Journal of Statistical Physics, Vol. 60, Nos. 5/6, 1990

Mutual Information Functions versus Correlation Functions

Wentian Li[1]

Received October 16, 1989; final March 13, 1990

This paper studies one application of mutual information to symbolic sequences: the mutual information function $M(d)$. This function is compared with the more frequently used correlation function $\Gamma(d)$. An exact relation between $M(d)$ and $\Gamma(d)$ is derived for binary sequences. For sequences with more than two symbols, no such general relation exists; in particular, $\Gamma(d)=0$ may or may not lead to $M(d)=0$. This linear, but not general, independence between symbols separated by a distance is studied for ternary sequences. Also included is the estimation of the finite-size effect on calculating mutual information. Finally, the concept of "symbolic noise" is discussed.

KEY WORDS: Mutual information function; correlation functions; linear and general dependence; symbolic noise.

1. INTRODUCTION

Mutual information is a measure of the dependence between two variables.[1] If the two variables are independent, the mutual information between them is zero. If the two are strongly dependent, e.g., one is a function of another, the mutual information between them is large. There are other interpretations of the mutual information; for example, the stored information in one variable about another variable, and the degree of the predictability of the second variable by knowing the first. Clearly, all these interpretations are related to the same notion of dependence and correlation.

The correlation function is another frequently used quantity to measure dependence. (The "function" in the name "correlation function" is

[1] Center for Complex Systems Research, Physics Department, Beckman Institute, University of Illinois, Urbana, Illinois 61801; Department of Physics, Columbia University, New York, New York 10027. Present address: Santa Fe Institute, Santa Fe, New Mexico 87501.

0022-4715/90/0900-0823$06.00/0

used because correlation is usually measured as a function of distance or time delay between two quantities.) It is now well understood that mutual information measures the general dependence, while the correlation function measures the linear dependence, and mutual information is a better quantity than the correlation function to measure the dependence. This difference leads to different methods in choosing the independent variables for constructing the phase trajectory in the study of chaotic dynamics.[2,3]

Another major difference between mutual information and correlation function is that the former can be applied to symbolic sequences as well as numerical sequences, but the latter can only be used on numerical sequences. This makes mutual information a natural alternative to the correlation function for symbolic sequences. In this context, I will use the name "mutual information function" to refer to the fact that it is the functional form—of the mutual information versus the distance between the two variables—that is emphasized.

Random signals, usually called "noise," are classified by their power spectra, and equivalently, correlation functions. Since a correlation function cannot be applied directly to symbolic sequences, the classification of the random symbolic sequences is rarely discussed. Most of the previous studies of letter sequences in natural languages or nucleotide sequences in DNA polymers are focused on entropy (starting from Shannon[4]). Sometimes, the nearest-neighbor correlations using conditional probabilities are also studied.[5,6] With the mutual information function, it is conceivable that a more complete characterization of the symbolic sequences can be accomplished.

Let me reintroduce the definition of the mutual information function and the correlation function for a finite sequence $\{x_i\}$ $(i=1, 2,..., N)$, where $x_i \in \{a_\alpha\}$ $(\alpha = 1, 2,..., K)$, the variable set. The correlation function is

$$\Gamma(d) \equiv \sum_{\alpha} \sum_{\beta} a_\alpha a_\beta P_{\alpha\beta}(d) - \left(\sum_{\alpha} a_\alpha P_\alpha \right)^2 \tag{1.1}$$

Both the single-site probabilities $\{P_\alpha\}$ and the joint probabilities for two sites $\{P_{\alpha\beta}(d)\}$ are accumulated from the single sequence to be analyzed. The (site-to-site) mutual information function is defined as

$$M(d)^{[1]} \equiv \sum_{\alpha}^{K} \sum_{\beta}^{K} P_{\alpha\beta}(d) \log \frac{P_{\alpha\beta}(d)}{P_\alpha P_\beta} \tag{1.2}$$

The block-to-block mutual information is defined as the mutual information between two L-blocks, i.e., blocks with L sites, separated by a distance of d sites. It is similar to the site-to-site mutual information function except

that P_α are the probabilities for L-blocks and $P_{\alpha\beta}(d)$ are the joint probabilities for two L-blocks:

$$M(d)^{[L]} \equiv \sum_{\alpha}^{K^L} \sum_{\beta}^{K^L} P_{\alpha\beta}(d) \log \frac{P_{\alpha\beta}(d)}{P_\alpha P_\beta} \tag{1.3}$$

If not specified, I will consider only the site-to-site mutual information functions (and the superscript $^{[1]}$ is dropped).

As an illustration for a comparison between correlation functions and mutual information functions, as well as the effects of block length L, Fig. 1 shows the $M(d)^{[L]}$ $(L=1, 2, 3, 4)$ and $\Gamma(d)$ for the binary sequences generated by nearest-neighbor cellular automaton rule 110.[7] Notice that $\Gamma(d)$ can either be positive or negative, whereas $M(d)$ is always non-negative. The peak at $d=14$ is an indication of the underlining periodicity of 14 in the sequences.

Mutual information has been applied to sequences in various fashions. Chaitin[8] proposes to split a system and calculate the mutual information among the components; then the maximum value of the mutual information in all conceivable partitions is used to mathematically define "life" or "organization." Shaw[9] and Grassberger,[10] among others, use the mutual

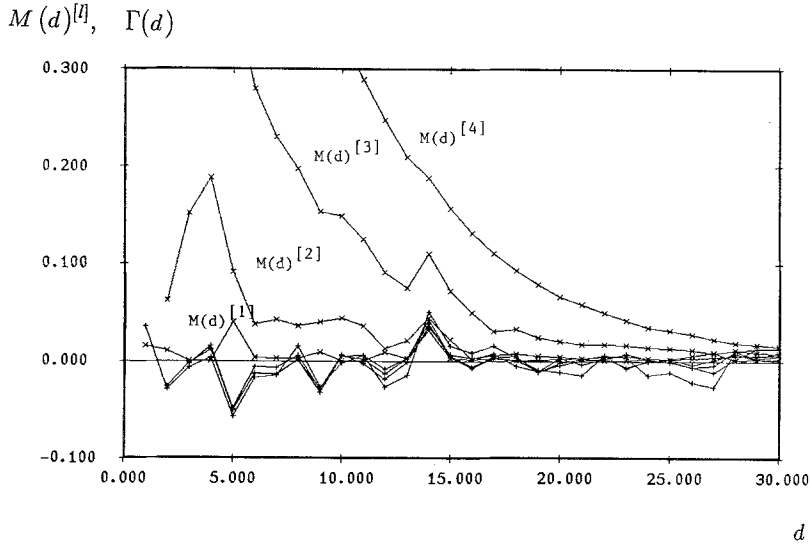


Fig. 1. $\Gamma(d)$ and $M(d)^{[L]}$ $(L=1, 2, 3, 4)$ for spatial sequences generated by cellular automaton rule 110 (or the following rule: $000 \to 0$, $001 \to 1$, $010 \to 1$, $011 \to 1$, $100 \to 0$, $101 \to 1$, $110 \to 1$, and $111 \to 0$). The sequence lengths are $N=400$, 1600, 6400, and 25,600, respectively, for increasing L.

information between two semi-infinite blocks in a sequence to define the "complexity." These applications of mutual information lead to a single value which characterizes the sequence. Our mutual information function gives a whole function.

In the definition of the correlation function, the probabilities are weighted by the variable values. As a result, correlation functions generally are not directly related to mutual information functions. For example, we can have zero value of correlation function at some distance d, while the mutual information function at that distance can be any value. In a special case that the joint probability distribution is Gaussian, it has been proved that the correlation function and mutual information are directly related to each other.[3] In this paper, the connection between the two functions in another special case (i.e., for binary sequences) will be given.

The paper is organized as the follows: Section 2 is about the relation between $\Gamma(d)$ and $M(d)$[1] for binary sequences; this relation is then illustrated by examples of regular languages in Section 3; Section 4 shows that for ternary sequences, the connection between the two functions fails to be true, and we discuss the case of "weak correlation"; Section 4 changes to a different topic of how to estimate the finite-size effect in a calculation of mutual information; and Section 5 discusses the concept of "symbolic noise."

2. RELATION BETWEEN $M(d)$ AND $\Gamma(d)$ FOR BINARY SEQUENCES

In this section, I will show that if the sequence is binary (i.e., there are only two symbols), the number of the independent joint probabilities for two sites is reduced from four to one. As a consequence, the correlation function can be directly related to the mutual information function.

Notice that only the nonzero variable values can contribute to the correlation function; we have the correlation function for binary sequences:

$$\Gamma(d) = P_{11}(d) - P_1^2 \tag{2.1}$$

where $P_{11}(d)$ is the joint probability for having two symbol 1's separated by distance d, and P_1 is the probability for having symbol 1.

At the first glance, one would be tempted to conclude that since the mutual information function needs all four joint probabilities to specify its value, whereas the correlation function needs only one, there should be no direct relation between the two functions. Nevertheless, because the two variables whose joint probabilities are to be calculated are extracted from the same stationary sequence, we have the following constraints.

First of all, suppose the sequence has no particular direction, i.e., the sequence is the same whether being looked at from left to right or from right to left. If this is true, we have the symmetry constraint on the joint probabilities:

$$P_{\alpha\beta}(d) = P_{\beta\alpha}(d) \tag{2.2}$$

for $\alpha, \beta \in (0, 1)$.

Second, by the very definition of the joint probability, we have

$$P_\alpha = \sum_{\beta=0}^{1} P_{\alpha\beta}(d) \tag{2.3}$$

There is otherwise nothing particularly interesting about this formula except that the right-hand side of the equation is a function of distance d, whereas the left-hand side is not! The implication is that the functional form of the two expressions $P_{\alpha\beta}(d)$ should be such that they cancel each others d-dependent term.

We also have the normalization condition

$$\sum_{\alpha\beta} P_{\alpha\beta}(d) = 1 \tag{2.4}$$

but it turns out that it is equivalent to the condition $\sum_\alpha P_\alpha = 1$ and will not provide more reductions to the number of independent $P_{\alpha\beta}(d)$.

For binary sequences, the first constraint provides one reduction, and the second provides two reductions. The number of independent joint probabilities is actually one!

Carrying out the details, we get

$$\begin{aligned} P_{01}(d) &= P_{10}(d) = P_1 - P_{11}(d) \\ P_{00}(d) &= (1 - 2P_1) + P_{11}(d) \end{aligned} \tag{2.5}$$

In term of correlation function, these become

$$\begin{aligned} P_{11}(d) &= \Gamma(d) + P_1^2 \\ P_{00}(d) &= \Gamma(d) + P_0^2 \\ P_{01}(d) = P_{10}(d) &= -\Gamma(d) + P_0 P_1 \end{aligned} \tag{2.6}$$

The relation between mutual information function and the correlation function for binary sequences is

$$\begin{aligned} M(d) = \Gamma(d) \log \frac{[1 + \Gamma(d)/P_1^2][1 + \Gamma(d)/P_0^2]}{[1 - \Gamma(d)/P_0 P_1]^2} + P_1^2 \log\left(1 + \frac{\Gamma(d)}{P_1^2}\right) \\ + P_0^2 \log\left(1 + \frac{\Gamma(d)}{P_0^2}\right) + 2P_0 P_1 \log\left(1 - \frac{\Gamma(d)}{P_0 P_1}\right) \end{aligned} \tag{2.7}$$

One approximation to the above equation is when the correlation function decays to zero at longer distances and both $\Gamma(d)/(P_\alpha P_\beta)$ are small. In this limit, we found that all the first-order terms of $\Gamma(d)$ are canceled, and only the second-order terms remain:

$$M(d) \approx \frac{\Gamma(d)^2}{2}\left(\frac{1}{P_0^2}+\frac{1}{P_1^2}+\frac{2}{P_0 P_1}\right)=\frac{1}{2}\left(\frac{\Gamma(d)}{P_0 P_1}\right)^2 \tag{2.8}$$

An interesting observation from this equation is that mutual information functions decay to zero at a faster rate than the corresponding correlation functions. For example, if $\Gamma(d) \sim 1/d^\beta$, then $M(d) \sim 1/d^{2\beta}$. This result is important in the study of symbolic noise to be discussed in the last section. For example, if we want to identify the symbolic sequences which are analogous to the numerical sequences with $1/f^\alpha$ ($\alpha \approx 1$) power spectra (called $1/f^\alpha$ noise), because we know that the correlation function for $1/f$ noise behaves like $\Gamma(d) \sim 1/d^{1-\alpha}$, we would expect that $M(d)$ for the symbolic sequence also behaves like a power law function with a different exponent $M(d) \sim 1/d^{2(1-\alpha)}$ *if the sequence is binary.*

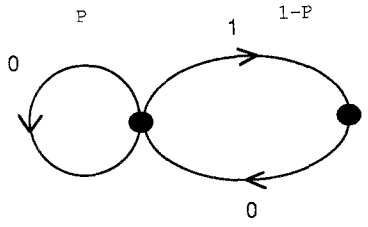
For sequences with more than two symbols, both the correlation function and the mutual information function receive contributions from more than one independent joint probability. One has to assume all the functional forms for these joint probabilities before making a connection between the two. Any relation between the exponents of the two power law functions for $\Gamma(d)$ and $M(d)$ will depend on a particular assumption made about the joint probabilities.

3. EXAMPLES OF MARKOV CHAIN AND REGULAR LANGUAGE

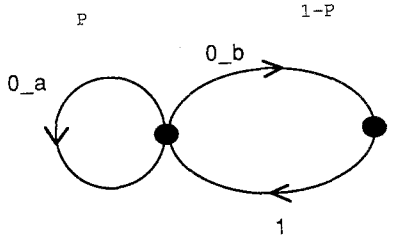
To illustrate the dependence among the joint probability $P_{\alpha\beta}(d)$'s for binary sequences, we include examples of a Markov chain and a regular language in this section. For Markov chains, the one-step transition probabilities $T_{\alpha \to \beta}$ are given, and all the d-step transition probabilities can be derived from the one-step transition probabilities. It is well known that the joint probabilities $P_{\alpha\beta}(d)$ decay exponentially with distance d, and so do the correlation functions (see, e.g., ref. 11).

Figure 2a shows a Markov chain with the one-step transition probabilities $T_{0\to 0}=p$, $T_{0\to 1}=1-p$, $T_{1\to 0}=1$, and $T_{1\to 1}=0$. These transition probabilities can be grouped into one matrix:

$$\mathbf{T}=\begin{array}{c} \\ 0 \\ 1 \end{array}\begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} p & 1-p \\ 1 & 0 \end{pmatrix} \end{array} \tag{3.1}$$



(A)



(B)

Fig. 2. (a) The graph which represents a simple Markov chain with transition probabilities $T_{0\to 0}=p$, $T_{0\to 1}=1-p$, $T_{1\to 0}=1$, and $T_{1\to 1}=0$. (b) A regular language similar to the Markov chain in (a).

where $T_{\alpha\beta}\equiv T_{\alpha\to\beta}$. The dth power of this matrix gives the d-step transition probabilities and the joint probabilities $P_{\alpha\beta}(d)=P_\alpha(T^d)_{\alpha\beta}$. The left eigenvector of the matrix for eigenvalue 1 gives the invariant probabilities (i.e., densities) of the symbols. They are

$$P_0=\frac{1}{2-p}, \qquad P_1=\frac{1-p}{2-p} \tag{3.2}$$

The dth power of the one-step transition matrix is

$$\mathbf{T}^d=\frac{1}{2-p}\begin{pmatrix}1 & 1-p\\ 1 & 1-p\end{pmatrix}+\frac{(-1+p)^d}{2-p}\begin{pmatrix}1-p & -(1-p)\\ -1 & 1\end{pmatrix} \tag{3.3}$$

Multiplying the first row by P_0 and the second row by P_1, we have the joint probabilities [all $P_{\alpha\beta}(d)$ are in one matrix $\mathbf{P}(d)$]:

$$\mathbf{P}(d)=\begin{pmatrix}\dfrac{1}{(2-p)^2}+\dfrac{1-p}{(2-p)^2}(-1+p)^d & \dfrac{1-p}{(2-p)^2}-\dfrac{1-p}{(2-p)^2}(-1+p)^d\\ \dfrac{1-p}{(2-p)^2}-\dfrac{1-p}{(2-p)^2}(-1+p)^d & \dfrac{(1-p)^2}{(2-p)^2}+\dfrac{1-p}{(2-p)^2}(-1+p)^d\end{pmatrix} \tag{3.4}$$

As expected, the formula $P_{\alpha\beta}(d)=(-1)^{1-\delta_{\alpha\beta}}\Gamma(d)+P_\alpha P_\beta$ in Eq. (2.6) indeed holds ($\delta_{\alpha\beta}=1$ if $\alpha=\beta$, and $=0$ if $\alpha\neq\beta$).

Figure 2b shows an example of a regular language[12] which is not exactly a Markov chain by the original definition, because the transition probability from symbol 0 to 1 can either be zero (if it is the symbol 0 on the left branch, or 0_a), or one (if it is the symbol 0 on the right branch, or 0_b). Nevertheless, it becomes a Markov chain if we consider it as a sequence with three symbols: 0_a, 0_b, and 1. Similar procedures can be carried out, but instead of dealing with a 2-by-2 matrix, we calculate the dth power of a 3-by-3 transition matrix. Notice that the joint probability, for example, $P_{00}(d)$ is determined by all four $P_{0_\alpha 0_\beta}(d)$ (0_α, 0_β are either 0_a or 0_b). A detailed calculation showed that the joint probabilities for sequences generated by Fig. 2b are exactly the same as those generated by Fig. 2a, and I will not include the details here. For more examples of calculating the joint probabilities for regular languages, see ref. 13.

4. WEAK CORRELATION IN TERNARY SEQUENCES

Call two variables $\{a_\alpha\}$ and $\{b_\beta\}$ *linearly independent* if $\sum_{\alpha\beta} a_\alpha b_\beta P_{\alpha\beta} = (\sum_\alpha a_\alpha P_\alpha)(\sum_\beta b_\beta P_\beta)$ for all α, β, and *generally independent* if $P_{\alpha\beta} = P_\alpha P_\beta$ for all α, β.[3] The linear independence is equivalent to the zero correlation (function), and the general independence is equivalent to the zero mutual information. For binary sequences, because $\Gamma(d)$ is related to $M(d)$ by Eq. (2.7), linear independence leads to general independence. Nevertheless, for sequences with more than two symbols, linear independence may or may not result in general independence. In this section, I will examine ternary sequences, i.e., sequences with three symbols, to see what constraints apply to $M(d)$ when $\Gamma(d) = 0$. I will call the two sites having zero correlation but nonzero mutual information *weakly correlated*, instead of using the long phrase "linearly independent but generally dependent."

Following a similar argument to Section 2, the nine joint probabilities for two-site ternary sequences are reduced to three independent functions by three symmetry conditions $P_{\alpha\beta}(d) = P_{\beta\alpha}(d)$ and three definitions of densities $P_\alpha = \sum_\beta P_{\alpha\beta}(d)$. Choose $P_{00}(d)$, $P_{11}(d)$, and $P_{22}(d)$ as the three independent functions.

It is easy to show that other joint probabilities become (for $\alpha \neq \beta$; γ is the third index, which is not equal to α or β):

$$P_{\alpha\beta}(d) = \tfrac{1}{2}[P_{\gamma\gamma}(d) - P_{\alpha\alpha}(d) - P_{\beta\beta}(d)] + \tfrac{1}{2}(-P_\gamma + P_\alpha + P_\beta) \qquad (4.1)$$

Set the correlation $\Gamma(d)$ equal to zero:

$$\begin{aligned} 0 = \Gamma(d) &= P_{11}(d) + 2P_{12}(d) + 2P_{21}(d) + 4P_{22}(d) - (P_1 + 2P_2)^2 \\ &= 2[P_{00}(d) - P_0^2] - [P_{11}(d) - P_1^2] + 2[P_{22}(d) - P_2^2] \end{aligned} \qquad (4.2)$$

Then $P_{11}(d)$ is no longer an independent function, and it is related to $P_{00}(d)$ and $P_{22}(d)$ by

$$P_{11}(d) = 2[P_{00}(d) - P_0^2] + 2[P_{22}(d) - P_2^2] + P_1^2 \tag{4.3}$$

Other joint probabilities in terms of the two independent functions $P_{00}(d)$ and $P_{11}(d)$ are

$$\begin{aligned} P_{01}(d) &= \tfrac{3}{2}[-P_{00}(d) + P_0^2] + \tfrac{1}{2}[-P_{22}(d) + P_2^2] + P_0 P_1 \\ P_{02}(d) &= \tfrac{1}{2}[P_{00}(d) - P_0^2] + \tfrac{1}{2}[P_{22}(d) - P_2^2] + P_0 P_2 \\ P_{12}(d) &= \tfrac{1}{2}[-P_{00}(d) + P_0^2] + \tfrac{3}{2}[-P_{22}(d) + P_2^2] + P_1 P_2 \end{aligned} \tag{4.4}$$

In order to see what possible mutual information values one can have when $\Gamma(d) = 0$, I did the following experiment: first randomly choose P_0 and $0 < P_2 < 1 - P_0$, then randomly choose $0 < P_{00} < P_0$ and $0 < P_{22} < P_2$, and calculate all the remaining joint probabilities by Eqs. (4.3) and (4.4). If all of them are nonnegative, calculate the mutual information. The $M(d)$ versus $P_{00} + P_{22}$ is plotted in Fig. 3.

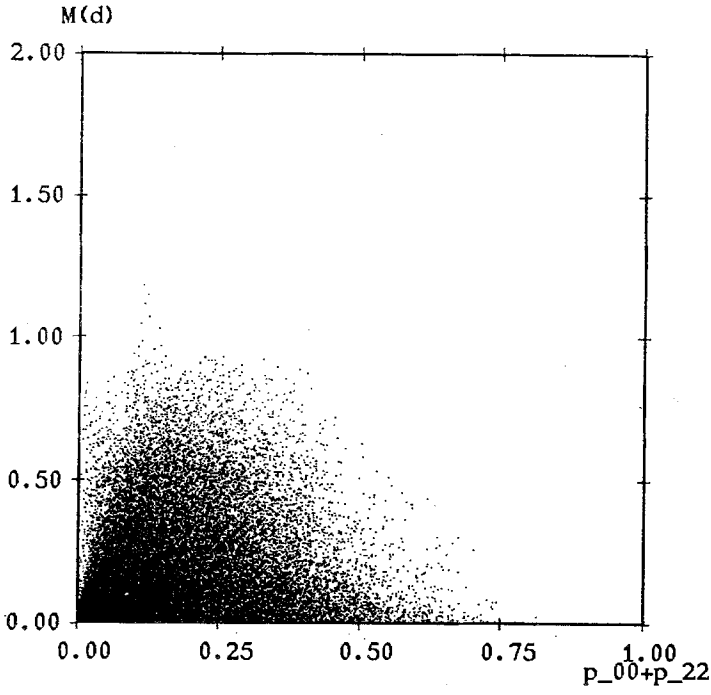


Fig. 3. Mutual information versus $P_{00} + P_{22}$ in ternary sequences when $\Gamma(d) = 0$.

5. FINITE-SIZE EFFECT: OVERESTIMATION OF $M(d)$

In order to calculate mutual information, one has to accumulate the joint probabilities $\{P_{\alpha\beta}\}$, which are the numbers of occurrence for the joint configuration $\{c_{\alpha\beta}\}$ divided by the total number of the counting N. In the case of the mutual information function, N is simply the length of the sequence. If we consider the joint probabilities calculated from an infinite number of countings as "true" values, those with finite statistics should deviate from the true value, or "a finite-size effect." In this section, I will show that the mutual information with a finite number of countings is almost always larger than the true value, and this overestimation is approximately $K(K-2)/2N$, where K is the number of states for each variable.

I use overbars to indicate the true values, and $P_{\alpha\beta}=c_{\alpha\beta}/N$, $P_\alpha=c_\alpha/N$. The fluctuation of the countings are

$$\delta c_{\alpha\beta}=c_{\alpha\beta}-\overline{c_{\alpha\beta}}, \qquad \delta c_\alpha=c_\alpha-\overline{c_\alpha} \tag{5.1}$$

$M(d)$ can be written in a form which takes contributions from both the true values and the fluctuations:

$$\begin{aligned} M(d)&=\sum_{\alpha\beta} P_{\alpha\beta}(d)\log\frac{P_{\alpha\beta}(d)}{P_\alpha P_\beta}=\frac{1}{N}\left(\sum_{\alpha\beta} c_{\alpha\beta}\log\frac{c_{\alpha\beta}}{c_\alpha c_\beta}\right)+\log(N)\\ &=\frac{1}{N}\sum_{\alpha\beta}(\overline{c_{\alpha\beta}}+\delta c_{\alpha\beta})\log\left(\frac{\overline{c_{\alpha\beta}}+\delta c_{\alpha\beta}}{(\overline{c_\alpha}+\delta c_\alpha)(\overline{c_\beta}+\delta c_\beta)}\right)+\log(N) \end{aligned} \tag{5.2}$$

Using short-hand notations for relative fluctuations

$$\alpha_1\equiv\delta c_{\alpha\beta}/\overline{c_{\alpha\beta}}, \qquad \beta_1\equiv\delta c_\alpha/\overline{c_\alpha}+\delta c_\beta/\overline{c_\beta}, \qquad \beta_2\equiv(\delta c_\alpha/\overline{c_\alpha})(\delta c_\beta/\overline{c_\beta}) \tag{5.3}$$

we find that Eq. (5.2) becomes

$$M(d)=\frac{1}{N}\sum_{\alpha\beta}\overline{c_{\alpha\beta}}(1+\alpha_1)\log\frac{\overline{c_{\alpha\beta}}(1+\alpha_1)}{\overline{c_\alpha}\cdot\overline{c_\beta}(1+\beta_1+\beta_2)}+\log(N) \tag{5.4}$$

The relative fluctuation terms can be approximated by (up to the second order of the relative fluctuations)

$$\begin{aligned} &(1+\alpha_1)\log\frac{1+\alpha_1}{1+\beta_1+\beta_2}\\ &\quad\approx(1+\alpha_1)\log(1+\alpha_1-\beta_1-\beta_2+\beta_1^2-\alpha_1\beta_1)\\ &\quad\approx(1+\alpha_1)(\alpha_1-\beta_1+\tfrac{1}{2}\beta_1^2-\tfrac{1}{2}\alpha_1^2-\beta_2)\\ &\quad\approx\alpha_1-\beta_1+\tfrac{1}{2}\alpha_1^2+\tfrac{1}{2}\beta_1^2-\beta_2-\alpha_1\beta_1 \end{aligned} \tag{5.5}$$

then the true mutual information $\overline{M(d)}$ can be separated from the $M(d)$:

$$M(d) = \overline{M(d)} + \frac{1}{N}\sum_{\alpha\beta} \overline{c_{\alpha\beta}}(\alpha_1 - \beta_1 + \tfrac{1}{2}\alpha_1^2 + \tfrac{1}{2}\beta_1^2 - \beta_2 - \alpha_1\beta_1)$$
$$+ \frac{1}{N}\sum_{\alpha\beta} \alpha_1 \overline{c_{\alpha\beta}} \log \frac{\overline{c_{\alpha\beta}}}{\overline{c_\alpha} \cdot \overline{c_\beta}} \tag{5.6}$$

Because more counting for one configuration means less counting for other configurations (the total number of the counting N is fixed), there is a conservation of fluctuations:

$$\sum_{\alpha\beta} \delta c_{\alpha\beta} = 0, \qquad \sum_{\alpha} \delta c_\alpha = 0 \tag{5.7}$$

This conservation simplifies the formula to

$$M(d) = \overline{M(d)} + \frac{1}{N}\sum_{\alpha\beta} \delta c_{\alpha\beta} \log \frac{\overline{c_{\alpha\beta}}}{\overline{c_\alpha} \cdot \overline{c_\beta}}$$
$$+ \frac{1}{N}\sum_{\alpha\beta} \overline{c_{\alpha\beta}}(\tfrac{1}{2}\alpha_1^2 + \tfrac{1}{2}\beta_1^2 - \beta_2 - \alpha_1\beta_1) \tag{5.8}$$

So far, we have not made any assumption on the joint probability distribution. If $\overline{c_{\alpha\beta}}/(\overline{c_\alpha} \cdot \overline{c_\beta})$ does not change with α and β very much, the second term on the right-hand side of the above equation is approximately zero. After making this approximation, and noting that $\sum_\beta c_{\alpha\beta} = c_\alpha$, $\sum_\beta \delta c_{\alpha\beta} = \delta c_\alpha$, we have

$$M(d) - \overline{M(d)} \approx \frac{1}{2N}\sum_{\alpha\beta} \frac{(\delta c_{\alpha\beta})^2}{c_{\alpha\beta}} - \frac{1}{N}\sum_{\alpha} \frac{(\delta c_\alpha)^2}{c_\alpha} \tag{5.9}$$

The typical fluctuation of a variable is of the magnitude of the square root of the variable value, or $\delta c_{\alpha\beta} \sim \sqrt{c_{\alpha\beta}}$ and $\delta c_\alpha \sim \sqrt{c_\alpha}$. Then

$$M(d) - \overline{M(d)} \approx \frac{K(K-2)}{2N} \tag{5.10}$$

where $K = \sum_\alpha 1$ is the total number of states for the variable. For example, if we want to calculate the $M(d)$ between two 3-blocks for binary sequences with length $N = 2400$, $K = 8$ and $M(d) - \overline{M(d)} \approx 24/2400 = 0.01$. Since K is always more than 2, the finite-size effect is always an overestimation of the mutual information. This is in contrast with the finite-size effect on the calculation of entropy H, which is always underestimated:

$$H - \bar{H} \approx -\frac{1}{N}\sum_{\alpha} \frac{(\delta c_\alpha)^2}{\overline{c_\alpha}} = -\frac{K}{N} \tag{5.11}$$

Again, first-order terms are discarded, assuming $\log(c_\alpha)$ does not change with α very much. (A similar discussion on the finite-size effect on the entropy calculation is given in ref. 14.)

6. SYMBOLIC NOISE

As emphasized in the introduction, the correlation function as defined by (1.1) does not apply to symbolic sequences. In practice, people sometimes calculate the correlation function for a particular symbol: the numerical value is one if that symbol is present and zero if not. For sequences with K symbols, there are K such correlation functions. This application of correlation to symbolic sequences is equivalent to $\Gamma_\alpha(d) \equiv P_{\alpha\alpha}(d) - P_\alpha^2$ for $\alpha = 1, 2, \ldots, K$. These functions will not measure the correlation between different symbols, since $P_{\alpha\beta}(d)$ are not used.

The mutual information function is a better quantity for measuring correlations in symbolic sequences, because mutual information is zero if and only if the two sites are generally independent, or $P_{\alpha\beta} = P_\alpha P_\beta$ for all α, β. It is conceivable that some other measure of correlation, such as $\Gamma_\alpha(d)$, will be zero if the two variables are generally independent, but the reverse may not be true.

Correlation functions as well as their Fourier transform, power spectra, play an important role in characterizing and classifying numerical random sequences, or *noise*. There are different types of noise, such as white noise, Brownian noise, and $1/f$ noise, based on the form of their correlation functions and power spectra (see, e.g., ref. 15). Few similar discussions exist in the literature for symbolic sequences, partly because there is no standard way to measure correlations in symbolic sequences. Such discussion is far from being useless, noting, for example, the important application to DNA molecules and other biopolymers.

Here I propose the name *symbolic noise* for those symbolic sequences with a large value of single-site entropy but many possible forms of mutual information functions. If the mutual information function for a symbolic sequence decays to zero even at the nearest neighbor, that sequence can be considered as the symbolic counterpart for white noise. On the other hand, if the mutual information function decays very slowly (power law function with very small exponent), the sequence might be something similar to the $1/f^\alpha$ noise, or can be called *symbolic* $1/f$ *noise*.

Of course, the classification of symbolic noise is more useful when we find some examples in the real world. The first things that come to mind are the letter sequences of natural language texts and the nucleotide sequences of DNA or RNA molecules. If the units of symbols are chosen differently, we can have, for example, word sequences in language and

codon sequences in DNA molecules. Suppose we consider only the smallest units; what types of symbolic noise are the natural language texts and the DNA sequences?

The mutual information functions for letter sequences in several English and German texts have been calculated.[16] One of the plots is reproduced in Fig. 4: the mutual information function for 28-symbol letter sequences (26 letters, one blank space, and one symbol incorporating all punctuations). Other ways of choosing symbols are also discussed in ref. 16, for example, using one symbol to represent all vowels, one symbol for all consonants, one for punctuations, and one for blank space. The mutual information function does not seem to be sensitive to the choice of symbol unit.

From Fig. 4, the mutual information functions decay somewhat between a power law function and an exponential function (for shorter distances). If we insist on using power law functions, the functional form can be approximated by

$$M(d) \sim 1/d^3 \tag{6.1}$$

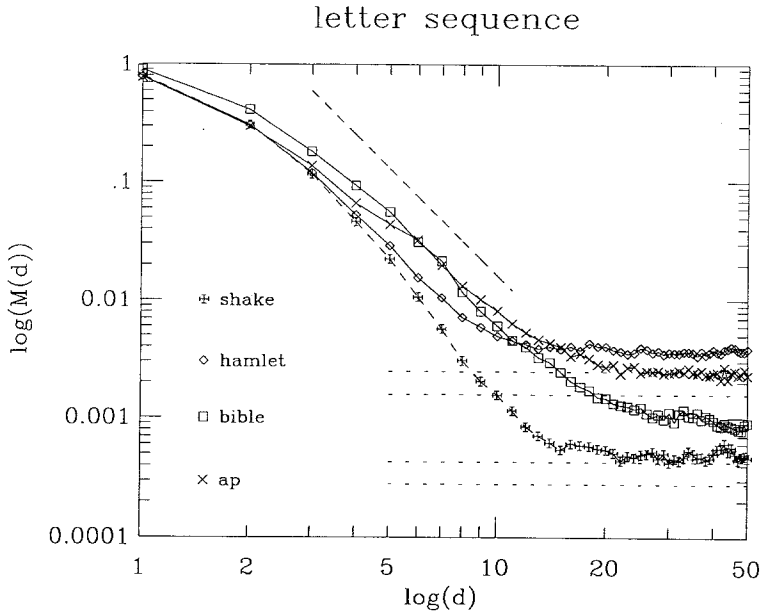


Fig. 4. Site-to-site $M(d)$ for letter sequences (28 symbols) from (1) Shakespeare's play *Hamlet*; (2) Associated Press news articles; (3) the five books of Moses from the Bible in German; (4) 11 plays by Shakespeare. The dashed line is an inverse power law function $1/d^3$. The dotted lines are the estimated residue values of $M(d)$ according to $K(K-2)/2N$.

For now, we do not know what a good name would be for this type of symbolic noise, but it is not symbolic $1/f$ noise, because the correlation measure decays too fast.

Another application of the classification of the symbolic noise is for sequences generated by formal languages,[12] discrete dynamical systems such as cellular automata,[7] and symbolic approximation of continuous dynamical systems (*symbolic dynamics*).[17] One topic being studied concerns the spatial correlations generated by cellular automata.[13] In ref. 13, only the two-state cellular automata are studied, and the correlation function should be as good as the mutual information function according to the discussion in Section 2. Nevertheless, if there are more symbols and if the numerical values for the symbols are not of substantial importance for the system, the mutual information function is preferred over the correlation function.

One way to generate $1/f^{\alpha}$ noise is to use a context-free L-system[18] called the expansion-modification system.[19,20] In this model, it is the symbolic operation that is responsible for the generation of the long-range correlation. Again, since only two symbols are used in refs. 19 and 20, it is possible to calculate the correlation function and the power spectrum. On the other hand, if L-systems are applied to more than two symbols and if the symbols do not have numerical values, we should use the mutual information function to characterize the correlation.

ACKNOWLEDGMENTS

I am grateful to Norman Packard for interesting me in this topic, and for many helpful suggestions. I also thank Stephen Eubank, Martin Casdagli, and Tom Meyer for discussions. The paper is revised from the Center for Complex Systems Research Technical Report (CCSR-89-01) with the same title. The research done at CCSR was supported by NSF grant PHY-86-58062 and ONR grant N00014-88-K-0293. Part of the work was done at Santa Fe Institute, and I acknowledge support from NSF grant PHY-87-14918 and DOE grant DE-FG05-88ER25054.

REFERENCES

1. C. E. Shannon, The mathematical theory of communication, *Bell Syst. Tech. J.* **27**:379–423 (1948).
2. A. Fraser and H. Swinney, Independent coordinates for strange attractors from mutual information, *Phys. Rev. A* **33**:1134–1140 (1986).
3. A. Fraser, Reconstructing attractors from scalar time series: A comparison of singular system and redundancy criteria, *Physics D* **34**:391–404 (1989).
4. C. E. Shannon, Prediction and entropy of printed English, *Bell Syst. Tech. J.* **1951**:50–64.

5. B. Hayes, A progress report on the fine art of turning literature into drivel, Computer Recreations, *Sci. Am.* **249**(5):18–28 (1983).
6. L. Gatlin, *Information Theory and the Living System* (Columbia University Press, 1972).
7. S. Wolfram, ed., *Theory and Application of Cellular Automata* (World Scientific, 1986).
8. G. J. Chaitin, Toward a mathematical definition of "life," in *The Maximum Entropy Formalism*, Levine and Tribus, eds. (MIT Press, 1979).
9. R. Shaw, *The Dripping Faucet as a Model Chaotic System* (Aerial Press, 1984).
10. P. Grassberger, Towards a quantitative theory of self-organized complexity, *Int. J. Theor. Phys.* **25**:907–938 (1986).
11. S. Karlin and H. Taylor, *A Second Course in Stochastic Processes* (Academic Press, 1981); S. Karlin, *A First Course in Stochastic Processes* (Academic Press, 1968).
12. J. E. Hopcroft and J. D. Ullman, *Introduction to Automata Theory, Languages, and Computation* (Addison-Welsey, 1979).
13. W. Li, Power spectra of regular languages and cellular automata, *Complex Syst.* **1**(1):107–130 (1987).
14. H. Herzel, Complexity of symbolic sequences, *Syst. Anal. Model. Simul.* **5**(5):435–444 (1988).
15. M. Gardner, Mathematical Games: White and brown music, fractal curves and $1/f$ fluctuations, *Sci. Am.* **238**(4):16–32 (1978).
16. W. Li, Mutual information functions of natural language texts, Santa Fe Institute preprint, SFI-89-008 (1989).
17. V. M. Alekseev and M. V. Yacobson, Symbolic dynamics and hyperbolic dynamical systems, *Phys. Rep.* **75**:287–325 (1981).
18. A. Lindenmayer, Mathematical models for cellular interactions in development I. Filaments with one-sided inputs, *J. Theor. Biol.* **18**:280–299 (1968).
19. W. Li, Spatial $1/f$ spectra in open dynamical systems, *Europhys. Lett.* **10**(5):395–400 (1989).
20. W. Li, Expansion-modification systems: Another model for $1/f$ spectra, preprint (1990).