A Space Efficient Algorithm for the Calculation of the Digit Distribution in the Kolakoski Sequence

Johan Nilsson

Fakultät für Mathematik, Universität Bielefeld
Postfach 100131, 33501 Bielefeld, Germany
jnilsson@math.uni-bielefeld.de

12th March 2012

Abstract

With standard algorithms for generating the classical Kolakoski sequence, the numerical calculation of the digit distribution requires a linear amount of space. Here, we present an algorithm for calculating the distribution of the digits in the classical Kolakoski sequence, that only requires a logarithmic amount of space and still runs in linear time. The algorithm is easily adaptable to generalised Kolakoski sequences.

1 Introduction

The classical Kolakoski sequence $K = (K_n)_{n=1}^{\infty}$ is the unique sequence on the alphabet $\{1, 2\}$ defined as the sequence of its own symbols' run lengths starting with a 1. The classical Kolakoski sequence is given in [7, 8], and is in the On-Line Encyclopedia of Integer Sequences [13] with entry number A000002. The first letters of K are

$$\begin{array}{ll} K = & 1 \quad 2 \quad 2 \quad 1 \quad 1 \quad 2 \quad 1 \quad 2 \quad 2 \quad \dots \\ & | \quad \wedge \quad \wedge \quad | \quad | \quad \wedge \quad | \quad \wedge \quad \wedge \\ K = & 1\; 2\; 2\; 1\; 1\; 2\; 1\; 2\; 2\; 1\; 2\; 2\; 1\; 1\; \dots \end{array} \tag{1}$$

There are several interesting questions, answered and unanswered, on the properties of the classical Kolakoski sequence; Kimberling presents several of these in [6]. One of the simplest, and yet unresolved, questions is that of the distribution of digits in K. If we let o_n be the number of 1s in K up to and including position n, that is $o_n = |\{i : K_i = 1, 1 \le i \le n\}|$, then the conjecture is

Conjecture 1. *The limit* $\lim_{n\to\infty} \frac{o_n}{n}$ *exists and equals* $\frac{1}{2}$.

Both parts of Conjecture 1, the existence and the value, are still open. Several aspects of the conjecture (along with other properties and questions regarding the Kolakoski sequence as well) are considered by Dekking in [3, 4, 5]; see also the survey by Sing [12] and further references therein.

In [14] Steinsky describes a recursion that generates the letters K_n and uses it to numerically calculate the distribution of the 1s up to $n = 3 \cdot 10^8$. It is worth noting that a straight-forward implementation of Steinsky's recursion leads to an algorithm that either runs in exponential time or requires a linear amount of space. For some time, Steinsky's result raised doubt as to the validity of Conjecture 1, however subsequent work by Monteil [9] suggested once again that the conjecture should hold. Monteil used a brute force method, requiring linear time and linear space in n, to push the calculation to $n = 10^{11}$.

The brute force, or straight-forward, method to find o_n generates a prefix of length n of the sequence K, using the intuitive method suggested by (1). That is, starting from a suitable initial sequence, we step through and read off the symbols one by one, with each letter telling us what to write in the sequence beneath, and thus what to append to the end of the current sequence.

We present here an algorithm which runs in linear time, yet only requires a logarithmic amount of space to find o_n. Using our algorithm, we can easily push the calculation further than the calculation made by Monteil; we present here values of o_n up to $n = 10^{13}$ (Table 1). Our calculation indicates that Conjecture 1 should hold, but once again gives no definite answer. We present our algorithm in Section 2 and state and prove the algorithm's run time performance in Section 3. In Section 4, we briefly remark on our algorithm's adaptability to more general Kolakoski sequences, and finally in Section 5 we present the results of our calculations.

2 The Algorithm

We present here an algorithm for calculating the number of 1s and 2s in the classical Kolakoski sequence K up to a position n. Our algorithm is more memory-efficient than the straight-forward algorithm for finding K_n; it requires only $O(\log n)$ amount of space (Proposition 4) compared to the $O(n)$ for a brute force algorithm. Here we use the standard asymptotic notation $O(n)$. That is, we write $f(n) = O(g(n))$ if there is a constant c such that $f(n) \leq c\, g(n)$ for all n. (For more of this see [2].) The run time of our algorithm is $O(n)$ to find o_n (Proposition 5); this is the same as for the brute force method.

The idea in our algorithm is that if we set out only to find o_n, we do not

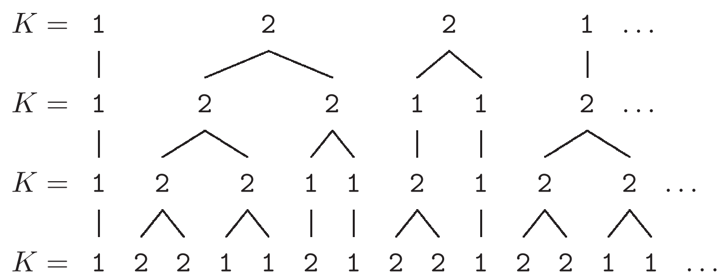

Figure 1: The tree structure in the Kolakoski sequence.

have to save the complete sequence up to position n when stepping through the sequence K. As in the intuitive way of generating K, we look back at a previous position to see which symbol run to append. However, this previous position is itself determined by a letter even further back, and so on. If we keep track only of these positions that we "look back at", we can drastically reduce the amount of space needed by the algorithm.

To get a hint of how this can be done, we take as a starting point a scheme, as in (1). We see that the upper row defines (or conversely, may be defined as) the run lengths of the symbols in the lower one. We expand this scheme by adding more rows above and connecting each symbol to the symbol in the row above that has (via run length) generated it. In this way, we obtain a tree structure, as illustrated in Figure 1.

We may thus interpret the letters in the classical Kolakoski sequence K as the leaves of a tree, (the leaves are the symbols in the bottom row in Figure 1). Each internal node in this tree structure is a symbol in in an upper row interpreted as a run length. Each letter is connected to the letter above that has generated it (called an ancestor), and also to the letter(s) below that it generates, termed children. This tree structure continues upwards without bound as we step through the symbols of the Kolakoski sequence. However, we only need to go up in the tree until we find an ancestor, to the leaf we are currently looking at, at a left most position.

From this point on we shall consider the sequence K', defined by $K = \mathtt{1}K'$. This simplifies matters somewhat, as we do not then have to deal with the left most 1s at each height in the tree.

The algorithm for finding o_n can concisely be described as an "in-order traverse" of this tree structure, where we start from the lower left, and where we keep track of the symbols we see in the leaves during the traverse. While traversing, we add new ancestors when needed; that is we build the tree as we traverse it. To reduce the memory requirement, we dynamically generate and keep track only of the part of the tree that we currently use for the traverse. While doing so, we store the ancestors along with an indicator that tells us which of its children we have already traversed. To this end, we introduce pointers P_k, which are assigned values from the set $S = \{\texttt{1}, \texttt{2}, \texttt{11}, \texttt{22}\}$. Note that here, a run is defined as word from the set S. At any given time, the pointer P_0 holds the current run in the leaves and P_1 holds the ancestor to P_0. Similarly, any P_k that has been initiated holds the ancestor to P_{k-1}.

We say here that pointers "hold" and not "are" a run because P_k may contain more than just the single-symbol ancestor of P_{k-1}, it may also contain a sibling of P_k. Here we refer to the single symbols (that is, `1`s or `2`s) of a two symbol run (`11` or `22`) as siblings.

The algorithm can now be described as follows.

Algorithm 2.

- To increase (or to assign a new value to) the pointer P_k we proceed as follows. Firstly, if P_k has not been initiated, let $P_k = \texttt{22}$. If P_k, for $k > 0$, contains two symbols then remove one of the symbols in P_k; otherwise (if $k = 0$), increase P_1.

  If, on the other hand, P_k contains only one symbol, then increase P_{k+1} recursively. When this increment is done, the new run to write in P_k is of the length given by the first symbol in P_{k+1} and the run to write has symbol(s) opposite to the symbol(s) previously held by P_k. Note that here we do not remove the first symbol of P_{k+1} when we return from the recursion.

- To step throw the sequence K (from its second symbol onwards) and calculate o_n, we repeatedly increment the pointer P_0 and keep track of the number of `1`s and `2`s that we see. ◇

Note that for a given run contained in P_0, the algorithm will generate only the pointers $P_1, \ldots, P_N$ to P_0, where the ancestor in P_N is at the left most position in the sequence K'. (And it is this height N that we shall shortly show is of the order of $\log n$ when P_0 holds the nth letter in the sequence). As we step through the algorithm, we shall see that the successive runs held by the pointer P_0 (and also for other P_k) are the symbols in the sequence K'. nI pseudo-code the increment of P_0, (or the step by step traverse of the leaves), would be done with the recursive call of the procedure `IncrementPointer` as presented below.

```
// Increments the pointer at height n.
// After initiating P[0] succesive calls to IncrementPointer(0)
// will yield the Kolakoski sequence from the second term onward.

IncrementPointer(int k)
{   if(P[k] has not been initiated)
    {   P[k] = 22
    }

    if(k == 0)
    {   IncrementPointer(1)
        if(P[0] == 1 or P[0] == 11)
        {   P[0] = (P[1] == 1) ? 2 : 22
        }else
        {   P[0] = (P[1] == 1) ? 1 : 11
        }
    }else if(P[k] == 1)
    {   IncrementPointer(k+1)
        P[k] = (P[k+1] == 1 or P[k+1] == 11) ? 2 : 22
    }else if(P[k] == 2)
    {   IncrementPointer(k+1)
        P[k] = (P[k+1] == 1 or P[k+1] == 11) ? 1 : 11
    }else if(P[k] == 11)
    {   P[k] = 1
    }else
    {   P[k] = 2
    }
}
```

To illustrate how the algorithm works, we now present through of its initial steps.

Example 3. Incrementing the pointer P_0 once is done through the following procedure;

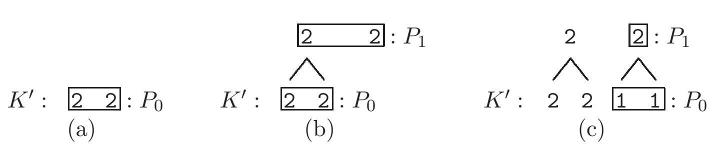

Figure 2: The first increment of the pointer P_0.

Figure 2 illustrates the first increment of the pointer P_0 in the algorithm. (a) The initiation of P_0. The framed symbols 22 are the contents of the pointer P_0. (b) To continue our leaf traverse we must generate the next leaf. This is done by looking at the ancestor of the run held by P_0. As this ancestor does not exist we have to generate it, that is we set $P_1 = \mathtt{22}$. (c) The first symbol of P_1 already has children (that is, it generated the initial run held by P_0). Therefore we step to the second symbol of P_1. The new run to assign to P_0 (that is, the new leaf we traverse) is then 11, since the current symbol in P_1 is 2 and P_0 currently holds the run 22.

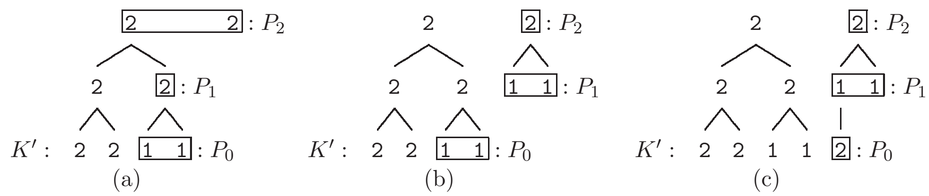

Figure 3: The second increment of the pointer P_0.

Figure 3 illustrates the second increment of the pointer P_0 in the algorithm. (a) To generate the next leaf we have to look at the ancestor of the run currently held by P_0. That is, we look at the pointer P_1. But since we have already used the symbol in P_1 we have to recursively look at the ancestor of P_1. This does not exist, so we initiate the ancestor and pointer $P_2 = \mathtt{22}$. (b) As the first symbol of P_2 already has children, we step to its second symbol. The new run to assign to P_1 is then 11, since the relevant ancestor in P_2 is 2 and P_1 currently holds the run 22. (c) We have not yet generated any of the children of any of the symbols held by P_1 and therefore the current one is the first one. This provides the new run of 2 in P_0, since the first symbol in P_1 is 1 and P_0 currently holds 11.

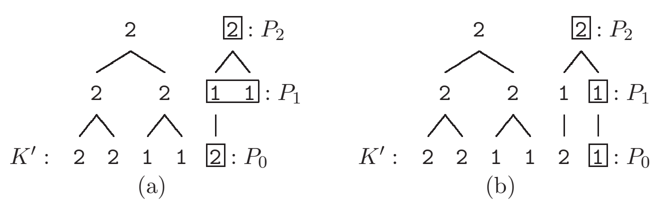

Figure 4: The third increment of the pointer P_0.

Figure 4 illustrates the third increment of the pointer P_0 in the algorithm. (a) The status of the pointers after the second increment of P_0. Note that we have only used the first symbol held by P_1. (b) To generate the next leaf we look at the ancestor of the run currently held by P_0, that is P_1, which contains the run 11. The first symbol already has a child, so we use the second symbol, 1, to generate the new run in P_0, which is 2, as P_0 currently holds the run 1.

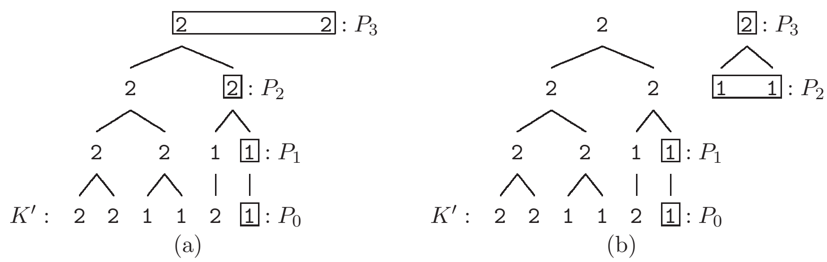

Figure 5: The first part of the fourth increment of the pointer P_0.

Figure 5 illustrates the first part of the fourth increment of the pointer P_0 in the algorithm. (a) To increase P_0 we have to look at the ancestors of the run held by P_0. We see that we have used all symbols in all of the ancestors, therefore we have to initiate the new pointer $P_3 =$ 22. (b) We have already used the first symbol held by P_3 and therefore we step to its second symbol. The new run to assign to P_2 is now 11 since $P_3 =$ 2 and $P_2 =$ 22.

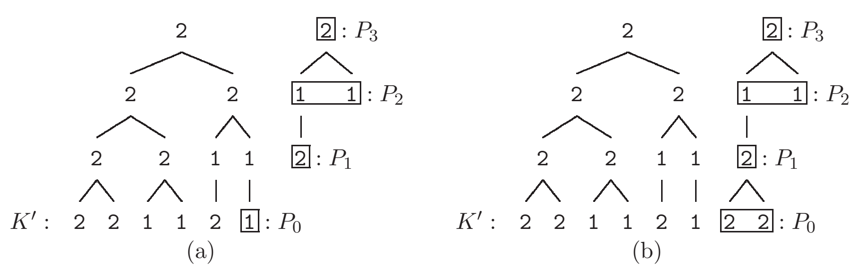

Figure 6: The second part of the fourth increment of the pointer P_0.

Figure 6 illustrates the second part of the fourth increment of the pointer P_0 in the algorithm. (a) We have not yet used any of the symbols held by P_2 and therefore the current one is the first. Then the new symbol in P_1 is 2 since the current symbol in P_2 is 1 and $P_1 =$ 1. (b) The new run to assign to P_0 is now 22 since the first symbol in P_1 is 2 and $P_0 = 1$. □

Note that the algorithm does not need to keep track of the tree structure that it steps through. The algorithm only keeps track of the current contents of the pointers P_k and how many of each symbol we have seen in P_0.

3 Run Time Analysis of the Algorithm

Let t_n be the number of 2s in K up to and including position n. That is $t_n = |\{i : K_i = 1, 1 \leq i \leq n\}|$. Recall that we have already similarly defined o_n as the number of ones. By considering words of the form

$$\texttt{11211} \quad \text{and} \quad \texttt{22122}$$

we see that we have the bounds

$$\frac{1}{4} \leq \frac{o_n}{t_n} \leq 4 \tag{2}$$

for $n \geq 2$. For the analysis, let $P(n)$ be the number of pointers used by Algorithm 2 to calculate o_n.

Proposition 4. *The amount of space used by Algorithm 2 to find o_n is logarithmic in n. That is, $P(n) = O(\log n)$.*

Proof. Let $w_0 = 122$ and $w_1 = 12211$ and similarly let w_k be the run length sequence defining w_{k+1}. Then w_k is a prefix of the sequence K for all $k \geq 0$. (The collection of the words w_k is known as the Kolakoski fan.) By the frequency bound (2) it follows that

$$\frac{6}{5} \leq \frac{|w_{k+1}|}{|w_k|} \leq \frac{9}{5}$$

whenever $k \geq 1$ and where $|\cdot|$ denotes the length of a word.

This implies that if pointer P_0 holds the symbol at position n in K' then the pointer P_1 is at most at position $\frac{5}{6}n$ and at least at position $\frac{5}{9}n$ in the Kolakoski sequence. This argument can now be applied to all pointers. Therefore we see that we have a bound on the number of pointers

$$P(n) \leq \left\lceil \frac{1}{\log \frac{6}{5}} \log n \right\rceil = O(\log n),$$

which completes the proof. □

If Conjecture 1 were shown to be true, it would follow that the number of pointers needed to find o_n is $P(n) \approx \frac{1}{\log \frac{3}{2}} \log n$.

Proposition 5. *The Algorithm 2 runs in (amortized) linear time. That is, to find o_n we have to do an amount of work of order $O(n)$.*

Proof. Let us consider the maximal amount of work we have to do to make n increments of the pointer P_0 (to generate n runs). Note that making n increments of P_0 will actually be enough to find at least $o_{\frac{6}{5}n}$, since in each step we generate a run of one or two symbols. Hence, as we seek a maximum, and including the factor $\frac{6}{5}$ would decrease the calculated amount of work by a constant factor, we may simplify our calculation by disregarding this factor.

Let $p_k(n)$ be the number of times we change the contents of pointer P_k under these n increments. Then the sum of the p_ks will give us the total amount of work we have to do. It is clear that $p_0(n) = n$, since we change P_0 at each increment, and from the algorithm we see directly that $p_1(n) = n$. The other pointers do not change every time; for $k \geq 2$ we make a change to P_k only when P_{k-1} consists of a single symbol.

Let $a_k(n)$ be the number of times the pointer P_k holds a single symbol under n increments of P_0. Similarly let $b_k(n)$ be the number of times that P_k holds two symbols under the n increments of P_0. From the algorithm we see that to find the maximal amount of work, we have to look for the maximal number of single-symbol pointer contents, since this is what forces us to go recursively higher in the tree. For the pointer P_0 it follows from (2) that we have the bounds

$$\frac{1}{4} \leq \frac{a_0(n)}{b_0(n)} \leq 4$$

for $n \geq 1$. For pointers higher up, we have that the number of times P_k holds a single symbol is at most four times the number of times it holds two symbols plus the number of times it holds two symbols, since in the latter case P_k will hold a single symbol in the next step of the algorithm. This gives

$$a_k(n) \leq 4b_k(n) + b_k(n) = 5b_n(k)$$

Therefore, our upper bound on the number of times a pointer holds a single symbol gives the bound on the amount of work we have to do with a pointer P_k compared to the amount of work for the pointer holding the children of P_k. This is

$$p_{k+1}(n) \leq \frac{5}{6}\, p_k(n) \tag{3}$$

for $k \geq 1$. The total amount of work we now have to do to increment the pointer P_0 n times is therefore bounded by the initial amount of work plus the convergent geometric series obtained from (3) We have

$$P(n) = \sum_{i=0}^{\infty} p_i(n) \leq n + c \log n + n \sum_{i=0}^{\infty} \left(\frac{5}{6}\right)^i \leq (7 + C)n, \tag{4}$$

where $c \log n$ is the initial amount of work for each pointer before we can apply our estimates above. □

4 Generalised Kolakoski Sequences

In this this section we remark that our algorithm is also applicable to a general Kolakoski sequence. By a generalised Kolakoski sequence we mean a sequence that is defined as its symbols' run length, as for the classical Kolakoski sequence, but the symbols may be taken from any alphabet $\{r, s\}$, where r and s are natural numbers, as discussed in [4]. We denote a generalised Kolakoski sequence over r and s with $K(r, s)$ and shall assume that $K(r, s)$ starts with the symbol r. The classical Kolakoski sequence is then $K = K(\mathtt{1}, \mathtt{2})$.

It is known that if $r + s$ is an even number, then the letter frequency in $K(r, s)$ can be calculated; see [1, 10, 11, 12]. When $r + s$ is odd, the existence and the value of the letter frequencies are still unknown, but are believed to exist and equal $\frac{1}{2}$.

Our algorithm easily adopts to count the letters in a generalised Kolakoski sequence; we may only have to change the initiation of new pointers. By applying the same idea as in the proof of Proposition 4 we see that the algorithm in this case with a generalised Kolakoski sequence uses fewer pointers than for the classical Kolakoski sequence, and therefore the space requirement must again be at most logarithmic.

Similarly, by looking at the proof of Proposition 5 we see that the number of times we use a pointer for a general Kolakoski sequence before having to consider its ancestor is longer than for the classical Kolakoski sequence. Therefore the bounding factor for the quoted amount of work between two consecutive levels (3) must be smaller than the $\frac{5}{6}$ given for the classical Kolakoski sequence. This gives then, by summing up as in (4), that the total amount of work for the generalised Kolakoski sequence is also linear in n.

5 Calculations

In Table 1 we present a short output from an implementation in Java of our Algorithm 2 for calculating the number of 1s in the classical Kolakoski sequence. The program was run on a standard PC. In Table 2 we present results of a calculation of the number of 2s in the generalised Kolakoski sequence $K(2,3)$, the sequence A071820 in the On-Line Encyclopedia of Integer Sequences [13].

We denote for the classical Kolakoski sequence the maximal deviation of the proportion of 1s from $\frac{1}{2}$ in a logarithmic decade by

$$D(n) = \max_{\frac{1}{10}n<i\leq n} \left| \frac{1}{2} - \frac{o_i}{i} \right| ,$$

where o_i is the number of 1s up to position i. We can similarly define the deviation for the generalised Kolakoski sequence $K(2,3)$.

6 Acknowledgement

The author wishes to thank M. Baake, P. Bugarin, V. Terauds and P. Zeiner at Bielefeld University, Germany, for our discussions of the problem and for reading drafts of the manuscript. This work was supported by the German Research Council (DFG), via CRC 701.

References

[1] M. Baake and B. Sing, Kolakoski(3; 1) is a (deformed) model set, *Canad. Math. Bull.* **47** (2), 168–190 (2004).

[2] T. H. Cormen, C. E. Leiserson, R. L. Rivest, Introduction to Algorithms, *MIT Press* (2001).

[3] F. M. Dekking, Regularity and irregularity of sequences generated by automata, (exposé no. 9), Sém. Th. Nombres Bordeaux 1979–1980, 901–910.

[4] F. M. Dekking, On the structure of self generating sequences, (exposé no. 31), Sém. Th. Nombres Bordeaux, 1980–1981, 3101–3106.

[5] F. M. Dekking, What is the long range order in the Kolakoski sequence?, *The Mathematics of Long-Range Aperiodic Order, NATO Adv.*

n	Number of 1s	$P(n)$	$D(n)$
1	1		
10^1	5	4	$1.667 \cdot 10^{-1}$
10^2	49	10	$8.333 \cdot 10^{-2}$
10^3	502	16	$1.351 \cdot 10^{-2}$
10^4	4 996	22	$3.588 \cdot 10^{-3}$
10^5	49 972	27	$5.481 \cdot 10^{-4}$
10^6	499 986	33	$2.800 \cdot 10^{-4}$
10^7	5 000 046	39	$3.892 \cdot 10^{-5}$
10^8	50 000 675	44	$2.054 \cdot 10^{-5}$
10^9	500 001 223	50	$8.586 \cdot 10^{-6}$
10^{10}	4 999 997 671	56	$2.152 \cdot 10^{-6}$
10^{11}	50 000 001 587	61	$4.453 \cdot 10^{-7}$
10^{12}	500 000 050 701	67	$2.140 \cdot 10^{-7}$
10^{13}	5 000 000 008 159	73	$6.774 \cdot 10^{-8}$

Table 1: The output of the calculation of the number of 1s in the classical Kolakoski sequence. Our result here exceeds the calculations made by Steinsky up to $n = 3 \cdot 10^8$, and Monteil up to $n = 10^{11}$, as mentioned in the introduction. The column with the number of 1s is the sequence A195206 in the On-Line Encyclopedia of Integer Sequences [13].

Sci. Inst. Ser. C Math. Phys. Sci. **489** Kluwer Acad. Publ., Dordrecht (1997), 115–125.

[6] C. Kimberling,
`http://faculty.evansville.edu/ck6/integer/index.html`

[7] W. Kolakoski, Problem 5304: Self Generating Runs, *Amer. Math. Monthly* **72** (1965), 674.

[8] W. Kolakoski, Problem 5304, *Amer. Math. Monthly* **73** (1966), 681–682.

[9] T. Monteil,
`http://www.lirmm.fr/~monteil/blog/BruteForceKolakoski/`

n	Number of 2s	$P(n)$	$D(n)$
1	1		
10^1	5	3	$2.143 \cdot 10^{-1}$
10^2	51	6	$8.333 \cdot 10^{-2}$
10^3	502	9	$2.459 \cdot 10^{-2}$
10^4	4 995	11	$3.318 \cdot 10^{-3}$
10^5	49 999	14	$6.353 \cdot 10^{-4}$
10^6	499 980	16	$8.448 \cdot 10^{-5}$
10^7	4 999 995	19	$2.464 \cdot 10^{-5}$
10^8	50 000 202	21	$7.936 \cdot 10^{-6}$
10^9	499 999 731	24	$3.279 \cdot 10^{-6}$
10^{10}	5 000 005 565	26	$8.382 \cdot 10^{-7}$
10^{11}	50 000 013 114	29	$5.606 \cdot 10^{-7}$
10^{12}	499 999 997 503	31	$1.430 \cdot 10^{-7}$
10^{13}	4 999 999 971 938	34	$3.744 \cdot 10^{-8}$

Table 2: The output of the calculation of the number of 2s in the generalised Kolakoski sequence $K(2,3)$. The column with the number of 2s is the sequence A195211 in the On-Line Encyclopedia of Integer Sequences [13].

[10] B. Sing, *Spektrale Eigenschaften der Kolakoski-Sequenzen* Diploma thesis, Universität Tübingen, (2002).

[11] B. Sing, Kolakoski-(2m; 2n) are limit-periodic model sets, *J. Math. Phys.* **44** (2) (2003), 899–912.

[12] B. Sing, More Kolakoski sequences, `INTEGERS #A14`, **11B** (2011).

[13] N. J. A. Sloane, The On-Line Encyclopedia of Integer Sequences. Published electronically at `http://oeis.org/`.

[14] B. Steinsky, A Recursive Formula for the Kolakoski Sequence A000002, *Journal of Integer Sequences*, Vol. **9** (2006), Article 06.3.7

2010 *Mathematics Subject Classification*: Primary 68R15; Secondary 68Q25.

Keywords: Combinatorics on words, Analysis of algorithms and problem complexity.