Analysis of Biomedical Signals by the Lempel-Ziv Complexity: the Effect of Finite Data Size

Jing Hu, Jianbo Gao, *Member, IEEE*, and Jose C. Principe, *Fellow, IEEE*

***Abstract*—The Lempel-Ziv (LZ) complexity and its variants are popular metrics for characterizing biological signals. Proper interpretation of such analyses, however, has not been thoroughly addressed. In this letter, we study the the effect of finite data size. We derive analytic expressions for the LZ complexity for regular and random sequences, and employ them to develop a normalization scheme. To gain further understanding, we compare the LZ complexity with the correlation entropy from chaos theory in the context of epileptic seizure detection from EEG data, and discuss advantages of the normalized LZ complexity over the correlation entropy.**

***Index Terms*—Biomedical signal analysis, epileptic seizure detection, Lempel-Ziv complexity.**

I. Introduction

THE LEMPEL-ZIV (LZ) complexity [1] and its variants are popular measures for characterizing the randomness of biomedical signals [2]–[6]. Despite its popularity, the issue of interpretation of the LZ complexity calculated from biomedical signals has not been thoroughly addressed. Along this line, recently an important step has been taken by Aboy *et al.* [7]. Unfortunately, most studies published so far assume that the LZ complexity normalized by the factor $n/\log_\alpha n$ [5] (where n is the sequence length and α is the number of alphabets in the symbolic sequence under study) is independent of sequence length. However, we find this is not the case (this point will be made clearer when we discuss Fig. 1 later). This issue can not be satisfactorily solved without an analytic understanding of the dependence of the LZ complexity on sequence length. Here, we derive analytic expressions for the LZ complexity for regular and random sequences of finite length, then develop a normalization scheme that makes the LZ complexity almost independent of sequence length, and finally compare the LZ complexity with another commonly used measure, the correlation entropy from chaos theory [8], through detection of epileptic seizures from electroencephalogram (EEG) data.

II. Lempel-Ziv (LZ) Complexity

To compute the LZ complexity, a numerical sequence has to be first transformed into a symbolic sequence. One popular approach is to convert the signal into a 0–1 sequence by comparing the signal with a threshold value S_d [3]. That is, whenever the signal is larger than S_d, one maps the signal to 1, otherwise, to 0. One good choice of S_d is the median of the signal [6]. After the symbolic sequence is obtained, it can then be parsed to obtain distinct words, and the words be encoded. Let $L(n)$ denote the length of the encoded sequence for those words. The LZ complexity can be defined as

$$C_{\mathrm{LZ}} = \frac{L(n)}{n}. \tag{1}$$

Note this is very much in the spirit of the Kolmogorov complexity [9].

There exist many different methods to perform parsing. One popular scheme is proposed by the original authors of the LZ complexity [1]. For convenience, we call this Scheme 1. Another attractive method is described by Cover and Thomas [10], which we shall call Scheme 2. For convenience, we describe them under the context of binary sequences.

- **Scheme 1**: Let $S = s_1 s_2 \cdots s_n$ denote a finite length 0–1 symbolic sequence; $S(i,j)$ denote a substring of S that starts at position i and ends at position j, that is, when $i \le j$, $S(i,j) = s_i s_{i+1} \cdots s_j$ and when $i > j$, $S(i,j) = \{\}$, the null set; $V(S)$ denote the vocabulary of a sequence S. It is the set of all substrings, or words, $S(i,j)$ of S, (i.e., $S(i,j)$ for $i = 1, 2 \cdots n$; $j \ge i$). For example, let $S = 001$, we then have $V(S) = \{0, 1, 00, 01, 001\}$. The parsing procedure involves a left-to-right scan of the

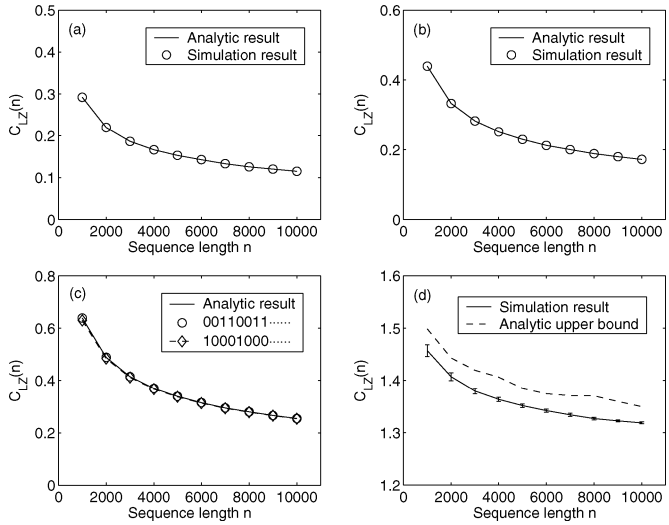

Fig. 1. The LZ complexity vs. the sequence length for (a) the constant sequence, (b) sequence with period 2, (c) sequences with period 4, and (d) random sequences. The vertical bars in (d) indicate the estimated standard errors of the mean.

Manuscript received June 23, 2006; revised July 22, 2006. *Asterisk indicates correpsonding author.*

*J. Hu is with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: jhu@ecel.ufl.edu).

J. Gao and J. C. Principe are with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: gao@ece.ufl.edu; principe@cnel.ufl.edu).

Digital Object Identifier 10.1109/TBME.2006.883825

sequence S. A substring $S(i,j)$ is compared to the vocabulary that is comprised of all substrings of S up to $j-1$, that is, $V(S(1,j-1))$. If $S(i,j)$ is present in $V(S(1,j-1))$, then update $S(i,j)$ and $V(S(1,j-1))$ to $S(i,j+1)$ and $V(S(1,j))$, respectively, and the process repeats. If the substring is not present, then place a dot after $S(j)$ to indicate the end of a new component, update $S(i,j)$ and $V(S(1,j-1))$ to $S(j+1,j+1)$ (the single symbol in the $j+1$ position) and $V(S(1,j))$, respectively, and the process continues. This parsing operation begins with $S(1,1)$ and continues until $j=n$, where n is the length of the symbolic sequence. For example, the sequence 1011010100010 is parsed as $1\cdot0\cdot11\cdot010\cdot100\cdot010\cdot$. By convention, a dot is placed after the last element of the symbolic sequence. In this example, the number of distinct words is 6.

- **Scheme 2**: The sequence $S = s_1 s_2 \ldots$ is sequentially scanned and rewritten as a concatenation $w_1 w_2 \ldots$ of words w_k chosen in such a way that $w_1 = s_1$ and w_{k+1} is the shortest word that has not appeared previously. In other words, w_{k+1} is the extension of some word w_j in the list, $w_{k+1} = w_j s$, where $0 \le j \le k$, and s is either 0 or 1. The above example sequence 1011010100010 is parsed as $1\cdot0\cdot11\cdot01\cdot010\cdot00\cdot10\cdot$. Therefore, a total of 7 distinct words are obtained. This number is larger than 6 of **Scheme 1** by 1.

The words obtained by Scheme 2 can be readily encoded. One simple way is as follows [10]. Let $c(n)$ denote the number of words in the parsing of the source sequence. For each word, we use $\log_2 c(n)$ bits to describe the location of the prefix to the word and 1 bit to describe the last bit. For our example, let 000 describe an empty prefix, then the sequence can be described as (000,1)(000,0)(001,1)(010,1)(100,0)(010,0)(001,0). The total length of the encoded sequence is $L(n) = c(n)[\log_2 c(n) + 1]$. Equation (1) then becomes

$$C_{\mathrm{LZ}} = c(n)\left[\log_2 c(n) + 1\right]/n \tag{2}$$

When n is very large, $c(n) \le n/\log_2 n$ [1], [10], thus, (2) can be further simplified as

$$C_{\mathrm{LZ}} = \frac{c(n)}{n/\log_2 n} \tag{3}$$

The commonly used definition of C_{LZ} takes the same functional form as (3), except that $c(n)$ is obtained by Scheme 1. Typically, $c(n)$ obtained by Scheme 1 is smaller than that by Scheme 2. However, encoding the words obtained by Scheme 1 needs more bits than that by Scheme 2. We surmise that the complexity defined by (1) is similar for both schemes. Indeed, numerically, we have observed that the functional dependence of C_{LZ} on n [based on (2) and (3)] is similar for both schemes. For ease of analysis, below, we shall employ Scheme 2.

III. LZ Complexity for Regular and Random Sequences of Finite Length

For convenience, we consider binary sources throughout this section. The results generalize easily to any finite alphabet.

A. LZ Complexity for Regular Sequences of Finite Length

We start with the constant sequence of length n(00000 ...). The sequence is parsed as 0,00,000,.... Denote the length of the longest word by k bits. It is clear that $(1+k)k/2+x=n$, where $x \le k$. The number of distinct words $c(n)$ is $k+\lceil x\rceil$ (where $\lceil x\rceil$ denotes the smallest integer that is not smaller than x). By (2), the LZ complexity is calculated as

$$C_{\mathrm{LZ}} = (k+\lceil x\rceil)\left[\log_2(k+\lceil x\rceil)+1\right]/n. \tag{4}$$

Next, we consider the sequence with period 2 (010101 ...). Denote the length of a parsed word by i bit(s). When $i=1$, the only possible words that appear in the parsing are 0 and 1; when $i=2$, the only possible words are 01 and 10; when $i=3$, the only possible words are 010 and 101; .... Generally, for an arbitrary i, there are always 2 different words. Let the length of the longest word be k bits. We have $(k-1)k + xk \le n$, where x is the number of words with length k bits, and it equals to 1 or 2 depending on the sequence length n. The total number of words is $c(n) = 2(k-1)+x$. Thus

$$C_{\mathrm{LZ}} = (2k-2+x)\left[\log_2(2k-2+x)+1\right]/n. \tag{5}$$

Now we generalize our results to sequences with arbitrary period $m \ge 2$. Let $m_1 = \lfloor \log_2 m \rfloor$ (where $\lfloor x\rfloor$ denotes the largest integer that is not greater than x), and denote the length of the longest word by k bits. When the length of the words $i \le m_1$, the number of possible words is 2^i (or less); when $i > m_1$, the number of possible words is m. Thus, we have

$$\sum_{i=1}^{m_1} i\cdot 2^i + \sum_{i=m_1+1}^{k-1} i\cdot m + x\cdot k = (m_1-1)2^{m_1+1} + 2 + m(k+m_1)(k-m_1-1)/2 + x\cdot k \le n \tag{6}$$

where $x \le m$. The total number of words is

$$c(n) = \sum_{i=1}^{m_1} 2^i + \sum_{i=m_1+1}^{k-1} m + x = 2^{m_1+1} - 2 + (k-m_1-1)m + x. \tag{7}$$

Therefore, the LZ complexity is

$$C_{\mathrm{LZ}} = \frac{1}{n}\left\{2^{m_1+1} - 2 + (k-m_1-1)m + x\right\} \cdot \left\{\log_2\left[2^{m_1+1} - 2 + (k-m_1-1)m + x\right] + 1\right\} \tag{8}$$

The constant sequence can be considered as the regular sequence with period $m=1$. Indeed, if we set $m=1$ and 2, (8) reduces to (4) and (5), respectively.

When $n \to \infty$, the right-hand side of (8) approaches 0. Hence, the LZ complexity for an infinite periodic sequence is 0. However, as we will show soon, for a finite periodic sequence, its LZ complexity can be much larger than 0.

Before verifying (8), we first discuss how to use it to analytically calculate the LZ complexity for regular sequences. The key is to get the values of k and x. They can be obtained as follows. Given a regular sequence with period m and of length n, first calculate $m_1 = \lfloor \log_2 m \rfloor$, then choose a not too large integer k that satisfies (6), and finally get $x = \lceil n - [(m_1-1)2^{m_1+1} + 2 + m(k+m_1)(k-m_1-1)/2]\rceil$. Check whether $x \le m$ holds or not, if not, then set $k = k+1$, and recalculate the x value.

To assess the goodness of (8) for estimating the LZ complexity of regular sequences, we simulate 0–1 sequences with different period and of different length, calculate their LZ complexity, and compare the simulation results with the analytic solutions obtained from (8). We find that (8) holds very accurately. A few examples are shown in Fig. 1(a)–(c), for the constant sequence, the sequences with period 2 and 4. We wish to emphasize that when the sequence length is finite, the LZ complexity can be considerably larger than 0: the shorter the sequence, the larger the value for the C_{LZ}. Note that the C_{LZ} calculated from (3) shows even stronger dependence on the sequence length.

We would like to make a comment on the LZ complexity of sinusoidal signals. As we have mentioned earlier, to calculate the LZ complexity, a sinusoidal signal has to be mapped to a symbolic sequence first. It is mapped to a patch of 0's and 1's in each half period, whose length, denoted by L, is equal to half of the ratio between the period and the sampling period. When the sampling period is small, L is large, the LZ complexity of the sequence is then small. On the other hand, when the sampling period is large, L is small, and the LZ complexity of the sequence is large. Now it is clear that the two limiting cases are the constant sequence and period-2 sequence discussed above. Therefore, (4) and (5) provide the lower and upper bounds of the LZ complexity for sinusoidal signals.

B. LZ Complexity for Random Sequences of Finite Length

Now we derive analytic expression for the LZ complexity for random sequences of finite length n. Denote the length of the longest word after parsing the sequence by k bits, and the number of words with length k bits by x. Note that the number of possible words with length $i(< k)$ bits is 2^i. Then

$$\sum_{i=1}^{k-1} i \cdot 2^i + x \cdot k = (k-2)2^k + 2 + x \cdot k \leq n \tag{9}$$

where $x \leq 2^k$. The total number of words is

$$c(n) = \sum_{i=1}^{k-1} 2^i + x = 2^k - 2 + x. \tag{10}$$

Therefore, the LZ complexity is

$$C_{\mathrm{LZ}} = \frac{(2^k - 2 + x)\left[\log_2(2^k - 2 + x) + 1\right]}{n}. \tag{11}$$

It is easy to prove as $n \to \infty$, the right-hand side of (11) approaches 1. This indicates that the LZ complexity for an infinite random sequence is 1. However, as we will show soon, for a finite random sequence, its LZ complexity can be considerably larger than 1.

To use (11) to calculate the LZ complexity of a random sequence, we need to know the values of k and x. They can be obtained as follows: first choose a not too large integer k that satisfies (9), and then get $x = \lceil n - [(k-2)2^k + 2] \rceil$. Check whether $x \leq 2^k$ is true or not, if not, then set $k = k + 1$, and recalculate x.

To verify (11), for each sequence length n, we simulate 20 random sequences uniformly distributed in the unit interval [0,1]. We convert each sequence into a 0–1 sequence by comparing the sequence with its median value, and calculate the LZ complexity. Then we calculate the mean and standard deviation of the LZ complexity from all the 20 sequences. Fig. 1(d) shows the averaged LZ complexity with the corresponding standard deviations. We notice that as the sequence becomes longer, the standard error of the LZ complexity becomes smaller. For comparison, the LZ complexity calculated from (11) is also shown in Fig. 1(d). We observe that the analytic results are on top of the simulation results, and the difference between these two is very small (<0.05). This suggests that (11) provides a quite accurate upper bound for the LZ complexity of random sequences.

From Fig. 1(d), we observe that when the sequence length is finite, the LZ complexity for a random sequence can be considerably larger than 1: the shorter the sequence, the larger the value for the C_{LZ}. Intuitively, we would expect the LZ complexity for a finite sequence not to change much with the sequence length, falling in the unit interval [0,1] just as the case of an infinite sequence. These motivate us to propose a normalization scheme as follows:

$$C_{\mathrm{normalizedLZ}}(n) = \frac{C_{\mathrm{LZ}}(n) - C_{\mathrm{constLZ}}(n)}{C_{randLZ}(n) - C_{\mathrm{constLZ}}(n)} \tag{12}$$

where $C_{\mathrm{constLZ}}(n)$ and $C_{\mathrm{randLZ}}(n)$ stand for the LZ complexity for the constant and random sequences of length n, respectively. These two values can be directly obtained from (8) and (11). It is easy to see that $0 \leq C_{\mathrm{normalizedLZ}}(n) \leq 1$. More interestingly, we find this normalization scheme makes the LZ complexity almost independent of the sequence length. This will be illustrated later when we study the EEG data.

IV. Comparison of LZ Complexity and Correlation Entropy

In this section, we compare the LZ complexity with the correlation entropy K_2 [8] from chaos theory, in the context of epileptic seizure detection from EEG data.

Let us introduce the correlation entropy first. It is a tight lower bound of the Kolmogorov-Sinai (KS) entropy. The KS entropy characterizes the rate of creation of information in a system. It is zero, positive, and infinite for regular, chaotic, and random motions, respectively. Let $x(1), x(2) \cdots x(n)$ denote the scalar time series under study. Before calculating K_2, one can use a time delay embedding [11] to form vectors of the form: $V_i = [x(i), x(i+L), \ldots, x(i+(m-1)L)]$, where the embedding dimension m and the delay time L are chosen according to optimization criteria [12]–[14]. K_2 can be readily computed from the correlation integral through the relation [8]

$$C(m, \epsilon) \sim \epsilon^{D_2} e^{-mL\tau K_2} \tag{13}$$

where m and L are the embedding dimension and the delay time, τ is the sampling period, D_2 is the correlation dimension, which quantifies the minimal number of variables needed to characterize the underlying dynamics of the signal, $C(m, \varepsilon) = (1/N^2) \sum_{i,j=1}^{N} \theta(\varepsilon - \|V_i - V_j\|)$ is the correlation integral, θ is the Heaviside step function, V_i and V_j are reconstructed vectors, N is the number of points in the time series, and ε is a prescribed small distance. Equation (13) means that in a plot of $\ln C(m, \varepsilon)$ vs. $\ln \varepsilon$ with m as a parameter, for truly low-dimensional chaos, one observes a series of parallel straight lines, with the slope being D_2, and the spacing between the lines estimating K_2 (where lines for larger m lie below those for smaller m).

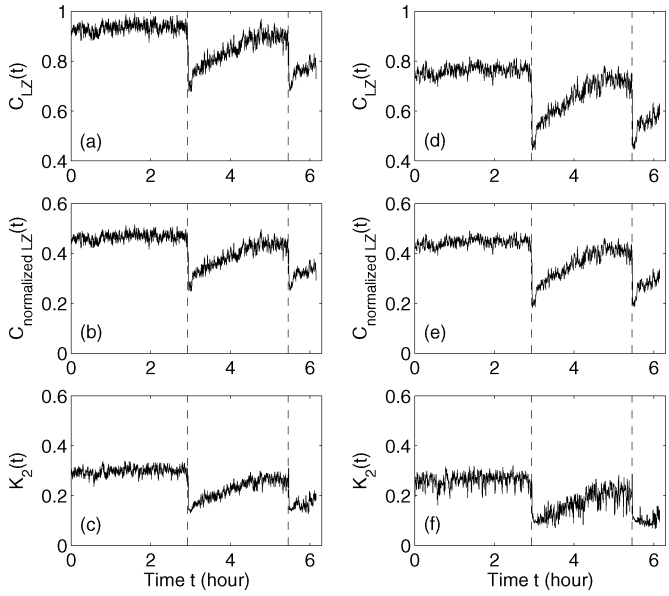

Fig. 2. The variation of (a) and (d) the LZ complexity, (b) and (e) the normalized LZ complexity, (c), (f) the K_2 entropy with time for the EEG signal of a patient. (a)–(c) are obtained by partitioning the EEG signals into short windows of length $W = 500$ points, (d)–(f) are obtained using $W = 2000$. The vertical dashed lines indicate seizure occurrence times determined by medical experts.

The EEG signals analyzed here were measured intracranially by the Shands hospital at the University of Florida. Such EEG data are also called depth EEG and considered cleaner and more free of artifacts than scalp (or surface) EEG. Altogether, we have analyzed 7 patients' multiple channel EEG data, each with a duration of a few hours, with a sampling frequency of 200 Hz. When analyzing EEG for epileptic seizure prediction/detection, it is customary to partition a long EEG signal into short windows of length W points, and calculate the measure of interest for each window. Here, we have tried two different W values, 500 and 2000. For calculating the K_2 entropy, we have chosen $m = 4$ and 5, $L = 1$ according to an optimization criterion of [12]–[14]. The relations between the LZ complexity and the correlation entropy are the same for all the 7 patients' EEG data. Here, we shall illustrate the results based on one patient's data. Fig. 2(a)–(c) shows the variation of the LZ complexity C_{LZ}, the normalized LZ complexity $C_{\mathrm{normalizedLZ}}$, and the correlation entropy K_2 with time for the EEG signal of a patient obtained with $W = 500$, respectively. Fig. 2(d)–(f) shows the variation of the same three measures with time for the same EEG signal obtained with $W = 2000$. We observe the following interesting features: 1) The pattern of variation of the three measures with time is quite similar: slightly after the seizure, all the measures have a sharp drop, followed by a gradual increase. This indicates that the brain dynamics first becomes more regular right after the seizure, then its irregularity increases as it approaches the normal state; 2) C_{LZ} varies a lot with the window size W, as can be seen by comparing Fig. 2(a) and (d). This can be readily understood by the fact that the LZ complexity for a finite sequence depends on the sequence length. This feature not only makes interpretation of the LZ complexity problematic, but also makes automated detection of seizure through thresholding very difficult. Fortunately, both problems can be readily overcome by the normalization scheme, as point 3) illustrates: $C_{\mathrm{normalizeLZ}}$ is almost independent of the window size W, as shown in Fig. 2(b) and (e). 4) The K_2 entropy is also almost invariant with respect to W. However, the curve shown in Fig. 2(f) is much more noisy than that shown in Fig. 2(c). This is quite counter intuitive, since usually for calculating measures based on chaos theory, the more data points we use, the better the result. This puzzling observation may be understood as follows: methods from chaos theory require the signals under study to be ergodic, and therefore stationary. However, the EEG data are nonstationary. As the window size W becomes larger, different nonstationary regions are included; thus, the result for the K_2 entropy becomes noisier. 5) Comparing Fig. 2(b), (c), (e) and (f), we find that the normalized LZ complexity provides better defined features than the K_2 entropy for detecting epileptic seizures. It should also be emphasized that the LZ complexity has the additional advantage that it is easy to implement and computationally very fast. Therefore, normalized LZ complexity provides a good and convenient characterization of epileptic seizure data.

V. Conclusion and Discussions

We have derived analytic expressions for the LZ complexity for regular and random sequences, and used them to study the effect of finite data size on the LZ complexity. We have also developed a normalization scheme that makes the LZ complexity almost independent of sequence length. Finally, we have compared the LZ complexity with the correlation entropy from chaos theory in the context of epileptic seizure detection from EEG data. We have found that the normalized LZ complexity appears to provide better defined features than the correlation entropy for detecting epileptic seizures.

References

[1] A. Lempel and J. Ziv, "On the complexity of finite sequences," *IEEE Trans. Inf. Theory,*, vol. IT-22, no. 1, pp. 75–81, Jan. 1976.

[2] N. Radhakrishnan and B. Gangadhar, "Estimating regularity in epileptic seizure time-series data," *IEEE Eng. Med. Biol. Mag.*, vol. 17, no. 3, pp. 89–94, May-Jun. 1998.

[3] X. S. Zhang, R. J. Roy, and E. W. Jensen, "EEG complexity as a measure of depth of anesthesia for patients," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 12, pp. 1424–1433, Dec. 2001.

[4] X. S. Zhang and R. J. Roy, "Derived fuzzy knowledge model for estimating the depth of anesthesia," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 3, pp. 312–323, Mar. 2001.

[5] X. S. Zhang, Y. S. Zhu, N. V. Thakor, and Z. Z. Wang, "Detecting ventricular tachycardia and fibrillation by complexity measure," *IEEE Trans. Biomed. Eng.*, vol. 46, no. 5, pp. 548–555, May 1999.

[6] R. Nagarajan, "Quantifying physiological data with Lempel-Ziv complexity—certain issues," *IEEE Trans. Biomed. Eng.*, vol. 49, no. 11, pp. 1371–1373, Nov. 2002.

[7] M. Aboy, R. Hornero, D. Abásolo, and D. Álvarez, "Interpretation of the lempel-ziv complexity measure in the context of biomedical signal analysis," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 11, pp. 2282–2288, Nov. 2006.

[8] P. Grassberger and I. Procaccia, "Estimation of the Kolmogorov entropy from a chaotic signal," *Phys. Rev. A*, vol. 28, pp. 2591–2593, 1983.

[9] A. N. Kolmogorov, "Three approaches to the quantitative definition of information," *Prob. Inf. Transmission*, vol. 1, pp. 1–7, 1965.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[11] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, "Geometry from a time series," *Phys. Rev. Lett.*, vol. 45, pp. 712–716, 1980.

[12] J. B. Gao and Z. M. Zheng, "Local exponential divergence plot and optimal embedding of a chaotic time series," *Phys. Lett. A*, vol. 181, pp. 153–158, 1993.

[13] ——, "Direct dynamical test for deterministic chaos and optimal embedding of a chaotic time series," *Phys. Rev. E*, vol. 49, pp. 3807–3814, 1994.

[14] ——, "Direct dynamical test for deterministic chaos," *Europhys. Lett.*, vol. 25, pp. 485–490, 1994.