Theoretical Mathematics & Applications, Vol. 10, No. 3, 2020, 1-25
ISSN: 1792-9687(print), 1792- 9709(online)
Scientific Press International Limited

Distribution of Frequencies of the Word Occurrence in a Random String in the Vicinity of the Word's Critical Length

V.I. Ilyevsky[1]

Abstract

This work explores features of the frequency distribution probability function for the preset word's occurrence in a random string. The recurrent formula that determines distribution function, which, in turn, depends on the word and the string lengths, as well as on the overlap coordinates, has been deduced based on the multitudes' properties and is being presented herewith in the form previously unknown. Asymptotic formulas have been drawn for minimum and maximum probabilities of the word's just for once occurrence in a random string. Critical distribution parameters have been determined: the word's critical length, whereby probability of its occurrence at least once is close to 0.5, and the lengths' critical interval, whereby probability of the word's just for once occurrence shifts from the value close to one, to the value close to zero. It has been shown, that in the long string case the critical interval's width does not depend on the lengths of either word or string, and meanwhile the word's critical length is linearly dependent on the string length's logarithm. Examples have been offered for the frequency probability distribution tabulation in different cases of overlaps and at different line lengths. The attached C - language SW application allows tabulation of the frequency distribution function at any word and string lengths' value.

Mathematics Subject Classification: 60C05, 0504, 6004
Keywords: Combinatorics on words, Overlaps, Probability extremes.

[1] Holon Institute of Technology, Israel.

Article Info: *Received:* March 29, 2020. *Revised:* April 23, 2020.
Published online: June 30, 2020.

1 Introduction

The problem of calculating frequency probabilities' distribution for the word occurrence in a random string has a long history, which can be traced back to the series of works: [1]-[6]. As is known, the function under scrutiny depends on the lengths of both the word and the string, as well as on the word overlaps' vector. In the context of the existing problem of such overlaps' accounting, major efforts had earlier been directed towards calculation of the probabilities distribution function in question. Mathematical expectation and variance of this function had been known as well (Gentleman and Mullin, 1989) [1]. In the meantime, the function's extreme properties and its dependence on the string - word length ratio had not been studied until very recently. Such a research, however, has been recently initiated in (Ilyevsky, 2019) [7]. This work produced a recurrent formula for the probability of the just for once word appearance in a random string. This formula made it possible to lay down and prove the extreme properties of the corresponding probability. The present article studies properties of the frequency distribution probability function, hereinafter referred to as $p_n^t(m)$, where n and m are, respectively, lengths of the word and of the string, t - frequency of the word occurrence in a random string. (Dependence on the overlaps' vector is not clearly shown).

Purposefulness of the present work is stemming from the following preliminary qualitative evaluations, substantiated by further detailed investigation. Apparently, at the preset string length n, there exists a certain (critical) word length value m_c, whereby probability that the word will never appear in the string equals $p_n^0(m_c) \cong 0.5$. It also stands to reason that at $m \geq m_c$ the frequency distribution function reaches maximum at $t = 0$. At a sufficiently large word length, compared to m_c, probability that the word will never occur in the string is close to one ($p_n^0(m) = 1 - \gamma$, $\gamma \ll 1$). In this case $p_n^t(m) \ll 1$ to all $t \geq 1$. In the sector of words with lengths smaller than m_c, maximum value of the distribution function $p_n^t(m)$ shifts towards larger values of t. At a sufficiently small word length, probability that the word will never occur in the string is small ($p_n^0(m) = \gamma, \gamma \ll 1$). Provided relevant values of m we shall have $p_n^t(m) \ll 1$ to all $t \geq 1$ again. It should be expected, that at the preset string length and the predetermined γ parameter there exists a certain word lengths' interval, wherein the word occurrence probabilities $p_n^t(m)$ are not very small values (as compared to 0.5) to all $t \geq 1$. Evaluation of this interval and study of the distribution's behavior taking overlaps into account is the principal objective of this research. In order to meet the target we needed a new presentation for the recurrent function of the frequency probabilities' distribution for the word occurrence in a random string, first obtained by Gentelman and Mullin (1989) [1] using the combinatorial enumeration method (Gulden and Jackson, 1983) [8]. The recurrent formula, obtained by Gentelman and Mullin , have form, whereby it is impossible to identify structural generality of terms, tied to the existence of overlaps. In the present work the recurrent formula for the distribution of frequencies of the word occurrence in a random string has been derived by the original method, exclusively based on the set operations' properties (Section 2). The recurrent formula thus obtained has a crucially new form, wherein all terms, tied to the overlaps, have a common structure for all overlap positions. From the recurrent formula explicit asymptotic formulas have been derived for the extreme values of the analyzed probability in case of at least one occurrence of the word in a random string (Section 3). In turn, perception of the extreme points makes it possible to calculate the critical word lengths' interval, wherein probabilities $p_n^t(m)$ are not small to all $t \geq 1$ (Section 4). Section 5 offers examples of tabulation of the frequency distribution probability function of the word occurrence inside the critical lengths' interval depending on the word length, overlap coordinates and string length. The Appendix contains a C - language SW application that allows tabulation of the frequency distribution function at any string - word length ratio.

2 Recurrent formula for distribution of the frequencies' probabilities for the word occurrence in a random string

2.1 Notation and method

Let there be a set of alphabet symbols, hereinafter referred to as $\mathcal{W}$, wherein the number of symbols equals $\mid \mathcal{W} \mid = k \geq 2$. Let there be given: random sequence R_n, which length is n symbols of the above mentioned alphabet, and the m-long preset sequence D. For convenience, we shall hereinafter refer to sequences D and R_n as the word and the string, respectively. Let us proceed from the model of equiprobable distribution of all alphabet symbols in the string R_n. Our objective is to find a probability that word D will occur in the string R_n t times exactly. Let us examine a set of k^n different sequences R_n. We shall denote this set as $\mathcal{R}_n$. We shall index the symbol positions in strings R_n right to left. We shall consider all strings of the set $\mathcal{R}_n$ equiprobable. Let us denote as $\mathcal{R}_n^t$ a subset of the set $\mathcal{R}_n$, in which word D occurs exactly t times. Each of the words D may begin in any position of the string R_n - from n to m. A set, where word D occurs not more than t times, we shall denote as $\mathcal{R}_n^{0:t}$. The string that belongs to set $\mathcal{R}_n^t$ shall be denoted as R_n^t. The number of sequences in sets $\mathcal{R}_n^t$ and $\mathcal{R}_n^{0:t}$ we shall denote as S_n^t and $S_n^{0:t}$, respectively. We shall also denote as $\mathcal{D}_n^t$ a subset of the set $\mathcal{R}_n^t$, in which there appears word D, that begins from the left end of the string R_n, i. e. that occupies positions from n to $n-m$. (It should be noted that $\mathcal{D}_n^0 = \emptyset$). To solve the problem we shall utilize the following idea, that leads to drawing out of the recursive formula for S_n^t. Set $\mathcal{R}_n^{0:t}$ may be tied to the set $\mathcal{R}_{n-1}^{0:t}$ in the following way. Let us attach to every string of the set $\mathcal{R}_{n-1}^{0:t}$ - from the left, one by one - all elements of the alphabet in question. We shall have a set hereinafter referred to as $\mathcal{W}\mathcal{R}_{n-1}^{0:t}$. Apparently, in this set $\mathcal{W}\mathcal{R}_{n-1}^{0:t}$ there are strings that have $t+1$ words D. In these strings, words D, having appeared in transit from $\mathcal{R}_{n-1}^{0:t}$ to $\mathcal{W}\mathcal{R}_{n-1}^{0:t}$, begin with the nth symbol of the string R_n . In order to receive set $\mathcal{R}_n^{0:t}$ from the set $\mathcal{W}\mathcal{R}_{n-1}^{0:t}$ we need to delete from the latter all strings of the set $\mathcal{D}_n^{t+1}$. Consequently, taking into account that $\mathcal{D}_n^{t+1} \subseteq \mathcal{W}\mathcal{R}_{n-1}^{0:t}$, we may write down the following formula:

$$|\mathcal{R}_n^{0:t}| = |\mathcal{WR}_{n-1}^{0:t}| - |\mathcal{D}_n^{t+1}|. \tag{2.1}$$

Formula (2.1) allows us to work out an equation, connecting S_n^t with S_n^{t-1}, S_{n-m}^t and $S_{n-s_i}^t$, where s_i - lengths of periods in the word D , detailed identification of which can be find in the following section.

2.2 Description of overlaps in the word D

Property of the D's overlap represents a certain type of the shift symmetry (see (Lotharie, 2001) [2], (Lotharie, 2004) [3], (Guibus and Odlyzko, 1981) [9]). Let us write D down as follows: $D = a_1a_2\ldots a_m$, where a_j represents characters of the given alphabet. Under the string D we shall write down an identical string, shifted to the right by s_i characters.

$$a_1a_2\ldots a_{s_i}a_{s_i+1}\ldots a_m$$

$$\qquad\qquad\qquad a_1\ldots a_{m-s_i}\ldots a_m$$

Provided all symbols in the top and bottom strings, located one under the other, coincide, we say there is an overlap in position $s_i + 1$.

Definition 2.1. *Word D involves an overlap position with the coordinate $s_i + 1$, provided there exists such a s_i in the range of $1 \le s_i \le m-1$, for which*

$$a_1a_2\ldots a_{m-s_i} = a_{s_i+1}a_{s_i+2}\ldots a_m. \tag{2.2}$$

Index i in (2.2) enumerate all overlaps in the word D from left to right. Word $a_1a_2\ldots a_{s_i}$ is being usually referred to as the D period.

Length of the period equals s_i . Henceforth, we shall assume that word D may have l nontrivial overlaps that correspond to the value of i in the range of $1 \le i \le l$. The value $s_0 = 0$ shall correspond to the trivial overlap of the word with itself. In [1] and [9] overlaps are being described by means of the overlap binary vector $\vec{Q}$. For $0 \le s_i \le m-1$, the following vector $\vec{Q}$ - overlap coordinates $s_i + 1$ relation exists:

$$\begin{cases} Q_j = 1, & j = m - s_i, \\ Q_j = 0, & j \ne m - s_i. \end{cases} \tag{2.3}$$

2.3 Auxiliary recursion formula for $S_n^{0:t}$

Lemma 2.2. *Let sequence D have $l \geq 1$ overlap areas. Overlap positions' coordinates are designated as follows:*

$$s_1+1, s_2+1, \dots, s_l+1,$$

where $1 \leq s_l \leq m-1$. Then, there exists the following formula that defines $S_n^{0:t}$:

$$S_n^0 = k^n, \quad 0 \leq n \leq m-1, \tag{2.4}$$

$$S_n^{0:t} = kS_{n-1}^{0:t} + \sum_{i=1}^{l}(kS_{n-s_i-1}^t - S_{n-s_i}^t) - S_{n-m}^t, \quad n \geq m. \tag{2.5}$$

Proof. Formula (2.4) is obvious. Considering that $|\mathcal{WR}_{n-1}^{0:t}| = kS_{n-1}^{0:t}$, in order to prove the ratio (2.5) by virtue of the formula (2.1) it would suffice to show that:

$$S_{n-m}^t - \sum_{i=1}^{l}(kS_{n-s_i-1}^t - S_{n-s_i}^t) = |\mathcal{D}_n^{t+1}|. \tag{2.6}$$

Let us consider set $\mathcal{R}_{n-s_i}^t$, presenting it as follows:

$$\mathcal{R}_{n-s_i}^t = \mathcal{D}_{n-s_i}^t \cup \mathcal{G}_{n-s_i}^t, \tag{2.7}$$

where $\mathcal{G}_{n-s_i}^t$ is the $n-s_i$ long strings that have t words D, but lack words D, beginning from the left end of the string. It is evident, that:

$$\mathcal{D}_{n-s_i}^t \cap \mathcal{G}_{n-s_i}^t = \emptyset. \tag{2.8}$$

Now, let us consider set $\mathcal{WR}_{n-s_i-1}^t$, that may be presented as follows:

$$\mathcal{WR}_{n-s_i-1}^t = \mathcal{D}_{n-s_i}^{t+1} \cup \mathcal{G}_{n-s_i}^t, \tag{2.9}$$

$$\mathcal{D}_{n-s_i}^{t+1} \cap \mathcal{G}_{n-s_i}^t = \emptyset. \tag{2.10}$$

Ratios (2.7)-(2.10) will give us

$$|\mathcal{WR}_{n-s_i-1}^t| - |\mathcal{R}_{n-s_i}^t| = |\mathcal{D}_{n-s_i}^{t+1}| - |\mathcal{D}_{n-s_i}^t|. \tag{2.11}$$

In the particular case of $t=0$, equation (2.11) describes the number of strings with words D, that appeared in transit from $\mathcal{R}_{n-s_i-1}^0$ to $\mathcal{WR}_{n-s_i-1}^0$ [7]. Considering that $|\mathcal{WR}_{n-s_i-1}^t| = kS_{n-s_i-1}^t$, $|\mathcal{R}_{n-s_i}^t| = S_{n-s_i}^t$, from (2.11) we have:

$$kS_{n-s_i-1}^t - S_{n-s_i}^t = |\mathcal{D}_{n-s_i}^{t+1}| - |\mathcal{D}_{n-s_i}^t|. \tag{2.12}$$

Let us present word D as a concatenation of words $u_0, u_1, \ldots, u_l$, where u_0 is the beginning of the word D s_1 characters long, u_1 is the next $s_2 - s_1$ characters long word etc., the last word u_l has length of $m - s_l$. The first character in each of the words u_i for $1 \leq i \leq l$ matches the i - numbered overlap position. As is known (Lotharie, 2001) [2], $|u_{i+1}| \leq |u_i|$. Because of overlaps, for each of the $1 \leq i \leq l$ word D may also be presented in the following way:

$$D = g_i f^i, \tag{2.13}$$

$$g_i = u_i u_{i+1} \ldots u_l, \tag{2.14}$$

where f^i is the corresponding s_i - long suffix. Let $\mathcal{F}^{t,(i)}_{n-m}$ for $1 \leq i \leq l$ denote a subset of the set $\mathcal{R}^t_{n-m}$, strings of which have prefix f^i, but lack longer prefixes $f^{i+1}, \ldots f^l$. Let $\mathcal{F}^{t,(0)}_{n-m}$ denote a subset of all strings of the set $\mathcal{R}^t_{n-m}$, having no of any f^i prefix. (Provided strings $\mathcal{R}^t_{n-m}$ are sufficiently short, sets $\mathcal{F}^{t,(i)}_{n-m}$ for predetermined i and t will be empty.) Now let us consider strings of the $D\mathcal{F}^{t,(i)}_{n-m}$ type, where each of the strings in the set $\mathcal{F}^{t,(i)}_{n-m}$ has word D attached from the left. For example, let $D = 1121122112112211211$, where, in accordance with the overlaps' coordinates, we have: $u_0 = u_1 = 1121122$, $u_2 = 112$, $u_3 = u_4 = 1$. In this case, strings $D\mathcal{F}^{t,(i)}_{n-m}$ have the following form:

$D\mathcal{F}^{t,(0)}_{n-m} = 1121122112112211211...$
$D\mathcal{F}^{t,(1)}_{n-m} = 1121122\hat{1}121122112112211211...$
$D\mathcal{F}^{t,(2)}_{n-m} = 1121122\hat{1}121122\hat{1}121122112112211211...$
$D\mathcal{F}^{t,(3)}_{n-m} = 11211221121122112\hat{1}121122112112211211...$
$D\mathcal{F}^{t,(4)}_{n-m} = 112112211211221121\hat{1}121122112112211211...$

Selected here are beginnings of words D that appear because of overlaps. Dots denote string continuation. For the set of strings $D\mathcal{R}^t_{n-m}$, in which every string of the set $\mathcal{R}^t_{n-m}$ has word D attached from the left, we have:

$$D\mathcal{R}^t_{n-m} = \bigcup_{i=0}^{l} D\mathcal{F}^{t,(i)}_{n-m}, \tag{2.15}$$

at that for $i \neq j$:

$$D\mathcal{F}^{t,(i)}_{n-m} \cap D\mathcal{F}^{t,(j)}_{n-m} = \emptyset. \tag{2.16}$$

Having attached to each of the strings in the sets $\mathcal{D}^{t+1}_{n-s_i}$ and $\mathcal{D}^t_{n-s_i}$ word $u_0 u_1 \ldots u_{i-1}$ and making use of the formulas (2.12) and (2.15), let us write the left part of the equation (2.6) down as follows:

$$\sum_{i=0}^{l} |D\mathcal{F}_{n-m}^{t,(i)}| + \sum_{i=1}^{l} (|u_0 u_1 \dots u_{i-1} \mathcal{D}_{n-s_i}^{t}| - |u_0 u_1 \dots u_{i-1} \mathcal{D}_{n-s_i}^{t+1}|). \qquad (2.17)$$

Let word D have $r+1$ complete periods, that is for $0 \le i \le r$ we have $u_0 = u_i$, whereas for $i > r$ we have $u_0 > u_i$. Let us consider two cases.

a. Let $1 \le i \le r$. Note that in this case due to overlaps, all strings in the sets $u_0 \dots u_i \mathcal{D}_{n-s_{i+1}}^{t}$ and $u_0 \dots u_{i-1} \mathcal{D}_{n-s_i}^{t+1}$ have the same number of D words and all of them begin with the word D. Then, starting from position $n-m$, suffixes of strings $u_0 \dots u_i \mathcal{D}_{n-s_{i+1}}^{t}$ and $u_0 \dots u_{i-1} \mathcal{D}_{n-s_i}^{t+1}$ may match; hence we have

$$u_0 \dots u_i \mathcal{D}_{n-s_{i+1}}^{t} \subseteq u_0 \dots u_{i-1} \mathcal{D}_{n-s_i}^{t+1}. \qquad (2.18)$$

Obviously the strings of set $u_0 \dots u_{i-1} \mathcal{D}_{n-s_i}^{t+1} \backslash u_0 \dots u_i \mathcal{D}_{n-s_{i+1}}^{t}$ have word f^i, that begins in position $n-m$, but lack longer words $f^{i+1}, \dots f^l$, beginning in this position. Therefor, by definition of $\mathcal{F}_{n-m}^{t,(i)}$ in this case under review we have:

$$|u_0 \dots u_{i-1} \mathcal{D}_{n-s_i}^{t+1}| - |u_0 \dots u_i \mathcal{D}_{n-s_{i+1}}^{t}| = |D\mathcal{F}_{n-m}^{t,(i)}|. \qquad (2.19)$$

b. Let $r \ge 0$ and $i > r$, then, starting from position $n-m$, suffixes of strings $u_0 \dots u_i \mathcal{D}_{n-s_{i+1}}^{t}$ and $u_0 \dots u_{i-1} \mathcal{D}_{n-s_i}^{t+1}$ cannot match. Indeed, should we admit that suffixes match, it would appear that the minimal period in the word D is less than u_0. Hence, in this case we have:

$$u_0 \dots u_i \mathcal{D}_{n-s_{i+1}}^{t} \cap u_0 \dots u_{i-1} \mathcal{D}_{n-s_i}^{t+1} = \emptyset. \qquad (2.20)$$

Because of $i > r$, in the first D word of strings $u_0 \dots u_i \mathcal{D}_{n-s_{i+1}}^{t}$ word D cannot begin in any of the overlap positions s_j $(1 \le j \le l)$. Therefore, each of the strings in the set $u_0 \dots u_i \mathcal{D}_{n-s_{i+1}}^{t}$ has exactly $t+1$ words D. Besides, by virtue of the condition (2.20), we have:

$$|u_0 \dots u_{i-1} \mathcal{D}_{n-s_i}^{t+1}| = |D\mathcal{F}_{n-m}^{t,(i)}|. \qquad (2.21)$$

Now let us apply results of both a. and b. cases to the expression (2.17). By virtue of the equations (2.19) and (2.21), identical summands in (2.17) are reduced. As a result, the left part of the equation (2.6) preserves only terms that describe the number of all possible strings, having exactly $t+1$ words D (these include all $D\mathcal{F}^{t,(0)}_{n-m}$ and $u_0\mathcal{D}^t_{n-s_i}$ type strings). Consequently, the equation (2.6) is valid. Provided no overlaps exist ($l=0$), a corresponding sum shall be eliminated from the equation (2.6). In this case, validity of the equation (2.6) is clear. Therefore, Lemma 2.2 has been proven. In the particular case of $t=0$, we have the result earlier received in (Ilyevsky 2019) [7]:

$$S^0_n = k^n, \quad 0 \le n \le m-1, \tag{2.22}$$

$$S^0_n = kS^0_{n-1} + \sum_{i=1}^{l}(kS^0_{n-s_i-1} - S^0_{n-s_i}) - S^0_{n-m}, \quad n \ge m. \tag{2.23}$$

2.4 Recursion formula for p^t_n

Theorem 2.3. *Probability p^t_n that word D would occur in a random string R_n exactly t times is being determined by the following recurrent formula:*

$$p^t_n = p^t_{n-1} + \sum_{i=1}^{l} k^{-s_i}[p^t_{n-s_i-1} - p^{t-1}_{n-s_i-1} - p^t_{n-s_i} + p^{t-1}_{n-s_i}] - \\ - k^{-m}(p^t_{n-m} - p^{t-1}_{n-m}). \tag{2.24}$$

In the formula (2.24) $t \ge 1, n \ge m$. The boundary condition for the formula (2.24) is the result for p^0_n, obtained in [7] also in the form of the recursion formula, resultant from (2.23):

$$p^0_n = 1, \quad 0 \le n \le m-1. \tag{2.25}$$

$$p^0_n = p^0_{n-1} + \sum_{i=1}^{l} k^{-s_i}(p^0_{n-s_i-1} - p^0_{n-s_i}) - k^{-m}p^0_{n-m}, \quad n \ge m. \tag{2.26}$$

Proof. Let us write down formula (2.5) for $S^{0:t-1}_n$.

$$S^{0:t-1}_n = kS^{0:t-1}_{n-1} + \sum_{i=1}^{l}(kS^{t-1}_{n-s_i-1} - S^{t-1}_{n-s_i}) - S^{t-1}_{n-m}, \quad n \ge m,\, t \ge 1. \tag{2.27}$$

Subtracting term by term (2.27) from (2.5), and taking into account that $S^{0:t}_n - S^{0:t-1}_n = S^t_n$, we have:

$$S^t_n = kS^t_{n-1} + \sum_{i=1}^{l}[k(S^t_{n-s_i-1} - S^{t-1}_{n-s_i-1}) - S^t_{n-s_i} + S^{t-1}_{n-s_i}] - \\ - S^t_{n-m} + S^{t-1}_{n-m}. \tag{2.28}$$

Having divided term by term formula (2.28) by k^n, we shall obtain the required ratio (2.24). In contrast to the recurrent formula, obtained in (Gentleman and Mullin, 1989)[1], equation (2.24) has a canonical form, in which the overlap positions' coordinates occur in the identical way.

3 Asymptotic for p_n^0 in two extreme cases

3.1 Asymptotic for p_n^0 in the case of zero overlaps

From the formula (2.23) in the case of zero overlaps we have:

$$S_n^0 = k^n, \quad 0 \le n \le m-1, \tag{3.1}$$

$$S_n^0 = kS_{n-1}^0 - S_{n-m}^0, \quad n \ge m. \tag{3.2}$$

Equation (3.2) represents the order m recursion. As is known, given the initial data (3.1), it is theoretically possible to define S_n^0 clearly. Solution will be expressed through the roots of the characteristic equation, having the following form for the recursion (3.2):

$$r^m - kr^{m-1} + 1 = 0. \tag{3.3}$$

In case of $m/k^m \ll 1$ we are interested in, the task of finding the clear approximation for S_n^0 , being defined by equations (3.1) and (3.2), simplifies essentially.

Theorem 3.1. *Let r be the real root of the characteristic equation (3.3), close to k , so that the following inequality holds true:*

$$\frac{(k-r)}{r} \ll 1. \tag{3.4}$$

Let us define function $\widehat{S_n^0}$ in the following way:

$$\widehat{S_n^0} = S_n^0 = k^n, \quad 0 \le n \le m-1, \tag{3.5}$$

$$\widehat{S_n^0} = (k^m - 1)r^{n-m}, \quad n \ge m. \tag{3.6}$$

Then, function $\widehat{S_n^0}$ gives the asymptotic approximation S_n^0 on the small parameter m/k^m at $n \ge 2m$ and $m \ge 3$. Correspondingly, provided the same conditions, the asymptotic approximation for p_n^0 has the following form:

$$\widehat{p_n^0} = (k^m - 1)\frac{r^{n-m}}{k^n}. \tag{3.7}$$

At that, calculation error of S^0_n by formula (3.6) is of the order $k^n(m/k^m)^2$:

$$0 < S^0_n - \widehat{S^0_n} \leq \left(1 + \frac{m(m-3)}{2}\right) k^{n-2m}. \tag{3.8}$$

Let us note, that (3.6) exactly satisfies the recurrent formula (3.2), and besides that

$$\widehat{S^0_m} = S^0_m = k^m - 1. \tag{3.9}$$

Proof. To begin with, let us show there exists a sought-after root of the characteristic polynomial and evaluate it approximately. Taking into account condition $m/k^m \ll 1$, we shall look for the relevant root of the equation (3.3) in the following form:

$$r = (k^m - m\alpha)^{\frac{1}{m}}. \tag{3.10}$$

where α is the parameter subject to definition. Then, we shall find that α satisfies to the following equation:

$$k^m - m\alpha = k(k^m - m\alpha)^{\frac{m-1}{m}} - 1. \tag{3.11}$$

We shall write equation (3.11) down in the following way:

$$\alpha = h(\alpha), \tag{3.12}$$

$$h(\alpha) = \alpha - (k^m - m\alpha)\Big[\Big(1 - \frac{m\alpha}{k^m}\Big)^{-\frac{1}{m}} - 1\Big] + 1. \tag{3.13}$$

In order to calculate parameter α in (3.12, 3.13), method of simple iterations can be utilized. Performing analysis in (3.13) by the small parameter m/k^m, it is easy to make sure that in zero order approximation $\alpha = 1$. In the vicinity of point $\alpha = 1$ we have $|h^{'}(\alpha)| \cong m/k^m \ll 1$. Therefore, the iteration process converges. At that, in expansion α on the small parameter, linear on m/k^m term remains unchanged in all iterations. To the accuracy of order m/k^m terms we shall have:

$$\alpha \cong 1 + \frac{m-1}{2k^m}, \tag{3.14}$$

$$r^m \cong k^m - m\Big(1 + \frac{m-1}{2k^m}\Big). \tag{3.15}$$

Considering r and r^m ties, that emerge from equation (3.3), we shall find as follows:

$$r = k\Big(1 - \frac{1}{1 + r^m}\Big) \approx k - \frac{1}{k^{m-1}} + \frac{1-m}{k^{2m-1}} + \dots. \tag{3.16}$$

(Real root r of the characteristic equation may be calculated from (3.3) with any required accuracy by means of the Newton's method). Making use of formulas (3.6) and (3.15), we shall obtain an expression for $\widehat{S^0_{2m}}$:

$$\widehat{S^0_{2m}} = (k^m - 1)r^m \cong (k^m - 1)\Big[k^m - m\Big(1 + \frac{m-1}{2k^m}\Big)\Big]. \tag{3.17}$$

Omitting terms that are small by parameter m/k^m , we shall have:

$$\widehat{S^0_{2m}} = k^{2m} - (m+1)k^m - 0.5m(m-3). \tag{3.18}$$

Alternatively, from the recurrent equations (3.1) and (3.2) the accurate expression for S^0_{2m} is readily available.

$$S^0_{2m} = k^{2m} - mk^m - k^m + 1. \tag{3.19}$$

From (3.18) and (3.19) we have:

$$S^0_{2m} - \widehat{S^0_{2m}} \cong 1 + 0.5m(m-3) = \delta. \tag{3.20}$$

At $m \geq 3$ we have $\delta > 0$. To begin with, by the mathematical induction method we shall prove that for $m \geq 3$ the following holds true:

$$S^0_n - \widehat{S^0_n} \geq S^0_{n-1} - \widehat{S^0_{n-1}} \geq 0. \tag{3.21}$$

For $n \leq m + 1$ conclusion (3.21) is valid, because

$$S^0_n - \widehat{S^0_n} = 0, \quad n \leq m, \tag{3.22}$$

$$S^0_{m+1} - \widehat{S^0_{m+1}} = (k^{m+1} - 2k) - (k^m - 1)r \cong \frac{m-2}{k^{m-1}} > 0. \tag{3.23}$$

Further, let us assume that (3.21) also holds true for $n, n-1, \ldots, m+2$ and prove its correctness for $n + 1$. As soon as both S^0_{n+1} and $\widehat{S^0_{n+1}}$ satisfy the recurrent equation (3.2), we have:

$$S^0_{n+1} - \widehat{S^0_{n+1}} = (k-1)(S^0_n - \widehat{S^0_n}) + (S^0_n - \widehat{S^0_n}) - (S^0_{n+1-m} - \widehat{S^0_{n+1-m}}). \tag{3.24}$$

From the inductive hypothesis (3.21) and definition (3.5), as well as taking into account that $k \geq 2$ and $m \geq 3$, from (3.24) we get $S^0_{n+1} - \widehat{S^0_{n+1}} \geq S^0_n - \widehat{S^0_n}$, that proves the required condition (3.21).

Now let us prove that at $n \geq 2m$

$$S^0_n - \widehat{S^0_n} \leq \delta k^{n-2m}. \tag{3.25}$$

At $n = 2m$, as it follows from (3.20), the conclusion is valid. Now let us assume conclusion (3.25) holds true for $n, n-1, \ldots, 2m+1$. Then, for $n+1$, taking into account both (3.21) and (3.25), just as required, we get:

$$S^0_{n+1} - \widehat{S^0_{n+1}} = k(S^0_n - \widehat{S^0_n}) - (S^0_{n+1-m} - \widehat{S^0_{n+1-m}}) \leq \delta k^{n+1-2m}. \tag{3.26}$$

Therefore, the Theorem 3.1 has been proven. For probability we have the asymptotic formula (3.7). At that, calculation error by formula (3.7) does not exceed the shown value:

$$\delta k^{-2m} = (1 + 0.5m(m-3))k^{-2m}. \tag{3.27}$$

Formula (3.7) offers sufficiently accurate results already at $m = 5$. For example, with $k = 2, n = 50, m = 5$ by the recursion formula (2.26) at zero overlaps we get $p^0_n = 0.186$, whereas by formula (3.7), in which root r had been found by the Newton's method, we shall have $\widehat{p^0_n} = 0.184$. Calculation error does not exceed expression (3.27):

$$p^0_n - \widehat{p^0_n} = 2 \cdot 10^{-3} < \delta \cdot 2^{-10} = 5.6 \cdot 10^{-3}$$

3.2 Asymptotic for p^0_n in case of maximum number of overlaps

At the maximum number of overlaps we have $s_i = i$ for all $1 \leq i \leq m-1$. In this case, from (2.23) we get:

$$S^0_n = (k-1)\sum_{i=1}^{m} S^0_{n-i}, \quad n \geq m. \tag{3.28}$$

$$p^0_n = (k-1)\sum_{i=1}^{m} \frac{p^0_{n-i}}{k^i}, \quad n \geq m. \tag{3.29}$$

Characteristic equation, that corresponds to the recursion equation (3.28), have the following form:

$$r^{m+1} = kr^m - k + 1. \tag{3.30}$$

Theorem 3.2. *Let r be the real root of the characteristic equation (3.30), close to k according to inequality (3.4). Then, at the maximum number of* *overlaps, function $\widehat{S_n^0}$ defined by formulas (3.5) and (3.6), gives an asymptotic approximation for S_n^0 on the small parameter $m/k^m \ll 1$ at $n \geq 2m$ and $m > (k+1)/(k-1)$. For the probability calculation error on the asymptotic formula (3.7), we have the evaluation as follows:*

$$0 < p_n^0 - \widehat{p_n^0} \leq \beta k^{-2m}, \tag{3.31}$$

where at large m we have $\beta \sim m^2$.

The proof is similar to the zero overlaps case. We shall cite it in the abridged version. We shall be looking for the real root of equation (3.30), close to k, same as earlier in the form of (3.10). We shall have:

$$r^m \cong k^m - m + \frac{m}{k} - \frac{m(1+m)\left(1-\frac{1}{k}\right)^2}{2k^m}, \tag{3.32}$$

$$r \cong k - \frac{k-1}{k^m} - \frac{m(k-1)^2}{k^{2m+1}} \tag{3.33}$$

Omitting terms of a higher order of smallness than m/k^m, from (3.6) and (3.32) we shall obtain an expression for $\widehat{S_{2m}^0}$:

$$\widehat{S_{2m}^0} \cong (k^m - 1)\left[k^m - m + \frac{m}{k} - \frac{m(1+m)\left(1-\frac{1}{k}\right)^2}{2k^m}\right] \tag{3.34}$$

The exact expression for S_{2m}^0 has the following form:

$$S_{2m}^0 = k^{2m} - (m+1)k^m + mk^{m-1}. \tag{3.35}$$

Omitting terms that are small by parameter m/k^m, we shall have:

$$S_{2m}^0 - \widehat{S_{2m}^0} \cong \frac{m(1+m)\left(1-\frac{1}{k}\right)^2}{2} - m + \frac{m}{k} = \beta. \tag{3.36}$$

By virtue of condition $m > (k+1)/(k-1)$, in (3.36) we have $\beta > 0$. Also, at these values of m, the next difference is positive:

$$S_{m+1}^0 - \widehat{S_{m+1}^0} = k^{m+1} - 2k + 1 - (k^m - 1)r \cong \frac{k-1}{k^{m+1}}(m(k-1) - k). \tag{3.37}$$

Further, carrying out stages of the proof, found in section 3.1, we make sure that at $n \geq m$ we have $S_n^0 - \widehat{S_n^0} > 0$; then we prove that at $n \geq 2m$ the following inequality takes place:

$$S_{2m}^0 - \widehat{S_{2m}^0} \leq \beta k^{n-2m}. \tag{3.38}$$

From (3.38) instantly follows conclusion (3.31). As an example, we hereby offer calculations for $m = 8$, $n = 100$, $k = 2$:

$$p_n^0 = 0.829792,\ \widehat{p_n^0} = 0.829725,\ p_n^0 - \widehat{p_n^0} \approx 5 \cdot 10^{-5},$$

$$\beta \cdot 2^{-16} = 5 \cdot 2^{-16} = 7.6 \cdot 10^{-5}.$$

4 Word - string lengths' critical ratio

We shall call ratio between lengths of the word m and the string n critical, provided length of the string for the given $\gamma \ll 1$ is within $m \epsilon [m_1, m_2]$ interval, where

$$p_n^0(m_1 - 1) < \gamma, \quad p_n^0(m_1) \geq \gamma, \tag{4.1}$$

$$p_n^0(m_2) \leq 1 - \gamma, \quad p_n^0(m_2 + 1) > 1 - \gamma. \tag{4.2}$$

We shall determine the critical length of the word m_c for the given n from the formula below:

$$p_n^0(m_c) < 0.5, \quad p_n^0(m_c + 1) > 0.5. \tag{4.3}$$

As soon as probability p_n^0 depends on overlaps, we shall evaluate critical intervals in two extreme cases. The first corresponds to zero overlaps, whereas the second one - to their maximum number. In both cases, as has been demonstrated in (Ilyevsky 2019) [7], we have minimum and maximum p_n^0 values, respectively. Union of relevant critical intervals will give us evaluation of the critical interval in the general case. The critical ratio of m and n lengths may be recognized by tabulation of the recurrently found function p_n^0. Nevertheless, asymptotic expressions for p_n^0, received as above at $m/k^m \ll 1$, simplify detection of the critical interval and the critical length for the word D.

4.1 String length as the p_n^0 probability function

Let us express n through p_n^0 at $n \gg m$ and $m/k^m \ll 1$. From the asymptotic formula (3.7), taking into account that $k - r \ll k$, we shall get:

$$n \cong \frac{k}{r - k} \ln p_n^0. \tag{4.4}$$

4.2 Critical parameters in the zero overlaps case

Being confined in asymptotic expression (3.16) by the first two terms and inserting r into (4.4), we shall get relation between p_n^0, m and n:

$$p_n^0 \cong e^{-\frac{n}{k^m}}. \tag{4.5}$$

$$m \cong \frac{1}{\ln k}\Big(\frac{n}{-\ln p_n^0}\Big). \tag{4.6}$$

From (4.6) and definitions (4.1),(4.2) we shell get:

$$a \leq m_1 < a+1, \quad b-1 < m_2 \leq b, \tag{4.7}$$

$$a = \frac{1}{\ln k}\ln\Big(\frac{n}{-\ln\gamma}\Big), b = \frac{1}{\ln k}\ln\Big(\frac{n}{-\ln(1-\gamma)}\Big). \tag{4.8}$$

Consequently for the critical interval we have:

$$[m_1, m_2] = [\lceil a \rceil, \lfloor b \rfloor] \tag{4.9}$$

For the critical interval length we shall have evaluation as follows:

$$\lceil b-a \rceil = \Big\lceil \frac{1}{\ln k}\ln\Big(\frac{\ln\gamma}{\ln(1-\gamma)}\Big)\Big\rceil. \tag{4.10}$$

For the critical word length from (4.3) and 4.6 we get:

$$m_c = \Big\lfloor \frac{1}{\ln k}\ln\Big(\frac{n}{\ln 2}\Big)\Big\rfloor. \tag{4.11}$$

4.3 Critical parameters at maximum number of overlaps

Being confined in asymptotic expression (3.33) by the first two terms and inserting r into (4.4), we shall get expression for p_n^0 through m and n:

$$p_n^0 \cong e^{-\frac{n(k-1)}{k^{m+1}}}. \tag{4.12}$$

From the expression (4.12) and definitions (4.1), (4.2) we get:

$$g-1 \leq m_1 < g, \quad h-2 < m_2 \leq h-1, \tag{4.13}$$

$$g = \frac{1}{\ln k}\ln\Big(\frac{n(k-1)}{-\ln\gamma}\Big), h = \frac{1}{\ln k}\ln\Big(\frac{n(k-1)}{-\ln(1-\gamma)}\Big). \tag{4.14}$$

Consequently for critical interval we have:

$$[m_1, m_2] = [\lceil g-1 \rceil, \lfloor h-1 \rfloor] \tag{4.15}$$

For length of critical interval we shell have next evaluation:

$$\lceil h-g \rceil = \left\lceil \frac{1}{\ln k} \ln \left(\frac{\ln \gamma}{\ln(1-\gamma)} \right) \right\rceil . \tag{4.16}$$

In this way, the critical interval length is independent of n and remains identical in both extreme cases (4.10) and (4.16) having been studied. However, boundaries of the critical interval fundamentally depend on both n and the overlaps' nature. For the critical word length at maximum number of overlaps we receive from (4.3):

$$m_c = \left\lfloor \frac{1}{\ln k} \ln \left(\frac{n(k-1)}{\ln 2} \right) - 1 \right\rfloor . \tag{4.17}$$

4.4 Expansion of the critical interval in the case of an arbitrary number of overlaps

In the general case, at arbitrary number of overlaps, it would be natural to expand the critical interval $[m_1, m_2]$ by union of relevant intervals (4.9), (4.15). We shall have:

$$[m_1, m_2] = [\lceil g-1 \rceil, \lfloor b \rfloor] \tag{4.18}$$

5 Distribution of probabilities in the vicinity of critical ratio

By calculating the critical interval through formulas (4.18), (4.14) and (4.8), as well as by using recurrent formulas (3.2) and (3.29), one might easily tabulate dependence $p_n^0(m)$ for any n value. As an example, such dependence is presented for $n = 100$ and $n = 10^4$ in Tables 1 and 2. In order to tabulate distribution of probabilities p_n^t (2.24), one should determine the interval that corresponds to the critical intervals' union both in case of zero overlaps and in case of their maximum number. Table 3 shows relevant calculations for high-symmetry binary words $(k = 2, n = 100, \gamma = 0.01)$. Table 4 shows distribution of probabilities for words with different periods in the immediate vicinity of the critical word - string lengths' ratio $(k = 4, n = 10^6, m_c = 10)$.

Table 1: Probability that the word will never appear in the binary $n = 100$ long string, depending on the word length m in case of maximum number of overlaps ($m_c = 6$) and in case of their nonexistence ($m_c = 7$). Given $\gamma = 0.01$. In case of zero overlaps $[m_1, m_2] = [5, 13]$, in case of their maximum number $[m_1, m_2] = [4, 12]$.

m	p_{100}^0 Maximum overlaps	p_{100}^0 Zero overlapsd
3	0.0003	0.0000
4	0.0273	0.0003
5	0.1899	0.0294
6	0.4539	0.1969
7	0.6825	0.4615
8	0.8298	0.6880
9	0.9124	0.8331
10	0.9559	0.9142
11	0.9780	0.9568
12	0.9891	0.9784
13	0.9946	0.9893
14	0.9973	0.9947

Table 2: Probability that the word will never appear in the binary $n = 10000$ long string, depending on the word length m in case of maximum number of overlaps ($m_c = 12, [m_1, m_2] = [11, 18]$) and in case of their nonexistence ($m_c = 13, [m_1, m_2] = [12, 19], \gamma = 0.01$).

m	p^0_{10000} Maximum overlaps	p^0_{10000} Zero overlapsd
10	0.0074	$5.28 \cdot 10^{-5}$
11	0.0867	0.0074
12	0.2948	0.0867
13	0.5436	0.2949
14	0.7372	0.5433
15	0.8586	0.7372
16	0.9266	0.8586
17	0.9626	0.9266
18	0.9811	0.9626
19	0.9905	0.9811
20	0.9952	0.9905

Table 3: Distribution of probabilities p^t_{100} for periodic binary words ($k = 2$), in which the number of ones and zeroes differs by one at most. Critical interval is $[m_1, m_2] = [4, 13]$.

D/t	1010	10101	101010	10101010	101010101	1010101010101
0	0.0029	0.0785	0.2985	0.7551	0.8716	0.9919
1	0.0162	0.1677	0.2924	0.1638	0.0915	0.0061
2	0.0461	0.2076	0.2000	0.0553	0.0265	0.0015
3	0.0889	0.1917	0.1124	0.0179	0.0075	0.0004
4	0.1305	0.1458	0.0554	0.0056	0.0021	0.0001
5	0.1552	0.0959	0.0248	0.0017	0.0006	0.0000
6	0.1556	0.0563	0.0103	0.0005	0.0002	0.0000
7	0.1352	0.0300	0.0040	0.0001	0.0000	0.0000
8	0.1037	0.0148	0.0015	0.0000	0.0000	0.0000
9	0.0713	0.0068	0.0005	0.0000	0.0000	0.0000
10	0.0444	0.0029	0.0002	0.0000	0.0000	0.0000

Table 4: Distribution of probabilities $p^t_{10^6}$ for words with different periods at the critical word - string lengths' ratio ($k = 4, n = 10^6, m_c = 10$).

D/t	1111111111	1212121212	1243312433	1243312433	1111111112
0	0.4891	0.4090	0.3868	0.3857	0.3853
1	0.2624	0.3428	0.3660	0.3671	0.3675
2	0.1360	0.1651	0.1746	0.1751	0.1752
3	0.0642	0.0594	0.0560	0.0558	0.0557
4	0.0284	0.0177	0.0136	0.0134	0.0133
5	0.0120	0.0046	0.0027	0.0026	0.0025
6	0.0049	0.0011	0.0004	0.0004	0.0004
7	0.0019	0.0002	0.0001	0.0001	0.0001
8	0.0007	0.0000	0.0000	0.0000	0.0000
9	0.0003	0.0000	0.0000	0.0000	0.0000
10	0.0001	0.0000	0.0000	0.0000	0.0000

6 Conclusion

New derivation of the recurrent formula for $p^t_n(m)$ and elaboration of the extreme $p^0_n(m)$ properties substantially add to the already known results on the problem under scrutiny. As a result of this research, the problem of calculation of the frequency distribution for the word occurrence in a random string becomes simple and accessible for the general user.

7 Appendix. C-language application to calculate distribution of probabilities p^t_n

The software provided herewith prompts for the alphabet size k, length of a random string n, parameter γ, that determines the critical interval, maximum frequency t and word D. Following input of n and γ , application informs of the critical interval boundaries. It is natural to input the word, length of which is within the critical interval. Otherwise, the calculation result for $t \geq 1$ will be a column of zeros or extremely small numbers. It shall be emphasise that critical interval is reliable only if $n \gg 1$, $n > 2m$ and $m/k^m \ll 1$. Otherwise, ignore critical interval message. Maximum length of the input word is 40 symbols. This length, provided alphabet length $k = 2$, corresponds to the string's critical length of $7.6 \cdot 10^{11}$ symbols.

```
#include < stdio.h >
#include < string.h >
#include < conio.h >
#include < malloc.h >
#include < math.h >
// Function to calculate the integer power of k:
int kpower(int x, int y)
{   int z, i;
z = 1;
if(y >= 1)
{
 for(i = 1; i <= y; i++)
 {
 z = z * x;
 };
}
 return(z);
}
int main()
{
int n, k, i, j, j1, j2, m, m1, m2, t, tmax, mmax, si, NumOverl, OverlapK[40];
double Gamma;
float g, b;
double S;
double **p t;
/*Word[40] -We look for the overlaps' coordinates in this word.
Its length does not exceed 40 symbols.
j-projected overlaps' coordinate;
m - word length;
k - alphabet size;
NumOverl - Number of nontrivial overlaps.
OverlapK[40]-array, where we write the overlaps' coordinates to;
```

```
n - random string length.*/
char Word[40];
printf("Enter alphabet size\n");
scanf s("%d", &k);
printf("Enter random string size\n");
scanf s("%d", &n);
printf("Enter small parameter describing proximity of probability to one\n");
scanf s("%lf", &Gamma);//Calculation of critical interval boundaries:
g = ((log(n * (1.0 − k)/log(Gamma)))/log(k)) − 1.0;
b = (log(−n/log(1.0 − Gamma)))/log(k);
g = ceil(g); b = floor(b);
printf("Critical interval boundaries : \n");
printf("m1 = %.0f, m2 = %.0f\n", g, b);
mmax = b;
printf("Enter maximum frequency\n");
scanf s("%d", &tmax);
getchar();
printf("Enter Word\n");
gets(Word);
m = strlen(Word); //We get length of word.
p t = (double **)malloc(mmax * sizeof(double*));
NumOverl = 0; //Initialization of the variable describing the number of overlaps.
//We check if any overlaps in j position exist.
for(j = 0; j <= m − 1; j ++)
{
 for(i = 0; Word[i] == Word[i + j]&&i + j <= m − 1; i ++)
 {
 if (i + j == m - 1)
  {
 //We write overlaps' coordinates to the OverlapK array:
 OverlapK[NumOverl] = j;
//We increase array index by 1 to write the next overlap coordinate:
 NumOverl ++;
  };
```

```
}
}
//Upon loop exit NumOverl variable equals the number of overlaps.
//Calculation of P_n^t on the recurrent formula:
for(j = 0; j <= m; j++)
{ //Here, j variable is the gradually growing string length.
 p t[j] = (double*)malloc(tmax * sizeof(double));
 for(t = 0; t <= tmax; t++)
 {
  if (t == 0) {
  p_t[j][t] = 1.;
  }
  else { p_t[j][t] = 0.; }
 }
 }
for(j = m; j <= n; j++)
{
for(t = 0; t <= tmax; t++)
 {
j1 = m;
 p t[j1][t] = 0.;
  for(i = 1; i <= NumOverl - 1; i++)
  {
  si = OverlapK[i];
  if(t == 0){
p_t[j1][t] = p_t[j1][t] + (p_t[j1 - si - 1][t] - p_t[j1 - si][t])/kpower(k, si);
  }
  else {
 p t[j1][t] = p t[j1][t] + (p t[j1 - si - 1][t] - p t[j1 - si - 1][t - 1]-
 p t[j1 - si][t] + p t[j1 - si][t - 1])/kpower(k, si);
   }
   } //Adding first and last term of the recurrent formula for p_n^t:
  if(t == 0){
p_t[j1][t] = p_t[j1 - 1][t] + p_t[j1][t] - (p_t[j1 - m][t])/kpower(k, m);
  }
```

```
else{
p t[j1][t] = p t[j1][t] + p t[j1 − 1][t] + (p t[j1 − m][t − 1]−
p t[j1 − m][t])/kpower(k, m);
 }
 }
 if(j<n){
  for(t = 0; t <= tmax; t + +)
  {
  for(j2 = 1; j2 <= m; j2 + +)
  {
  p_t[j2 − 1][t] = p_t[j2][t];
  }//Probability calculation results for lengths from n-m+1 to n
  }//are stored in the array cells from 0 to m-1
  }
  }
for(t = 0; t <= tmax; t + +){ //Distribution function print out
printf("t = %d, p t = %f\n", t, p t[m][t]);
}
S = 0; //Check. Probabilities' sum should equal one.
for(t = 0; t <= tmax; t + +)
{
 S = S + p t[m][t];
  }
printf("S = %f", S);
free(p t); }
```

References

[1] Gentleman J.F., Mullin R.C. (1989). The Distrribution of the Frequency of Occurrence of Nucleotide Subsequence, Based on Their Overlap Capability. *Biometrics*, **vol.45**, pp.35 - 52.

[2] Lotharie (2001). *Algebric Combinatorics on Words.* Cambridge University Press.

[3] Lotharie (2004) *Applied Combinatorics on Words*. Cambridge University Press.

[4] Fu, J.C., Koutras M.V. (1994). Theory of Runs: A Markov Chain Approach. *Journal of the American Statistical Association*, **89**(427), pp. 1050 - 1058.

[5] Fu, J. C. and Lou, W. (2003). *Distribution Theory of Runs and Patterns and its Applications: A Finite Markov Chain Imbedding Approach.* Singapore: World Scientific.

[6] Chufang Wu. (2005). The Distribution of the Frequency of Occurrence of Nucleotide Subsequence. *Methodology and Computing in Applied Probability*, **7**, pp. 325-334.

[7] Ilyevsky V. I. (2019). Recursive Formula for the Random String Word Detection Probability, Overlaps and Probability Extremes. *Journal of Mathematics Research*, **Vol. 11**(2), pp.171-180.

[8] Gulden I.P. and Jackson D.M. (1983). *Combinatorial Enumeration.* Wiley, New York.

[9] Guibus L.J., Odlyzko A.M. (1981). Periods in Strings. *Journal of Combinatorial Theory*, **A 30**, pp. 19-42.