeling DNA Sequences as Automata," *Physica D*, **10** (1984) 157–167.

[2] D. Borsnitk *et al.*, "Analysis of Apparent $1/f^a$ Spectrum in DNA Sequences," *Europhysics Letters*, **23** (1993) 389–394.

[3] A. L. Buldyrev *et al.*, "Long-range Correlation Properties of Coding and Noncoding DNA Sequences: GenBank Analysis," *Physical Review E*, **51** (1995) 5084–5091.

[4] C. A. Chatzidimitriou-Dreismann and D. Larhamar, "Long-range Correlation in DNA," *Nature*, **361** (1993) 212–213.

[5] S. Karlin and V. Brendel, "Patchiness and Correlations in DNA Sequences," *Science*, **259** (1993) 677–680.

[6] W. Lie, "Power Spectra of Regular Languages and Cellular Automata," *Complex Systems*, **1** (1987) 107–130.

[7] W. Li, "Spatial 1/*f* Spectra in Open Dynamical Systems," *Europhysic Letters*, **10** (1989) 395–400.

[8] W. Li, "Generating Nontrivial Long-range Correlation and 1/*f* Spectra by Replication and Mutation," *International Journal of Bifurcation and Chaos*, **2** (1992) 137–154.

[9] W. Li and K. Kaneko, "Long-range Correlation and Partial $1/f^a$ Spectrum in a Noncoding DNA Sequence," *Europhysics Letters*, **17** (1992) 655–660.

[10] W. Li and K. Kaneko, "DNA Correlations," *Nature*, **360** (1992) 635–636.

[11] R. Mansilla and R. Mateo-Reig, "On the Mathematical Modeling of Intronic Sectors of the DNA Molecule," *International Journal of Bifurcation and Chaos*, **5** (1995) 1235–1241.

[12] P. Miramontes, "Cellular Automatas, Genetic Algorithms and DNA Evolution," Ph. D. Thesis, UNAM, 1992.

[13] S. Nee, "Uncorrelated DNA Walks," *Nature*, **357** (1992) 450.

[14] C. K. Peng *et al.*, "Long-range Correlation in Nucleotide Sequences," *Nature*, **356** (1992) 168–170.

[15] C. K. Peng *et al.*, "Finite Size Effects on Long-range Correlations: Implications for Analyzing DNA Sequences," *Physical Review E*, **47** (1993) 3730–3733.

[16] V. V. Prabhu and J. M. Claverie, "Correlations in Intronless DNA," *Nature*, **359** (1992) 782.

[17] R. Voss, "Evolution of Long-range Correlations and 1/*f* Noise in DNA Base Sequences," *Physical Review Letters*, **68** (1992) 3805–3808.

[18] W. Li, "Absence of 1/*f* Spectra in Dow Jones Daily Average," *International Journal of Bifurcation and Chaos*, **1** (1991) 583–597.

[19] W. Li, "Expansion-modification Systems: A Model for Spatial 1/*f* Spectra," *Physical Review A*, **43** (1991) 5240–5260.

[20] R. Mansilla *et al.*, "Energetical Regularities of Introns in HUMHBB," Technical Report, Physical Institute, UNAM, Mexico, 1993.

[21] W. Li, "Mutual Information Function versus Correlation Function," *Journal of Statistical Physics*, **60** (1990) 823–837.

[22] B. Genedenko, *The Theory of Probability* (MIR Publisher, Moscow, 1980).

[23] R. Mansilla and G. Cocho, "Expansion-modification Systems in Sequences Over Three and Four Letter Alphabets: Implication to Intergenic Sequences," submitted to *Physical Review E*.