Measurement of disorder in non-periodic sequences

1991 J. Phys. A: Math. Gen. 24 3979

(http://iopscience.iop.org/0305-4470/24/16/032)

J. Phys. A: Math. Gen. 24 (1991) 3979–3987. Printed in the UK

Measurement of disorder in non-periodic sequences

B L Burrows†§ and K W Sulston‡

† Mathematics Department, Staffordshire Polytechnic, Beaconside, Stafford ST18 0AD, UK
‡ Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

Received 22 April 1991

Abstract. An information theoretic measure is introduced to compare the disorder in non-periodic sequences. It is shown that the measure correctly distinguishes quasiperiodic and aperiodic sequences which have been deduced from earlier studies using diffraction patterns, although it is often necessary to use a set of measures, depending on the order of the source used. The particular sequences studied are the Thue–Morse sequence and the generalizations of the golden mean sequence commonly studied in connection with quasicrystals.

1. Introduction

The discovery by Shechtman *et al* (1984) of an alloy with apparent quasicrystalline structure, and the work of Merlin *et al* (1985) on Fibonacci superlattices, have inspired a great deal of theoretical research concerning systems with order intermediate between periodic and random. Quite generally there is interest in non-periodic structures with long-range positional order, although mathematical constraints have restricted the majority of studies to one-dimensional systems. The most popular lattice studied has been that defined by the Fibonacci sequence, because it is the analogue of the three-dimensional Penrose tiling (Penrose 1974) generally thought to describe the positions of atoms in quasicrystals. Also of interest are the generalized Fibonacci (GF) sequences (Gumbs and Ali 1988a, b, 1989, Ali and Gumbs 1988, Holzer 1988a, b, Kolar and Ali 1990), the Thue–Morse (TM) sequence (Thue 1906, 1912, Morse 1921) especially its recent generalization (Kolar *et al* 1991) and any sequence generated in the same manner using some type of deterministic substitution rule.

An important consideration is the question as to which non-periodic sequences are the more 'disordered'. One possible way to answer this question is in terms of diffraction patterns: if the non-periodic sequence is ordered to the extent that the diffraction pattern shows Bragg peaks, similar to those seen for periodic systems, then the structure is called quasiperiodic (QP); otherwise the structure is usually called aperiodic (AP). The Fibonacci sequence and certain of its generalizations are in the set of quasiperiodic structures (Bombieri and Taylor 1986, 1987), whereas the TM sequence is aperiodic (Cheng *et al* 1988, Kolar *et al* 1991) and, by this criterion, can be thought of as a link between QP and random sequences. On the other hand, electronic spectrum calculations (Riklund *et al* 1987, Cheng *et al* 1988, Qin *et al* 1990) of a one-dimensional tight-bending TM chain indicate an electronic structure more like a periodic chain compared with the structure of a Fibonacci chain (Niu and Nori 1986, Ma and Tsai 1988) so from

§ Corresponding author.

0305-4470/91/163979+09$03.50

this perspective, the TM sequence is intermediate between the peridic and Fibonacci sequences. This apparent contradiction shows the need for further investigation of these lattices and for the development of ways of measuring the similarities and differences between the sequences.

In order to compare the different non-periodic chains of atoms with each other and with the periodic and random chains (the latter of which model amorphous material), we can introduce a measure from information theory. For simplicity we can consider the atoms to be either A or B so that a chain of period 2 is $ABABABAB\ldots$; however, the ideas presented could be extended to any number of types of atoms and could be applied to any one-dimensional sequence of atoms in a three-dimensional structure. For a given positive integer k we can calculate the conditional probabilities

$$P(x_{k+1}|x_1x_2\ldots x_k)$$

where each $x_i = A$ or B. From these proababilities an entropy H_k can be defined (see the next section) which will have the maximum value of 1 for a random chain and the minimum value of 0 for the period 2 chain. For chains such as those defined by the Fibonacci and TM sequences, H_k will generally have a value between 0 and 1 so that the entropy can be used as a measure of disorder in the sequence, allowing us to compare quantitatively the disorder in different lattices. We should point out that these ideas from information theory have been applied to binary sequences appearing in other circumstances (Hamming 1980), but not, to our knowledge, to those most closely connected to the phenomenon of quasicrystals. Furthermore, since the measure only depends on the sequences $x_1\ldots x_k$ it can be applied in other situations where, for example, A and B may represent two different bond lengths or two different layers of atoms (as for example in superlattices).

In section 2 of this paper we shall give a general overview of the mathematics of entropies and their use as a measure of disorder in binary sequences. In section 3 we present calculations for the specific sequences (GF, TM, random, periodic) of interest here and in section 4 we discuss our numerical results.

2. The measure of disorder

We consider here a semi-infinite sequence composed of the letters A and B, which can be taken to represent the two types of atoms in a one-dimensional lattice. If we regard this sequence as a kth-order Markov source, then we can calculate the conditional probabiliites

$$P(x_{k+1}|x_1\ldots x_k) \qquad (x_i = A \text{ or } B)$$

i.e. the probability that a particular string of letters $x_1\ldots x_k$ appearing somewhere in the sequence is immediately followed by the letter x_{k+1}. For example, with a first-order source, the four conditional probabiliites are $P(A|A)$, $P(B|A)$, $P(A|B)$ and $P(B|B)$. For the completely random sequence these probabiliites are all $\frac{1}{2}$, whereas for the sequence of period 2 they are

$$P(A|A) = P(B|B) = 0 \qquad \text{and} \qquad P(A|A) = P(B|A) = 1.$$

From the above probabilities, the conditional entropy is defined by

$$\begin{aligned} H(x|x_1\ldots x_k) &= -P(A|x_1\ldots x_k)\log_2 P(A|x_1\ldots x_k) \\ &\quad - P(B|x_1\ldots x_k)\log_2 P(B|x_1\ldots x_k) \end{aligned} \tag{1}$$

which leads to the following definition of the kth-order entropy (or measure of uncertainty):

$$H_k = \sum P(x_1 \ldots x_k) H(x | x_1 \ldots x_k) \tag{2}$$

where the sum is over all possible strings of length k occurring in the sequence. The use of entropy to measure information and uncertainty was first introduced by Shannon and Weaver (1949) and it has been developed extensively. We will employ (2) to measure the uncertainty or disorder in the sequences considered. (For a recent discussion, see Burrows 1989.)

Clearly the practicality of (2), for some particular value of k, depends upon how easy it is to calculate all the required probabilities associated with the sequence being studied. For a purely random sequence, all the conditional probabilities are $\frac{1}{2}$, so that each conditional entropy becomes 1 and hence from (2) we get $H_k = 1$ for all values of k, signifying of course complete uncertainty about the occupation of each lattice position. On the other hand, each conditional probability for a sequence of period 2 is all either 0 or 1, making the conditional entropy 0, so that $H_k = 0$ for all k, indicating complete order and no uncertainty. Between these two extremes, we can expect to find those sequences which are neither periodic nor random. In this paper we will consider (2) with $k = 1$ or 2 as a measure of disorder in such sequences as the generalized Fibonacci and TM sequences and we will focus on this application in the next sections.

3. Calculation of the entropies

In this section we examine the first-order and second-order entropies (2) for various sequences for which we will need to calculate the probabilities $P(x)$, $P(xy)$, $P(x|y)$ and $P(x|yz)$ where x, y and z are either A or B.

3.1. The Fibonacci sequence

The Fibonacci (or golden-mean) sequence can be generated by the substitution rule

$$A \to AB \qquad B \to A$$

which gives rise to the chain

$$ABAABABAABAABABAABABA\ldots.$$

As the length of the sequence goes to infinity the ratio of the As to Bs approaches the golden mean

$$\sigma = (1+\sqrt{5})/2$$

and it is easy to show that $P(A) = \sigma^{-1}$ and $P(B) = \sigma^{-2}$ with σ satisfying

$$\sigma^2 - \sigma - 1 = 0. \tag{3}$$

To calculate the other probabilities involves more work and we will illustrate the general approach by deriving $P(AB)$. By referring back to the substitution rule it can be seen that the number of occurrences of AB after the substitution is equal to the number of occurrences of A in the string before the transformation:

$$\bar{N}(AB) = N(A).$$

Also the number of As and Bs after the substitution is

$$\bar{N} = 2N(A) + N(B).$$

Thus

$$\bar{N}(AB)/\bar{N} = N(A)/(2N(A)+N(B)) = (N(A)/N)/(1+N(A)/N)$$

where $N = N(A)+N(B)$. In the limit of the infinite sequence, we obtain

$$P(AB) = P(A)/(1+P(A)) = 1/(1+\sigma) = \sigma^{-2} \qquad \text{(using (3))}.$$

Similarly, it is found that $P(AA) = \sigma^{-3}$ and $P(BA) = \sigma^{-2}$ and we have the obvious result $P(BB) = 0$. Hence it is easy to calculate the conditional probabilities for a first-order source, such as

$$P(B|A) = P(AB)/P(A) = \sigma^{-1}$$

and likewise

$$P(A|A) = \sigma^{-2},\ P(A|B) = 1 \qquad \text{and} \qquad P(B|B) = 0.$$

Some of the conditional probabilities for a second order source are readily apparent, namely

$$P(A|AA) = P(B|AB) = 0 \qquad \text{and} \qquad P(B|AA) = P(A|AB) = 1.$$

The remaining probabilities can be calculated using relationships such as

$$P(A|BA) = P(BAA)/P(BA)$$

which, following the evaluation of $P(BAA)$ (in the same manner as for $P(AB)$), leads to $P(A|BA) = \sigma^{-1}$ and in the same way $P(B|BA) = \sigma^{-2}$.

The probabilities given above are sufficient to evaluate the entropies H_1 and H_2 via (1) and (2). Obviously, explicit analytic expressions for the entropies can be obtained in terms of σ, but as they are quite complicated and not particularly illuminating, they are omitted here. The actual numerical values will be given in section 4.

3.2. The generalized Fibonacci sequences

The Fibonacci sequence of section 3.1 can be generalized (Gumbs and Ali 1988a, b, 1989) in terms of the substitution rule

$$A \to A^m B^n \qquad B \to A$$

where A^m is a string of m successive As and B^n n successive Bs, m and n being positive integers. The mean governing a particular GF sequence is

$$\sigma = \tfrac{1}{2}[m + (m^2+4n)^{1/2}]$$

which is the positive root of

$$\sigma^2 - m\sigma - n = 0. \tag{4}$$

The case $m = n = 1$ is just the standard golden-mean lattice considered in section 3.1. Other special cases of particular interest are $m = 2$, $n = 1$ (silver mean), $m = 3$, $n = 1$ (bronze mean), $m = 1$, $n = 2$ (copper mean), $m = 1$, $n = 3$ (nickel mean). At this point, it should be mentioned that Bombieri and Taylor (1986, 1987) have studied the Fourier spectra of these lattices, and their conclusions show that a GF sequence has a spectrum with discrete δ-function peaks (i.e. quasiperiodic), only if $m > n$ or $m = n$; if $m < n$ then the sequence is aperiodic. (The main idea in the work of Bomberi and Taylor is that a one-dimensional 'tiling' is quasiperiodic if the characteristic equation (4) of the substitution rule has just one root of absolute value greater than one, and the rest of absolute value less than one.) From a different perspective, Kolar and Ali (1990) separated the GF sequences into what they called class I ($n=1$) and class II ($n>1$) according to whether the corresponding dynamical trace map is volume-preserving and invertible. Unfortunately the connection (if any) between the Kolar–Ali classification and the division according to quasiperiodicity is unclear.

In turning to the calculation of the entropies for the GF sequences we first note that it is a fundamental property of the infinite sequences that the ratio of the As to the Bs is σ/n, from which is obtained

$$P(A)=\sigma/(\sigma+n) \qquad \text{and} \qquad P(B)=n/(\sigma+n).$$

For general values of m and n, it is lengthy, but straightforward, to calculate the remaining probabilities needed, in the same manner as was done for the golden-mean lattice, so we just present the final values. The probabilities for the various strings of length 2 are

$$P(AA)=(\sigma-1)/(\sigma+n) \qquad P(AB)=P(BA)=1/(\sigma+n)$$
$$P(BB)=(n-1)/(\sigma+n).$$

The conditional probabilities for a first-order source are

$$P(A|A)=(\sigma-1)/\sigma \qquad P(B|A)=\sigma^{-1}$$
$$P(A|B)=n^{-1} \qquad P(B|B)=1-n^{-1}.$$

The conditional probabilities for a second-order source are

$$P(A|AA)=\begin{cases}(n-1)/n & m=1\\(\sigma-2)/(\sigma-1) & m>1\end{cases}$$

$$P(B|AA)=\begin{cases}1/n & m=1\\1/(\sigma-1) & m>1\end{cases}$$

$$P(A|AB)=\begin{cases}1 & n=1\\0 & n>1\end{cases}$$

$$P(B|AB)=\begin{cases}0 & n=1\\1 & n>1\end{cases}$$

$$P(A|BA)=\begin{cases}1/\sigma & m=1\\1 & m>1\end{cases}$$

$$P(B|BA)=\begin{cases}(\sigma-1)/\sigma & m=1\\0 & m>1\end{cases}$$

$$P(A|BB)=\begin{cases}\text{—} & n=1\\1/(n-1) & n>1\end{cases}$$

$$P(B|BB)=\begin{cases}\text{—} & n=1\\(n-2)/(n-1) & n>1.\end{cases}$$

(Note the symbol '—' denotes that the conditional probability is undefined.) With these probabilities the first-order and second-order entropies can be readily calculated, using (1) and (2), for arbitrary values of m and n.

3.3. The Thue–Morse sequence

One way of generating the TM sequence is with the substitution rule $A \to AB$ and $B \to BA$ which leads to the infinite sequence

$$ABBABAABBAABABBA$$

and it is easy to see that, at any stage, the number of As and Bs are equal so that $P(A) = P(B) = \frac{1}{2}$. To calculate a probability such as $P(AB)$, it must first be noticed from the substitution rule that the number of occurrences of AB after the substitution is equal to the number of occurrences of A plus the number of occurrences of BB before substitution, i.e.

$$\bar{N}(AB) = N(A) + N(BB).$$

Coupled with this, we have the fact that the substitution doubles the number of letters, i.e.

$$\bar{N} = 2N$$

so that we obtain

$$P(AB) = \lim_{N\to\infty} \bar{N}(AB)/\bar{N} = \tfrac{1}{2}P(A) + \tfrac{1}{2}P(BB) = \tfrac{1}{4} + \tfrac{1}{2}P(BB) \tag{5}$$

using $P(A) = \frac{1}{2}$. Now we have

$$\bar{N}(BB) = N(AB)$$

so that

$$P(BB) = \lim_{N\to\infty} N(AB)/\bar{N} = \tfrac{1}{2}P(AB). \tag{6}$$

Together equations (6) and (5) imply

$$P(AB) = \tfrac{1}{3} \qquad \text{and} \qquad P(BB) = \tfrac{1}{6}$$

and by symmetry

$$P(BA) = \tfrac{1}{3} \qquad \text{and} \qquad P(AA) = \tfrac{1}{6}.$$

The conditional probabilities for a first-order source are now found immediately to be

$$P(A|A) = P(B|B) = \tfrac{1}{3} \qquad \text{and} \qquad P(A|B) = P(B|A) = \tfrac{2}{3}.$$

After some further calculations, similar to those carried out for the Fibonacci sequence, the conditional probabilities needed for a second-order source turn out to be

$$P(A|AA) = P(B|BB) = 0 \qquad P(B|AA) = P(A|BB) = 1,$$
$$P(A|AB) = P(B|AB) = P(A|BA) = P(B|BA) = \tfrac{1}{2}.$$

All the probabilities needed to calculate H_1 and H_2 are now at hand.

4. Results and discussion

We have used the probabilities given in section 3 to calculate the first-order and second-order entropies for the Fibonacci sequence, a number of its generalizations, and the Thue–Morse sequence. The results are presented in table 1, along with those for the periodic and random chains, for comparative purposes. The sequences considered to be QP (i.e. those with $m \geqslant n$) can be clearly seen to be generally more ordered (i.e. to have lower entropy) than those which are not QP, particularly the TM and nickel mean sequences. Thus the TM sequence, according to the measures used here, does constitute a link between the QP and random sequences as implied by the Fourier spectra. (One other point regarding the TM sequence is that H_1 and H_2 are identical with those for the GF sequence with $m=2$, $n=3$, perhaps indicating simularities between the physical properties of the two systems.) More specifically, the golden, silver and bronze mean sequences (all QP) possess the entropies which are lowest to first order and get even lower to second order. The other QP sequence ($m=n=2$) in the table has high H_1 but much lower H_2 because it actually consists entirely of strings of AAs and BBs, so that a first-order measure is inadequate to show how well ordered this sequence is. To first order, the silver and bronze mean sequences have about the same entropy, but to second order, that for the silver mean is much lower. The reason for this becomes apparent by writing out the sequences explicitly, from which one can see that the string AAB predominates in both sequences, so that H_2 depends most strongly on $P(A|AA)$ and $P(B|AA)$ which are nearer to the random values of $\frac{1}{2}$ and $\frac{1}{2}$ for the bronze mean giving it the large H_2.

The copper mean sequence ($m=1$, $n=2$) is completely random to first order due to the fact that the ratio of As to Bs is 1 and the conditional probabilities are all $\frac{1}{2}$, making it indistinguishable from the random sequence. However, the eigenvalues for the characteristic equation (4) of the sequence are 2 and -1 so that the Bombieri–Taylor criterion is almost satisfied. Thus the sequence can be regarded as almost QP and it is actually quite ordered, as can be seen from the low value of 0.5 for H_2.

Table 1. First-order and second-order entropies for various one-dimensional sequences.

Sequence	H_1	H_2
periodic	0.0	0.0
$m=1$, $n=1$ (golden)	0.593	0.366
$m=3$, $n=1$ (bronze)	0.679	0.528
$m=2$, $n=1$ (silver)	0.692	0.361
Thue–Morse	0.918	0.667
$m=1$, $n=3$ (nickel)	0.948	0.789
$m=2$, $n=2$	0.970	0.360
$m=1$, $n=2$ (copper)	1.0	0.5
random	1.0	1.0

The division of GF sequences, by Kolar and Ali (1990), into class I ($n = 1$) and class II ($n > 1$) does not seem to be strongly represented in the entropies calculated here although the value of H_1 is generallly lower for $n = 1$ than for $n > 1$. This may be connected with the fact that a first-order source is used rather than a reflection of the degrees of disorder in the two classes.

One general comment we will make is that no single entropy of a particular order k is sufficient to compare the relative disorder in all these different sequences. One needs to consider two (or more) of them in concert, or perhaps utilize instead some more complicated comparison such as a rank correlation for a set of H_k. For example, H_1 is quite informative about many of the sequences, but cannot distinguish the copper mean from the random one; a higher-order measure is needed. Consequently it is more meaningful to qualify a comparison with the order of the Markov source used so that the copper mean is identical with the random sequence to first order but much more ordered than the random sequence to second order. Also H_1 tends towards 0 for large m or n, because such sequences consists of long strings of As and/or Bs, so that their structure cannot be represented faithfully using a first-order source. One might expect that the most useful measures for a particular GF sequence would be those whose order is about the same as $\max\{m, n\}$.

To summarize, we have presented a measure of disorder in non-periodic sequences using the idea of entropy for a first-order or second-order Markov source. We have applied this measure specifically to the Fibonacci, GF and TM sequences and found that the TM and non-QP GF sequences are more disordered than the QP ones, so that they can be thought of as a 'link' between the QP and random sequences. With the increasing importance of non-periodic sequences in condensed matter theory, it is hoped that the relationships among these sequences can be further clarified, using the tools of information theory.

Acknowledgments

We are grateful to Professors S G Davison and F O Goodman for helpful discussions and financial support through the Natural Sciences and Engineering Research Council of Canada.

References

Ali M K and Gumbs G 1988 *Phys. Rev.* B **38** 7091
Bombieri E and Taylor J E 1986 *J. Physique* **47** C3-19
—— 1987 *Contemp. Math.* **64** 241
Burrows B L 1989 *Int. J. Math. Ed. Sci. Tech.* **20** 913
Cheng Z, Savit R and Merlin R 1988 *Phys. Rev.* B **37** 4375
Gumbs G and Ali M K 1988 *J. Phys. A: Math. Gen.* **21** L517
—— 1988 *Phys. Rev. Lett.* **60** 1081
—— 1989 *J. Phys. A: Math. Gen.* **22** 951
Hamming R W 1980 *Coding and Information Theory* (Englewood Cliffs, NJ: Prentice-Hall)
Holzer M 1988 *Phys. Rev.* B **38** 1709
—— 1988 *Phys. Rev.* B **38** 5756
Kolar M and Ali M K 1990 *Phys. Rev.* B **41** 7108
Kolar M, Ali M K and Nori F 1991 *Phys. Rev.* B **43** 1034
Ma H R and Tsai C H 1988 *J. Phys. C: Solid State Phys.* **21** 4311

Merlin R, Bajema K, Clarke R, Juang F-Y and Bhattacharya P K 1985 *Phys. Rev. Lett.* **55** 1768
Morse M 1921 *Trans. Am. Math. Soc.* **22** 84
—— 1921 *Am. J. Math.* **43** 35
Niu Q and Nori F 1986 *Phys. Rev. Lett.* **57** 2057
Penrose R 1974 *Bull IMA* **10** 266
Qin M G, Ma H R and Tsai C H 1990 *J. Phys: Condens. Matter* **2** 1059
Riklund R, Severin M and Liu Y 1987 *J. Mod. Phys.* B **1** 121
Shannon C E and Weaver M 1949 *The Mathematical Theory of Communication* (Chicago, IL: University of Illinois Press)
Shechtman D, Blech I, Gratias D and Cahn J W 1984 *Phys. Rev. Lett.* **53** 1951
Thue A 1906 *Norske Vididensk. Selsk. Skr* **I 7** 1
—— 1912 *Norske Vididensk. Selsk. Skr.* **I 1** 1