# STRUCTURE IN THE KOLAKOSKI SEQUENCE

Robert Steacy

DMS-725-IR January 1966

# Structure in the Kolakoski Sequence

Robert Steacy

Department of Mathematics and Statistics

University of Victoria

P.O. Box 3045

Victoria, Canada

V8W 3P4

e-mail:*rsteacy@smart.math.uvic.ca*

# Contents

# 1 Introduction

A classic unsolved problem of theoretical computer science, cf. Dekking, is the following: We have a machine with a single infinite tape and two heads R and W, where R reads $d$ symbols off the tape at each time instant $n$, and W writes a string of symbols at time $n$, this string only depending on the $d$ symbols read by R at time $n$. Now suppose that R reads two symbols at a time, and W writes according to the table:

| R reads | W writes |
|---|---|
| 11 | 21 |
| 12 | 211 |
| 21 | 221 |
| 22 | 2211 |

To begin, the tape contains only 22. The write head writes 2211, overwriting the 22 with 2211. The read head then reads the next two symbols, 11, and the write head concatenates 21 onto the 2211, so the tape now contains 221121. This procedure continues without limit.The open question is, does the fraction of 1 (or 2) in the resulting unique infinite string equal $\frac{1}{2}$?

In **Ergodic Theory, Symbolic Dynamics and Hyperbolic Spaces** Keane poses the following question concerning the sequence which is also widely known as the Kolakoski sequence:

> "Let x = 221121221221121122... where symbols 2 and 1 occur each in groups of length two or one, and the sequence of group lengths is the same as the sequence x itself. Open question: In x = 221121221221121122... as above, is the frequency of 1 (or of 2) equal to 1/2?"

The purpose of this article is to examine the structure of the Kolakoski sequence, particularly how the sequence is self-replicating. Rather than take the lemma/theorem/proof approach, the reader will find it entertaining and instructive to verify results with pencil and paper, or by writing a short computer program.

# 2 Constructing the Sequence

The sequence, or *string*, starts with a two. To satisfy the definition, there must be two of them:

string 2 2

substring 2

The sequence of group lengths, or *substring*, is the same as the string itself, so the next element of the substring must also be a two. In order that this second two is the group length of the second group in the string, the two two's must be followed by two ones:

string 2211

substring 22

This means that the substring must continue with two ones as well:
string 2211
substring 2211
The string must now continue with a two followed by a one:
string 221121
substring 2211
It is by now apparent that the string is being constructed by forming the *overstring* of the substring, and extending the substring by identical elements. Some writers prefer to start with a one instead of a two, which results in the same string, only with a one in front of it.

Here are the first three hundred or so elements of the sequence:
221121221221121122121121221121121221221121221121121221211
221221121221221121121221211212212211211212211211221211212211221221211
211221221121221211211221211212211211212212112212211212212211211221221
211212211211221211211221222121121221121121221221121221211211221221121
221221121122121121221122121221121221121121212221121221211221...

Overstrings and substrings are all identical, and can be formed without limit.

We may regard the string as being made up of four distinct elements, *singleton one*, *singleton two*, *doubleton one*, and *doubleton two*, i.e. 1, 2, 11, and 22.

A widely known result, cf. Paun, is that any string of ones and twos can be identified as not possibly part of the Kolakoski sequence if it or any of the substrings of its *surely complete* block contains three ones or three twos consecutively. *Surely complete* means that we discard a leading or final singleton one or two, since we do not know for certain if it is a singleton, or part of a doubleton. The substring of a substring must also be taken with respect to the surely complete block of the substring. We shall refer to this result as the *substring lemma*.

By process of elimination, it is found that the string consists of only the following fourteen pieces of string, which we shall refer to as *letters*. Any other pieces of string violate the substring lemma.

A 112212112
B 112212212
C 112212212112
D 1122121121
E 1122121121221
F 1122122121121
G 11221
H 221121221
I 221121121
J 221121121221
K 2211212212
L 2211212212112
M 2211211212212
N 22112

The entire Kolakoski sequence can be expressed in the letters by parsing it in front of every occurrence of either 1122 or 2211.

We shall be interested in the difference between the number of ones and twos *digits* at any point in the string. If we stop the string after any occurrence of a singleton or doubleton one, we must have that 1's + 11's = 2's + 22's, counting a singleton or doubleton as one *group*. We then have that the difference between the number of one and twos *digits* equals 1's + (2 x 11's) - 2's - (2 x 22's). Since after any occurrence of a singleton or doubleton one we have 1's + 11's = 2's + 22's, we obtain the simplification that the difference between the number of one and twos *digits* equals 11's - 22's. We then have that if we stop the string after a singleton or doubleton one, either at the end of a letter or one digit before it, that the difference between the number of one and twos *digits* equals A's + D's + I's - B's - H's - K's. For convenience, we shall refer to the difference between one and twos *digits* simply as 1's - 2's, not to be confused with the number of singleton ones minus the number of singleton twos.

We can represent the overstring and the string:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Overstring | HN | D | J | LG | HM | G | HN | EI | N | AG | ... |
| String | H | N | D | J | L | G | H | M | G | H | ... |

Notice how the initial H produces HN in the overstring, and then the N, which is preceded in the overstring by the HN, produces a D. The D, preceded in the overstring by a D, produces a J. The J, preceded in the overstring by J, produces LG. The reader will notice that the ones and twos of the overstring letters are not always aligned precisely above the letters of the string, but that like pieces of a jigsaw puzzle, the overstring letters sometimes borrow or lend a 12 or a 21 from or to the next overstring letter. This entire procedure by which the string produces the overstring, which is by definition identical to the string, can be summarized in a table.

Overstring letters for each string letter, given that the previous letter in the overstring is:

| | **BCKLMN** | **DEFGIJ** |
|---|---|---|
| A | F | M |
| B | GI | NB |
| C | GM | NF |
| D | C | J |
| E | CG | JN |
| F | GJ | NC |
| G | G | N |
| H | AG | HN |
| I | E | L |
| J | EN | LG |
| K | AB | HI |
| L | AF | HM |
| M | EI | LB |
| N | D | K |

Note that neither A nor H can ever be the previous letter in the overstring.

We shall find it convenient to refer to the first column of the table as *Type 1 transitions* and the second column as *Type 2 transitions*. They are so named because the first overstring letter in a Type 1 transition begins with a one.

We shall say that, for example, the first H in the sequence *undergoes a Type 2 transition* to produce the HN in the overstring.

Carpi's definition of the sequence can be summarized by saying that one starts with a singleton two, and subjects it to an infinite sequence of Type 2 transitions: 2,22,2211,221121,221121221,22112122122112

# 3 Syntax of the String

Using the substring lemma, we can establish a syntax table for the string. We could, if we wanted, do the same for the twenty-eight overstring characters.

| Always preceded by: | String letter | Always followed by: |
|---|---|---|
| N | A | B,G,F |
| A,L,N | B | G |
| K,M,N | C | D,E,G |
| C,L,N | D | J,M,N |
| C,N | E | I,K,N |
| A,K,N | F | J,N |
| A,B,C,K,L,M | G | H,I,J,K,L,M |
| G | H | I,M,N |
| E,G,H | I | N |
| D,F,G | J | K,L,N |
| E,J,G | K | C,F,G |
| G,J | L | B,D,G |
| D,G,H | M | C,G |
| D,E,F,H,I,J | N | A,B,C,D,E,F |

Of course, the preceding table does not tell the whole story. There are many combinations of letters which would be permissible according to the table, but which are disallowed by the substring lemma.

We have the following general structure for the string:
Every occurrence of a D,E,F or G is followed by an
H,I,J,HI, or none of these, followed by a
K,L,M, or N which is followed by an
A,B,C,AB, or none of these, followed by a
D,E,F, or G ...

An immediate consequence of this is that if we stop the string after any occurrence of D,E,F or G we have *exact* equality of D's + E's + F's + G's = K's + L's + M's + N's.

# 4 More on 1's - 2's

It is a simple process to verify that every occurrence in the string of A,B,E, F,G,H,I,L,M, or N which undergoes a Type 1 transition induces a Type 2 transition in the following letter in the string. Similarly, a Type 2 transition among these letters induces a Type 1 transition in the following letter.

Every occurrence in the string of C,D,J, or K which undergoes a Type 1 transition induces a Type 1 transition in the following letter in the string. Similarly, a Type 2 transition among these letters induces a Type 2 transition in the following letter.

We have then:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Transition type: | 2 | 1 | 2 | 2 | 2 | 1 | 2 | 1 | 2 | 1 | ... |
| String letter: | H | N | D | J | L | G | H | M | G | H | ... |

So if we were to delete all the C's,D's,J's and K's in the string, we would be left with:

Transition type: 2 1 2 1 2 1 2 1 ...
String letter: H N L G H M G H ...

That is, there is strict alternation of Type 1 and Type 2 transitions among all letters except C,D,J and K.

Every Type 1 transition of B,H,L,M or N in the string produces an increase in 1's - 2's in the overstring of 2. Every Type 2 transition among these letters produces a decrease of 2.

Every Type 1 transition of A,E,F,G or I in the string produces an increase in 1's - 2's in the overstring of 1. Every Type 2 transition among these letters produces a decrease of 1.

Type 1 and 2 transitions of C,D,J or K in the string produce no change in 1's - 2's in the overstring.

Since every Type 2 transition among A,B,E,F,G,H,I,L,M and N is paired with a Type 1 transition, each pair produces a change of only at most 1 in the 1's -2's in the overstring.

A Type 2 transition of any letter is the *complement* of the Type 1 transition for that letter, i.e. with each 1 replaced by a 2 and each 2 by a 1. It is the strict alternation between Type 1 and Type 2 transitions among A,B,E,F,G,H,I,L,M and N (and that it doesn't matter what happens to C,D,J and K) which is the mechanism providing the tightness between the number of 1's and 2's. Chvátal observed that among the first billion 1's and 2's in the string, there is no point at which the number of 1's - 2's differs by more than 4,933.

We can now write another expression for the difference of 1's - 2's.

1's - 2's in overstring = (Type 1 - Type 2)(B's,H's,L's,M's,N's) in string

That is, if we stop at some point in the string and overstring, and add up the number of the five letters in the string which undergo Type 1 transition, and subtract from that the number of the five letters which undergo Type 2 transition, we obtain the difference of 1's - 2's in the overstring.

Since there is strict pairing of Type 1 and Type 2 transitions among A,B,E,F, G,H,I,L,M and N we could also write:

1's - 2's in overstring = (Type 2 - Type 1)(A's,E's,F's,G's,I's) in string.

In both cases, we need to stop the string with an even number of the ten letters, so that pairing has taken place.

# 5 A Proposed Model for the String

Computed result for the transitions of the first million letters:

| | Type 1 | Type 2 |
|---|---|---|
| A | 27,617 | 27,732 |
| B | 27,704 | 27,914 |
| C | 27,736 | 27,817 |
| D | 28,052 | 27,750 |
| E | 27,833 | 27,806 |
| F | 27,538 | 27,847 |
| G | 83,376 | 83,155 |
| H | 27,819 | 27,865 |
| I | 27,796 | 27,721 |
| J | 27,783 | 27,782 |
| K | 27,649 | 27,700 |
| L | 27,605 | 27,981 |
| M | 28,011 | 27,651 |
| N | 83,567 | 83,193 |

**Transition Type Conjecture:** Over the length of the string, each letter undergoes equal transitions of Type 1 and Type 2.

If this conjecture is true, then from the table of transitions in Section 2, we have immediately that the frequency of complementary letters must be the same, i.e. A's = H's, B's = I's, ...G's = N's. We also obtain immediately that the overstring is 1.5 times as long as the substring, in the limit over the length of the string.

From the table of transitions, the conjecture and the identity of the string and the overstring, and using A to represent the proportion of A's found in the string, we have:

$$
\begin{aligned}
1.5\ \mathrm{A} &= 0.5\ \mathrm{A} + 0.5\ \mathrm{D} + 0.5\ \mathrm{E} \\
1.5\ \mathrm{B} &= 0.5\ \mathrm{B} + 0.5\ \mathrm{D} + 0.5\ \mathrm{F} \\
1.5\ \mathrm{C} &= 0.5\ \mathrm{D} + 0.5\ \mathrm{E} + 0.5\ \mathrm{F} \\
1.5\ \mathrm{D} &= 0.5\ \mathrm{G} \\
1.5\ \mathrm{E} &= 0.5\ \mathrm{B} + 0.5\ \mathrm{C} + 0.5\ \mathrm{F} \\
1.5\ \mathrm{F} &= 0.5\ \mathrm{A} + 0.5\ \mathrm{C} + 0.5\ \mathrm{E} \\
1.5\ \mathrm{G} &= 0.5\ \mathrm{A} + 0.5\ \mathrm{B} + \mathrm{C} + 0.5\ \mathrm{E} + 0.5\ \mathrm{F} + 0.5\ \mathrm{G} \\
0.5 &= \mathrm{A} + \mathrm{B} + \mathrm{C} + \mathrm{D} + \mathrm{E} + \mathrm{F} + \mathrm{G}
\end{aligned}
$$

This system of eight linear equations in seven unknowns contains one equation which is linearly dependent and can be discarded. We are then left with seven independent equations in seven unknowns which can be solved very simply. Combining the solution for A to G with the equality of complementary letters, we obtain the following frequencies for the letters:

| | | | |
|---|---|---|---|
| A | 1/18 | H | 1/18 |
| B | 1/18 | I | 1/18 |
| C | 1/18 | J | 1/18 |
| D | 1/18 | K | 1/18 |
| E | 1/18 | L | 1/18 |
| F | 1/18 | M | 1/18 |
| G | 3/18 | N | 3/18 |

The frequency of G and N has been left as 3/18 to emphasize that these two letters occur with three times the frequency of any others. The first million letters, for example, can be tabulated as follows:

| | | |
|---|---|---|
| A | 55,349 | 1 in 18.067 |
| B | 55,618 | 1 in 17.980 |
| C | 55,553 | 1 in 18.001 |
| D | 55,802 | 1 in 17.921 |
| E | 55,639 | 1 in 17.973 |
| F | 55,385 | 1 in 18.055 |
| G | 166,531 | 3 in 18.015 |
| H | 55,684 | 1 in 17.958 |
| I | 55,517 | 1 in 18.013 |
| J | 55,565 | 1 in 17.997 |
| K | 55,349 | 1 in 18.067 |
| L | 55,586 | 1 in 17.990 |
| M | 55,662 | 1 in 17.966 |
| N | 166,760 | 3 in 17.990 |

Another way to demonstrate the closeness of the model to reality is to consider the first 526,200,068 digits of the string. Chvátal's results have been adjusted by one because his string starts 12211.... At this point, 1's - 2's = 4,932.

If the model is valid, then Keane's question can be answered in the affirmative.

# 6 Other Approaches

If we start at the beginning of the string, it is possible to work backwards and obtain a mirror image, separated from the string by a single one:

$\ldots, 22122121122122112122122 \ldots$

The question of whether there is long–run equality between ones and twos in this *doublestring* is the same as for the string itself. The doublestring, however, possesses many features of structure which the string alone does not.

Secondly, it would be possible to regard each letter as being composed of 1,2,11 and 22 elements and form a 4-vector containing this information. Then, we could use some measure of entropy to examine what happens to this entropy as we proceed from a short piece of the string through

successive overstrings, this procedure being carried on without limit. If the proposed model is correct, we will not only have long–run equality of ones and twos *digits* but of all four elements, 1, 2, 11 and 22 as well.

Lastly, the string could be parsed at every point where a Type 1 transition is followed by a Type 2 transition. This would give us HN, whose overstring is HND, then DJLG, whose overstring is JLGHMG, and so on. This approach has the advantage that every piece, instead of undergoing one of two transition types, has a unique overstring. The number of different such string pieces is finite for the reason that C,D,J and K cannot occur three in a row. There would be at most a few hundred of them, which would be amenable to computer analysis.

# 7 Acknowledgements

I learned about Keane's question at a seminar at the University of Victoria. Michael Keane referred me to F. M. Dekking and Vašek Chvátal. Chris Bose, Reinhard Illner, R.R. Davidson, Sean Bohun, Jed Chapin, Rob Bures, David Feldman, and Eugene Neufeld of the University of Victoria made valuable suggestions as to how to proceed. A sincere thank–you to you all.

# 8 References

[1] A. Carpi, Repetitions in the Kolakovski sequence, *Bull. of the EATCS* **50** (1993), 194–196.

[2] V. Chvátal, Notes on the Kolakoski sequence, *DIMACS Technical Report* **93–84** *(revised)* March 1994.

[3] F. M. Dekking, Regularity and irregularity of sequences generated by automata, *Séminaire de Théorie des Nombres de Bordeaux*, 1979–1980, exposé no. 9.

[4] F. M. Dekking, On the structure of selfgenerating sequences, *Séminaire de Théorie des Nombres de Bordeaux*, 1980–1981, exposé no. 31.

[5] M. S. Keane, Ergodic theory and subshifts of finite type, in: *Ergodic Theory, Symbolic Dynamics and Hyperbolic Spaces.* T. Bedford, M. Keane, C. Series (Eds.), Oxford University Press, Oxford, 1991, pp. 35–70.

[6] C. Kimberling, Advanced Problem 6281*, *American Math. Monthly* **86** (1979), 793.

[7] W. Kolakoski, Self Generating Runs, Problem 5304, *American Math. Monthly* **72** (1965), 674. Solution: *American Math. Monthly* **73** (1966), 681–682.

[8] G. Paun, How much Thue is Kolakoski?, *Bull. of the EATCS* **49** (1993), 183–185.