Recursive Formula for the Random String Word Detection Probability, Overlaps and Probability Extremes

V. I. Ilyevsky

Correspondence: V.I. Ilyevsky, Holon Institute of Technology (HIT), 52 Golomb St., Holon 5810201, Israel

Received: February 17, 2019 Accepted: March 18, 2019 Online Published: March 24, 2019

doi:10.5539/jmr.v11n2p171 URL: https://doi.org/10.5539/jmr.v11n2p171

Abstract

In this paper, for the first time ever, the properties of the word detection probability in a random string have been investigated. The formerly known methods led to numerical evaluation of the researched probabilities only. The present work derives the simplest algorithm for calculation of the word's at least once detection probability in a random string. A recursive formula that considers the overlap capability has been deduced for the probability under study. This formula is being used for the proposition on comparison of the word detection probabilities in a random string for the words with different periods. The result allows determining the structure of words that have maximum and minimum detection probabilities. In particular, words having equal number of alphabetic characters have been studied. It has been established, that for the words in question detection probability is minimal for the ideally symmetrical words that have irreducible period - and maximal for the words devoid of the overlap feature. These results will be useful for molecular genetics, as well as for students studying discrete mathematics, probability theory and molecular biology.

Keywords: word occurrence, probability, combinatorics, overlaps, probability extremes

1. Introduction

Calculation of detection probability of a given word in a random string is of the great interest, most of all in connection with the molecular genetics research activities. The substantial breakthrough in the problem solution had been made by Gentleman and Mullin (1989). The work (Gentleman & Mullin 1989) established distribution of the subsequence's occurrence frequencies within the nucleotide sequence, taking into account possible overlaps within the equiprobable distribution model. Such success was stipulated by employment of the enumerative combinatory analysis and generating functions (Gulden, 1983). Chufang (2005) offered a different scheme to solve the problem in question, based on the finite Markov chains imbedding technique. A common approach to the similar problems, based on the Markov chains model, had been also developed in works (Robin & Daudin, 1999), (Robin & Daudin, 2001), (Lotharie, 2004), (Rigner, 1995). The principal problem in calculation of the word occurrence probability in a random string is the overlap capability. Although the known solution methods give reliable numerical results, they do not allow investigating the extremal properties of the word detecting probability in a random string that relates to its symmetry, stipulated by presence of overlaps. It should be emphasized that the symmetry factor plays an important part in the debate on the degree of order and the amount of information (Ilyevsky, 2014), (Ilyevsky, 2017). This paper offers an original method of deriving a recursion formula for the word detection probability to obtain recurrence relations in a previously unknown elegant form. The proposed recursive formula is very simple for computer programming. By means of the derived recurrence formula the theorem on the extremal properties of the probabilities under study, associated with the presence of overlaps, has been proved.

Section 2 of the present paper considers a random string, n characters of $k \geq 2$ alphabetical elements long, within the equiprobable distribution model. Problem to calculate the probability p_n of detecting a specified m characters long word in a random string at least once, has been solved. The solution has been obtained in the form of a recursive formula that connects p_n with probabilities p_{n-1}, p_{n-s_i} and p_{n-m}, where $s_i + 1$ are coordinates of the word overlap positions, numbered with the i index. The result allows to calculate the precise p_n meaning under any value of m and n. In section 2.3 the explicit formula for p_n has been received for the zero overlap cases. Section 3 offers proposition on comparison of the word occurrence probabilities in a random string for the words with different periods. Section 4 shows, that for words with equal number of each of the k alphabetic characters the p_n value is maximal at zero overlaps and is minimal at the ideal symmetry of the word, when it have the irreducible period.

2. Recursive Formula for the Given Word's at Least Once Occurrence Probability in a Random String

2.1 Basic Idea and Method

In contrast to the known methods, the approach offered below allows us to derive necessary formulas without invoking the enumerated combinatorics.

Let us assume there are: an alphabet of $k \geq 2$ characters, R_n - a random sequence n characters long and D - a preset sequence m characters long. For the sake of D and R_n sequence convenience we shall hereinafter denote them a word and a string, respectively. Let us draw on the model of equiprobable distribution of all alphabetic characters in the R_n string. Our objective is to find a probability of the D word's at least once occurrence in the R_n string. Let us consider a set, that consists of k^n different R_n sequences of all kinds. We shall denote this set as $\mathcal{R}_n$. All strings within the $\mathcal{R}_n$ set are equiprobable. Let us denote a subset of the $\mathcal{R}_n$ set, in which D does not occur even once, as $\mathcal{R}'_n$. Correspondingly, we shall denote strings that belong to the $\mathcal{R}'_n$ set, as R'_n. The number of sequences in $\mathcal{R}'_n$ we shall designate by Q_n. Now let us construct a recursive formula for Q_n. To this end, we shall employ the idea as follows. The $\mathcal{R}'_n$ set could be deduced from the $\mathcal{R}'_{n-1}$ set using the following procedure.

Let us denote a set of all possible alphabet characters as $\mathcal{W}$. Let us add all possible $w \epsilon \mathcal{W}$ words to the left of every R'_{n-1} string. We shall have a set of strings, hereinafter referred to as $\mathcal{W}\mathcal{R}'_{n-1}$. The number of such strings shall be:

$$|\mathcal{W}\mathcal{R}'_{n-1}| = kQ_{n-1}.$$

Now let us cross out all strings that have word D from the $\mathcal{W}\mathcal{R}'_{n-1}$ set. We shall have a $\mathcal{R}'_n$ set. Hereinafter we shall refer to the procedure described above as the $\mathcal{R}'_{n-1}$ to $\mathcal{R}'_n$ transition. (We shall also use such transitions step by step to describe the transition from $\mathcal{R}'_{n-m}$ to $\mathcal{R}'_n$.) The number of crossed-out strings in the transition $\mathcal{R}'_{n-1} \rightarrow \mathcal{R}'_n$ will be:

$$|\mathcal{W}\mathcal{R}'_{n-1}| - |\mathcal{R}'_n| = kQ_{n-1} - Q_n.$$

On the other hand, the number of crossed-out strings in transition from $\mathcal{R}'_{n-1}$ to $\mathcal{R}'_n$ can be expressed through Q_{n-m}. Indeed, since length of the word D equals m, in crossed-out strings it occupies positions from $n - m + 1$ to n. Should we add all possible words of length m to the left of each of the $\mathcal{R}'_{n-m}$ set strings, then D will be one of these words. Therefore, the number of strings with length n and prefix D that are deleted in transition from $\mathcal{R}'_{n-1}$ to $\mathcal{R}'_n$, must equal Q_{n-m} minus number of strings crossed out in the overlap positions. The corresponding detailed analysis of the overlap accounting is given below in sections 2.2, 2.5. Equating the number of crossed out strings obtained in two ways, we arrive at the required recurrence relation.

2.2 Overlap Description

As already noted, when constructing a recursive formula for Q_n, one shall take into account all possible overlaps of the word D in the string R_n. D's overlap feature represents a certain type of the shift symmetry. Let us write D down as follows: $D = a_1 a_2 \dots a_m$, 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.

$$a_1 a_2 \dots a_{s_i} a_{s_i+1} \dots a_m$$

$$a_1 \dots a_{m-s_i} \dots a_m$$

If all the characters of the upper and lower strings, one above the other, are the same, then equivalent definitions are introduced by different authors, such as: the notion of autocorrelation of the D word (Guibus & Odlyzko, 1981), D's overlap capability (Gentleman, 1989), or periodicity (Lotharie, 2001). For our purposes it would be more convenient to define overlaps as follows.

Definition. A word D has an overlap position with the coordinate $s_i + 1$ if within the range $1 \leq s_i \leq m - 1$ there exists such s_i, to which

$$a_1 a_2 \dots a_{m-s_i} = a_{s_i+1} a_{s_i+2} \dots a_m. \tag{1}$$

Index i in (1) enumerate possible overlaps from left to right. A $a_1 a_2 \dots a_{s_i}$ word is usually referred as the D period. Length of the period equals s_i. As an example, let us consider the following word from the alphabet 1,2,3:

$$1121123112112311211 \tag{2}$$

In the above example (2) we have: $s_1 = 7$, $s_2 = 14$, $s_3 = 17$, $s_4 = 18$, and, correspondingly, coordinates of the overlap positions are :8, 15, 18, 19. In the discussions that follow, we omit the case of the D's trivial coincidence with himself, that corresponds to $s_0 = 0$. Let us note here, that s_i values can not be arbitrary, but obey certain rules, derived in work (Guibus & Odlyzko, 1981). To begin with, let us consider a case, when a word D does not have any overlap positions.

2.3 Recursive Formula for Zero Overlaps in D

Let a m - long word D have no overlaps, meaning that D does not have any such s_1 within the $1 \le s_1 \le m-1$ range, for which (1) holds. We shall hereinafter refer to a word D, having such property, as D_0. For the Q_n number of strings in the set $\mathcal{R}_n$, that have no D_0 words, we have the recurrence relation as follows:

$$Q_n = k^n, \qquad 0 \le n \le m-1, \tag{3}$$

$$Q_n = kQ_{n-1} - Q_{n-m}, \quad n \ge m. \tag{4}$$

Proof. For $0 \le n < m$ the proposition (3) is obvious, since in the string R_n, which is shorter then m, a word D_0 does not occur even once. Let $n \ge m$. Let us consider the set $\mathcal{R}'_{n-1}$, which means there are $n-1$ - long strings, in which D_0 does not occur even once. The $\mathcal{R}'_{n-1}$ set is being derived from $\mathcal{R}_{n-1}$ by means of elimination from the latter of all strings, in which D_0 occurs. According to the designation introduced as above, number of sequences in $\mathcal{R}'_{n-1}$ equals Q_{n-1}. To each of the sequences in the set $\mathcal{R}'_{n-1}$ we shall add (to the left, in turn) all characters from the alphabet. Such a procedure will generate a $\mathcal{W}\mathcal{R}'_{n-1}$ set that contains kQ_{n-1} new strings. Apparently, among strings in the $\mathcal{W}\mathcal{R}'_{n-1}$ set there are $R_n = D_0R'_{n-m}$ strings that represent concatenation of the word D_0 with the string R'_{n-m}. These strings generate a set of D_0 - prefixed strings that we shall refer to as $D_0\mathcal{R}'_{n-m}$. Let us represent a word D_0 in the form of $D_0 = u_s v_{m-s}$ where u_s and v_{m-s} shall be parts of the word D_0, s and $m-s$ long, correspondingly ($s\epsilon[1, m-1]$). Now let us analyze the set $v_{m-s}\mathcal{R}'_{n-m}$, in which every R'_{n-m} string has a v_{m-s} word, attached from the left. Since the word D_0 does not have any overlaps, the word v_{m-s} shall not represent a D_0 prefix. That is the reason why none of the $v_{m-s}\mathcal{R}'_{n-m}$ set elements shall be crossed out during transition from $\mathcal{R}'_{n-m}$ to $\mathcal{R}'_{n-1}$ for all $s\epsilon[1, m-1]$. This means that $D_0\mathcal{R}'_{n-m} \subset \mathcal{W}\mathcal{R}'_{n-1}$. Consequently, following transition from $\mathcal{R}'_{n-m}$ to $\mathcal{W}\mathcal{R}'_{n-1}$ there appear exactly $|D_0\mathcal{R}'_{n-m}| = Q_{n-m}$ D_0-prefixed strings that belong to $\mathcal{W}\mathcal{R}'_{n-1}$ set. In this way we arrive at the equation (4).

Having divided equation (4) by k^n, we get the recursive relation for a $q_n(D_0)$ probability that not a single word D_0 shall ever occur in the $\mathcal{R}_n$ set:

$$q_n = \begin{cases} 1, & 0 \le n \le m-1, \\ q_{n-1} - k^{-m}q_{n-m}, & n \ge m. \end{cases} \tag{5}$$

For the probability that word D_0 would occur in the set $\mathcal{R}_n$ at least once we have:

$$p_n = 1 - q_n. \tag{6}$$

The appropriate recurrence relation for the probability $p_n(D_0)$ will be:

$$p_n = \begin{cases} 0, & 0 \le n \le m-1, \\ p_{n-1} + k^{-m}(1 - p_{n-m}), & n \ge m. \end{cases} \tag{7}$$

2.4 Explicit Formula Fore the Probability $p_n(D_0)$

Based on the inclusion - exclusion principle one may derive an explicit formula for the probability $p_n(D_0)$. At $n \ge m$ we get:

$$p_n(D_0) = \sum_{t=1}^{h(n)} (-1)^{t-1} \binom{n - tm + t}{t} k^{-tm}, \qquad h(n) = \left\lfloor \frac{n}{m} \right\rfloor. \tag{8}$$

Let us show, that result (8) satisfies the recurrence relation (7). Substituting the result (8) to the right side of the expression (7), we obtain for $n \ge m$:

$$p_n = A + B, \tag{9}$$

$$A = \sum_{t=1}^{h(n-1)} (-1)^{t-1} \binom{n - 1 - tm + t}{t} k^{-tm},$$

$$B = -\sum_{t=0}^{h(n-m)} (-1)^{t-1} \binom{n - (t+1)m + t}{t} k^{-(t+1)m}.$$

Performing the substitution $t + 1 = \widetilde{t}$ in the sum B, and omitting the tilde sign over t, we get:

$$B = \sum_{t=1}^{h(n-m)+1} (-1)^{t-1} \binom{n - tm + t - 1}{t - 1} k^{-tm}. \tag{10}$$

Let $n \geq m$. Let us denote $n = um + v$, where u and v are natural numbers, and v satisfies an inequality $0 \leq v < m$. Thereupon, we obtain as follows:

$$\left\lfloor \frac{n}{m} \right\rfloor = u, \qquad \left\lfloor \frac{n-1}{m} \right\rfloor = \begin{cases} u & 1 \leq v < m, \\ u - 1 & v = 0, \end{cases} \tag{11}$$

$$\left\lfloor \frac{n-m}{m} \right\rfloor + 1 = u. \tag{12}$$

From (11) and (12) it follows that, provided $v \neq 0$, the upper limits in the sum A and sum B , written down in the form of (10), equal u. Using well-known property of binomial coefficients, we transform sums of corresponding combination terms in expressions A and B for the case $v \neq 0$ in the following way:

$$\binom{n - 1 - tm + t}{t} + \binom{n - tm + t - 1}{t - 1} = \binom{n - tm + t}{t}. \tag{13}$$

Then for the case $v \neq 0$ we get:

$$A + B = \sum_{t=1}^{h(n)} (-1)^{t-1} \binom{n - tm + t}{t} k^{-tm} = p_n, \qquad h(n) = \left\lfloor \frac{n}{m} \right\rfloor. \tag{14}$$

When $v = 0$, the the sum A (9) has one summand less than the sums (10) and (8). In order to convert the first corresponding $u - 1$ terms in the sums A and B (10) we have expression (13), and the last summand in the B (10) equals, as it is easily seen, the last summand in the sum (8). In this way, we have proven formula (8) on the basis of the recursive relation (7).

2.5 Recurrence Formula for the D Overlaps Case

2.5.1 Recurrence Formula for Q_n

Theorem 1. *Let a word D have $l \geq 1$ overlap positions. We shall denote these overlap positions' coordinates as follows:*

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

where $1 \leq s_l \leq m - 1$. Then, the following recursive formula is valid:

$$Q_n = k^n, \qquad 0 \leq n \leq m - 1, \tag{15}$$

$$Q_n = kQ_{n-1} + \sum_{i=1}^{l} (kQ_{n-s_i-1} - Q_{n-s_i}) - Q_{n-m}, \quad n \geq m. \tag{16}$$

Proof. Expression (15) is obvious. Let us prove equation (16). Similar to equation (4) we shall write:

$$Q_n = kQ_{n-1} - U_{n-m}, \tag{17}$$

where U_{n-m} is a number of D words crossed out during $\mathcal{R}'_{n-1}$ to $\mathcal{R}'_n$ transition. For now, however, due to the presence of overlaps, $U_{n-m} < Q_{n-m}$. Let us express U_{n-m} through Q_{n-m} and Q_{n-s_i}. A word D will be presented as concatenation of words $u_0, u_1, \ldots, u_l$: $D = u_0u_1 \ldots u_l$, where u_0 is the beginning of the D, s_1 characters long, u_1 is the next word $s_2 - s_1$ characters long etc. The last word u_l will be $m - s_l$ - long. The first character of every u_i word for $1 \leq i \leq l$ matches the i-numbered overlap position. Alternatively, due to overlaps, for every $1 \leq i \leq l$ a word D may be presented as follows:

$$D = u_iu_{i+1} \ldots u_lf_i, \tag{18}$$

where f_i is the corresponding suffix. The $\mathcal{R}'_{n-m}$ set does not have any D words, but its strings may incorporate f_i prefixes. Let us consider a $\mathcal{F}_i$ subset of the set $\mathcal{R}'_{n-m}$, strings of which have prefix f_i, but lack prefixes $f_{i+1}, \ldots, f_l$. Having extended

the $\mathcal{R}'_{n-m}$ set strings to the left, step by step according to the procedure described earlier and as far as the n position, we shall get $\mathcal{WR}'_{n-1}$ set. Let us single out subsets of the strings, that have prefix D: $D\mathcal{R}'_{n-m}$ and $D\mathcal{F}_i$. During transition from $\mathcal{R}'_{n-m}$ to $\mathcal{R}'_{n-1}$ all strings, that have prefix D (18), are being crossed out consequentially, starting with $i = l$ and ending with $i = 1$. Therefore, the set $\mathcal{WR}'_{n-1}$ has no strings containing D words (18), that begin at $n - s_i$ position for all $1 \le i \le l$ ($Df_i \notin \mathcal{WR}'_{n-1}$). All such strings, however, are present in the $D\mathcal{R}'_{n-m}$ set, because $D\mathcal{F}_i \subseteq D\mathcal{R}'_{n-m}$. Subsets $D\mathcal{F}_i$ do not intersect for different i. Therefore, the number of D - prefixed strings, that should be crossed out from the $\mathcal{WR}'_{n-1}$ set, shall be as follows:

$$U_{n-m} = |D\mathcal{R}'_{n-m}| - \sum_{i=1}^{l} |D\mathcal{F}_i|. \tag{19}$$

The number of strings, crossed out from each of the $n - s_i$ overlap positions, may be written down in the following way:

$$|D\mathcal{F}_i| = |\mathcal{F}_i| = J_{n-s_i}, \tag{20}$$

$$J_j = -Q_j + kQ_{j-1}, \quad j \ge 1. \tag{21}$$

From (19) and (20) we get:

$$U_{n-m} = Q_{n-m} - \sum_{i=1}^{l} J_{n-s_i}. \tag{22}$$

Substituting expression (22) into equation (17) we get the recurrence formula (16). In this way the Theorem 1 is proven.

2.5.2 Recurrence Formula for the p_n Probability

Recurrence formula for the probability p_n is the Theorem 1 corollary. Having divided equation (16) by k^n, we get the recurrence formula for the probability q_n that not a single word D shell never occur in the set $\mathcal{R}_n$:

$$q_n = 1, \qquad 0 \le n \le m-1, \tag{23}$$

$$q_n = \sum_{i=0}^{l} \left(\frac{q_{n-s_i-1}}{k^{s_i}} - \frac{q_{n-s_{i+1}}}{k^{s_{i+1}}} \right), \qquad n \ge m. \tag{24}$$

In equation (24) and in the following text, we denote $s_0 = 0$, $s_{l+1} = m$. From equation (24) we get a recursive formula for $p_n = 1 - q_n$:

$$p_n = 0, \qquad 0 \le n \le m-1, \tag{25}$$

$$p_n = \sum_{i=0}^{l} \left(\frac{p_{n-s_i-1}}{k^{s_i}} - \frac{p_{n-s_{i+1}}}{k^{s_{i+1}}} \right) + \frac{1}{k^m}, \qquad n \ge m. \tag{26}$$

3. Comparison of the Given Word Occurrence Probability in a Random String for Two Words With Different Periods

3.1 Lemma on the Number of Crossings out

Lemma 1. *The number of crossings out J_j, performed during transition from the set $\mathcal{R}'_{j-1}$ to the set $\mathcal{R}'_j$, is a nondecreasing function of j, that is for any $j \ge m$:*

$$J_j \ge J_{j-1}. \tag{27}$$

Proof by induction. By virtue of the fact that we are only interested in cases of nontrivial overlaps in D, let $m \ge 2$. For $j = m$ conclusion (27) holds, because $J_{m-1} = 0$, $J_m = 1$. Let us assume, that (27) holds for $m + 1 \le j \le n$ and show, that it also holds for $j = n + 1$. Using the formula of J_n (21) and recurrence relation (16), we get:

$$J_{n+1} = (k-1)J_n + \sum_{i=1}^{l} (k-1)J_{n-s_i} + \sum_{i=1}^{l+1} (J_{n-s_{i-1}} - J_{n+1-s_i}). \tag{28}$$

In equation (28) we have: $J_{n-s_i} \geq 0$, $k \geq 2$. Further at $s_{i-1} = s_i - 1$ we have $J_{n-s_{i-1}} - J_{n+1-s_i} = 0$, and if $s_{i-1} < s_i - 1$, then by inductive hypothesis we have: $J_{n-s_{i-1}} - J_{n+1-s_i} \geq 0$. Consequently, from (28) it follows that:

$$J_{n+1} \geq J_n. \tag{29}$$

In this way the Lemma 1 has been proven.

3.2 Theorem on Comparison of Probabilities

Theorem 2. *Let there be given two words of equal length $m \geq 2$, hereinafter referred to as D and E . Let a word D have l overlaps ($l \geq 0$), described dy periods t_i, whereas E has $l + r$ overlaps , described dy periods s_i. Let l first periods of E be smaller then l corresponding periods of D, i.e. for all $1 \leq i \leq l$ we have $s_i < t_i$. Then at $n > m + s_1$, probability p_n to detect word E at least once in a random string R_n shall be smaller, than the corresponding probability g_n for the word D :*

$$p_n < g_n. \tag{30}$$

Let us give an example of two words D and E that satisfy conditions of the theorem:

$$E : 112\tilde{1}12\tilde{1}12\tilde{1}12\tilde{1}\tilde{1}$$

$$D : 111121\tilde{1}1112\tilde{1}\tilde{1}\tilde{1}$$

The highlighted characters mark the overlap positions. Calculations for
$k = 2$, $n = 5000$, $m = 14$ give us $p_n(E) = 0.2340$, $g_n(D) = 0.2590$.

3.3 Proof of the Theorem 2

Let us denote the number of words E and D in the sets $\mathcal{R}'_n(E)$, $\mathcal{R}'_n(D)$ by Q_n and G_n correspondingly. We shell also denote $\delta_n = Q_n - G_n$. In this notation inequality (30) is equivalent to inequality $\delta_n > 0$.

To begin with, let us consider a particular case of the theorem, when number of overlaps in both E and D is the same and equals l. Primarily, we shall make sure that:

$$\delta_n = 0, \quad m \leq n < m + s_1. \tag{31}$$

Then, by induction, we shall prove correctness of the following inequality:

$$\delta_n > \delta_{n-1} \geq 0, \quad n \geq m + s_1. \tag{32}$$

(In the (32) the $\delta_{n-1} = 0$ case holds only at $n = m + s_1$). According to the recurrence formula for Q_n (15) and (16) we have:

$$Q_n = G_n = k^n, \quad 0 \leq n \leq m - 1, \tag{33}$$

$$Q_m = G_m = k^m - 1, \tag{34}$$

Context of the formula (34) is evident - in transition from $\mathcal{R}_m$ to $\mathcal{R}'_m$ exactly one string D and one string E are being crossed out from $\mathcal{R}_m$. Further, for Q_n we have an equation (16) and for G_n - a similar equation as follows:

$$G_n = kG_{n-1} + \sum_{i=1}^{l}(kG_{n-t_i-1} - G_{n-t_i}) - G_{n-m}, \quad n \geq m. \tag{35}$$

Let $n < m + s_1$. Then, from (33-35), taking into account that $t_i > s_i \geq 1$, $m \geq 2$, we get:

$$kQ_{n-s_i-1} - Q_{n-s_i} = 0, \quad 1 \leq i \leq l, \tag{36}$$

$$kG_{n-t_i-1} - G_{n-t_i} = 0, \quad 1 \leq i \leq l, \tag{37}$$

$$\delta_n = k\delta_{n-1}. \tag{38}$$

Since at $n = m$, based upon (34), we have $\delta_n = 0$, then from (38) there follows conclusion (31). Let us calculate δ_{m+s_1}. From (16) and (35)-(37) we get:

$$Q_{m+s_1} = kQ_{m+s_1-1} - Q_m + kQ_{m-1} - Q_{s_1}. \tag{39}$$

$$G_{m+s_1} = kG_{m+s_1-1} - G_{s_1}. \tag{40}$$

If $s_1 = 1$ we have:

$$Q_{m+1} = k^{m+1} - 2k + 1. \tag{41}$$

$$G_{m+1} = k^{m+1} - 2k. \tag{42}$$

At $s_1 > 1$ we get:

$$Q_{m+s_1} = kQ_{m+s_1-1} - k^{s_1} + 1. \tag{43}$$

$$G_{m+s_1} = kQ_{m+s_1-1} - k^{s_1}. \tag{44}$$

From equations (41-44), taking into account (31), we get:

$$\delta_{m+s_1} = 1. \tag{45}$$

Thus, we obtain:

$$\delta_{m+s_1} > \delta_{m+s_1-1}. \tag{46}$$

In order to prove inequality (32), let us make use of the result (46) as the induction beginning on the variable n . Let us suppose, that for all j values in the interval of $m + s_1 \le j \le n - 1$ the following inequality has been fulfilled:

$$\delta_j > \delta_{j-1} \ge 0 \tag{47}$$

Let us prove that for $j = n$ inequality (47) has also been fulfilled. From (16) and (35) we have:

$$\delta_n = k\delta_{n-1} - \sum_{i=1}^{l}[(Q_{n-s_i} - G_{n-t_i}) - k(Q_{n-s_i-1} - G_{n-t_i-1})] - (Q_{n-m} - G_{n-m}). \tag{48}$$

Let us represent equation (48) in the next form:

$$\delta_n = \sum_{i=0}^{l}[(k-1)\delta_{n-s_i-1} + (\delta_{n-s_i-1} - \delta_{n-s_{i+1}})] + \sum_{i=1}^{l}(J^D_{n-s_i} - J^D_{n-t_i}). \tag{49}$$

In the formula (49) J^D_j is a number of crossings out of the strings, that have D while in transition from $\mathcal{R}'_{j-1}$ to $\mathcal{R}'_j$:

$$J^D_j = -G_j + kG_{j-1}. \tag{50}$$

According to the inductive hypothesis (47), and taking into account that $s_{i+1} \ge s_i + 1$, in equation (49) we have:

$$\delta_{n-s_i-1} \ge 0. \tag{51}$$

$$\delta_{n-s_i-1} - \delta_{n-s_{i+1}} \ge 0. \tag{52}$$

In accordance with Lemma 1 we get:

$$J^D_{n-s_i} - J^D_{n-t_i} \geq 0. \tag{53}$$

In addition to inequalities (51-53), we have $k \geq 2$. Also, for $n \geq m + s_1 + 1$ we have a strict inequality $\delta_{n-1} > 0$. By virtue of the above mentioned, from (49) we obtain, that for all $n \geq m + s_1 + 1$:

$$\delta_n > \delta_{n-1}. \tag{54}$$

In this way, Theorem 2 for the equal number of overlaps positions has been proven. If $r > 0$, then next additional sum appears in the equation (49):

$$J^E_{n-s_{l+1}} + J^E_{n-s_{l+2}} + \cdots + J^E_{n-s_{l+r}}, \tag{55}$$

where J^E_j is the number of crossing out of strings that contain E during transitions from $\mathcal{R}'_{j-1}$ to $\mathcal{R}'_j$. All terms in the sum (55) are not negative, therefore conclusions (32) and (30) remain valid.

3.4 Corollary of the Theorem 2

The word D at least once occurrence probability in a random string is maximal if the word has no overlaps and is minimal if all m characters of the word are identical.

4. Theorem on Extreme Probabilities for the Words With Equal Number of All Alphabetic Characters

4.1 Statement of the Theorem

Theorem 3. *We shall consider the set $\mathcal{D}$ that consists of $m = kd$ - long D words, in which every alphabetic character occurs exactly d times, where $d \geq 2$. As above, let us denote probability, that a word $D \epsilon \mathcal{D}$ would be detected in a random string R_n at least once, as $p_n = p_n(D)$. In particular, let us single out an ideally symmetrical word $E \epsilon \mathcal{D}$, in which a single set of all alphabetic characters circulates exactly d times. (For example, for $d = 3,\ k = 3$: $E = abcabcabc$.) In this case, two conclusions hold true:*

1. *Probability p_n is maximal in the case of overlaps absence, i.e. when word D lacks the shift symmetry.*
2. *Probability p_n is minimal in case of the D word's ideal symmetry, that is:*

$$p_n(E) \leq p_n(D). \tag{56}$$

4.2 Proof of the Theorem 3

Proposition 1 follows directly from the Theorem 2. In order to prove Proposition 2, let us consider an ideally symmetrical word E and word $\widetilde{D} \epsilon \mathcal{D}$ with a minimal period $T > k$. Let us show, that pair of words E and $\widetilde{D}$ satisfy Theorem 2 conditions. Let us denote number of overlaps in E and $\widetilde{D}$ as l_0 and l correspondingly. To begin with, let us establish relation between l_0 and l, required for the proof. Let us write words E and $\widetilde{D}$ down in the following way:

$$E = v_0 v_1 \ldots v_{l_0}, \tag{57}$$

$$\widetilde{D} = w_0 w_1 \ldots w_l. \tag{58}$$

In (57) all v_i words are identical and include one of each of the alphabetic elements. The number of nontrivial overlaps in the ideally symmetric word E is, apparently, $l_0 = d - 1$. In (58) we have $|w_0| = T > k$; all overlap positions in $\widetilde{D}$ correspond to the first character of the word w_i $(1 \leq i \leq l)$. For $\widetilde{D}$ words with equal number of all alphabetic elements and the minimal period $T > k$ the number of overlaps l shall not exceed l_0. It can be argued that:

$$\begin{cases} l \leq l_0, & d = 2, \\ l < l_0, & d > 2. \end{cases} \tag{59}$$

Indeed, if there is one full minimum period in $\widetilde{D}$, that is in the case $\lfloor m/T \rfloor = 1$ we have $l \leq \lfloor 0.5d \rfloor$, at that, if $d = 2$ we have $l \leq 1 = l_0$. Let the word $\widetilde{D}$ have more than one complete minimal period. Subsequently, we need to consider two cases. The first one is when all w_i words in $\widetilde{D}$ are identical. Then $m/T = u$, where u is a natural number and $u \geq 2$. In this case, $l = kd/T - 1 < l_0$. The second one is when m/T is not a natural number and $\lfloor m/T \rfloor \geq 2$. In this case the word $\widetilde{D}$ has two or more complete periods T, i.e. $w_0 = w_1$ at least. Then, because of $T > k$, one of the alphabetic elements, for

example, character α, occurs in the period w_0 at least twice. Consequently, in the word $w_1 \dots w_l$ character α occurs $d-2$ times at most. With the provision as above, number of overlaps in the word $\widetilde{D}$ cannot exceed $d-2 < l_0$.

Statement of the Theorem 3 for $d = 2$ is being verified directly, so we shall concentrate on the case $d > 2$. As above, we shall denote coordinates of overlaps in E by means $s_i + 1$, whereas coordinates of overlaps in $\widetilde{D}$ - by means $t_i + 1$. It would be sufficient for the proof to show, that for both E and $\widetilde{D}$ Theorem 2 statement had been fulfilled, i.e. for all $1 \le i \le l$:

$$s_i < t_i. \tag{60}$$

We shall write down the periods of words E and $\widetilde{D}$ correspondingly for $1 \le i \le l$ in the following way:

$$s_i = ik = m - (l_0 + 1 - i)k, \quad 1 \le i \le l, \tag{61}$$

$$t_i = \sum_{j=0}^{i-1} |w_j| = m - \sum_{j=i}^{l} |w_j|, \quad 1 \le i \le l. \tag{62}$$

From the results, presented in (Guibus & Odlyzko 1981), (Lotharie, 2001), it follows, that sequence of the w_i words' lengths does not increase:

$$|w_{i+1}| \le |w_i|. \tag{63}$$

If for all i we have $|w_i| > k$, then from (61), (62) it follows, that inequality (60) is satisfied. Let us assume now, that for $1 \le i \le i_1$ we have $|w_i| > k$, whereas for $i_1 < i \le l$ we have $|w_i| \le k$. Then, due to the fact, that $l < l_0$ holds at $d > 2$, for $i_1 < i \le l$ we get:

$$t_i = m - \sum_{j=i}^{l} |w_j| \ge m - k(l + 1 - i) > m - k(l_0 + 1 - i) = s_i. \tag{64}$$

In this way for all i condition (60) is satisfied and, consequently, Theorem 2 conditions had been fulfilled for words E and $\widetilde{D}$. Hence it follows, that for words E and $\widetilde{D}$ we have: $p_n(E) < p_n(\widetilde{D})$ or $p_n(E) \le p_n(D)$.

As an example, let us present three following words that illustrate conclusion of the Theorem 3.

$$E : 123123123123$$

$$\widetilde{D}_1 : 112233112233$$

$$\widetilde{D}_2 : 123123123132$$

In the above example, the word $\widetilde{D}_1$ has one overlap, whereas $\widetilde{D}_2$ has none. Calculations for $n = 40000$ give us: $p_n(E) = 0.0699$, $p_n(\widetilde{D}_1) = 0.0724$, $p_n(\widetilde{D}_2) = 0.0725$. Let us emphasize a certain point of interest. The ideally symmetrical word E and the word $\widetilde{D}_2$ differs only in rearrangement of the two last characters. Nevertheless, such rearrangement deprives the word $\widetilde{D}_2$ of the shift symmetry and, therefore, probability of its occurrence in a random string is the same as for any other word devoid of overlaps.

5. Conclusion

The main result of this work is establishment of the extremal properties of the word's at least once occurrence probability in a random string. The method used to derive the necessary formula for the probability under study can be generalized in the event of calculating a given word's occurrence frequency distribution in a random string. The corresponding recursion relations and further study of the discussed probabilities' properties have been obtained by the author and will be offered for publication in the nearest future.

In conclusion, let us cite a number of open questions stemming from the research undertaken above.

- It seems very likely that Theorem 2 conditions may be substantially weaken by replacing condition $s_i < t_i$ for all $1 \le i \le l$ by single requirement $s_1 < t_1$.
- Obviously, along with increase of the alphabet's size k , difference between $P_n(E)$ and $P_n(\widetilde{D})$ shrinks. It would be interesting to establish the relevant asymptotic dependence.
- It would be interesting to locate words that possess extreme detection probabilities in a random string, generated by the Markov process.

References

Chufang, W. (2005). The distribution of the frequency of occurrence of nucleotide subsequence. *Methodology and Computing in Applied Probability*, 325-334.

Gentleman, J. F., & Mullin, R. C. (1989). The distrribution of the frequency of occurrence of nucleotide subsequence, based on their overlap capability. *Biometrics*, *45,* 35-52. https://doi.org/10.2307/2532033

Gregory, N. (2008). Pattern Markov chains: Optimal Markov chain embedding through deterministic finite automata. *Journal of Applied Probability*, *45,* 226-243. https://doi.org/10.1239/jap/1208358964

Guibus, L. J., & Odlyzko, A. M. (1981). Periods in strings. *Journal of Combinatorial Theory*, *Serias A 30*, 19-42. https://doi.org/10.1016/0097-3165(81)90038-8

Gulden, I. P., & Jackson, D. M. (1983). *Combinatorial enumaration*. New York: Wiley.

Ilyevsky, V. I. (2014). Degree of order criteria of the elements' deterministic chains with relations between the closest neighbors. *British Journal of Mathematics & Computer Science*, *4*(19), 2752-2764. https://doi.org/10.9734/BJMCS/2014/11490

Ilyevsky, V. I. (2017). Order, chaos and symmetry in the deterministic chains of elements. *Theoratical Mathematics & Applications*, *7*, 27-49.

Lotharie. (2001). *Algebric combinatorics on words*. Cambridge University Press.

Lotharie. (2004). *Applied Combinatorics on words*. Cambridge University Press.

Regnier, M., & Szpankovsky, W. (1995). Frequency of pattern occurrence in a (DNA) sequence. *Purdue University, Department of Computer Science, Technical Reports*, *1227*, 1-11.

Regnier, M. (2000) A unified approach to word occurrence probabilities. *Discrete Applied Mathematics*, *104*, 259-280. https://doi.org/10.1016/S0166-218X(00)00195-5

Robin, S., & Daudin, J. J. (1999). Exact distribution of words occurrence in a ranndom sequence of letters. *Journal of Applied Probability*, *36*(1), 179-193. https://doi.org/10.1239/jap/1032374240

Robin, S., & Daudin, J. J. (2001). Exact dstribution of the distances between any occurances of a set of words. *Annals of the Institute of Statistical Mathematics*, *53*(4), 895-905. https://doi.org/10.1023/A:1014633825822

Copyrights