Elemente der Mathematik 78 (2023), 1–8
DOI 10.4171/EM/468

© 2021 Swiss Mathematical Society
Published by EMS Press
This work is licensed under a CC BY 4.0 license

Intriguing infinite words composed of zeros and ones

Clark Kimberling

Clark Kimberling received his Ph.D. from Illinois Institute of Technology under the direction of Abe Sklar. He recently celebrated the completion of fifty years of teaching at the University of Evansville, where he manages the Encyclopedia of Triangle Centers: https://faculty.evansville.edu/ck6/encyclopedia/etc.html.

1 Thue–Morse word, TM = 0110100110010110...

The rules for generating TM are $0 \to 01$ and $1 \to 10$, starting with 0. Let us take this one step at a time:

Step 1: $0 \to 01$,

Step 2: apply the rules to 01 to get 01 and 10, hence 0110,

Step 3: $0110 \to 01101001$,

Step 4: $01101001 \to 0110100110010110$,

and so on, so that, for every $n \geq 1$, a word of length 2^{n-1} morphs to one of length 2^n. If you apply these rules to TM, the result is TM! (Try it.)

Loosely speaking, half of the digits (or letters, or symbols) of TM are 1, which is to say that a randomly chosen digit of TM is 1 is $1/2$. In order to make such notions more precise, we introduce some notation. Let $T(n)$ denote the initial n-length word of TM, and let $s(n)$ denote the number of 1s in $T(n)$, so that $s(n)$ is the sum of the first n digits of TM; consequently, the mean of $s(n)$ is simply $s(n)/n$. The claim that "half the digits of

Bestimmte Folgen oder Sequenzen, die ausschliesslich aus Nullen und Einsen bestehen, sind bemerkenswert einfach zu definieren, bieten sich jedoch für überraschende und anspruchsvolle Fragestellungen an. Drei dieser Folgen werden hier vorgestellt: die Thue-Morse-Folge (die unter dem Morphismus $0 \to 01$, $1 \to 10$ invariant ist), die Fibonacci-Sequenz (die unter $0 \to 01$, $1 \to 0$ invariant ist) und die Kolakoski-Folge (die ihre eigene Lauflängenkodierung ist, wenn die Nullen und Einsen durch Einsen und Zweien ersetzt werden). Diese Folgen und ihre vielen Varianten sind in der Kombinatorik und Zahlentheorie von Interesse. Der Autor stellt auch kurze Mathematica-Programme zur Erzeugung dieser Sequenzen vor und verweist auf zahlreiche verwandte Folgen, die in der *Online Encyclopedia of Integer Sequences* aufgeführt sind.

Figure 1. Axel Thue (1863–1922) used the TM sequence in his ground-breaking study of the combinatorics of words.

Figure 2. Marston Morse (1892–1977) applied the TM sequence to differential geometry. Shown here is an image of Morse as a graduate student in 1915. Accession 13660 Box 2, Harvard University Archives.

TM are 1" leads to the fraction $s(n)/n$ as n increases. It is easy to see that if n is even, then $s(n)/n = 1/2$, and if n is odd, then $s(n) - n/2 = \pm 1/2$, so that

$$\left| \frac{s(n)}{n} - \frac{1}{2} \right| \leq \frac{1}{2n}$$

for all n. This shows that, as n takes larger and larger values, the mean $s(n)/n$ comes closer and closer to $1/2$. Taking the limit as $n \to \infty$ gives this conclusion: the limiting density of 1 in TM is $1/2$, and the limiting density of 0 is also $1/2$.

Having established $1/2$ as the probability that a randomly chosen digit of TM is 1, we ask this: if two consecutive digits of TM are randomly chosen, and the first is 1, what is the probability that the second is also 1? A bit of experimenting should convince you that this probability is $1/3$.

So far, we have treated TM as a population (with limiting mean and standard deviation both equal to $1/2$) from which samples can be regarded statistically. It is easy and fun to compose statistical questions about TM and to carry out corresponding sampling experiments using a computer.

Instead, however, we turn to some combinatorial observations. Consider the tree in Figure 3, which shows how subwords in TM generate longer subwords. Note that there are six subwords of length 3; four of them yield two 4-length subwords, but only two yield only one 4-length subword. These two, which we call singular, are 001 and 100, and they yield 0110 and 1001, respectively. Of course, for any subword w of TM, if w ends with 00, then $w1$ is a subword of TM but $w0$ is not. One might expect that all singular subwords end with 00 or 11, but a bit of experimentation should lead to other conclusions – and also a basis for original explorations.

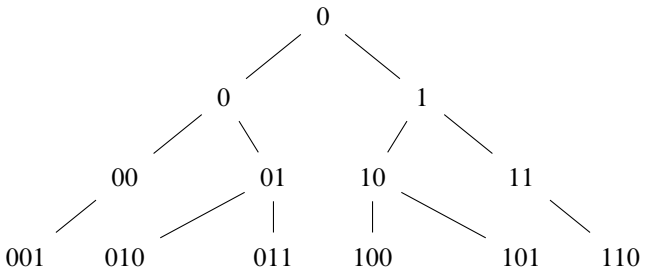

Figure 3. Subwords of TM

2 Infinite Fibonacci word, IFW = 01001010010...

The rules for generating IFW are $0 \to 01$ and $1 \to 0$, starting with 0. Here are the first four steps:

Step 1: $0 \to 01$,

Step 2: $01 \to 010$,

Step 3: $010 \to 01001$,

Step 4: $01001 \to 01001010$,

and so on. In order to examine IFW, recall the Fibonacci sequence, defined as follows:

$$F(1) = 1, \quad F(2) = 1, \quad F(3) = 2, \quad F(4) = 3, \quad F(5) = 5,$$

and so on. Here, the rule is

$$F(n) = F(n-1) + F(n-2)$$

for $n \geq 3$, starting with $F(1) = 1$ and $F(2) = 1$. It is well known that, as n increases, the fractions $F(n)/F(n-1)$ approach the golden ratio, τ. The exact value of τ is $(1+\sqrt{5})/2$, which is approximately 1.618. Now, returning to the chain

$$0 \to 01 \to 010 \to 01001 \to 01001010 \to \ldots,$$

note that each word, beginning with the third, is simply a concatenation of the two immediately preceding words. Consequently, the lengths of the initial subwords of IFW are the Fibonacci numbers. For comparison with the Thue–Morse word, IFW is the unique word that starts with 1 and remains unchanged when the rules $0 \to 01$ and $1 \to 0$ are applied.

For further comparison of IFW with TM, the limiting mean of the first n digits of IFW is $\tau - 1$ and the limiting variance is τ.

Next, we generate two more sequences of numbers using simple rules and then show their connection to the IFW. Start by writing $1, 2, 3, \ldots$ in a row. To create the two sequences, let $A(n)$ be the least number in row 1 that has not yet appeared in row 2 or row 3, and then let $B(n) = A(n) + n$, as shown in Table 1.

Figure 4. Postage stamp honoring Fibonacci. For a gateway to the vast literature about the Fibonacci sequence and other recurrence sequences, see The Fibonacci Association website: https://mathstat.dal.ca/fibonacci/.

n	1	2	3	4	5	6	7	8	9	10
$A(n)$	1	3	4	6	8	9	11	12	14	16
$B(n)$	2	5	7	10	13	15	18	20	23	26

Table 1. Wythoff sequences $(A(n))$ and $(B(n))$

The numbers in row 2 give the positions of 0 in IFW, and the numbers in row 3 give the positions of 1 in IFW. The two sequences, known as the lower and upper Wythoff sequences, are given by $A(n) = \lfloor n\tau \rfloor$ and $B(n) = \lfloor n\tau^2 \rfloor$, where $\lfloor\ \rfloor$ denotes the floor function, i.e., $\lfloor x \rfloor$ is the greatest integer less than or equal to x.

3 Kolakoski word, KW = 011001011011...

The classical Kolakoski sequence [3] is defined using 1s and 2s, and the zero-one word KW results from changing the 1s and 2s to 0s and 1s, respectively, and deleting the commas. We shall begin with the sequence (as a word) and then explore the corresponding zero-one word. First, write 1, and then write 12, and then 122, and then 12211. Here is the rule for transforming 122 into 12211: the initial 1 in 122 tells how many 1s to write; then the next digit, 2, tells how many 2s to write, and the final 2 tells how many 1s to write. Next, apply the same procedure to 12211 to get 1221121, and then again to get 1221121221, and so on. The limiting infinite word is KW.

A *run* is a subword of identical digits; for example, the runs in 1221121221 are 1, 22, 11, 2, 1, 22, 1; that is 1 one, 2 twos, 2 ones, 1 two, 1 one, 2 twos, and 1 one, so that the runlengths, written consecutively as a word, are 1221121, which was the word used to generate 122112122. This illustrates the remarkable fact that KW is the unique binary word that starts with 1 and is its own runlength word. As a zero-one word, we have

$$\text{KW} = 01100101101100100110100101100100101101100101101001001 1 \ldots.$$

Does it seem that on the average, about half the digits are 0? Perhaps this is so, but no proof is known!

Michel Dekking found quite a different way to construct KW, without reference to runlengths. Start with 01, and then repeatedly apply these five substitution rules:

$$00 \to 01, \quad 01 \to 011, \quad 10 \to 001, \quad 11 \to 001,$$

and whenever those first four rules are applied to a word of odd length, the last digit of the word remains unchanged. Following those five rules gives this chain of successive initial subwords of KW:

$$01 \to 011 \to 0111 \to 011001 \to 011001011 \to 011001011011 \to \cdots .$$

4 Amazing Mathematica programs and OEIS

The preceding sections have briefly introduced three intriguing infinite words, about which much more is known and much more remains to be discovered. A quick way to find what is out there, including conjectures and the correct pronunciation of Thue's surname, is the Online Encyclopedia of Integer Sequences (OEIS) [4]. For example, the Thue–Morse word is indexed as A010060, where you can find short Mathematica programs that quickly generate thousands of terms of TM. Here is one of those programs:

```
Nest[ Flatten[ # /. {0 -> {0, 1}, 1 -> {1, 0}}] &, {0}, 7]
```

In OEIS, the infinite Fibonacci word, IFW, appears in several guises, of which the main one is A003849, generated by this short program:

```
Nest[ Flatten[ # /. {0 -> {0, 1}, 1 -> {0}}] &, {0}, 10]
```

Variants of IFW include A003842, A005614, and A005614.

The Kolakoski sequence, indexed in OEIS as A000002, is generated by

```
n=8; Prepend[ Nest[ Flatten[ Partition[#, 2] /.
   {{2, 2} -> {2, 2, 1, 1}, {2, 1} -> {2, 2, 1}, {1, 2} ->
   {2, 1, 1}, {1, 1} -> {2, 1}}] &, {2, 2}, n], 1]
```

This little program implements Dekking's construction mentioned earlier. His comments, and access to many puzzlements regarding KW, are given at A000002.

5 Subwords of the infinite Fibonacci word

For $m \geq 1$, let $L(m)$ be the set of m-length subwords of IFW, e.g., $L(2) = \{00, 01, 10\}$. It is well known [1, p. 17] that the number of words in $L(m+1)$ is $m+1$. Indeed, for each word w in $L(m)$, at least one of the words $w0$ and $w1$ must be in $L(m+1)$, and there must be exactly one such w for which both $w0$ and $w1$ are in $L(m+1)$. We call the latter a *branchword*. Successive sets $L(m)$ appear as levels in the graph in Figure 5, where each branchword has two outgoing edges.

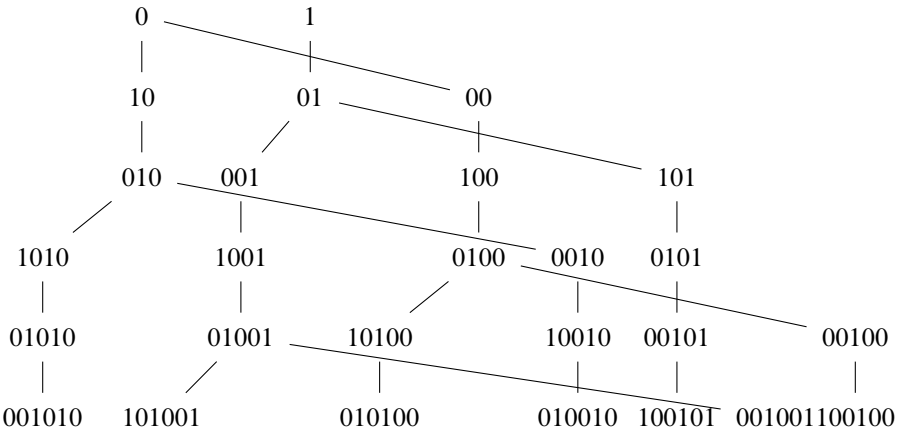

Figure 5. Subwords of IFW

As already noted, the positions in IFW of subwords beginning with 0 are given by $A(n) = \lfloor \tau \rfloor$, and the positions of words beginning with 1, by $B(n)\lfloor \tau^2 \rfloor$. Consequently, subwords beginning with 00 or (01) have positions given by subsequences of $(A(n))$, and subwords beginning with 10 and 11 have positions given by subsequences of $(B(n))$. This is a good time to look carefully at the Wythoff sequences [2] and their composites.

$$\begin{aligned}
A(n) &= (1, 3, 4, 6, 8, 9, 11, 12, 14, 16, 17, 19, 21, \dots) = \text{A000201},\\
B(n) &= (2, 5, 7, 10, 13, 15, 18, 20, 23, 26, 28, 31, \dots) = \text{A001950},\\
AA(n) &= A(A(n)) = (1, 4, 6, 9, 12, 14, 17, 19, \dots) = \text{A003622},\\
AB(n) &= A(B(n)) = (3, 8, 11, 16, 21, 24, 29, \dots) = \text{A003623},\\
BA(n) &= B(A(n)) = (2, 7, 10, 15, 20, 23, 28, \dots) = \text{A035336},\\
BB(n) &= B(B(n)) = (5, 13, 18, 26, 34, 39, 47, \dots) = \text{A101864}.
\end{aligned}$$

Next, we shall order the set of *all* the composites of A and B. Let $S(1) = (A)$ and $S(2) = (B, AA)$. Let $S(3)$ be the set of composites formed by suffixing B to every composite in $S(1)$ and suffixing A to every composite in $S(2)$, so that $S(3) = (AB, BA, AAA)$. The procedure just begun is now formalized inductively. For $n \geq 3$, we form F_{n-1} composites CB as C ranges through $S(n-2)$, followed by F_n composites CA as C ranges through $S(n-1)$, thereby obtaining a total of F_{n+1} composites in $S(n)$.

Next, we transform composites to sums of Fibonacci numbers, beginning with an example. For $AABABB$, replace each A by 1 and each B by 2 to get 112122. Then write

$$F_2 + F_3 + F_5 + F_6 + F_8 + F_{10} = 1 + 2 + 5 + 8 + 21 + 55 = 92,$$

where the subscripts satisfy

$$(3-2, 5-3, 6-5, 8-6, 10-8) = (1, 1, 2, 1, 2, 2);$$

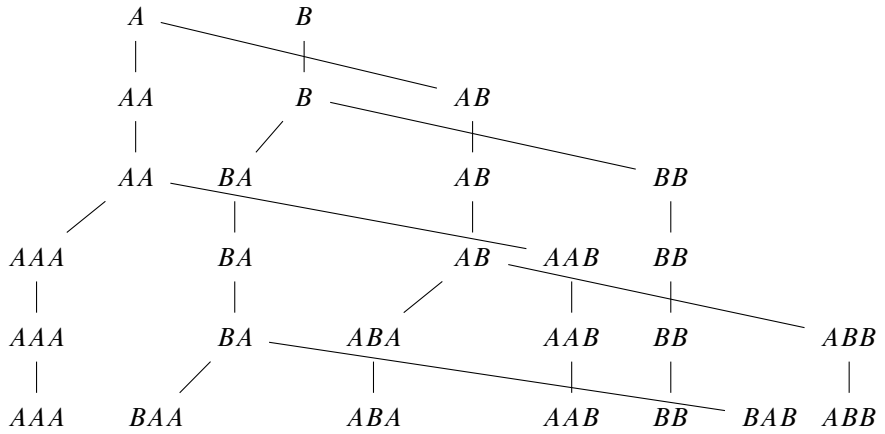

Figure 6. Wythoff composites matching the subwords of IFW

n	C	$M(C)$	Branchpoint
1	A	1	0
2	B	2	10
3	AA	$1+2=3$	010
4	AB	$1+3=4$	0010
5	BA	$2+3=5$	10010
6	AAA	$1+2+3=6$	010010
7	BB	$2+5=7$	1010010
8	AAB	$1+2+5=8$	01010010
9	ABA	$1+3+5=9$	001010010
10	BAA	$2+3+5=10$	1001010010
11	$AAAA$	$1+2+3+5=11$	01001010010
12	ABB	$1+2+3+8=12$	001001010010
13	BAB	$2+3+8=13$	1001001010010
14	$AAAB$	$1+2+3+8=14$	01001001010010

Table 2. First 14 composites, Fibonacci sums, and branchpoints

we write $M(AABABB) = 92$. Table 2 shows the sum $M(C)$ for selected composites C, along with the corresponding branchwords, which [1, p. 17] are simply the reversals of initial subwords of IFW.

As indicated by Table 2, every positive integer appears as a sum of Fibonacci numbers. This type of representation is known by two names: lazy Fibonacci and maximal Fibonacci, in contrast to the more widely studied Zeckendorf (alias minimal Fibonacci) representation. Regarding the former, note that the number of As and Bs in a composite, C, is the number of summands in the maximal Fibonacci representation of $M(C)$, given by A095791.

n	$(n)_{\text{base2}}$	Maximal Fibonacci representation
1	1	1
2	10	$1\cdot 2+0\cdot 1=2$
3	~~11~~	
4	100	$1\cdot 3+0\cdot 2+0\cdot 1=3$
5	101	$1\cdot 3+0\cdot 2+1\cdot 1=4$
6	~~110~~	
7	~~111~~	
8	1000	$1\cdot 5+0\cdot 3+0\cdot 2+0\cdot 1=5$
9	1001	$1\cdot 5+0\cdot 3+0\cdot 2+1\cdot 1=6$
10	1010	$1\cdot 5+0\cdot 3+1\cdot 2+0\cdot 1=7$
11	~~1011~~	
12	~~1100~~	
13	~~1101~~	
14	~~1110~~	
15	~~1111~~	
16	10000	$1\cdot 8+0\cdot 5+0\cdot 3+0\cdot 2+0\cdot 1=8$

Table 3. Maximal Fibonacci representations

Before leaving the minimal and maximal Fibonacci representations, we note a nifty way that both arise from base 2 representations. First, list the positive integers in base 2. Delete all of those that contain consecutive 1s; what is left are the maximal Fibonacci representations using $\{1, 2, 3, 5, 8, \dots\}$ as base, as exemplified in Table 3.

Minimal Fibonacci representations (i.e., Zeckendorf representations) are similarly obtained by deleting all base 2 numbers that contain consecutive 0s.

As a final point of interest, suppose that $C = (C(n))$ is a Wythoff composite, as in Table 2. Is $C(n+1) - C(n)$ a Fibonacci number for every n?

References

[1] J. Berstel, Fibonacci words – A survey, in: *The Book of L*, edited by G. Rozenberg and A. Salomaa, Springer, Berlin (1986), 13–27.

[2] E. Duchêne, A. Fraenkel, V. Gurvich, N. Ho, C. Kimberling and U. Larsson, Wythoff visions, in: *Games of No Chance 5*, edited by U. Larsson, MSRI Publ. 70, Cambridge University Press, Cambridge (2017), 35–87.

[3] W. Kolakoski, Advanced problem 5304, *Amer. Math. Monthly* **72** (1965), 674.

[4] N. J. A. Sloane, Online Encyclopedia of Integer Sequences, https://oeis.org/.

Clark Kimberling
Department of Mathematics
University of Evansville
1800 Lincoln Avenue
Evansville, IN 47722, USA
ck6@evansville.edu