On the spectral criteria of disorder in nonperiodic sequences: application to inflation models, symbolic dynamics and DNA sequences

1994 J. Phys. A: Math. Gen. 27 4875

(http://iopscience.iop.org/0305-4470/27/14/016)

J. Phys. A: Math. Gen. 27 (1994) 4875–4898. Printed in the UK

On the spectral criteria of disorder in non-periodic sequences: application to inflation models, symbolic dynamics and DNA sequences

V R Chechetkin and A Yu Turygin

Troitsk Institute of Innovation and Thermonuclear Investigations—TRINITI, 142092, Troitsk, Moscow Region, Russia

Received 11 October 1993

Abstract. The spectral representation gives an effective approach to the analysis of statistical characteristics of symbolic sequences. We derive the corresponding criteria for the random case. The criteria ensure the dichotomic classification (random–non-random) for relatively short sequences of about several thousand symbols. The theory is applied to inflation models, symbolic dynamics and DNA sequences.

1. Introduction

Sequences of symbols are the usual tool of communications and appear in various physical and non-physical problems. The physical applications† are related to quasicrystals and inflation models [1–3], as well as symbolic dynamics [4–6]. The analysis of DNA sequences attracts interest both from physicists and specialists in genetics [7–11], while the symbolic sequences in natural languages and in the transmission of signals are the objects of investigation in linguistics and information theory. Very often *a priori* information on the underlying algorithms is absent. In this case the first step consists of the simple dicohotomic classification: is a sequence random or non-random? If a sequence contains both regular and chaotic contributions, the second step consists of the separation of different parts.

In [12, 13] the relevant criteria have been discussed from the information theoretic viewpoint [14]. Here we consider an alternative approach based on the spectral representation of a symbolic sequence. The method is a slight modification of the technique used in the theory of quasicrystals and substitutional sequences [2, 3]. The layout of our paper is as follows. The general formulation of the problem is presented in section 2. The expressions for the characteristic and probability distribution functions for spectral harmonics are derived in section 3, while the particular criteria of disorder are given in section 4. These results are applied to the analysis of inflation models, symbolic dynamics and DNA sequences in section 5. The final section contains some concluding remarks.

† The literature devoted to the problems mentioned below is very vast. For this reason we will mention only monographs, reviews and papers with elements of review, where further detailed references can be found.

0305-4470/94/144875+24$19.50

2. Spectral representation of sequences

2.1. Structure factor of sequence

The general idea of our approach consists of the Fourier representation of a symbolic sequence and derivation of the statistical criteria for the distributions of the spectral harmonics. In this and the following sections we present the main results of the general theory.

Let us consider an abstract sequence of length M and l different symbols $\{A_l\}$. A sequence can equivalently be described in terms of the position function:

$$\rho_{m,\alpha} = \begin{cases} 1 & \text{if the } \alpha\text{th symbol occupies a position } m \\ 0 & \text{otherwise} \end{cases} \qquad (2.1)$$

$$\alpha \in (A_1, \ldots, A_l) \qquad m = 1, \ldots, M.$$

Then, the Fourier harmonics corresponding to a subsequence of the αth symbols are determined according to

$$\rho_\alpha(q_n) = M^{-1/2} \sum_{m=1}^{M} \rho_{m,\alpha} \exp(-iq_n m) \qquad q_n = 2\pi n/M \qquad n = 0, 1, \ldots, M-1. \quad (2.2)$$

The reciprocal transformation is given by

$$\rho_{m,\alpha} = M^{-1/2} \sum_{n=0}^{M-1} \rho_\alpha(q_n) \exp(iq_n m) \qquad m = 1, \ldots, M. \qquad (2.3)$$

The zeroth Fourier harmonic does not contain any positional information and is related only to the total number of the αth symbols, N_α:

$$\rho_\alpha(0) = N_\alpha / M^{1/2}. \qquad (2.4)$$

The reality of $\rho_{m,\alpha}$ leads to the condition

$$\rho_\alpha^*(q_n) = \rho_\alpha(2\pi - q_n) \qquad (2.5)$$

(the asterisk denotes complex conjugation).

The pair correlations (see below) are characterized by the structure factors:

$$F_{\alpha\beta}(q_n) = \rho_\alpha(q_n)\rho_\beta^*(q_n). \qquad (2.6)$$

As is seen from equation (2.5),

$$F_{\alpha\beta}(q_n) = F_{\beta\alpha}(2\pi - q_n) \qquad (2.7)$$

and, in particular, a diagonal structure factor $F_{\alpha\alpha}(q_n)$ is a symmetric function of q_n with the centre of symmetry at $q_n = \pi$.

The pair correlations can also be described in terms of the circular correlation functions:

$$K^c_{\alpha\beta}(m_0) = M^{-1} \sum_{m=1}^{M} \tilde{\rho}_{m,\alpha} \tilde{\rho}_{m+m_0,\beta} \qquad (2.8)$$

$$\tilde{\rho}_{m,\alpha} = \begin{cases} \rho_{m,\alpha} & 1 \leqslant m \leqslant M \\ \rho_{m-M,\alpha} & M+1 \leqslant m \leqslant 2M-1 \end{cases} \qquad (2.9)$$

where $1 \leqslant m_0 \leqslant M-1$. Both characteristics are mutually connected through the analogue of the Wiener–Khinchin relationship [15]:

$$K^{c}_{\alpha\beta}(m_0) = M^{-1} \sum_{n=0}^{M-1} F_{\alpha\beta}(q_n) \exp(-iq_n m_0). \tag{2.10}$$

It follows from definitions (2.8) and (2.9) that

$$K^{c}_{\alpha\beta}(m_0) = K^{c}_{\beta\alpha}(M - m_0). \tag{2.11}$$

The higher products of Fourier harmonics are related to the higher correlation functions.

2.2. Sum rules

The statistical criteria are expressed through a set of universal (independent of a particular distribution of symbols) mean spectral parameters determined by the exact sum rules. The first relationship can be derived directly from equations (2.1)–(2.7), and is given by

$$\sum_{n=0}^{M-1} F_{\alpha\beta}(q_n) = \sum_{m=1}^{M} \rho_{m,\alpha}\rho_{m,\beta} = \delta_{\alpha\beta} N_\alpha \tag{2.12}$$

where $\delta_{\alpha\beta}$ is the Kronecker symbol. Taking into account equation (2.4) for the zeroth harmonics, one obtains

$$\bar{F}_{\alpha\beta} = \left(\sum_{n=1}^{M-1} F_{\alpha\beta}(q_n) \right)(M-1)^{-1} = (\delta_{\alpha\beta} N_\alpha - N_\alpha N_\beta / M)/(M-1). \tag{2.13}$$

Analogously, one derives the more general sum rule

$$\sum_{\substack{0 \leqslant q_n \leqslant 2\pi(M-1)/M \\ (q_{n_1} + \cdots + q_{n_r}) \bmod 2\pi = 0}} \rho_{\alpha_1}(q_{n_1}) \ldots \rho_{\alpha_r}(q_{n_r}) = M^{(r-2)/2} \sum_{m=1}^{M} \rho_{m,\alpha_1} \ldots \rho_{m,\alpha_r}$$

$$= \delta_{\alpha_1\alpha_2} \ldots \delta_{\alpha_1\alpha_r} M^{(r-2)/2} N_{\alpha_1}. \tag{2.14}$$

Equation (2.10) relates the mutual deviations of circular correlation functions and structure factor harmonics with $n \neq 0$ from their respective mean values:

$$\sum_{m_0=1}^{M-1} (K^{c}_{\alpha\beta}(m_0) - \bar{K}^{c}_{\alpha\beta})(K^{c}_{\gamma\delta}(m_0) - \bar{K}^{c}_{\gamma\delta}) = M^{-1} \sum_{n=1}^{M-1} (F_{\alpha\beta}(q_n) - \bar{F}_{\alpha\beta})(F^{*}_{\gamma\delta}(q_n) - \bar{F}^{*}_{\gamma\delta}) \tag{2.15}$$

where

$$\bar{K}^{c}_{\alpha\beta}(m_0) = (M-1)^{-1} \sum_{m_0=1}^{M-1} K^{c}_{\alpha\beta}(m_0) = (N_\alpha N_\beta - \delta_{\alpha\beta} N_\alpha)/M(M-1). \tag{2.16}$$

Similar expressions can be derived for the higher correlation functions as well.

2.3. Excluded volume effects

One position is occupied by only one symbol and there are no voids in a sequence. This condition imposes the additional restrictions on the two-valued position functions (2.1)

$$\sum_{\alpha=1}^{l} \rho_{m,\alpha} = 1 \tag{2.17}$$

i.e. the positions of subsequences for any $(l-1)$ different symbols determine unambiguously that of the remaining subsequence. In terms of Fourier harmonics the same restriction is formulated by

$$\sum_{\alpha=1}^{l} \rho_{\alpha}(q_n) = 0 \qquad (n \neq 0). \tag{2.18}$$

Condition (2.17) leads to the specific correlations (called traditionally 'excluded volume effects') even for the random sequences.

3. Statistical characteristics of random sequences

3.1. Characteristic function

The statistical properties of Fourier harmonics can be determined by the averaging of the characteristic (or generating) function [15, 16]:

$$Z = \exp\left(\mathrm{i}\sum_{\alpha=1}^{l}\sum_{n=1}^{M-1} u_{\alpha}(q_n)\rho_{\alpha}(q_n)\right). \tag{3.1}$$

It is convenient to impose on the auxiliary variables $u_{\alpha}(q_n)$ the same condition as in equation (2.5), i.e.

$$u_{\alpha}^{*}(q_n) = u_{\alpha}(2\pi - q_n). \tag{3.2}$$

Using definitions (2.1) and (2.2), Z is rewritten in the form

$$Z = \prod_{\alpha=1}^{l}\prod_{m=1}^{M}(1 + \rho_{m,\alpha}z_{m,\alpha}) \tag{3.3}$$

where

$$z_{m,\alpha} = \exp\left(\mathrm{i}M^{-1/2}\sum_{n=1}^{M-1} u_{\alpha}(q_n)\exp(-\mathrm{i}q_n m)\right) - 1. \tag{3.4}$$

Thus, the problem is reduced to the averaging of various products of position functions $\rho_{m,\alpha}$.

The averaging should be performed over the ensemble of random realizations with the same symbol content, and is determined by the simple combinatorics

$$\underbrace{\langle \rho_{m_1,\mathrm{A}_1} \cdots \rho_{m_{L_1},\mathrm{A}_1}}_{L_1} \cdots \underbrace{\rho_{m_{L_1+\ldots+L_{l-1}+1},\mathrm{A}_l} \cdots \rho_{m_{L_1+\ldots+L_l},\mathrm{A}_l}\rangle}_{L_l} = C^{M-L_1-\ldots-L_l}_{N_1-L_1,\ldots,N_l-L_l} / C^{M}_{N_1,\ldots,N_l} \tag{3.5}$$

$$C^{m}_{n_1,\ldots,n_l} = m!/n_1!\ldots n_l!$$
$$n_1+\ldots+n_l = m \qquad 0! = 1. \tag{3.6}$$

Here the angular brackets denote averaging over the ensemble, L_k is the number of position functions corresponding to the symbols A_k, and N_k is the total number of symbols A_k in a sequence with total length M. All position subscripts on the LHS of equation (3.5) must be different. The RHS of equation (3.5) is equal to the ratio of two combinatorial factors: $C^{M}_{N_1,\ldots,N_l}$ is the total number of different random realizations, while $C^{M-L_1-\ldots-L_l}_{N_1-L_1,\ldots,N_l-L_l}$ is the total number of realizations with respective $L_1,\ldots,L_l$ positions for different symbols held fixed (these are just the positions filled by symbols corresponding to the LHS of equation (3.5)). After collection of various random products and symmetrization over identical symbols, the final expression for the averaged characteristic function $\langle Z\rangle$ takes the form

$$\langle Z\rangle = 1/C^{M}_{N_1,\ldots,N_l}\Bigg(\sum_{L_1,\ldots,L_l=0}^{L_1\leqslant N_1,\ldots,L_l\leqslant N_l}{\sum_{m_1,\ldots,m_{L_1+\ldots+L_l}=1}^{M}}' C^{M-L_1-\ldots-L_l}_{N_1-L_1,\ldots,N_l-L_l}/L_1!\ldots L_l!\, z_{m_1,\mathrm{A}_1} \ldots z_{m_{L_1+\ldots+L_l},\mathrm{A}_l}\Bigg). \tag{3.7}$$

The prime on the sum over different positions means that all terms with at least two (or more) coincident position subscripts should be discarded.

The consideration shows that in the limit $N_\alpha \gg 1$, $M \gg 1$ the asymptotic (up to $\sim 1/N_\alpha$, $1/M$) cumulant expansion for $\ln\langle Z\rangle$ is given by

$$\ln\langle Z\rangle \approx \sum_{\{r\}}\sum_{\{q_{n_r}\}} \mathrm{i}^{r_1+\ldots+r_l}/r_1!\ldots r_l!\langle\!\langle \underbrace{\rho_{\mathrm{A}_1}(q_{n_1})\ldots\rho_{\mathrm{A}_1}(q_{n_{r_1}})}_{r_1}$$
$$\ldots \underbrace{\rho_{\mathrm{A}_1}(q_{n_{r_1+\ldots+r_{l-1}+1}})\ldots\rho_{\mathrm{A}_l}(q_{n_{r_1+\ldots+r_l}})}_{r_l}\rangle\!\rangle u_{\mathrm{A}_1}(q_{n_1})$$
$$\ldots u_{\mathrm{A}_1}(q_{n_{r_1}})\ldots u_{\mathrm{A}_l}(q_{n_{r_1+\ldots+r_{l-1}+1}})\ldots u_{\mathrm{A}_l}(q_{n_{r_1+\ldots+r_l}}). \tag{3.8}$$

Here the summations are restricted by the inequalities $0 \leqslant r_k \ll M$, $r_1+\cdots+r_l \ll M$, $2\pi/M \leqslant q_{n_r} \leqslant 2\pi(M-1)/M$. Besides that, in each cumulant $\langle\!\langle \rho_{\mathrm{A}_1}(q_{n_1})\ldots\rho_{\mathrm{A}_1}(q_{n_{r_1+\ldots+r_l}})\rangle\!\rangle$ a partial sum for any subset of wavenumbers except the sum over all wavenumbers cannot be equal to $2\pi p$ (where p is an integer), i.e.

$$\Big(\sum q_{n_r}\Big) \bmod 2\pi \neq 0 \qquad q_{n_r} \in (q_{n_1},\ldots,q_{n_{r_1+\ldots+r_l}}) \tag{3.9a}$$

$$(q_{n_1}+\cdots+q_{n_{r_1+\ldots+r_l}}) \bmod 2\pi = 0. \tag{3.9b}$$

The expressions for cumulants can be derived explicitly either directly from equation (3.7) or (this way is much simpler) by the recurrent extraction of the corresponding contributions from the sum rule (2.14).

3.2. Probability distribution functions

Below we restrict ourselves mainly to the statistical properties of harmonics (and their complex conjugates) with coincident wavenumbers q_n, which are in the leading terms with respect to $M^{-1/2}$ determined by the characteristic function

$$\langle Z\rangle_n = \exp\left(-\sum_{\alpha,\beta=1}^{l} \bar{F}_{\alpha\beta} u_\alpha u_\beta^*\right). \tag{3.10}$$

The same statistics can be described in terms of the probability distribution function for arbitrarily chosen $(l-1)$ symbols (since the remaining harmonic is expressed through equation (2.18)). It is obtained by reciprocating the characteristic function (3.10) for any $(l-1)$ variables (and equalizing the remaining one to zero),

$$P_{l-1}(|\rho_{\alpha_1}|,\ldots,|\rho_{\alpha_{l-1}}|;\varphi_{\alpha_1},\ldots,\varphi_{\alpha_{l-1}}) = (\det\|\bar{F}_{\alpha\beta}^{(l-1)}\|)^{-1}\exp\left(-\sum_{\alpha,\beta=1}^{l-1}\bar{R}_{\alpha\beta}^{(l-1)}\rho_\alpha\rho_\beta^*\right) \tag{3.11}$$

where

$$\sum_{\beta=1}^{l-1}\bar{R}_{\alpha\beta}^{(l-1)}\bar{F}_{\beta\gamma}^{(l-1)} = \sum_{\beta=1}^{l-1}\bar{F}_{\alpha\beta}^{(l-1)}\bar{R}_{\beta\gamma}^{(l-1)} = \delta_{\alpha\gamma} = \begin{cases} 1 & \alpha=\gamma \\ 0 & \alpha\neq\gamma. \end{cases} \tag{3.12}$$

$\bar{F}_{\alpha\beta}^{(l-1)}$ is an $(l-1)\times(l-1)$ submatrix of $\bar{F}_{\alpha\beta}$ (see equation (2.13)) for the chosen $(l-1)$ symbols, $\det\|\bar{F}_{\alpha\beta}^{(l-1)}\|$ is its determinant, the complex Fourier harmonics ρ_α are described in terms of modulus $|\rho_\alpha|$ and phase φ_α,

$$\rho_\alpha = |\rho_\alpha|\exp(\mathrm{i}\varphi_\alpha) \qquad \rho_\alpha^* = |\rho_\alpha|\exp(-\mathrm{i}\varphi_\alpha) \tag{3.13}$$

and any function $\Phi^{(l-1)}$ of variables $|\rho_\alpha|$, φ_α is averaged according to

$$\langle\Phi^{(l-1)}\rangle = \int \Phi^{(l-1)}P^{(l-1)}\,\mathrm{d}\Omega_{l-1} \tag{3.14}$$

$$\mathrm{d}\Omega_{l-1} = |\rho_{\alpha_1}|\mathrm{d}|\rho_{\alpha_1}|\,\mathrm{d}\varphi_{\alpha_1}/\pi\ldots|\rho_{\alpha_{l-1}}|\mathrm{d}|\rho_{\alpha_{l-1}}|\,\mathrm{d}\varphi_{\alpha_{l-1}}/\pi \tag{3.15}$$

with integrations over $0\leqslant|\rho_\alpha|\leqslant\infty$, $0\leqslant\varphi_\alpha\leqslant 2\pi$.

The distribution functions for the lower subsets of symbols can be obtained from P_{l-1} by integrations over the remaining variables. In particular, the one-symbol distribution function is given by

$$P_1(F_{\alpha\alpha}) = \exp(-F_{\alpha\alpha}/\bar{F}_{\alpha\alpha})/\bar{F}_{\alpha\alpha} \tag{3.16}$$

and coincides with the well known Rayleigh distribution [15].

4. Criteria of disorder in non-periodic sequences

4.1. Distribution of amplitudes of harmonics

The results of the previous section present some particular criteria of disorder that are useful in applications. The first test concerns the amplitudes of harmonics.

Due to the symmetry property (2.7) only half of the $(M-1)$ structure factor harmonics $F_{\alpha\alpha}(q_n)$ is statistically independent. Consider for definiteness the left half of the spectrum with $0 < q_n \leqslant \pi$. As is seen from equation (3.16), the probability of finding a given harmonic with an amplitude exceeding a fixed value $F_{\alpha\alpha}^{(0)}$ is equal to

$$\text{Prob}\{F_{\alpha\alpha} > F_{\alpha\alpha}^{(0)}\} = \int_{F_{\alpha\alpha}^{(0)}}^{\infty} \mathrm{d}F'_{\alpha\alpha}\, P_1(F'_{\alpha\alpha}) = \exp(-F_{\alpha\alpha}^{(0)}/\bar{F}_{\alpha\alpha}). \tag{4.1}$$

This means by definition that the average number of harmonics with heights exceeding $F_{\alpha\alpha}^{(0)}$ is given by

$$\langle n_\alpha \rangle = (M/2)\exp(-F_{\alpha\alpha}^{(0)}/\bar{F}_{\alpha\alpha}). \tag{4.2}$$

The condition $\langle n_\alpha \rangle = 1$ determines the characteristic maximum value:

$$F_{\alpha\alpha,\max} = \bar{F}_{\alpha\alpha}\ln(M/2). \tag{4.3}$$

Since the probability that all the $M/2$ harmonics would have heights less than $F_{\alpha\alpha}^{(0)}$ is equal to $[1-\exp(-F_{\alpha\alpha}^{(0)}/\bar{F}_{\alpha\alpha})]^{M/2}$, the probability that at least one from the $M/2$ harmonics exceeds $F_{\alpha\alpha}^{(0)}$ is defined as

$$\begin{aligned}\text{Prob}\{F_{\alpha\alpha} > F_{\alpha\alpha}^{(0)}; M/2\} &= 1 - [1-\exp(-F_{\alpha\alpha}^{(0)}/\bar{F}_{\alpha\alpha})]^{M/2} \\ &\approx 1 - \exp[-\exp(-(F_{\alpha\alpha}^{(0)} - F_{\alpha\alpha,\max})/\bar{F}_{\alpha\alpha})].\end{aligned} \tag{4.4}$$

Analogously, the probability that at least one from the $M/2$ harmonics has an amplitude less than $F_{\alpha\alpha}^{(0)}$ is determined as

$$\text{Prob}\{F_{\alpha\alpha} < F_{\alpha\alpha}^{(0)}; M/2\} = 1 - \exp(-F_{\alpha\alpha}^{(0)}/F_{\alpha\alpha,\min}) \tag{4.5a}$$

$$F_{\alpha\alpha,\min} = \bar{F}_{\alpha\alpha}/(M/2). \tag{4.5b}$$

The values $F_{\alpha\alpha,\max}$ and $F_{\alpha\alpha,\min}$ characterize the influence of mesoscopic fluctuations related to the particular random realizations.

4.2. Smoothed spectra

In many cases the order and disorder coexist with each other. It is very important to extract the underlying long-range correlations (if they exist). For example, $1/q_n^{\nu}$-like fluctuations in $F_{\alpha\alpha}(q_n)$ determine the (anti)persistent variations in the coarse-grained local density of the αth symbols depending on the sign of ν (i.e. the retaining or reversing in the tendency of variations during motion along a sequence from the beginning to the end) [17]. There are several methods for obtaining a solution to this problem [18]. The simplest consists of the partial smoothing of spectra in order to display the underlying trends. The smoothing means movable averaging over s left and right neighbours,

$$\tilde{F}_{\alpha\alpha}(q_n) = (2s+1)^{-1}\sum_{n'=n-s}^{n+s} F_{\alpha\alpha}(q_{n'}). \tag{4.6}$$

If $s \ll M$, then the correlations between harmonics with different q_n may be neglected and the approximate probability distribution function for the amplitudes of $\tilde{F}_{\alpha\alpha}(q_n)$ is given by the Nakagami distribution [19]:

$$\tilde{P}_s(\tilde{F}_{\alpha\alpha}) = \int_0^\infty \dots \int_0^\infty \delta\left(\tilde{F}_{\alpha\alpha} - (2s+1)^{-1} \sum_{n'=n-s}^{n+s} F_{\alpha\alpha}(q_{n'})\right) P_1(F_{\alpha\alpha}(q_{n-s}))$$
$$\dots P_1(F_{\alpha\alpha}(q_{n+s}))\, \mathrm{d}F_{\alpha\alpha}(q_{n-s}) \dots \mathrm{d}F_{\alpha\alpha}(q_{n+s})$$
$$= [(2s+1)/\bar{F}_{\alpha\alpha}]^{2s+1} \tilde{F}_{\alpha\alpha}^{2s} / \Gamma(2s+1) \exp[-(2s+1)\tilde{F}_{\alpha\alpha}/\bar{F}_{\alpha\alpha}] \tag{4.7}$$

where $P_1(F_{\alpha\alpha}(q_n))$ is defined by equation (3.16) and $\Gamma(2s+1)$ is a gamma function. The mean values of the height and standard deviation corresponding to the Nakagami distribution (4.7) are equal respectively to

$$\langle \tilde{F}_{\alpha\alpha} \rangle = \bar{F}_{\alpha\alpha} \qquad \sigma^2(\tilde{F}_{\alpha\alpha}) = \tilde{F}_{\alpha\alpha}^2/(2s+1). \tag{4.8}$$

In the limit $M \gg s \gg 1$ the distribution (4.7) tends to the Gaussian function, with the mean value and standard deviation determined by equation (4.8).

4.3. Mutual correlations

The excluded volume effects lead to the specific correlations even for the random sequences. The mutual correlations are characterized by the cross-correlation coefficients [15, 16]:

$$k(F_{\alpha\beta}|F_{\gamma\delta}) = \sum_{n=1}^{M-1} (F_{\alpha\beta}(q_n) - \bar{F}_{\alpha\beta})(F^*_{\gamma\delta}(q_n) - \bar{F}^*_{\gamma\delta})/(M-1)\sigma(F_{\alpha\beta})\sigma(F_{\gamma\delta}) \tag{4.9}$$

$$\sigma^2(F_{\alpha\beta}) = \sum_{n=1}^{M-1} (F_{\alpha\beta}(q_n) - \bar{F}_{\alpha\beta})(F^*_{\alpha\beta}(q_n) - \bar{F}^*_{\alpha\beta})/(M-1) \tag{4.10}$$

(see equations (2.6) and (2.13) for clarification of the nomenclature). The same characteristics can be calculated by the circular pair correlation functions (see equations (2.8)–(2.10), (2.15) and (2.16)):

$$\sigma(K^{\mathrm{c}}_{\alpha\beta}) = \sigma(F_{\alpha\beta})/M^{1/2} \tag{4.11}$$

$$k(K^{\mathrm{c}}_{\alpha\beta}|K^{\mathrm{c}}_{\gamma\delta}) = k(F_{\alpha\beta}|F_{\gamma\delta}). \tag{4.12}$$

Since the zeroth harmonics or correlation functions for $m_0 = 0$ do not contain any positional information, they are always discarded in the corresponding spectral sums.

Assuming the equivalence of the averaging over spectra to that over the ensemble of random realizations, and using equation (3.10) (or equation (3.11)), one obtains

$$\sigma^2(F_{\alpha\beta}) = \bar{F}_{\alpha\alpha}\bar{F}_{\beta\beta} \tag{4.13}$$

$$k(F_{\alpha\beta}|F_{\gamma\delta}) = \bar{F}_{\alpha\gamma}\bar{F}_{\delta\beta}/(\bar{F}_{\alpha\alpha}\bar{F}_{\beta\beta}\bar{F}_{\gamma\gamma}\bar{F}_{\delta\delta})^{1/2} \tag{4.14}$$

in particular, for $\alpha \neq \beta$,

$$k(F_{\alpha\alpha}|F_{\beta\beta}) \equiv k_{\alpha\beta} = N_\alpha N_\beta/(M-N_\alpha)(M-N_\beta). \tag{4.15}$$

Equation (4.15) has a simple probabilistic meaning. The cross-correlation coefficient $k_{\alpha\beta}$ is equal to the probability of simultaneously finding the αth symbols in positions free from the βth symbols and vice versa. The standard deviation $\sigma(F_{\alpha\alpha}) = \bar{F}_{\alpha\alpha}$ coincides with its counterpart for the spectrum of white noise [18].

At the end of this subsection we give an estimate for the cross-correlations between two uncorrelated random sequences (1 and 2) with the same lengths M:

$$k(F^{(1)}_{\alpha\beta}|F^{(2)}_{\gamma\delta}) = \sum_{n=1}^{M-1}(F^{(1)}_{\alpha\beta}(q_n) - \bar{F}^{(1)}_{\alpha\beta})(F^{*(2)}_{\gamma\delta}(q_n) - \bar{F}^{*(2)}_{\gamma\delta})/(M-1)\sigma(F^{(1)}_{\alpha\beta})\sigma(F^{(2)}_{\gamma\delta}).$$

As can easily be shown, after the independent averagings over two sequences the corresponding moments are approximately equal to

$$\langle k(F^{(1)}_{\alpha\beta}|F^{(2)}_{\gamma\delta})\rangle \approx 0 \tag{4.16a}$$

$$\langle k^2(F^{(1)}_{\alpha\beta}|F^{(2)}_{\gamma\delta})\rangle \approx 1/M(\bar{F}^{(1)2}_{\alpha\beta}\bar{F}^{(2)2}_{\gamma\delta}/\bar{F}^{(1)}_{\alpha\alpha}\bar{F}^{(1)}_{\beta\beta}\bar{F}^{(2)}_{\delta\delta}\bar{F}^{(2)}_{\gamma\gamma} + 1). \tag{4.16b}$$

Thus, the mesoscopic random cross-correlations do not exceed an order of magnitude $\sim M^{-1/2}$. The relative mesoscopic fluctuations in equations (4.13)–(4.15) are of the same order.

4.4. Structural entropy of sequence

Let a function $f(x)$ have monotonous first and smooth second derivatives. Then, an equation $f'(x) = \text{const}$ has a unique solution for any given value of the constant. Consider, now, a spectral sum,

$$S_\alpha = \sum_{n=1}^{M-1} f(F_{\alpha\alpha}(q_n)) \tag{4.17}$$

under the additional restriction

$$\sum_{n=1}^{M-1} F_{\alpha\alpha}(q_n) = \text{const} \tag{4.18}$$

(corresponding to the sum rule (2.13)). Using the standard technique of Lagrange multipliers, it is easy to see that the conditional extremum of S_α is attained for the strictly uniform distribution of structure factor harmonics over the spectrum:

$$F_{\alpha\alpha}(q_n) = \bar{F}_{\alpha\alpha}. \tag{4.19}$$

Since the heights of Fourier harmonics for the random sequences are distributed over wavenumbers q_n more uniformly than for an ordered state (cf., for example, the crystal and quasicrystal spectra with sets of sharp Bragg peaks), a sum S_α may serve as a structural entropy. The quantitative criterion for the random sequences can be obtained by averaging S_α with the probability distribution function (3.16):

$$\langle S_\alpha\rangle = (M-1)\langle f(F_{\alpha\alpha})\rangle \tag{4.20}$$

$$\langle f(F_{\alpha\alpha})\rangle = \int_0^\infty f(F_{\alpha\alpha})P_1(F_{\alpha\alpha})\,\mathrm{d}F_{\alpha\alpha}. \tag{4.21}$$

The typical choices of $f(F_{\alpha\alpha})$ are given by

$$f(F_{\alpha\alpha}) = \begin{cases} \ln F_{\alpha\alpha} & (4.22a) \\ -F_{\alpha\alpha} \ln F_{\alpha\alpha} & (4.22b) \\ F_{\alpha\alpha}^{r} & (4.22c) \end{cases}$$

with respective mean values equal to

$$\langle f(F_{\alpha\alpha}) \rangle = \begin{cases} \ln \bar{F}_{\alpha\alpha} - \mathbf{C} & (4.23a) \\ -\bar{F}_{\alpha\alpha} \ln \bar{F}_{\alpha\alpha} - \bar{F}_{\alpha\alpha}(1 - \mathbf{C}) & (4.23b) \\ \Gamma(r+1)\bar{F}_{\alpha\alpha}^{r} & (4.23c) \end{cases}$$

where $\bar{F}_{\alpha\alpha} = N_\alpha(M - N_\alpha)/M(M-1)$, $\mathbf{C} = 0.577\,215\ldots$ is the Euler constant, and $\Gamma(r+1)$ is a gamma function. Function (4.22*a*) corresponds to the standard spectral definition of information entropy [18] (the various definitions are equivalent only for quasi-Gaussian statistics). The particular calculations show, however, two serious drawbacks in this case. First, the weak logarithmic dependence is not sensitive enough and, secondly, S_α tends to infinity for any hidden periodicity (this choice does not discern, for example, between a doubled random sequence and a two-periodic sequence). In practical applications the choice of the function $f(F_{\alpha\alpha})$ in equation (4.17) is restricted by the requirement that the mesoscopic fluctuations related to the particular random realizations are small.

Functions (4.22*a*) and (4.22*b*) correspond to the conditional maximum, while function (4.22*c*) gives the conditional maximum for $0 < r < 1$ and the minimum for $r > 1$.

The more universal criteria can be obtained by using the relative variables $F_{\alpha\alpha}(q_n)/\bar{F}_{\alpha\alpha}$ or the local spectral probabilities introduced in [3],

$$p_\alpha(q_n) = F_{\alpha\alpha}(q_n)/(M-1)\bar{F}_{\alpha\alpha}. \tag{4.24}$$

Generally, the multidimensional fractal formalism [20] should be applied to the many-symbolic sequences with scaling invariance:

$$\sum_{n=1}^{M-1} p_{\alpha_1}^{r_1}(q_n) \ldots p_{\alpha_{l-1}}^{r_{l-1}}(q_n) \sim (M-1)^{-\tau(r_1,\ldots,r_{l-1})}. \tag{4.25}$$

As is seen from equations (4.5) and (4.23*c*), the random mesoscopic fluctuations dominate in the range of exponents $r < -1$. In the range of positive values $r > 0$ the finite-size effects related to the singular outbursts (4.3) start to become important since $r \gtrsim \ln M / \ln\ln M$, and special measures should be taken to avoid the artefact multifractality [3].

5. Particular applications

5.1. Substitutional sequences

In this section, several applications of the spectral criteria will be illustrated. In order to demonstrate the sensitivity of various criteria, one of the sequences is chosen to be nearly random, while the other is distinctly non-random but has a relatively broad spectrum. We begin with the substitutional sequences. Their growth is determined by consecutive iterations according to the inflation rule:

$$\mathrm{A}_1 \to \sigma_1(\mathrm{A}_1, \ldots, \mathrm{A}_l), \ldots, \mathrm{A}_l \to \sigma_l(\mathrm{A}_1, \ldots, \mathrm{A}_l) \tag{5.1}$$

where $\sigma_k(A_1, \ldots, A_l)$ is some combination of the symbols $A_1, \ldots, A_l$. The corresponding inflation rules may be both deterministic (as in equation (5.1)) and probabilistic (e.g. see [21]).

Figure 1 shows typical results for the four-symbol Rudin–Shapiro substitution

$$A \to AC \qquad B \to DC \qquad C \to AB \qquad D \to DB. \tag{5.2}$$

The Rudin–Shapiro inflation is known to be strongly randomizing [2, 3]. For definiteness we consider the iterational growth from the symmetric seed ABCD. As can be checked, during subsequent iterations, A and D occupy invariably only the odd positions, while B and C are placed in even sites. This gives sharp coherent Bragg peaks at $q_n = \pi$,

$$\rho_A(0) = \rho_B(0) = \rho_C(0) = \rho_D(0) = -\rho_A(\pi) = \rho_B(\pi) = \rho_C(\pi) = -\rho_D(\pi) = N/M^{1/2} \tag{5.3}$$

$$N = 2^p \qquad M = 2^{p+2} \qquad p \geqslant 1 \tag{5.4}$$

where p is the number of iterations. For this reason such a sequence can be represented either by the intersite merging of two quasi-random binary sequences or by the partial destruction of exact two-periodicity. The coherent harmonics at $q_n = \pi$ should be subtracted from the corresponding criteria for the random sequences, e.g. (see equations (2.7), (2.12) and (5.3))

$$\tilde{F}_{\alpha\alpha} = 2\left(\sum_{n=1}^{M/2-1} F_{\alpha\alpha}(q_n)\right)(M-2)^{-1} = 2^{p-2}/(2^{p+1}-1) \tag{5.5}$$

and analogously for equations (4.9) and (4.10).

The positions of A and D (or B and C) are completely correlated, $k_{AD} = k_{BC} = 1$ (as should be the case for the binary sequences (cf equations (2.18) and (4.15))), while the correlations between odd and even symbols are nearly absent, e.g. $k_{AB} = k_{AC} = k_{DB} = k_{DC} = 1.47 \times 10^{-3}$ for $p = 10$ ($M = 4096$). The numerical values of standard deviations coincide practically with the respective mean values of harmonics (5.5) (cf equation (4.13)). The values of structural entropies (4.17), (4.22*b*) (without harmonics with $q_n = \pi$) are equal to 7.89×10^2 for $p = 10$ (in comparison to 8.48×10^2 predicted by equations (4.20) and (4.23*b*) with $(M-1)$ replaced by $(M-2)$). The significant deviations in figure 1(*c*) are observed for $\sim$ 30–40 harmonics from 2047. All these results support the suggestion of strong randomization in Rudin–Shapiro inflation (preserving, however, the partially destroyed underlying two-periodicity).

Figure 2 illustrates the other example of structural ordering, the binary Thue–Morse sequence

$$A \to AB \qquad B \to BA. \tag{5.6}$$

This sequence has a relatively broad structural spectrum corresponding to the fractal filling of the various periodic positions. The dominant role is played by the fractal three-periodicity. The even harmonics are equal to zero, and harmonics with small wavenumbers ($n \ll M$) are asymptotically suppressed (the analytical treatment of these results can be found in [2, 3, 22]).

Taking into account the equality to zero of even harmonics in the Thue–Morse sequence, a more correct comparison would be performed with the doubled random sequence (with

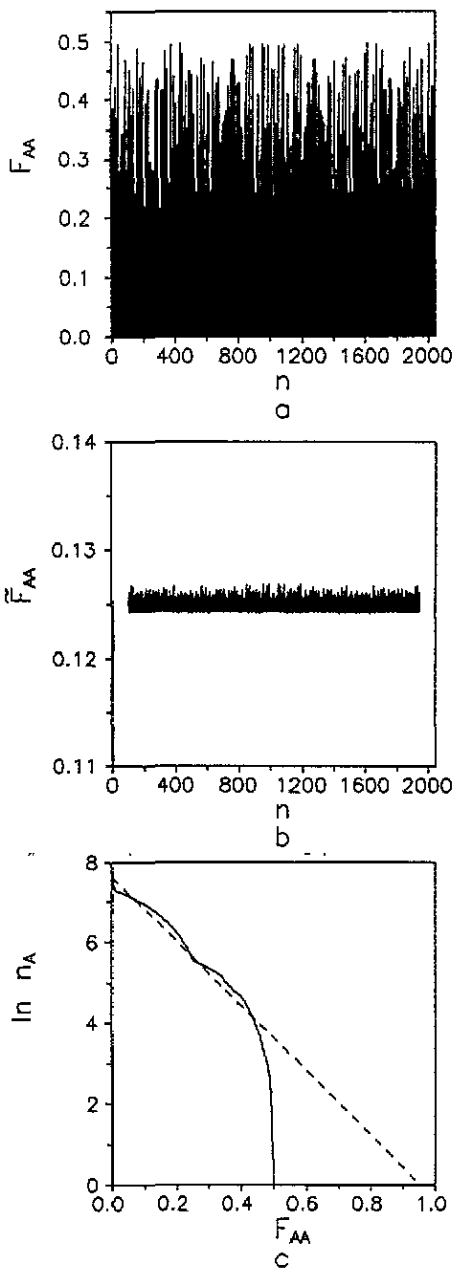

Figure 1. The spectral characteristics for the Rudin–Shapiro sequence generated from the ABCD seed after $p = 10$ iterations ($M = 4096$). (*a*) The structure factor harmonics $F_{AA}(q_n)$ with $1 \leqslant n \leqslant M/2 - 1$. (*b*) The smoothed spectrum for $s = 100$ (equation (4.6)). (*c*) Plot of the logarithm of the number of structure factor harmonics exceeding a given value F_{AA} (full curve). The broken line corresponds to the theoretical prediction for random sequences (4.2) with $\bar{F}_{\alpha\alpha}$ determined by equation (5.5).

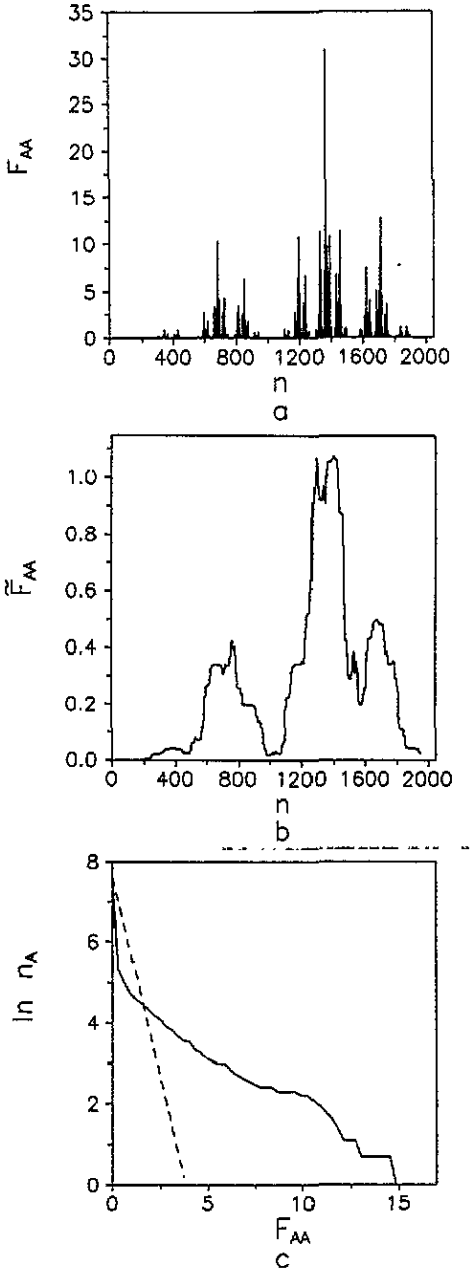

Figure 2. The spectral characteristics for the Thue–Morse sequence generated from the A-seed after $p = 12$ iterations ($M = 4096$). (*a*) The structure factor harmonics $F_{AA}(q_n) = F_{BB}(q_n)$ with $1 \leqslant n \leqslant M/2 - 1$. (*b*) The smoothed spectrum for $s = 100$. (*c*) Plot of the logarithm of the number of odd structure factor harmonics exceeding a given value F_{AA} (full curve). The broken line corresponds to the theoretical distribution of even harmonics for the doubled random counterpart with $\bar{F} = \frac{1}{2}$.

zero odd harmonics), which gives $\bar{F} = \frac{1}{2}$ for iterations started either from A or B. All summations and averages are in this case calculated only for non-zero harmonics. The values for the standard deviation and structural entropy (4.17), (4.22*b*) are respectively equal to 1.74 and -1.26×10^3 for $p = 12$ iterations started from A ($M = 4096$), in comparison to 0.5 and 2.77×10^2 for the doubled random counterpart. The dependences of the logarithm of the number of non-zero harmonics exceeding a given value F_{AA} shown in figure 2(*c*) are also quite different. The maximum amplitude of harmonics $F_{AA}(q_n)$ is equal to 31.0 ($n = 1365$, $F_{max}/\bar{F} = 62.0$), and is much higher than the singular random outbursts (equation (4.3)).

Figure 3 illustrates the corresponding variations in the circular pair correlation function (equation (2.8)) $\Delta K_{AA}(m_0) = K_{AA}(m_0) - \bar{K}_{AA}$ and the current standard deviations coarse grained over a scale of 100 sites according to

$$\sigma_{m_0}(K^c_{AA}) = \left((100)^{-1} \sum_{m_0'=m_0}^{m_0+99} (K^c_{AA}(m_0') - \bar{K}^c_{AA})^2 \right)^{1/2} \tag{5.7}$$

where $\bar{K}^c_{AA} = N_A(N_A - 1)/M(M-1)$ and $m_0 = 1, 101, 201, \ldots, M/2$. As is seen from figure 3, besides the dominating three-periodicity there are additional long-range correlations over $\sim M/2$ sites.

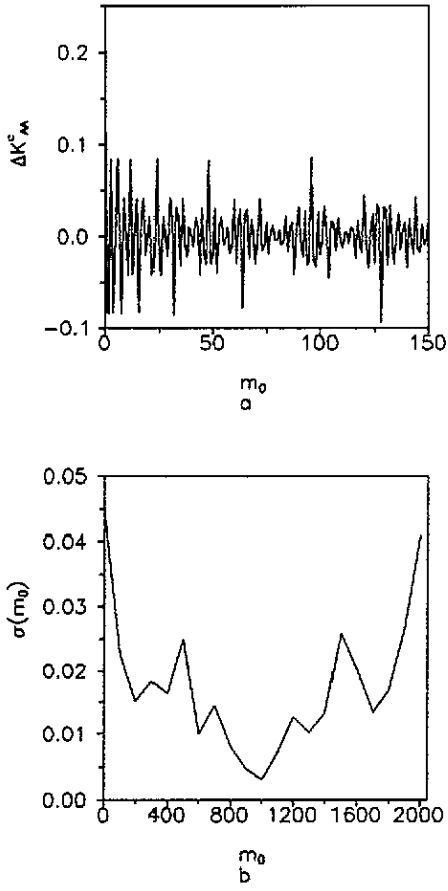

Figure 3. The variations in (*a*) the pair correlation function (equation (2.8)) $\Delta K_{AA}(m_0) = K_{AA}(m_0) - \bar{K}_{AA}$ and (*b*) the coarse-grained current standard deviation determined by equation (5.7).

The results illustrate the strongly non-random character of the Thue–Morse sequence. Although this fact is well known, the example gives a useful test for the criteria concerned.

5.2. Symbolic dynamics

The symbolic dynamics describes the coarse-grained behaviour of a dynamical system [4–6]. A phase space of a system is subdivided by the net of finite cells marked by symbols, and the consecutive visiting of the cells during evolution generates the symbolic sequence. The natural partitioning exists only for the one-dimensional maps. As an illustration we consider the logistic map (e.g. see [23]):

$$x_{n+1} = rx_n(1 - x_n) \tag{5.8}$$

with $1 < r \leqslant 4$. The partitioning determined by the zero of the derivative of the function $f(x) = rx(1 - x)$ is defined as follows. If $0 < x_n < \frac{1}{2}$, then symbol L is placed at the nth position, while if $\frac{1}{2} < x_n < 1$, then symbol R must be substituted. Since the sequence is binary, its structural characteristics are determined by the positions of L (or R) symbols only (equation (2.18)). The mean value of spectral harmonics is equal to

$$\bar{F}_{LL} = \bar{F}_{RR} = N_L N_R / M(M - 1) \tag{5.9}$$

where N_L and N_R are the total numbers of L- and R-symbols, and $M = N_L + N_R$ is the total length of a sequence.

Figures 4 and 5 present the results for the fully chaotic evolution ($r = 4$) and approximate three-periodic regime with intermittency ($r_c - r = 0.002$, $r_c = 1 + \sqrt{8}$) [23, 24]. The numerical parameters for $x_1 = 0.4$ after $p = 4095$ iterations ($M = 4096$) are: (i) $r = 4$; $N_L = 2040$, $N_R = 2056$; $\bar{F}_{LL} = \bar{F}_{RR} = 0.250$; $\sigma(F_{LL}) = \sigma(F_{RR}) = 0.251$; the values of the structural entropies defined by equations (4.17), (4.22*a*) and (4.22*b*) are -8.05×10^3 and 9.81×10^2; the theoretical values calculated by equations (4.20), (4.23*a*) and (4.23*b*) are -8.04×10^3 and 9.86×10^2; (ii) $r_c - r = 0.002$, $r_c = 1 + \sqrt{8}$; $N_L = 1263$, $N_R = 2833$; $\bar{F}_{LL} = \bar{F}_{RR} = 0.213$; $\sigma(F_{LL}) = \sigma(F_{RR}) = 0.370$; the values of the structural entropies defined by equations (4.17), (4.22*a*) and (4.22*b*) are -1.03×10^4 and 6.27×10^2; the theoretical values calculated by equations (4.20), (4.23*a*) and (4.23*b*) are -8.69×10^3 and 9.80×10^2. These results demonstrate clearly the sensitivity of various criteria. We have not observed the $1/f^{\alpha}$ tails at small wavenumbers typical for intermittency [23] in figures 5(*a*) and 5(*b*), probably because of the relative shortness of the sequence.

5.3. Structural analysis of DNA sequences

In this subsection we consider an object of greater primary scientific importance, DNA sequences. The recent observations of the long-range correlations in a variety of genomes [9–11, 25] attach additional interest to this problem. Our considerations will follow mainly the lines discussed above. The discussion of the other aspects of Fourier analysis for DNA sequences can be found in [7, 26] (and the references therein). For the convenience of the reader we recall the main facts from genetics (e.g., see, [27]).

The information for the development of an organism is stored in long DNA sequences (called genomes) consisting of four different nucleotides: adenine (A), cytosine (C), guanine (G) and thymine (T). The whole genome is subdivided by segments (genes) with different genetic functions. The transmission of information from DNA to proteins is governed by the triplet genetic code: a triplet of nucleotides corresponds to one amino acid, a sequence

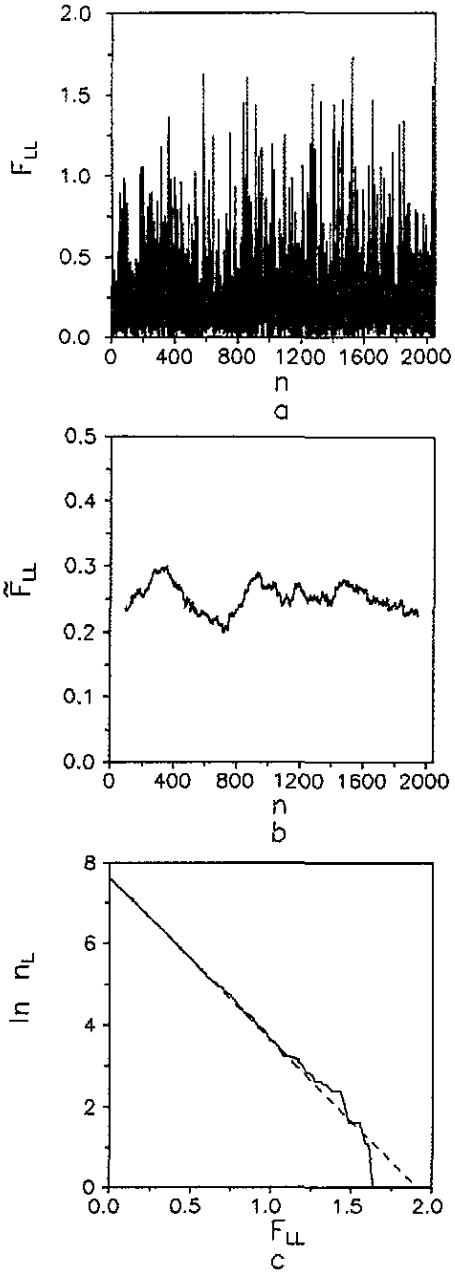

Figure 4. The spectral characteristics for the symbolic sequence generated by the logistic map (5.8) with $r = 4$ and $x_1 = 0.4$ after $p = 4095$ iterations ($M = 4096$). (*a*) The structure factor harmonics $F_{LL}(q_n) = F_{RR}(q_n)$ with $1 \leqslant n \leqslant M/2$. (*b*) The smoothed spectrum for $s = 100$ (equation (4.6)). (*c*) Plot of the logarithm of the number of structure factor harmonics exceeding a given value F_{LL} (full curve). The broken line corresponds to the theoretical prediction for random sequences (4.2) with $\bar{F}_{LL}$ determined by equation (5.9).

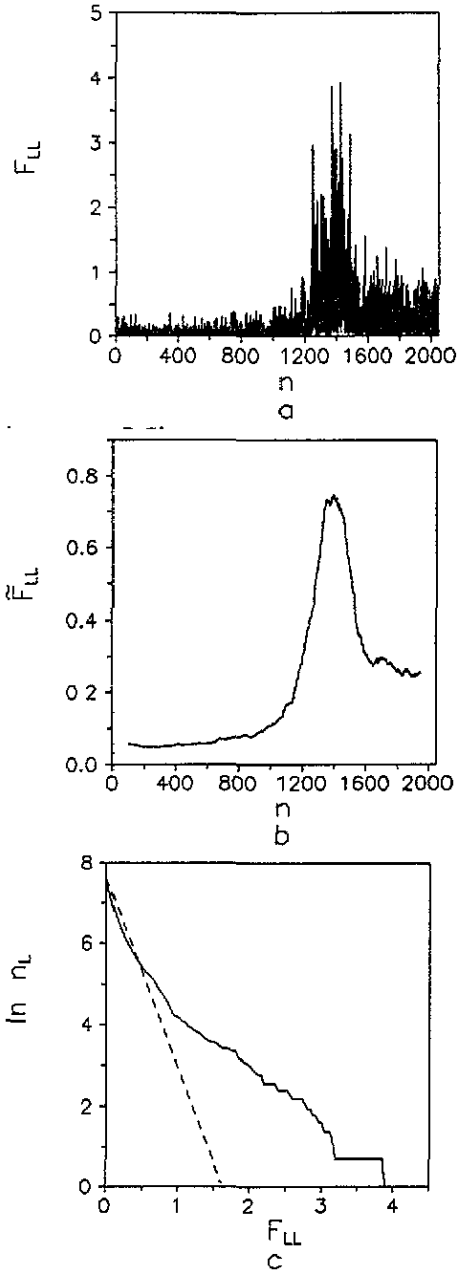

Figure 5. The structural characteristics for the symbolic sequence generated by the logistic map with the same parameters as in figure 4 except for $r_c - r = 0.002$, $r_c = 1 + \sqrt{8}$.

of amino acids forms a protein. In 1976 Crick *et al* [28] (see also [29]) suggested that the primitive progenomes in the earliest stages of evolution were formed by RNY coding triplets (where R=(A,G) is purine, Y=(C,T) is pyrimidine, and N is any base). The underlying three-periodicity has since been identified in various natural genomes and used for determination of the protein-coding stretches [30]. Here we shall show how these features are displayed via structural symbolic analysis.

As a particular example we shall use the well studied genome of the bacteriophage PHIX174 [31]. Similar behaviour has been observed in five other genomes of viruses (INXXF1, MIG4XX, INIKE, PPR and TOEAV) [32]. The left halves of the symmetrical diagonal structure factor spectra for PHIX174 are shown in figure 6. Two peculiar features deserve special attention and are separately reproduced in inserts to figure 6. The first one is the higher initial harmonics in the A- and T-series (with amplitudes about 5–10 times the mean level), indicating the underlying superstructure and long-range correlations discussed below. The detailed calculations by Voss [10] show the ubiquitous nature of such a behaviour (see also [9, 11, 25]). The other feature is the very high peaks near $q_n = 2\pi/3$, giving evidence for the strong contribution of three-periodic constituents. Their relative heights and other structural characteristics are summarized in table 1. We should remember that for the random sequences the corresponding standard deviations are equal to $\sigma(F_{\alpha\alpha}) = \bar{F}_{\alpha\alpha}$ (equation (4.13)) and they may also be considered as structural entropies (cf equations (4.20), (4.22*c*) and (4.23*c*)). The peaks corresponding to three-periodicity are usually the highest ones in the spectra, but not always. For instance, the maximum for the $F_{CC}(q_n)$ structure factor is attained at $n = 2116 (F_{CC}/\bar{F}_{CC} = 9.32)$. Criterion (4.4) excludes safely the occasional origin of three-periodicity in the A-, G- and T-series, which are distinctly more ordered that the C-sequence (see table 1 and figure 9 below).

The partially destroyed three-periodicity does not exhaust the underlying long-range correlations. Figure 7 shows the additional persistent variations for the A- and T-nucleotides and mixed antipersistent–persistent behaviours depending on the length scale for the C- and G-series displayed through the smoothed spectra (cf section 4.2). The study of Hurst's curves [17] supports this conclusion. Discussion of such a method is, however, outside the scope of our present paper and the corresponding results will be published separately [32].

The three-periodicity and additional long-range correlations over $\sim 10^3$ sites are also clearly seen in the variations of circular pair correlation functions (equations (2.8), (2.9) and (2.16)) shown in figure 8. The current standard deviations in figure 8(*b*) have been coarse grained over a scale of 100 sites analogously to equation (5.7) with $\bar{K}^c_{\alpha\alpha} = N_\alpha(N_\alpha - 1)/M(M-1)$ (equation (2.16)).

The positions of maximum harmonics near $q_n = 2\pi/3$ give the key to understanding the origin of long-range correlations. The number of harmonics closest to the exact three-periodicity is equal to $n = 1795$. The small but distinct deviations from this value in table 1 are evidently related to the long-range modulations of the nucleotide densities.

It is easy to prove the following result. Let $k = 2, 3, \ldots$ be a series of integers. Then, the number of modulations in the envelope of the maxima of $\cos(q_n m)$ (where $q_n = 2\pi n/M$) during changing m from 1 to M and n/M lying near $1/k$ is equal to the modulus of the difference:

$$n_s = |kn - M|. \tag{5.10}$$

The crossover from a period k to $k+1$ occurs for harmonics with n determined by the relationship

$$(k+1)n - M = M - kn. \tag{5.11}$$

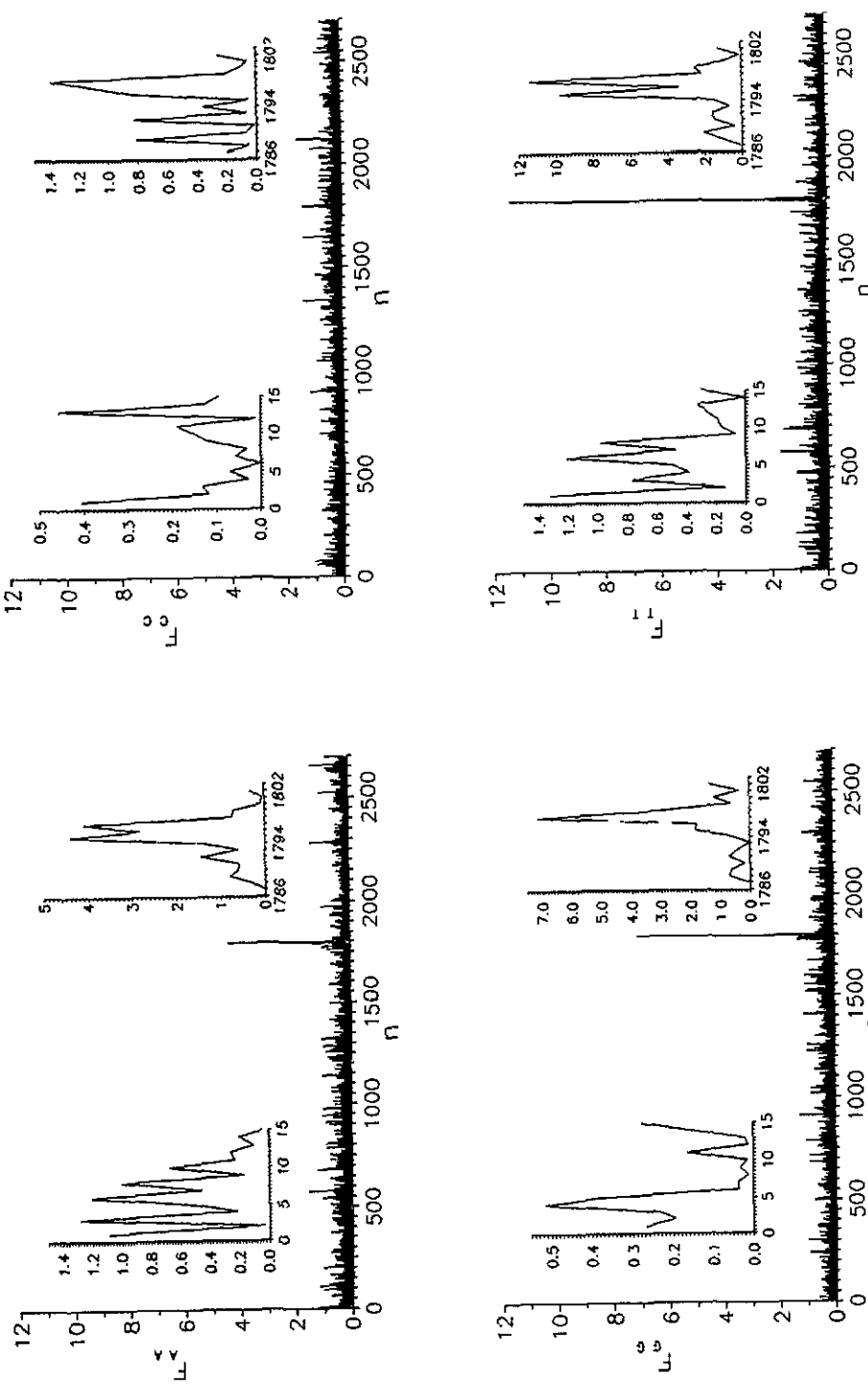

Figure 6. The structure factor spectra for the different nucleotide sequences in the genome of phage PHIX174.

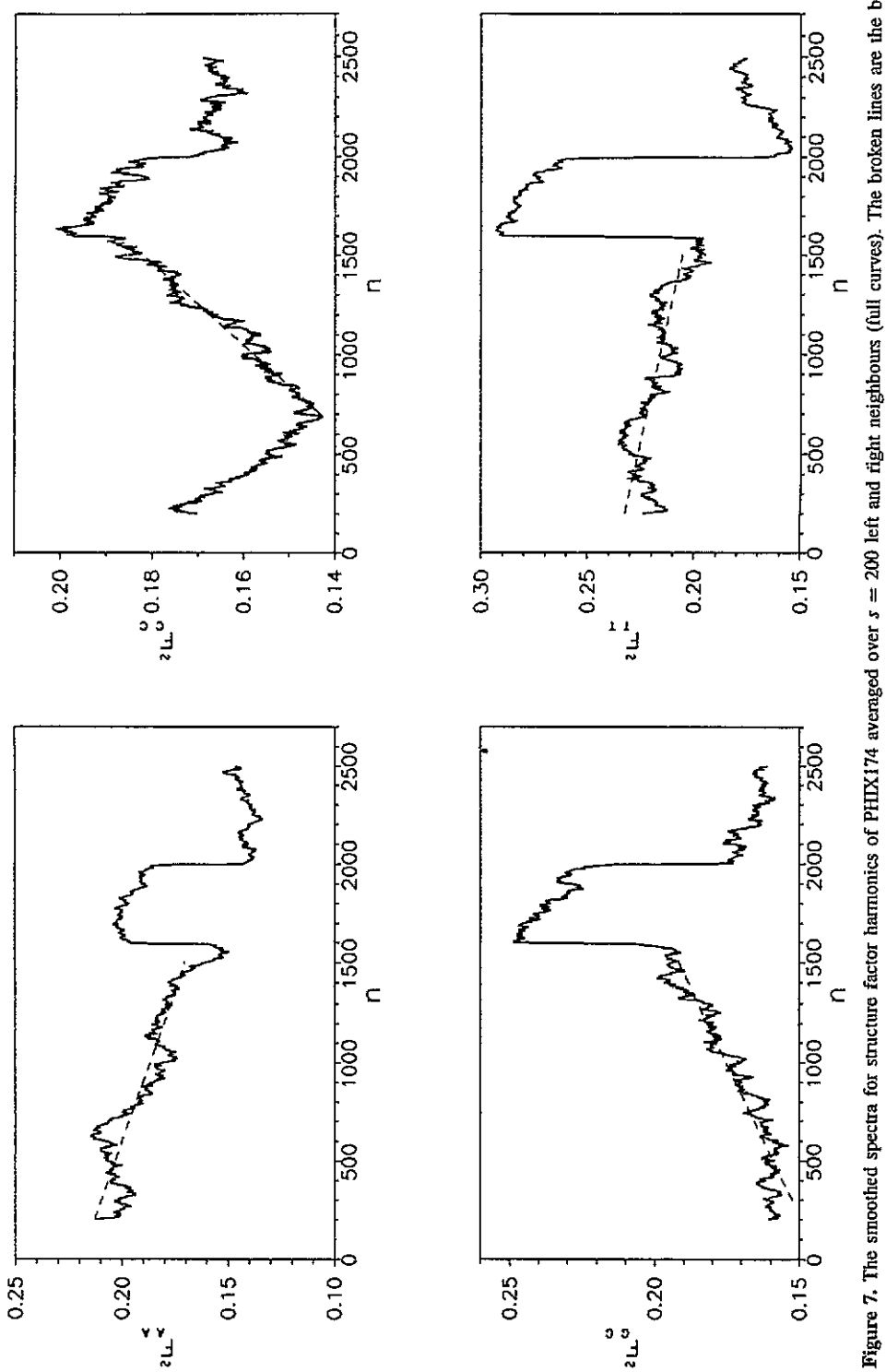

Figure 7. The smoothed spectra for structure factor harmonics of PHIX174 averaged over $s = 200$ left and right neighbours (full curves). The broken lines are the best linear fits in the corresponding intervals and are shown as a guide for the eye.

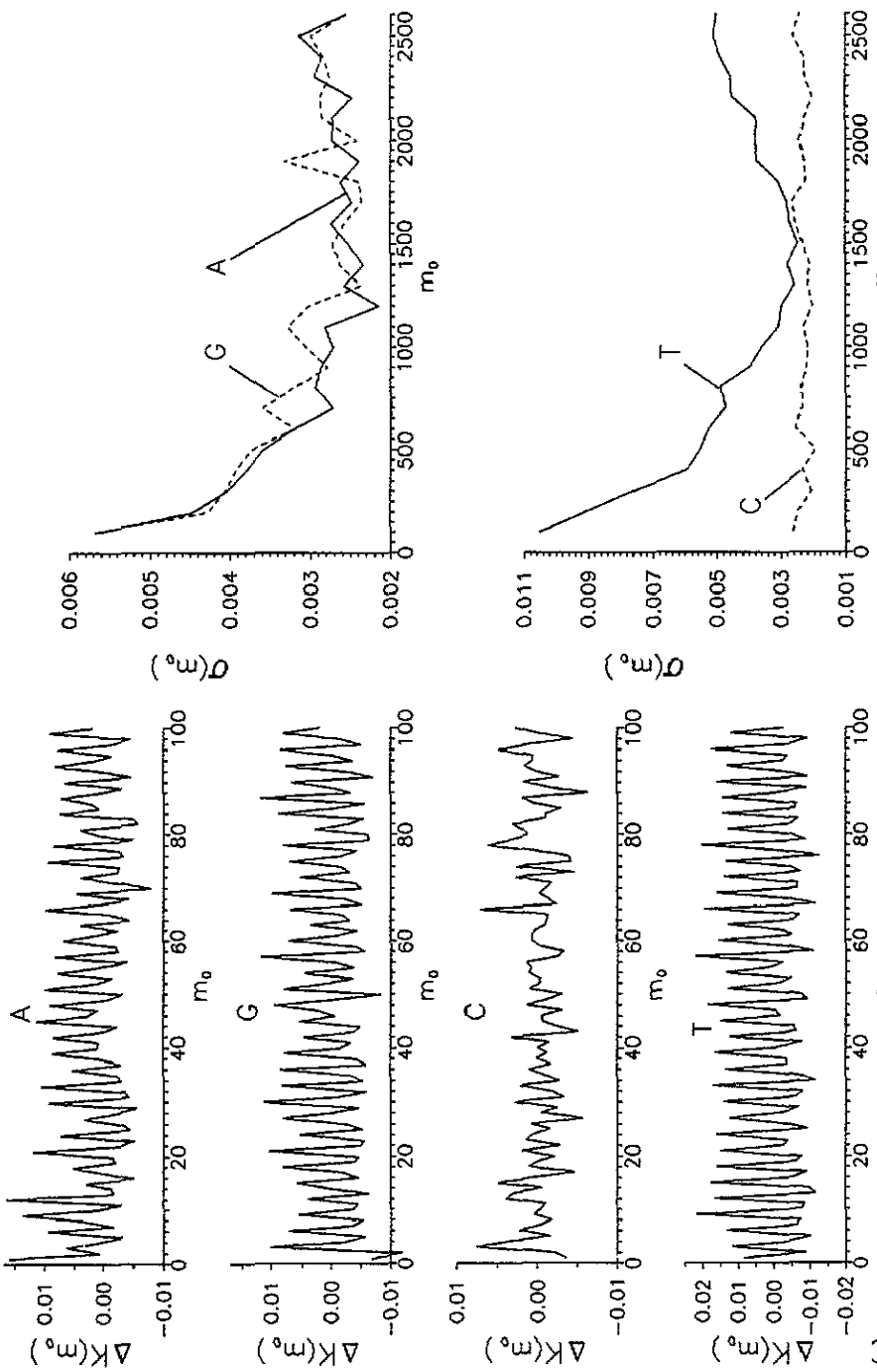

Figure 8. (*a*) Variation in the deviation of pair correlation functions, $\Delta K_{\alpha\alpha}(m_0) = K_{\alpha\alpha}(m_0) - \bar{K}_{\alpha\alpha}$, for the genome of PHIX174. (*b*) Current standard deviations, $\sigma_\alpha(m_0)$, for $\Delta K_{\alpha\alpha}(m_0)$ calculated for consecutive intervals of 100 sites of m_0 (see equation (5.7)).

Table 1. Summary of structural analysis for genome of bacteriophage PHIX174, $M = 5386$.

Structural characteristics	Nucleotides			
	A	C	G	T
Nucleotide content	1291	1157	1254	1684
Mean value of harmonics, $\bar{F}_{\alpha\alpha}$ (equation (2.13))	0.182	0.169	0.179	0.215
Standard deviation, $\sigma_\alpha \equiv \sigma(F_{\alpha\alpha})$ (equation (4.10))	0.226	0.170	0.234	0.365
Relative height of maximum harmonics $(F_{\max}/\bar{F})_{\alpha\alpha}$ near $q_n = 2\pi/3$ and their number	24.2 ($n = 1795$)	8.24 ($n = 1798$)	40.2 ($n = 1797$)	53.2 ($n = 1797$)
Information entropy (equations (4.17) and (4.22*a*))	-1.25×10^4	-1.26×10^4	-1.25×10^4	-1.17×10^4
Relative information entropy $(\langle S_\alpha\rangle_{\text{random}} - S_\alpha)/\lvert S_\alpha\rvert$ (equations (4.20) and (4.23*a*))	1.92×10^{-2}	-4.58×10^{-3}	9.50×10^{-3}	2.54×10^{-2}
Structural entropy for the choice defined by (equations (4.17) and (4.22*b*))	1.19×10^3	1.23×10^3	1.19×10^3	1.15×10^3
Relative structural entropy $(\langle S_\alpha\rangle_{\text{random}} - S_\alpha)/S_\alpha$ (equations (4.20) and (4.23*b*))	5.74×10^{-2}	-5.15×10^{-4}	4.64×10^{-2}	1.24×10^{-1}

	k_{AC}	k_{AG}	k_{AT}	k_{CG}	k_{CT}	k_{GT}
Cross-correlation coefficients $k_{\alpha\beta} \equiv k(F_{\alpha\alpha}\lvert F_{\beta\beta})$ (equations (4.9) and (4.10))	0.140	0.348	0.545	0.179	0.160	0.535
Relative cross-correlation coefficients $(k_{\alpha\beta} - k_{\alpha\beta,\text{random}})/k_{\alpha\beta}$ (equations (4.9) and (4.15))	0.384	0.725	0.737	0.537	0.221	0.742

Equation (5.11) corresponds to the coincidence of modulational superperiods at the crossover value n. The modulations of the maxima are the longest for the harmonics with ratios n/M near $1/k$, and shorten in the crossover regions. The modulational superperiods can be observed only when within the interval $(2\pi/(k+1), 2\pi/k)$ there are at least two different wavenumbers q_n, or

$$1/k - 1/(k+1) > 1/M. \tag{5.12}$$

If $k(k+1) > M$ and $n < M/k$, then the realization of superperiods is impossible. The shortest superperiods correspond to the transition from two-periodicity ($k = 2$) to three-periodicity ($k = 3$). As is seen from equations (5.10) and (5.11), in this case $n_s = [M/5]$ and $n = [2M/5]$ (the square brackets denote the whole parts of the quotients).

Returning now to the genome of phage PHIX174, it can easily be checked that harmonics with $n = 1795$ and 1797 generate respectively, 1 and 5 long superperiods overlapping the three largest ($\sim 10^3$ sites) and several shorter (~ 400 sites) genes. These superperiods are responsible for long-range correlations and are related to the segmentation of the genome by the separate genes.

The distributions of structure factor harmonics according to their heights are shown in figure 9.

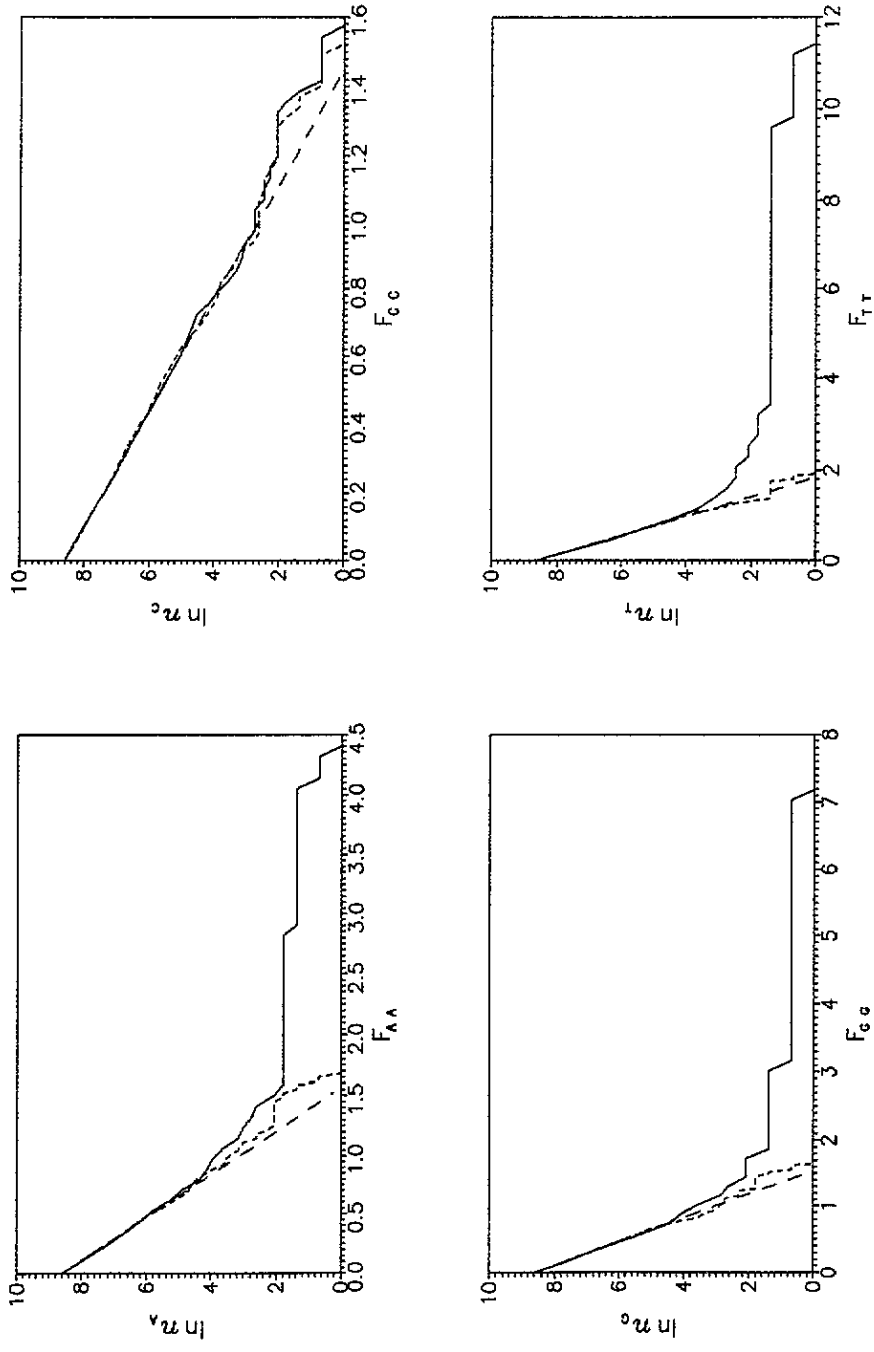

Figure 9. Plots of the logarithm of the number of structure factor harmonics $n_{\alpha\alpha}$ with heights exceeding a given value $F_{\alpha\alpha}$ for the different nucleotide sequences of PHIX174. The full curves describe the distributions of harmonics for PHIX174, the broken lines with long dashes correspond to the theoretical prediction for the random sequences (equation (4.2)), while the broken curves with short dashes correspond to the particular control random sequence with the same nucleotide composition.

The formation of three-periodicity is performed on the binary rather that tetrary level (see also [28, 29]). This means that subdivision of four different nucleotides into two various classes and the subsequent identification of nucleotides within each class retains the underlying three-periodicity. The three different subdivisions include: (i) R = (A, G) and Y = (C, T); (ii) W = (A, T) and S = (C, G); (iii) M = (A, C) and K = (G, T). For the binary sequences the positions of one subsequence determine fully that of the other, and therefore $F_{RR}(q_n) = F_{YY}(q_n)(n \neq 0)$, etc. The number of maximum harmonics and their relative heights for the binary sequences in the genome of PHIX174 are equal to: $n = 1795$, $F_{RR,max}/\bar{F}_{RR} = 33.8(R - Y)$; $n = 1797$, $F_{WW,max}/\bar{F}_{WW} = 35.0(W - S)$; $n = 1797$, $F_{MM,max}/\bar{F}_{MM} = 35.7(M - K)$. The underlying three-periodicity is thus nearly degenerate with respect to various subdivisions. The harmonic $n = 1795$ for the R–Y sequence gives, however, the closest value to the exact three-periodicity. In this sense the R–Y sequence produces the most complete three-periodicity.

The formation of genomes invokes the general problem of automatic generation of sequences with dominating three-periodicity. It is worth noting that the Thue–Morse model may have some interest not only for arithmetic or the theory of multifractal structural spectra but for genetics as well. The canonical mechanism of genomic growth (e.g., see, [8]) is assumed to be performed schematically by the replication–duplication $\begin{matrix}\cdots\mathrm{R}\cdots\mathrm{Y}\cdots\\ \cdots\mathrm{Y}\cdots\mathrm{R}\cdots\end{matrix} \rightarrow \ldots\mathrm{R}\ldots\mathrm{Y}\ldots\mathrm{R}\ldots\mathrm{Y}\ldots$. The Thue–Morse inflation R $\rightarrow$ RY, Y $\rightarrow$ YR can be considered as the intersite merging of two complementary strands $\begin{matrix}\cdots\mathrm{R}\cdots\mathrm{Y}\cdots\\ \cdots\mathrm{Y}\cdots\mathrm{R}\cdots\end{matrix} \rightarrow \ldots\mathrm{RY}\ldots\mathrm{YR}\ldots$. The doubled Thue–Morse inflation $\begin{matrix}\cdots\mathrm{RY}\cdots\mathrm{YR}\cdots\\ \cdots\mathrm{YR}\cdots\mathrm{RY}\cdots\end{matrix} \rightarrow \ldots\mathrm{RYYR}\ldots\mathrm{YRRY}\ldots$ is morphologically equivalent to the following canonical duplication: $\begin{matrix}\cdots\mathrm{RYYR}\cdots\\ \cdots\mathrm{YRRY}\cdots\end{matrix} \rightarrow \ldots\mathrm{RYYR}\ldots\mathrm{YRRY}\ldots$. Thus, the Thue–Morse zipper mechanism unites elegantly the dominating three-periodicity with variability of a sequence. The alternative approach [28, 29] consists of the consecutive growth of three-periodicity $\ldots\mathrm{RRY}\ldots$ (or $\ldots\mathrm{RYY}\ldots$) with subsequent canonical replication–duplication.

6. Conclusion

The examples considered above show that the spectral representation is a versatile and powerful tool for the identification of random constituents in symbolic sequences. The additional advantage of the symbolic Fourier transformation is related to the numerous fast computational algorithms known in the literature [18, 33]. The technique of analysis of Fourier spectra is also well developed.

The spectral representation permits identification of finite-memory effects as well. The qualitative picture may be described by using the standard Ornstein–Uhlenbeck approximation [15] with exponential decay both for the pair correlation functions *(equation (2.8) with $m_0 < M/2$) and p-periodic oscillations,*

$$\Delta K_{\alpha\alpha}(m_0) = K_{\alpha\alpha}(m_0) - \tilde{K}_{\alpha\alpha} \sim \exp(-m_0/r_c) \tag{6.1a}$$

$$\Delta K_{\alpha\alpha}(m_0) \sim \cos(2\pi m_0/p)\exp(-m_0/r_c) \tag{6.1b}$$

and by the reciprocating Wiener–Khinchin relationship (2.10). An elementary consideration shows that in the case (6.1*a*) the small wavenumber range of the spectrum with $q_n \lesssim 1/r_c$ will be enriched by the higher harmonics (if $\Delta K > 0$) or have a deepening (if $\Delta K < 0$), while in the case (6.1*b*) the finite-memory effects will produce the typical Lorentzian-like

smearing of Bragg peaks with widths $\Delta q \sim 1/r_c$. In this sense the randomly destroyed three-periodicity in DNA sequences (or two-periodicity in the Rudin–Shapiro substitution) can easily be differentiated from the finite Markovian memory with $r_c = 3$ or from the damping three-periodicity with finite decay. The extension of statistical criteria to the random sequences with memory needs, however, separate investigation.

Acknowledgments

The authors are indebted to A A Ezhov and V A Kutvitskii for advice during computations, to A L Chernjakov, E B Levchenko and A A Vedenov for useful discussions, and to Yu A Sprizhitskii and M V Malysheva for permission to use the databank of the Moscow Institute of Molecular Genetics. Financial support from the emergency programme of the International Science Foundation is also acknowledged.

References

[1] Janssen T and Janner A 1987 *Adv. Phys.* **36** 519
[2] Queffelec M 1987 *Substitution Dynamical Systems: Spectral Analysis (Springer Lecture Notes in Mathematics 1294)* (Berlin: Springer)
[3] Godreche C and Luck J M 1990 *J. Phys. A: Math. Gen.* **23** 3769
[4] Alexeev V M and Yakobson M V 1981 *Phys. Rep.* **75** 290
[5] Hao B-L 1989 *Elementary Symbolic Dynamics and Chaos in Dissipative Systems* (Singapore: World Scientific)
[6] Hao B-L 1991 *Physica* D **51** 161
[7] Tavare S and Giddings B M 1989 *Mathematical Methods for DNA Sequences* ed M S Waterman (Boca Raton, FL: CRC Press) pp 116–32
[8] Nussinov R 1987 *J. Theor. Biol.* **125** 219
[9] Li W and Kaneko K 1992 *Europhys. Lett.* **17** 655
[10] Voss R F 1992 *Phys. Rev. Lett.* **68** 3805
[11] Peng C-K, Buldyrev S V, Goldberger A L, Havlin S, Sciortino F, Simons M and Stanley H E 1992 *Nature* **356** 168
[12] Burrows B L and Sulston K W 1991 *J. Phys. A: Math. Gen.* **24** 3979
[13] Kluiving R, Capel H W and Pasmanter R A 1992 *Physica* A **183** 67, 96; **186** 405
[14] Shannon C E and Weaver M 1949 *The Mathematical Theory of Communication* (Chicago, IL: University of Illinois Press)
[15] Feller W 1970 *An Introduction to Probability Theory and Its Applications* (New York: Wiley)
[16] Van Campen N G 1984 *Stochastic Processes in Physics and Chemistry* (Amsterdam: North-Holland)
[17] Feder J 1988 *Fractals* (New York: Plenum)
[18] Marple S L Jr 1987 *Digital Spectral Analysis with Applications* (Englewood Cliffs, NJ: Prentice Hall)
[19] Nakagami M 1960 *Statistical Methods in Radio Wave Propagation* ed W C Hoffman (New York: Pergamon) p 3
[20] Meneveau C, Sreenivasan K R, Kailasnath P and Fan M S 1990 *Phys. Rev.* A **41** 894
[21] Li W 1991 *Phys. Rev.* A **43** 5240
[22] Cheng Z, Savit R and Merlin R 1988 *Phys. Rev.* B **37** 4375
Cheng Z and Savit R 1991 *Phys. Rev.* A **44** 6379
[23] Schuster H G 1984 *Deterministic Chaos* (Weinheim: Physik)
[24] Hirsch J E, Huberman B E and Scalapino D 1981 *Phys. Rev.* A **25** 519
[25] Karlin S and Brendel V 1993 *Science* **259** 677
[26] Silverman B D and Linsker R J 1986 *J. Theor. Biol.* **118** 295
Deev A A, Edintsov I M, Ivanitsky G R and Kunisky A S 1989 *Biofizika* **34** 564 (in Russian)
Benson D C 1990 *Nucleic Acids Res.* **18** 3001, 6305
[27] Alberts B, Bray D, Lewis J, Raff M and Watson J D 1983 *Molecular Biology of the Cell* (New York: Garland)

[28] Crick F H C, Brenner S, Klug A and Pieczenik G 1976 *Orig. Life* **7** 389
[29] Eigen M, Gardiner W, Schuster P and Winkler-Oswatitsch R 1981 *Sci. Am.* **244** 88
[30] Pieczenik G 1980 *Proc. Natl Acad. Sci USA* **77** 3539
Shepherd J C W 1981 *J. Mol. Biol.* **17** 94; *Proc. Natl Acad. Sci. USA* **78** 1596
Fickett J W 1982 *Nucleic Acids Res.* **10** 5303
[31] Sanger F, Air G M, Barrel B G, Brown N L, Coulson A R, Fiddes J C, Hutchison C A III, Slocombe P M and Smith N 1977 *Nature* **265** 687
[32] Chechetkin V R, Knizhnikova L A and Turygin A Yu 1994 *J. Biomol. Struct. Dyn.* **11** in press
[33] Cheever E A, Overton G C and Searls D B 1991 *Comput. Appl. Biosci.* **8** 143