Fractals, Vol. 14, No. 1 (2006) 27–35

World Scientific
www.worldscientific.com

APPLICATION OF CHAOS GAME REPRESENTATION TO NONLINEAR TIME SERIES ANALYSIS

TOMOYA SUZUKI
Department of Electronic Engineering, Tokyo Denki University
2-2 Kanda-Nishiki-cho, Chiyoda-ku, Tokyo, 101-8457, Japan
suzuki@d.dendai.ac.jp

TOHRU IKEGUCHI
Graduate School of Science and Engineering, Saitama University
225 Shimo-Ohkubo, Sakura-ku, Saitama-city, 338-8570, Japan
tohru@nls.ics.saitama-u.ac.jp

MASUO SUZUKI
Graduate School of Science, Tokyo University of Science
1-3 Kagurazaka, Shinjuku-ku, Tokyo, 162-8601, Japan
msuzuki@rs.kagu.tus.ac.jp

Received May 19, 2005
Accepted August 29, 2005

Abstract

Iterative function systems are often used for investigating fractal structures. The method is also referred as Chaos Game Representation (CGR), and is applied for representing characteristic structures of DNA sequences visually. In this paper, we proposed an original way of plotting CGR to easily confirm the property of the temporal evaluation of a time series. We also showed existence of spurious characteristic structures of time series, if we carelessly applied the CGR to real time series. We revealed that the source of spurious identification came from non-uniformity of the frequency histograms of the time series, which is often the case of analyzing real time series. We also showed how to avoid such spurious identification by applying the method of surrogate data and introducing conditional probabilities of the time series.

Keywords: Chaos Game Representation; Method of Surrogate Data; Time Series Analysis; Conditional Probability.

1. INTRODUCTION

Today, most DNA sequences of genomes, including human genomes, are decoded, then it becomes important to understand functional structures of DNA sequences. It is well known that DNA sequences contain short sequences, which are related to evolution process and properties of lives. The short sequences are sometimes called "motif,"[1] which indicate important functionality embedded iteratively in DNA sequences. For visualizing characteristic structures of such motifs, the method of Chaos Game Representation (CGR)[2,3] is reported to be effective. The CGR is based on Iterative Function System.[2,3] In Ref. Jeffrey,[3] it was reported that CGR can represent a characteristic structure hidden behind DNA sequences by a density of dots in a two-dimensional unit square.

We found that the method works well under some conditions. However, we also confirmed that the method of CGR often led to misinterpretation of result analysis. In the present paper, we reported how and why such misinterpretations can occur. The source of such misidentification came from the non-uniformity of the frequency histograms of the data.

To avoid such spurious identification, we proposed two statistical tests in the present paper. The first one is based on the method of surrogate data,[4,5] which is often used in the field of chaotic time series analysis. Although the surrogate tests are effective for judging whether or not a finite sequence is produced by random, linear or nonlinear dynamics, it does not always mean that we can understand the detail of the dynamics, even if the existence of a dynamical structure is indicated by the surrogate tests. If there exist any dynamical relations, the present value of the time series might depend on the previous values. From this viewpoint, we also introduced the second test; calculating conditional probabilities to evaluate occurrences of each symbol in the sequence, we visualized the difference from random processes.

To validate our method, we first analyzed human-DNA sequences. Moreover, we applied our tests to analyze interbank-exchange rates derived from a financial system. It is acknowledged that financial time series often has a fractal structure.[6] Thus, it is an interesting issue to apply the proposed tests to financial data to reveal hidden characteristic structures.

Interbank-exchange rates are described by bid and ask prices. Several financial models[7,8] are used to discuss influence of bid and ask prices on these future prices and interaction with other variables, such as dealing time intervals. In our previous study,[8] we discussed a nonlinear response of such a financial system against past price movements, introducing the spreads between bid and ask prices, which represent benefits and risks of a dealer. Such spreads are also an important variable in discussing a complex mechanism of a financial system. Thus, we analyzed the movements of the spreads by using CGR.

2. CHAOS GAME REPRESENTATION

2.1. How to Draw CGR[2,3]

First, let us assume that time series $X(t)$ is already quantized into four integers, $X(t) \in \{1, 2, 3, 4\}$. Next, we prepare a two-dimensional unit square U, whose four vertices are denoted by V_i $(i = 1, 2, 3, 4)$ which correspond to the values of $X(t)$. Then, an initial point $A(0)$ is randomly plotted in the square. The second point $A(1)$ is defined as the midpoint between $A(0)$ and $V_{X(1)}$. In general, $A(t)$ is recursively plotted as the midpoint between $A(t-1)$ and $V_{X(t)}$. If $X(t)$ has a characteristic structure, it was reported[3] that the plotted image shows a characteristic fractal pattern on the unit square.

2.2. A Theory of CGR

Let us describe a theory that is useful for predicting a region where a sequence $\{X(t), X(t-1), \ldots, X(t-l)\}$ will be converged. First, we introduced a selecting function $f_i(T)$: which corner of a square T is nearest to the vertex V_i. We also denote the sequence $\{X(t), X(t-1), \ldots, X(t-l)\}$ as $\{X(t:t-l)\}$ and denote a region corresponding to the sequence $\{X(t:t-l)\}$ as $S(\{X(t:t-l)\})$.

Then, we can express the region S as follows:

$$\begin{aligned} &S(\{X(t:t-l)\}) \\ &\quad = f_{X(t-l)}(f_{X(t-l+1)}(\cdots(f_{X(t-1)}(f_{X(t)}(U)))\cdots)) \\ &\quad = (f_{X(t-l)} \circ f_{X(t-l+1)} \circ \cdots \circ f_{X(t-1)} \circ f_{X(t)})(U), \end{aligned}$$

where U is the initial unit square, and the scale of the region where the sequence is included is $(1/2)^{l+1}$ times of U as shown in Fig. 1.

However, because an order of functions for selecting the regions is opposite to temporal evolution,

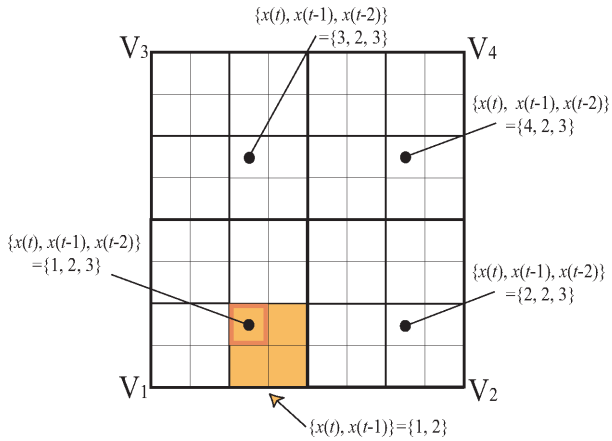

Fig. 1 In the case of the original CGR. It is difficult to understand which regions correspond to each sequence according to temporal evolution.

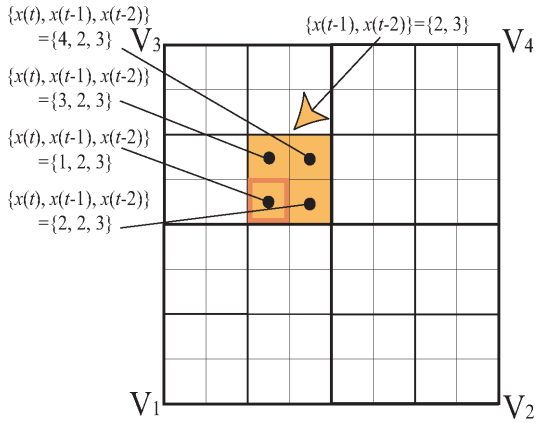

Fig. 2 In the case of using a reversed time series. As time evolves, the regions where each sequence is plotted are specified. The sequences with the same previous parts are plotted in neighboring regions.

a set of sequences with the same previous parts is plotted in separate regions. For example, let us consider four sequences, $\mathbf{X}_1 = \{1, 2, 3\}$, $\mathbf{X}_2 = \{2, 2, 3\}$, $\mathbf{X}_3 = \{3, 2, 3\}$ and $\mathbf{X}_4 = \{4, 2, 3\}$. They have the same part $\{2, 3\}$. Although these four sequences have the same previous part, they are plotted separately as shown in Fig. 1. It indicates that this rule made it hard to understand the essence of CGR. To treat the issue and understand the essence more easily, we proposed a reversed time series $X^R(t)$ of $X(t)$ for drawing CGR. Then, the above representation of S for $X^R(t)$ could be rewritten as

$$\begin{aligned} &S(\{X^R(t : t - l)\}) \\ &\quad \equiv S(\{X(t - l : t)\}) \\ &\quad = f_{X(t)}(f_{X(t-1)}(\cdots(f_{X(t-l)}(U))\cdots)) \\ &\quad = (f_{X(t)} \circ f_{X(t-1)} \circ \cdots \circ f_{X(t-l)})(U). \end{aligned}$$

In Fig. 2, since each function for selecting the region have the same order with the temporal evolution, the sequences with the same previous parts are plotted in neighboring regions of each other. Namely, the process of the reversing time series swapped the regions where each sequence is plotted without changing the direction of temporal evolution.

3. THE METHOD OF SURROGATE DATA

The method of surrogate data[4,5] is frequently used in the field of chaotic time series analysis,[4] because it is useful for obtaining reliable results in nonlinear time series analysis. Since we must avoid any spurious identification of a dynamical structure underlying time series data, it is well known that the method of surrogate data is important to avoid careless estimation of nonlinear indices such as fractal dimensions,[9] or Lyapunov exponents.[10] In chaotic time series analysis,[4] a set of surrogate data is constructed by a null hypothesis based on existence of linear and stochastic process. Thus, the nonlinear (dynamical) structure is completely destroyed. By comparing statistics of the original data and surrogate data sets, it is possible to reject the null hypothesis in the case that the statistics of the original are far different from the surrogates. If the original data does not have a dynamical structure, the calculated statistics become almost the same as those of its surrogates. On the other hand, if the original data is deterministic, its properties are completely different from surrogates.

In the present paper, we used an algorithm for making random shuffle (RS) surrogate and iterative amplitude adjusted Fourier transformed (IAAFT) surrogate data. The null hypothesis of RS surrogate is that the original data is generated by a random process. The algorithm only shuffles sampled values of the original time series randomly. Thus, RS surrogates preserve the empirical histogram but completely destroy the correlation structure of the original time series.

On the other hand, IAAFT surrogate is an improved version of amplitude adjusted Fourier transformed (AAFT) surrogate, whose null hypothesis is that the original data is generated by a linear stochastic process, then it is transformed through a static monotonic nonlinear transformation in observation. The IAAFT surrogate preserves the empirical histogram of the original time series because the algorithm is composed of shuffling the data point. It

can also preserve the temporal correlation structure well. Namely, the algorithm is a controlled shuffling of the data point. We described the algorithm for making IAAFT surrogate data in the Appendix.

4. SURROGATE TEST FOR DATA SEQUENCE

4.1. Propositions

It is difficult to analyze the time series with CGR if the occurrence probabilities of each symbol are not the same. In this case, even if the time series was generated by random numbers, the result of CGR showed a characteristic pattern. In this section, we first show the evidence and we proposed a surrogate test for CGR in order to avoid such spurious identification.

First, we generated s sets of RS surrogates $\hat{X}_k(t)$ $(k = 1, 2, \ldots, s)$ of an original $X(t)$ and drew each CGR $\hat{C}_k$ of $\hat{X}_k^R(t)$. In the present paper, we fixed $s = 199$. Then, we compared $\hat{C}_k$ to the original CGR using $X^R(t)$. It should be noted that RS surrogate preserves the frequency distribution but destroys the auto-correlation structure of $X(t)$.

We divided an initial unit square to $N \times N$ subsquares, $S(m, n)$ $(m, n = 1, 2, \ldots, N)$, which has a 1-to-1 correspondence to a sequence whose length is $\log_2 N$.[2,3] Then, we counted the number of plotted dots $\hat{L}_k(m, n)$ of the k^{th} surrogate CGR $\hat{C}_k$ in $S(m, n)$. If the number of plotted dots $L(m, n)$ by the original CGR C_k in $S(m, n)$ was smaller than the $\alpha(s + 1)/2^{\text{th}}$ smallest value of $\hat{L}_k(m, n)$ or was larger that the $\alpha(s + 1)/2^{\text{th}}$ largest value of $\hat{L}_k(m, n)$, we drew each region $S(m, n)$ in blue (smaller) or red (larger). Here, α is a significance level for hypothesis testing. In general, it is often used that $\alpha = 0.05$. In the following, we show the results of $\alpha = 0.05$, but we also conducted the case that $\alpha = 0.01$, which is almost similar to the case of $\alpha = 0.05$. In such colored regions, we could reject the null hypothesis that the time series has a random structure and does not have any characteristic structures. If the null hypothesis of RS surrogate was not rejected, we should consider that the reason why such characteristic patterns appeared might be that the time series has non-uniform frequency distribution.

Next, we used IAAFT surrogates instead of RS surrogates in order to investigate the existence of auto-correlation structures of the time series, since IAAFT surrogate preserves both of the auto-correlation and the frequency distribution of the time series. If the null hypothesis of IAAFT surrogate was rejected, we could consider that the characteristic pattern shown by CGR was not caused by the correlation structure of time series.

4.2. Simulations

We used the following three data sets for simulation.

- **[Data A]**
  We analyzed tick data of interbank-exchange rates between the US dollars and the Swiss francs.[12] As introduced in Sec. 1, interbank-exchange rates are composed of ask prices $a(t)$ and bid prices $b(t)$. Combining these price movements, we analyzed the behavior of the financial system by CGR. Then, we defined the time series $X(t)$ as follows:

  $X(t) = 1$, if $\Delta a(t) > 0$ and $\Delta b(t) > 0$;
  $X(t) = 2$, if $\Delta a(t) = 0$ and $\Delta b(t) = 0$;
  $X(t) = 3$, if $\Delta a(t) < 0$ and $\Delta b(t) < 0$;
  $X(t) = 4$, otherwise.

  Each frequency is $119,852$, $43,385$, $117,496$ and $2,222$.
- **[Data B]**
  Next, we also used the spreads $S(t)$ between ask prices $a(t)$ and bid prices $b(t)$. Combining the movements of spreads, we expect that it is possible to analyze fluctuation of dealers motivation against the behavior of such financial system because the spreads implicitly reflect benefits and greed of the dealer. Thus, we defined data B as follows:

  $X(t) = 1$, if $\Delta a(t) = 0$ and $\Delta b(t) = 0$, that is, $\Delta S(t) = 0$;

  $X(t) = 2$, if $\Delta a(t) \neq 0$ and $\Delta b(t) \neq 0$ and $\Delta S(t) = 0$;

  $X(t) = 3$, if $\Delta a(t) \neq 0$ and $\Delta b(t) \neq 0$ and $\Delta S(t) > 0$;

  $X(t) = 4$, if $\Delta a(t) \neq 0$ and $\Delta b(t) \neq 0$ and $\Delta S(t) < 0$.

  Although we excluded the case that $\Delta a(t) \neq 0$ or $\Delta b(t) \neq 0$, we considered that it does not affect the results, since the frequency is very low (0.4%). Then, each frequency is $35,603$, $80,967$, $41,371$ and $41,282$.

- **[Data C]**
  In order to consider a general case that each frequency is almost uniform, we used real DNA sequence HUMHBB (human beta globin region, chromosome 11)[2] for analysis. The HUMHBB is composed from four letters which represent four acids "A" (Adenine), "C" (Cytosine), "G" (Guanine) and "T" (Thymine). We defined the time series $X(t)$ as follows:

  $X(t) = 1$, if $X(t)$ was the letter "A";
  $X(t) = 2$, if $X(t)$ was the letter "T";
  $X(t) = 3$, if $X(t)$ was the letter "C";
  $X(t) = 4$, if $X(t)$ was the letter "G".

  Each frequency is $22,068$, $22,309$, $14,146$ and $14,785$.

In addition, we prepared the fourth case that data A were generated by random numbers. Namely, we prepared a new data set A$'$, which has the same frequency distribution as the data set A, but whose temporal structure was destroyed. Then, we drew CGR of the data and applied the proposed surrogate tests for the data. Figure 3 shows the results of the test using RS surrogates.

Here, we could confirm the danger of drawing CGR. Namely, although data A$'$ are random, the CGR showed a fractal structure. However, by applying the proposed statistical test, we could have a possible conjecture that the structure of data A$'$ might be random since the null hypothesis of RS surrogate was not rejected in almost regions. If we did not apply the statistical test, we were misled to the wrong conjecture that data A$'$ had a fractal structure.

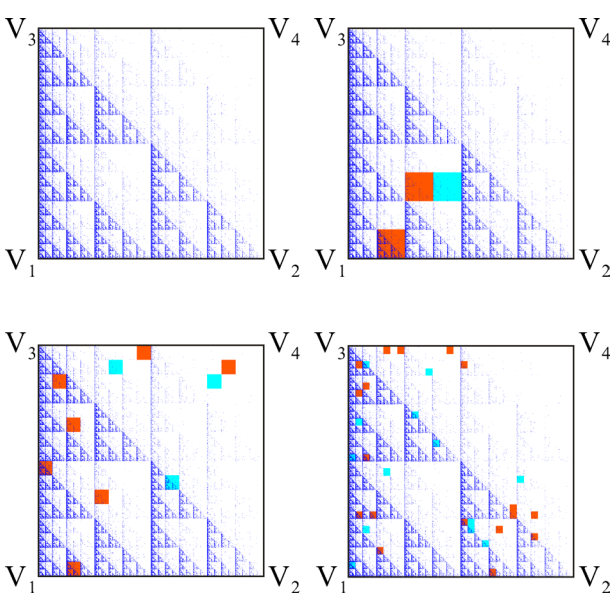

Fig. 3 The results of RS-surrogate test for data A$'$. Dotted patterns were drawn by CGR. The upper-left, upper-right, lower-left and lower-right figures are the cases that the initial number of dividing $N = 4, 8, 16$ and 32, respectively.

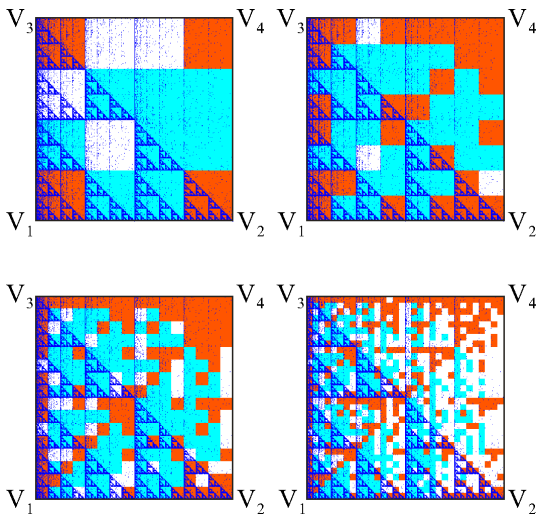

Fig. 4 The same as Fig. 3, but for data A.

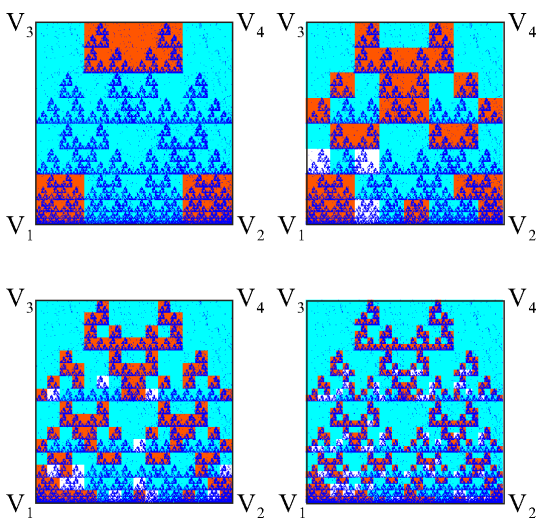

Fig. 5 The same as Fig. 3, but for data B.

Next, in Figs. 4 to 6, we show the results for the data sets A, B and C using RS surrogates by changing N. In these figures, we could confirm that the null hypothesis of RS surrogate was rejected in almost regions. Namely, each frequency distribution was not essential for each characteristic structure shown by CGR. However, we could not deny a possibility that each correlation structure of the data sets A, B and C was essential for the existence of characteristic structure shown by CGR. Thus, we should examine the possibility with the IAAFT surrogate method. In Figs. 7 to 9, we show the results of

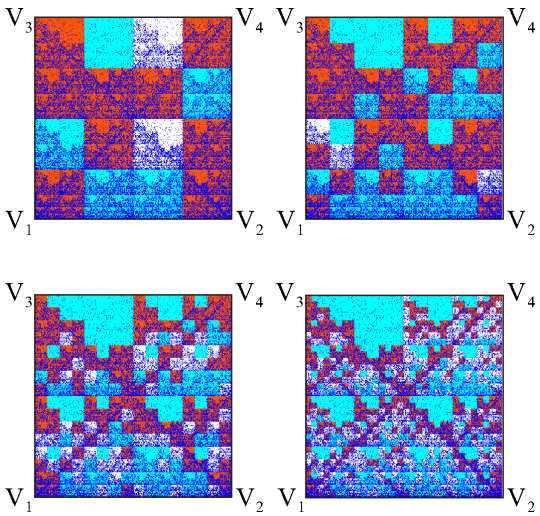

Fig. 6 The same as Fig. 3, but for data C.

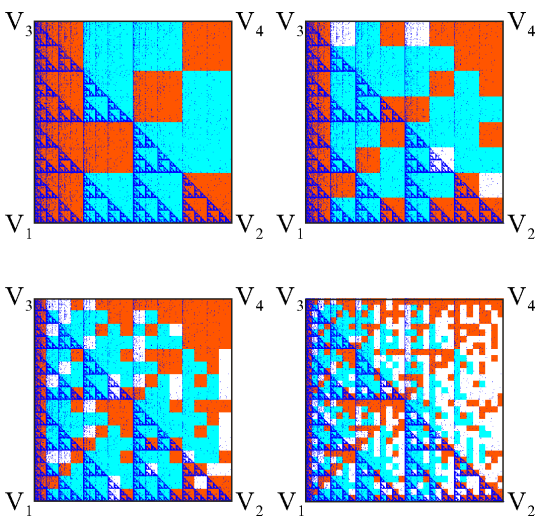

Fig. 7 The same as Fig. 4, but using IAAFT.

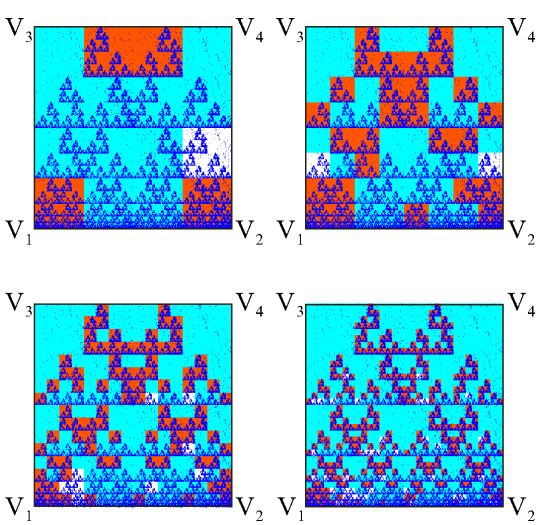

Fig. 8 The same as Fig. 5, but using IAAFT.

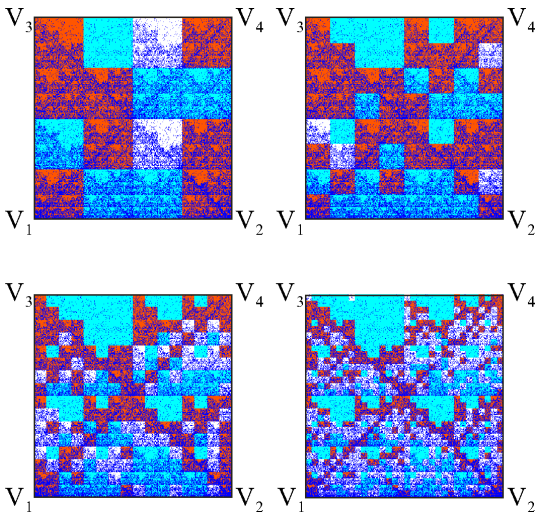

Fig. 9 The same as Fig. 6, but using IAAFT.

the data A, B and C using IAAFT surrogates. Since these results showed almost the same as the results of using RS surrogates, we could confirm that each correlation structure was not essential for each pattern shown by CGR as well. In these simulations, we only showed the case of $\alpha = 0.05$. We confirmed that a similar tendency is observed for the case of $\alpha = 0.01$.

5. VISUALIZATION OF THE CONDITIONAL PROBABILITY

5.1. Propositions

We proposed the method of statistical testing for the frequency of occurrence of any sequences in the former section. Now, we have to raise a question. If the characteristic structures by CGR are independent of the correlation structure of the time series, what is a source for producing such characteristic structures? One possible answer is that there exists a dynamical structure. A high-order dependency might cause characteristic structures. In this section, we proposed the method to evaluate how a current value $X(t)$ depends on previous series by the conditional probability, $P(X(t)|X(t-1), X(t-2), \ldots, X(t-l))$.

As introduced in Sec. 2, a region corresponding to the sequence $\{X(t:t-l)\}$ is decided by the theory of CGR. We drew the region $S(\{X(t:t-l)\})$ in a color according to the value calculated by $P(X(t)|X(t-1:t-l)) - P(X(t))$. If a sequence is generated by a random process, $P(X(t)|X(t-1), \ldots, X(t-l)) = P(X(t))$. In such a case, the color of the region $S(\{X(t:t-l)\})$ is 0, which means that

values of the time series are independent of the past. If the frequency of the sequence $\{X(t-1:t-l)\}$ is very low, the value of $P(X(t)|X(t-1:t-l))$ becomes unreliable. In such a case, we did not color the region.

5.2. Simulations

To examine the high-order dependency and visualize the conditional probabilities, we applied the method proposed in Sec. 5.1 to the same data sets as Sec. 4.2.

In Figs. 10 to 12, we show the results of the data sets A, B and C. Each color bar shows the value of $P(X(t)|X(t-1:t-l)) - P(X(t))$. Then, each range of color bars corresponds to the values from the maximum to the minimum of $P(X(t)|X(t-1:t-l)) - P(X(t))$. As results, we could confirm that the frequency of the present $X(t)$ depended on the previous occurrence patterns strongly. If each datum obeys a dynamical structure, every region was separated from 0. Namely, the method could quantify the dynamical structure of data. Moreover, the method was also very useful for confirming what will be produced as a next occurrence if we have the information of the present situation.

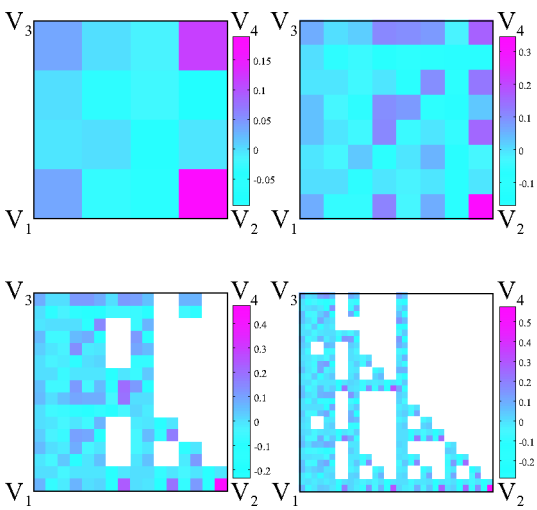

Fig. 10 Visualization of conditional probabilities for data A. Each figure shows the same case as Fig. 3.

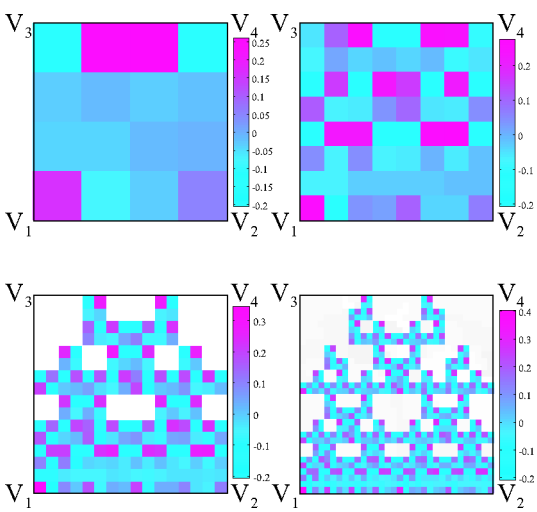

Fig. 11 The same as Fig. 10, but for data B.

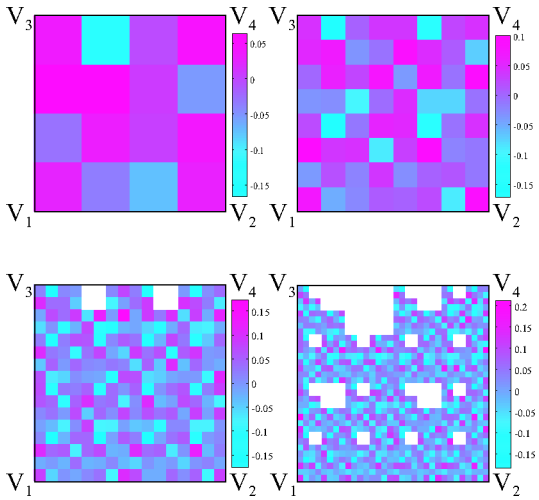

Fig. 12 The same as Fig. 10, but for data C.

6. CONCLUSIONS

We proposed two statistical tests for analyzing which structure of the analyzed time series caused the characteristic patterns of CGR. In addition, for practical usage of CGR, we proposed the original way of plotted CGR in order to specialize the region where each sequence is plotted according to the temporal evolution of each sequence. In particular, we showed that a careless usage of CGR in the case that the frequency of data was not uniform led to a spurious existence of a fractal structure. In such a case, we were asked to carefully interpret the results of CGR. In the present paper, we confirmed that the proposed methods were very effective for avoiding such identification. Moreover, we could conclude that our methods could be a universal support for CGR, since we could use not only DNA sequences but also financial time series, real time series in the other field, whose frequency is not uniform.

ACKNOWLEDGMENTS

The research was partially supported by Grant-in-Aids for Scientific Research (C) (No. 13831002) from the Japan Society for the Promotion of Science (JSPS) to T. Ikeguchi.

REFERENCES

1. D. W. Mount, *Bioinformatics*: *Sequence and Genome Analysis* (Cold Spring Harbor Laboratory Press, 2001).
2. M. Barnsley, *Fractals Everywhere* (Academic Press Inc., London, 1988).
3. H. J. Jeffrey, Chaos game representation of gene structure, *Nucleic Acids Res.* **18**(8) (1990) 2163–2170.
4. T. Ikeguchi *et al.*, in *Fundamentals and Applications of Chaotic Time Series Analysis*, ed. K. Aihara (Sangyo-Tosho Ltd., 2000) (in Japanese).
5. J. Theiler *et al.*, Testing for nonlinearity in time series: the method of surrogate data, *Physica D* **58** (1992) 77–94.
6. E. Mantegna and H. Stanley, Scaling behaviour in the dynamics of an economic index, *Nature* **376** (1995) 46–49.
7. A. Sato and H. Takayasu, Dynamical models of stock market exchanges: from microscopic determinism to macroscopic randomness, *Physica A* **250** (1998) 231–252.
8. T. Suzuki, T. Ikeguchi and M. Suzuki, A model of complex behavior of interbank exchange markets, *Physica A* **337**(1–2) (2004) 196–218.
9. P. Grassberger and I. Procaccia, Measuring the strangeness of strange attractors, *Physica D* **9** (1983) 189–208.
10. J. P. Eckmann, S. Oliffson Kamphorst, D. Ruelle and S. Ciliberto, Lyapunov exponents from time series, *Phys. Rev. A* **34**(6) (1986) 4971–4979.
11. T. Schreiber and A. Schmitz, Improved surrogate data for nonlinearity tests, *Phys. Rev. Lett.* **77**(4) (1996) 635–638.
12. A. S. Weigend and N. A. Gershenfeld (eds.), *Time Series Prediction* (Addison-Wesley, 1993).

APPENDIX: MAKING THE AAFT AND THE IAAFT SURROGATE DATA

- The AAFT surrogate is generated by the following algorithm:
  (1) We make the Gaussian time series, which is reordered to have the same rank order as the observed original time series. Here, the rank order is an order of state values in the time series. In this process, the reordered Gaussian time series corresponds to a linear stochastic process and is an inverse by the same static monotonic nonlinear transformation for the original time series.
  (2) In order to make a time series preserving a correlation structure, we apply Fourier transformed (FT) algorithm to the reordered Gaussian time series. An FT surrogate making procedure is as follow: the Fourier transform is applied to the reordered Gaussian time series and its power spectrum is obtained. In order to preserve the power spectrum (a correlation function) of the reordered Gaussian time series, the phase of the spectrum is randomized by Gaussian random numbers, and the randomized spectrum is symmetrized to obtain a real time series. Then, the inverse Fourier transform produces an FT surrogate which preserves the power spectrum of the reordered Gaussian time series.
  (3) We shuffle the original data to have the same rank order as the FT surrogate data in the process (2). This shuffled data is the AAFT surrogate, which completely preserves the histogram and approximately preserves the correlation structure of the original data.
- Next, IAAFT surrogate is generated by the following algorithm.
  (1) We make Fourier shuffled (FS) surrogate data of the original time series. The procedure for making an FS surrogate is as follow: we make an FT surrogate data of the original time series. Since the FT surrogate does not preserve the empirical histogram, we shuffle the original data to have the same rank order as the FT surrogate data. Although the FS surrogate completely preserves the empirical histogram, it cannot completely preserve the power spectrum as the FT surrogate.
  (2) The Fourier transform is applied to the FS surrogate. Here, the power spectrum of the FS surrogate is replaced by that of the original time series, but the phase of the power spectrum is kept to remain the same.
  (3) The inverse Fourier transform is applied to the data obtained at the second process. Although the generated time series has the same power spectrum as the original time

series, its empirical histogram is different from the original one.

(4) In order to preserve the histogram of the original data, the original time series is reordered to have the same rank order with the generated time series by the third process. The reordered original time series is called an IAAFT surrogate.

(5) If the discrepancy of the power spectrum between the original and the IAAFT surrogate obtained in the previous step is not smaller than a threshold, we repeat the above step by replacing the FS surrogate to the IAAFT surrogate in the second process. Large numbers of repeating the processes preserve more accurately the power spectrum of the original data than AAFT surrogate does. In Ref. Schreiber and Schmitz,[11] for the repeating process of IAAFT surrogate, seven-times repeating is enough from the viewpoint of false rejection. In the present paper, we used ten-times iterated IAAFT surrogate. As a result, IAAFT surrogate completely preserves the empirical histogram and preserves the correlation structure more accurately than the AAFT surrogate.