CDMTCS Research Report Series

On the Number of Occurrences of All Short Factors in Almost All Words

I. Tomescu
Bucharest University, Romania

CDMTCS-177
February 2002

Centre for Discrete Mathematics and
Theoretical Computer Science

On the Number of Occurrences of All Short Factors in Almost All Words

Ioan Tomescu
Faculty of Mathematics,
University of Bucharest,
Str. Academiei, 14,
R-70109 Bucharest, Romania
e-mail: ioan@math.math.unibuc.ro

Abstract

We previously proved that almost all words of length n over a finite alphabet A with m letters contain as factors all words of length $k(n)$ over A as $n \to \infty$, provided $\limsup_{n\to\infty} k(n)/\log n < 1/\log m$.
In this note it is shown that if this condition holds, then the number of occurrences of any word of length $k(n)$ as a factor into almost all words of length n is at least $s(n)$, where $lim_{n\to\infty} \log s(n)/\log n = 0$. In particular, this number of occurrences is bounded below by $C \log n$ as $n \to \infty$, for any absolute constant $C > 0$.

Keywords: Word; Factor; Occurrence; Random string

1 Notation and preliminary results

Let A be a finite alphabet of cardinality $|A| = m$. A word $b \in A^*$ is said to be a *factor* of $a \in A^*$ if there exist $p, q \in A^*$ such that $a = pbq$ [1]. A factor b of a word a can occur in a in different *positions*, each of those being uniquely determined by the length of the prefix of a preceding b. For example, abc occurs in $abcababc$ in positions 0 and 5. If $\alpha_1 \in A$, let $\alpha = \alpha_1 \dots \alpha_1 \in A^*$ be the word of length $|\alpha| = k \geq 1$ having all letters equal to α_1. Let $L(n)$ denote the number of words $a \in A^*$ such that $|a| = n$ and a does not contain the factor α. We need the following properties of the numbers $L(n)$ [2]:

Lemma 1.1 *We have*

$$L(n) \leq 8k(m - 1/m^k)^n$$

and the number of words $a \in A^*$ such that $|a| = n$ and a does not contain a fixed factor $\beta = \beta_1 \dots \beta_k$ of length k over A is less than or equal to $L(n)$.

From [2,3] we also deduce

Lemma 1.2 *If* $\limsup_{n\to\infty} k(n)/\log n < 1/\log m$, *then almost all words of length* n *over* A *contain as factors all words of length* $k(n)$ *over* A *as* $n \to \infty$.

Here the notion "almost all" has the following meaning: If $\mathcal{W}(n,k,A)$ denotes the set of words w of length n over A having the property that each word of length k over A is a factor of w, then $\lim_{n\to\infty} |\mathcal{W}(n,k,A)|/m^n = 1$ holds. Note that in [3] it is also shown that if $\lim_{n\to\infty} |\mathcal{W}(n,k,A)|/m^n = 1$ then $\limsup_{n\to\infty} k(n)/\log n \le 1/\log m$ holds.

If b is a factor of a, i.e., $a = pbq$ occurring in position $|p| = r$, $p = p_1 \dots p_r$, $q = q_1 \dots q_s$ and $b = b_1 \dots b_k$ ($|a| = r + k + s$), let
$u(a,b,|p|) = \{r + i - 1 : 2 \le i \le k \text{ and } b_i b_{i+1} \dots b_k q_1 \dots q_{i-1} = b\}$;
$l(a,b,|p|) = \{r - k + j : 1 \le j \le k - 1 \text{ and } p_{r-k+j+1} \dots p_r b_1 \dots b_j = b\}$
Note that $u(a,b,|p|)$ and $l(a,b,|p|)$ is the set of positions of the occurrences of b in a overlapping the occurrence of b in a with position $|p|$ and which are greater (resp. less) than $|p|$.
If $u(a,b,|p|) \ne \emptyset$ let $r + i_0 - 1 = max\, u(a,b,|p|)$ and denote

$$UW(a,b,|p|) = b_{i_0} b_{i_0+1} \dots b_k q_1 \dots q_{i_0-1}$$

the rightmost occurrence of b in a (having position $r+i_0-1$), that overlaps the occurrence of b in a with position $|p| = r$.

The occurrences of b in a appear in *blocks*, which are maximal factors of a consisting of overlapping occurrences of b in a.
A block B of occurrences of b in a ($|b| = k$) is a factor with a position r in a such that:
(i) $B = b$, $u(a,B,r) = l(a,B,r) = \emptyset$, or
(ii) $|B| \ge k + 1$; the prefix γ_1 of length k of B and the suffix $\gamma_t (t \ge 2)$ of length k of B satisfy $\gamma_1 = \gamma_t = b$, $l(a,\gamma_1,r) = u(a,\gamma_t, r + |B| - k) = \emptyset$; there exists a sequence of factors of B: $\gamma_2, \dots, \gamma_{t-1}$ having positions $r_2, \dots, r_{t-1}$ such that $\gamma_i = b$ for every $2 \le i \le t-1$ and $UW(a,\gamma_1,r) = \gamma_2$; $UW(a,\gamma_i,r_i) = \gamma_{i+1}$ for every $2 \le i \le t-1$.

Lemma 1.3 *If* $a \in A^*$ *contains at least one occurrence of* $b \in A^*$,*then*

$$a = A_1 B_1 A_2 B_2 \dots A_q B_q A_{q+1}, \tag{1}$$

where $q \ge 1$, $A_1, \dots, A_{q+1} \in A^*$ *do not contain occurrences of* b *and* $B_1, \dots, B_q$ *are blocks of occurrences of* b *in* a.

Proof: Consider an occurrence of b in a having the minimum position denoted by $l_1 \ge 0$. It follows that $a = A_1 b C$, where $|A_1| = l_1$ and $l(a,b,l_1) = \emptyset$. If we also have $u(a,b,l_1) = \emptyset$ then by denoting this occurrence by B_1 we get $a = A_1 B_1 C$ and apply the same argument to the word C if a has at least two occurrences of b; otherwise, by denoting $A_2 = C$ we get (1) for $q = 1$. If $u(a,b,l_1) \ne \emptyset$ we consider $UW(a,b,l_1)$ and so on by producing a sequence of occurrences of b in a having positions $l_1, \dots, l_m$ such that $UW(a,b,l_i)$ has position l_{i+1} for every $1 \le i \le m-1$ and $u(a,b,l_m) = \emptyset$. The factor of a with position l_1 and length $l_m - l_1 + |b|$ will be denoted by B_1 and it follows that B_1 is a block of occurrences of b in a satisfying (ii). We can write $a = A_1B_1C$. If the set of occurrences of b in a coincides with the set of occurrences of b in B_1, then by denoting $A_2 = C$ we obtain (1) for $q = 1$. Otherwise, by applying an inductive argument to C instead of a we get (1).

□

Let u be a word of length k in A^*, say $u = a_1 \dots a_k$ and $L_s(u,n)$ be the number of words $a \in A^*$ such that $|a| = n$ and the factor u of length k occurs exactly s times in a.
Our purpose is to evaluate the numbers $L_s(u,n)$. This will be done in the next section.

2 Main results

Lemma 2.1 *If n, k, s are positive integers, the following inequalities hold:*

$$L_s(u,n) < (n+k)^s L_0(u,n) \leq (n+k)^s L(n)$$

Proof: The inequality $L_0(u,n) \leq L(n)$ follows from Lemma 1.1. It remains to prove that

$$L_s(u,n) < (n+k)^s L_0(u,n) \tag{2}$$

Let $a \in A^*$ be a word such that $|a| = n$ and the factor u of length k occurs s times in a. Let B be the rightmost block of occurrences of u in a. Suppose that the position of B in a is r. We shall consider two subcases: I. $|B| = k$ and II. $|B| \geq k+1$.
I. If $|B| = k$, by deleting the factor B from a we get a word of length $n-k$ with $s-1$ occurrences of u.
II. If $|B| \geq k+1$, it is clear that $l(a,b,r+|B|-k) \neq \emptyset$. The suffix of length k of B is a factor equal to u and let

$$h = max\, l(a,b,r+|B|-k)$$

It follows that by deleting the factor $\delta = a_{h+k+1} \dots a_{r+|B|}$ from a (this factor is a suffix of B), we get a word of length $n-(r+|B|-h-k)$ having exactly $s-1$ occurrences of u. Since

$$r+|B|-2k+1 \leq h \leq r+|B|-k-1$$

it follows that $1 \leq r+|B|-h-k \leq k-1$, hence $1 \leq |\delta| \leq k-1$. If $s=1$ we can write

$$L_1(u,n) \leq (n-k+1)L_0(u,n-k) \leq nL_0(u,n) < (n+k)L_0(u,n)$$

because all words $a \in A^*$ of length n having a single occurrence of u can be generated by inserting (in $n-k+1$ ways) the factor u between consecutive letters in all words of length $n-k$ over A which do not contain any occurrence of u. Eventually, some words generated in this way contain more occurrences of u and the inequality between $L_1(u,n)$ and $(n-k+1)L_0(u,n-k)$ may be strict for some words u. Hence (2) is proved for $s=1$.

Now let $s \geq 2$. If the word $c = c_1 \dots c_{n-k} \in A^*$ contains $s-1$ occurrences of $u = a_1 \dots a_k$, let U be a block of occurrences of u in c with position r such that r is

maximum. It follows that the number of letters $c_{r+|U|}, c_{r+|U|+1}, \ldots, c_{n-k}$ occurring in c at the right of B is less than or equal to $n-k-(k+s-2) = n-2k-s+2$. Equality holds if and only if $a_1 = a_2 = \ldots = a_k$ and B is the unique block of occurrences of u in c, of length $k+s-2$, which is a prefix of c, i.e., $r = 0$.
Hence the number of ways of inserting the factor u of length k between consecutive letters at the right of the block B is at most equal to $n-2k-s+3$. In this way we produce at most $(n-2k-s+3)L_{s-1}(u, n-k)$ words of length n and this set of words contains (strictly for some words u) the set X of words $a \in A^*$ of length n containing the factor u s times and having the property that the block B of occurrences of u with maximum position has $|B| = k$. If this block B with maximum position has its length $|B| \geq k+1$, we have seen that there exists a suffix δ of B such that $1 \leq |\delta| \leq k-1$ and by deleting δ from a, a word of length $n-\delta$ with $s-1$ occurrences of u is produced. Because the suffix of length k of B is a word equal to u, it follows that the set Y of all words $a \in A^*$ of length $|a| = n$ containing s occurrences of u, with the property that the block B of occurrences of u with maximum position has $|B| \geq k+1$, can be generated by the following procedure:
For $i = 1, \ldots, k-1$, consider the set of words in A^* of length $n-i$ having $s-1$ occurrences of u. For each such word one inserts the factor $a_{k-i+1}a_{k-i+2}\ldots a_k$ at the right of the block of occurrences of u with the maximum position. In this way one generates at most

$$L_{s-1}(u, n-1) + L_{s-1}(u, n-2) + \ldots + L_{s-1}(u, n-k+1)$$

words. Of course, this set of words may contain some words which do not belong to Y.
It follows that for $s \geq 2$ we have: $L_s(u, n) = |X \cup Y| =$
$|X| + |Y| \leq (n-2k-s+3)L_{s-1}(u, n-k) + \sum_{i=1}^{k-1} L_{s-1}(u, n-i) \leq nL_{s-1}(u, n-k) + (k-1)L_{s-1}(u, n-1) < (n+k)L_{s-1}(u, n)$. Since $L_1(u, n) < (n+k)L_0(u, n)$ and $L_s(u, n) < (n+k)L_{s-1}(u, n)$ for every $s \geq 2$, (2) is proved.

□

This inequality can be used to estimate the number of words $a \in A^*$ with $|a| = n$ which contain at most $s-1$ occurrences of $u = a_1 \ldots a_k$.
Let $\mathcal{W}(n, k, s, A)$ denote the set of words w of length n over the alphabet A with m letters, having the property that each word of length $k(n)$ over A has at least $s(n)$ occurrences in w.

Theorem 2.2 *If the following two conditions are fulfilled:*
(i) $\limsup_{n\to\infty} k(n)/\log n < 1/\log m$*;*
(ii) $\lim_{n\to\infty} \log s(n)/\log n = 0$,
then $\lim_{n\to\infty} |\mathcal{W}(n, k, s, A)|/m^n = 1$*, i.e., almost all words of length* n *over* A *belong to* $\mathcal{W}(n, k, s, A)$.

Proof: For every $i \geq 0$ let $\mathcal{L}_u^i$ be the set of words of length n over A having exactly i occurrences of the word $u = a_1a_2\ldots a_k$. It follows that $|\mathcal{L}_u^i| = L_i(u, n)$ and $|\mathcal{W}(n, k, s, A)| = |\mathcal{W}(n, k, A)| - |\bigcup_{i=1}^{s-1} \bigcup_{u=a_1\ldots a_k} \mathcal{L}_u^i|$.
By Lemmas 1.1 and 2.1 we deduce
$|\bigcup_{i=1}^{s-1} \bigcup_{u=a_1\ldots a_k} \mathcal{L}_u^i| \leq \sum_{i=1}^{s-1} \sum_{u=a_1\ldots a_k} L_i(u, n) \leq m^k \sum_{i=1}^{s-1} L_i(u, n) \leq m^k \sum_{i=1}^{s-1}(n +$

$k)^i L(n) < m^k(n+k)^s L(n)$.
Since $L(n) \leq 8k(m-1/m^k)^n$ it follows that $\lim_{n\to\infty} m^k(n+k)^s L(n)/m^n = \lim_{n\to\infty}(n+k)^s L(n)/m^{n-k} = \lim_{n\to\infty} n^s L(n)(1+o(1))/m^{n-k}$, and
$\lim_{n\to\infty} n^s k(m-1/m^k)^n/m^{n-k} = e^{\lim_{n\to\infty} g(n)}$, where

$$g(n) = -n/m^{k+1} + k\ln\, m + s\ln\, n + \ln\, k < -n/m^{k+1} + s\ln\, n + 2k\ln\, m$$

Because (i) and (ii) hold, it follows that $\log\, n/m^{k+1} = \log\, n(1-(k+1)\log\, m/\log\, n) \to \infty$ as $n \to \infty$ because $\liminf_{n\to\infty}(1 - k\log\, m/\log\, n) = 1 - \limsup_{n\to\infty} k\log\, m/\log\, n > 0$; also
$\log\, km^{k+1}/n = \log\, k + (k+1)\log\, m - \log\, n \to -\infty$ and $\log\, m^{k+1}s\ln\, n/n = -\log\, n(1 - \log\, s/\log\, n - (k+1)\log\, m/\log\, n - \log\,\ln\, n/\log\, n) \to -\infty$ as $n \to \infty$.
Consequently,
$\lim_{n\to\infty} g(n) = -\infty$, which implies $\lim_{n\to\infty}(n+k)^s L(n)/m^{n-k} = e^{-\infty} = 0$. $\square$

Note that (ii) is verified if we take $s(n) = C\log\, n$, for any absolute constant $C > 0$.

Note

The paper will appear in *Theoret. Comput. Sci.*

References

[1] C. Choffrut, J. Karhumäki, *Combinatorics of Words*, Rapport LITP 97/12, Institut Blaise Pascal, Paris, 1997, 110p.

[2] I. Tomescu, On words containing all short subwords, *Theoretical Computer Science* 197(1998) 235-240.

[3] I. Tomescu, A threshold property concerning words containing all short factors, *Bulletin of the EATCS* no. 64(1998) 166-170.