On the Word Fragment Length for Unambiguous Reconstruction of a Periodic Word from a Complete Multiset of Fragments of Fixed Length

V. A. Alekseev and Yu. G. Smetanin

Presented by Academician of the RAS K.V. Rudakov November 18, 2021

Received March 23, 2021; revised January 27, 2022; accepted February 17, 2022

Abstract—We consider the problem of reconstructing a word from a multiset of its fragments of fixed length. Words consist of symbols from a finite alphabet. The word to be reconstructed is assumed to be periodic or contain a periodic word as a subword. It is shown that a periodic word with a period $p$ can be reconstructed from a multiset of its fragments of length $k$, where $k \geq \frac{16}{7} \sqrt{p} + 5$. For a word consisting of a $q$-periodic prefix repeated $m$ times and a $p$-periodic suffix repeated $l$ times, if $l \geq mq\frac{16}{7}\sqrt{p} + 5$, then the estimate becomes $k \geq \frac{16}{7} \sqrt{P} + 5$, where $P = \max(p, q)$.

Keywords: word, fragment, subword, periodic word, reconstruction from incomplete information, word reconstruction, multiset of subwords

DOI: 10.1134/S106456242202003X

In the general case, the problem of reconstructing an object of some nature from incomplete information on its “parts” can be treated as a pattern recognition problem [1]. Specifically, when objects are words in the form of sequences of symbols over an alphabet, the question arises as to whether a word can be reconstructed from a set or multiset of its subwords. This problem is concerned with an area of discrete mathematics known as combinatorics on words [2–6], which studies the relationship between sequences of symbols and sets of their subsequences.

The problem of reconstructing words from subwords has a number of applications. For example, in character encoding, arbitrary information is represented in the form of sequences of symbols over an alphabet, the question arises as to whether a word can be reconstructed from a set or multiset of its subwords. This problem is concerned with an area of discrete mathematics known as combinatorics on words [2–6], which studies the relationship between sequences of symbols and sets of their subsequences.

In this paper, on the words to be reconstructed, we impose the constraint that they contain a periodic subword. This makes it possible to reduce the fragment length sufficient for word reconstruction. For example, a result concerning the reconstruction of a word with a periodic suffix and an aperiodic prefix was presented in [17]. Below, this result of [17] is strengthened in Theorem 3, which deals with the reconstruction of a word with a periodic suffix and a periodic prefix. Theorems 1 and 2 concern the reconstruction of periodic words and repeat their counterparts from [17]. Additionally, some more cases of words that are not periodic, but contain a periodic word or subwords are considered in this paper. Specifically, we state and prove theorems on the reconstruction of a word with an aperiodic subword contained between periodic prefix and suffix (Theorem 4) and for a periodic word with a constraint on a repeated subword (Theorem 5).
In turn, each of these theorems is included in Theorem 3 as a special case.

1. FORMULATION OF THE PROBLEM

We introduce the following notation and definitions. In what follows, Greek letters denote words, and lowercase Latin letters denote alphabetic symbols. Let $E_n^2$ denote the set of words of length $n$ from the alphabet $\{0, 1\}$. For a word $\alpha = (a_1a_2 \ldots a_n) \in E_n^2$, the symbol $|\alpha|$ denotes the sum of its elements; $|\alpha| = a_1 + a_2 + \ldots + a_n$. Let $\lambda$ denote the empty word, i.e., the word of length zero. The word $\alpha$ raised to power $s$ denotes the word consisting of $s$ repeats of the word $\alpha$, i.e., $\alpha^s = \beta_1 \beta_2 \ldots \beta_s$, where $\beta_i = a_i, i \in \{1, 2, \ldots, s\}$.

Given a word $\alpha = (a_1a_2 \ldots a_n) \in E_n^2$ and a reference vector $v = (v_1v_2 \ldots v_n)$, where $v_i \in \{0, 1\}, i \in \{1, 2, \ldots, n\}$, the fragmentation operation $(\alpha \cdot v)$ constructs a word of length $|v|$ according to the following rule:

$$\langle \alpha \cdot v \rangle = \begin{cases} a_i, & v_i = 1 \\ \lambda, & v_i = 0. \end{cases}$$

A fragment, or subword, of a word $\alpha = (a_1a_2 \ldots a_n) \in E_n^2$ is a word $\widehat{\alpha} \in E_\Omega^2$ of the form

$$\widehat{\alpha} = (a_1a_i \ldots a_n), \quad 1 \leq i_1 < i_2 < \ldots < i_k \leq n.$$ 

Let $f^*(\alpha)$ be the smallest fragment length $k$ for which the word $\alpha$ of length $n$ is uniquely reconstructed from a multiset of its subwords of length $k$.

**Problem.** In the general form, the problem of reconstructing a word from a multiset of its subwords is formulated as follows. Given

- a set of reference vectors $V = \{v^1, v^2, \ldots, v^N\}, v^i \in E_n^2, |v^i| = k, i \in \{1, 2, \ldots, N\}$, and
- a set of words $X = \{\chi^1, \chi^2, \ldots, \chi^N\}, \chi^i \in E_k^2, i \in \{1, 2, \ldots, N\}$,

the task is to check whether $X$ is a set of fixed-length fragments of some unknown word $\alpha \in E_n^2$ constructed with the help of fragmentation operations with vectors from $V$ and to find all possible solutions.

Note that all estimates obtained in the case of the binary alphabet remain valid for an arbitrary alphabet [3], since, in the case of the alphabet $\{0, 1, \ldots, k\}$, the problem is reduced to a set of $k$ problems by the binary alphabet with the help of the mappings

$$\varphi_\beta : x \mapsto \begin{cases} 1, & x = i \\ 0, & x \neq i, \end{cases} \quad i \in \{0, 1, \ldots, k\}.$$ 

For this reason, in what follows, we consider only binary words $\alpha = (a_1a_2 \ldots a_n), a_i \in \{0, 1\}$. Moreover, we are interested in periodic words.

A word $\alpha = (a_1a_2 \ldots a_n), a_i \in \{0, 1\}$ is called periodic with a period $p$ if

$$\alpha = (a_1a_2 \ldots a_p)^p, \quad a_i \in \{0, 1\}. \quad (1)$$

Here, the subword $\widehat{\alpha} = a_1a_2 \ldots a_n$ is called the generating subword for a word $\alpha$ of form (1).

2. STATE OF THE ART IN THE PROBLEM

It is known that the problem of word reconstruction from subwords can be reduced to verifying the uniqueness of solutions to Diophantine equations of certain type [2]. It has been shown that the problem of existence and uniqueness of a solution is NP-complete.

In the case of a complete multiset of fragments ($V = E_n^2$), it was established [14] that a word $\alpha$ of length $n$ can be uniquely reconstructed if the fragment length satisfies $k \geq \exp(\Omega(\log n))$.

For the same case of a complete multiset of fragments, the following estimates for the sufficient fragment length are known [3, 13, 12]: $k \geq \left\lfloor \frac{n}{2} \right\rfloor, k \geq (1 + o(1))\sqrt{n} \log n, k \geq \left\lfloor \frac{16}{7} \sqrt{n} \right\rfloor + 5$.

Several special cases were studied. For example, it was shown in [15] that, for words consisting of $l$ series, it is sufficient to use fragments of length $l$.

3. RESULTS

The first theorem is as follows.

**Theorem 1.** A periodic word of length $n$ and period $p$ can be uniquely reconstructed from the multiset of all its fragments of length $k$ if

$$k \geq (1 + o(1))\sqrt{p \log p}.$$ 

Before proving the theorem, we introduce the following quantity.

**Definition 1.** For a word $\alpha$, let $N_\beta(\alpha)$ denote the number of fragments equal to $\beta$ in $\alpha$.

It was shown in [2] that, given $N_\beta(\alpha)$ for all binary words $\beta$ of length $k$, the numbers of fragments of all lengths shorter than $k$ can be uniquely reconstructed. Next, we formulate and prove a lemma that directly follows from the system of equations in [2].

**Lemma 1.** For any word $\alpha$, the set of its moments of the form $\sum_{i=1}^n a_i r_i$ is uniquely determined by the set of its fragments of the form $x^i 1$. 

DOKLADY MATHEMATICS  Vol. 105  No. 2  2022
Proof. Let us derive formulas for \( N_{x^{k+1}}(\alpha) \) for various \( k \):

\[
N_1(\alpha) = \sum_{r=1}^{n} a_r,
\]

\[
N_{x^k}(\alpha) = \sum_{r=1}^{n} (r - 1) a_r
\]

\[
= \sum_{r=1}^{n} a_r - N_1(\alpha) = \sum_{r=1}^{n} a_r - f_1(N_1(\alpha)),
\]

\[
\ldots
\]

\[
N_{x^{k+1}}(\alpha) = \sum_{r=1}^{n} \left( r - 1 \right) \frac{(k - 1)!}{(k - 1)!} a_r
\]

\[
- f_{k-1}(N_1(\alpha), N_{x^k}(\alpha), \ldots, N_{x^{k+1}}(\alpha)).
\]

The functions \( f_2, f_3, \ldots, f_{k-1} \) are computed using combinatorial relations. For example, we find an expression for \( f_p(N_1(\alpha), \ldots, N_{x^{k+1}}(\alpha)) \), \( 2 \leq p \leq k - 1 \), in terms of \( f_1, \ldots, f_{p-1} \) obtained at the preceding steps (here, it is possible to set \( f_0 \equiv 0 \)):

\[
N_{x^p}(\alpha) = \sum_{r=1}^{n} \left( p - 1 \right) \frac{(r - 1)!}{p!} a_r
\]

\[
= \sum_{r=1}^{n} (r - 1) \cdot (r - 2) \cdots (r - p) a_r
\]

\[
= \frac{1}{p!} \sum_{r=1}^{n} a_r - \sum_{i=1}^{p-1} \left( \begin{array}{l} r \end{array} \right) (i-1) \sum_{j=1}^{n} \left( \begin{array}{l} r \end{array} \right) (i-j) a_r
\]

\[
= \frac{1}{p!} \sum_{r=1}^{n} a_r - \sum_{i=1}^{p-1} \left( \begin{array}{l} r \end{array} \right) (i-1) \sum_{j=1}^{n} a_r (i-j)
\]

\[
= \frac{1}{p!} \sum_{r=1}^{n} a_r - \sum_{i=1}^{p-1} \left( \begin{array}{l} r \end{array} \right) (i-1) \sum_{j=1}^{n} a_r (i-j)
\]

The estimate of \( f^*(\alpha) \) in [13] was obtained by analyzing the system of equations

\[
\begin{cases}
\sum_{r=1}^{n} a_r r^j = s_j(\alpha) \\
0 \leq j \leq k - 1,
\end{cases}
\]

for which the solution is unique if \( f^*(\alpha) \leq (1 + o(1)) \sqrt{n \log n} \).

Suppose that the word \( \alpha \) consists of \( l = \frac{n}{p} \) periods:

\[ \alpha = (a_1, a_2, \ldots , a_p)^l. \]

Then the zero equation \( \sum_{r=1}^{n} a_r = s_0(\alpha) \) can be rewritten as \( \sum_{r=1}^{p} a_r = s_0(\alpha) \). For an arbitrary index \( k' \) ranging from 1 to \( k - 1 \) inclusive, we have

\[
\sum_{r=1}^{n} a_r r^{k'} = \sum_{r=1}^{p} a_r r^{k'} + \sum_{r=p+1}^{2p} a_r r^{k'} + \cdots + \sum_{r=(l-1)p+1}^{lp} a_r r^{k'}
\]

\[
= \sum_{r=1}^{p} a_r r^{k'} + \sum_{r=p+1}^{2p} a_r (p + 1)^k' + \cdots + \sum_{r=(l-1)p+1}^{lp} a_r ((l-1)p + r)^k'.
\]

where \( f_k' \) is a linear function. The rest of the proof is the same as in [13].

Let us formulate a stronger result.

Theorem 2. A periodic word of length \( n \) with period \( p \) is uniquely reconstructed from the multiset of all its fragments of length \( k \) if

\[ k \geq \left\lfloor \frac{16}{7} \sqrt{p} \right\rfloor + 5. \]

Proof. It is based on [12], where the possibility of unique reconstruction of a word from its subwords is proved not by analyzing system (2) used in [13], but rather by analyzing a similar system of the form

\[
\begin{cases}
\sum_{r=1}^{n} a_r r^j = s_j(\alpha) \\
0 \leq j \leq k - 1,
\end{cases}
\]
written for two words, \( \alpha = a_1 \ldots a_q \) and \( \beta = b_1 \ldots b_p \). It is proved that \( \alpha \) and \( \beta \) have identical multisets of subwords of length \( k \) if and only if system (3) has a non-trivial solution \((a_1, \ldots, a_q, b_1, \ldots, b_p)\). Thus, the proof is reduced to a search for a condition under which the system of Diophantine equations (3) has only the trivial solution. Next, the authors of [12] refer to [16], where the following result is obtained:

\[
\begin{align*}
    k &\geq \left\lfloor \frac{16}{7} \sqrt{m} \right\rfloor + 5. \\
\end{align*}
\]

For the case of periodic words, in that work, it is shown in the proof of Theorem 1 how system (3) can be rewritten in terms of only \( a_1, \ldots, a_q, b_1, \ldots, b_p \), where \( p \) is the word period. Thus, the original word can be uniquely reconstructed from subwords if their length satisfies

\[
\begin{align*}
    k &\geq \left\lfloor \frac{16}{7} \sqrt{p} \right\rfloor + 5. \\
\end{align*}
\]

Word reconstruction is also possible when the word itself is aperiodic, but consists of several subwords, at least one of which is periodic. Below, we consider several cases of such words. For each case, we state a separate theorem, which is proved relying on what was proved previously for a periodic word (Theorem 2).

**Theorem 3.** Suppose that a word \( \alpha \) consists of two subwords: a prefix and a suffix, which are both periodic:

\[
\begin{align*}
    \alpha &= (a_1a_2 \ldots a_q)^m (a_{q+1}a_{q+2} \ldots a_{q+p})', \\
\end{align*}
\]  

Then, for \( l \geq mq \left\lfloor \frac{16}{7} \sqrt{P} \right\rfloor + 5 \), where \( P = \max (p, q) \), word (4) can be uniquely reconstructed from fragments of length \( k \) if \( k \geq \left\lfloor \frac{16}{7} \sqrt{P} \right\rfloor + 5 \).

**Proof.** Let us find \( s_0(\alpha) \):

\[
\sum_{r=1}^{n} a_r = m \sum_{r=1}^{q} a_r + l \sum_{r=q+1}^{q+p} a_r = s_0(\alpha). 
\]

For \( mq < l \), we obtain \( \sum_{r=1}^{q+p} a_r < 1 \), so

\[
\begin{align*}
    \sum_{r=q+1}^{q+p} a_r &= s_0(\alpha) \left\lfloor \frac{l}{m} \right\rfloor, \\
    \sum_{r=1}^{q} a_r &= \frac{s_0(\alpha)}{l} \cdot \frac{l}{m}. \\
\end{align*}
\]

Consider the case \( s_1(\alpha) \):

\[
\begin{align*}
    s_1(\alpha) &= \sum_{r=1}^{q} a_r + \sum_{r=q+1}^{q+p} a_r. \\
\end{align*}
\]

The first sum in the expression for \( s_1(\alpha) \) (5) can be represented as follows:

\[
\begin{align*}
    \sum_{r=1}^{q} a_r r &= a_1 \cdot (qj + 1) + \ldots + a_q \cdot (qj + q) \\
    &= \frac{q}{r} \sum_{r=1}^{q} a_r \left( \frac{mr + q(m-1)}{2} \right) = \frac{q}{r} \sum_{r=1}^{q} a_r \left( \frac{q(m-1)}{2} \right) \\
\end{align*}
\]

Similarly, for the second sum in (5), we obtain

\[
\begin{align*}
    \sum_{r=q+1}^{q+p} a_r r &= a_{q+1} \cdot (qm + p) + \ldots + a_{q+p} \cdot (qm + p + p) \\
    &= \sum_{r=q+1}^{q+p} a_r \left( \frac{mirror + pl(l-1)}{2} \right) \\
    &= \frac{1}{l} \sum_{r=q+1}^{q+p} a_r r + \left( \frac{qml + pl(l-1)}{2} \right) \sum_{r=q+1}^{q+p} a_r. \\
\end{align*}
\]

Finally,

\[
s_1(\alpha) = m \sum_{r=1}^{q} a_r + l \sum_{r=q+1}^{q+p} a_r r + f_1(s_0). 
\]

If \( mq(q + 1)/2 < l \), then the equations for the prefix and the suffix split, as in the case with \( s_0(\alpha) \). Now, we consider the case of an arbitrary \( k^* \) such that \( 2 \leq k^* \leq k - 1 \):

\[
s_k^*(\alpha) = \sum_{r=1}^{q} a_r r^{k^*} + \sum_{r=q+1}^{q+p} a_r r^{k^*}. 
\]

The first sum is expanded as

\[
\begin{align*}
    \sum_{r=1}^{q} a_r r^{k^*} &= \sum_{r=1}^{q} a_r \left( a_1a_2 \ldots a_q \right) \left( r + q(m-1) + r \right)^{k^*} \\
    &= \sum_{r=1}^{q} a_r \sum_{h=1}^{m} \left( r + q(m-h) \right)^{k^*} \\
    &= \sum_{r=1}^{q} a_r \sum_{j=0}^{m} \left( k^* \right)^j \left( q(m-h) \right)^{k^* - j} \\
    &= m \sum_{r=1}^{q} a_r r^{k^*} + g_k \left( s_{k^* - 1}, \ldots, s_0 \right). \\
\end{align*}
\]

In the last transition, the term with \( r \) raised to power \( k^* \) (for \( j = k^* \)) was separated from the others. The terms with \( r \) raised to lower powers were represented in the form of a linear function \( g_k \).
Similarly, for the second sum, we have
\[
\sum_{r=qn+1}^{qm} a_r r^{k'} = \sum_{r=qn+1}^{qm} a_r ((q(m-1)+r)^{k'} + \ldots + (q(m-1)+p(l-1)+r)^{k'})
\]
\[
+ \sum_{r=qn+1}^{qm} a_r \sum_{j=0}^{l-1} \binom{k'}{j} (q(m-1)+r)^{(p(l-1)+r)^{k'-j}}
\]
\[
= \sum_{r=qn+1}^{qm} a_r \sum_{j=0}^{l} \sum_{s=0}^{k'} \binom{k'}{j} (q(m-1)+r)^{(p(l-1)+r)^{k'-j}}
\]
\[
= l \sum_{r=qn+1}^{qm} a_r r^{k'} + h_k^*(s_{k-1}, \ldots, s_0).
\]
Here, in the last transition, the term with \( r \) raised to power \( k' \) (for \( j = k' \) and \( s = 0 \)) was again separated from the others. The terms with \( r \) raised to lower powers were represented in the form of a linear function \( h_k^* \).

Thus, combining the results obtained for both sums, we can return to the computation of \( s_k^* (\alpha) \):
\[
s_k^* (\alpha) = \sum_{r=1}^{qm} a_r r^{k'} + \sum_{r=qn+1}^{qm} a_r r^{k'}
\]
\[
= \left( m \sum_{r=1}^{q} a_r r^{k'} + g_k^* (s_{k-1}, \ldots, s_0) \right)
\]
\[
+ l \sum_{r=qn+1}^{qm} a_r r^{k'} + h_k^*(s_{k-1}, \ldots, s_0)
\]
\[
= m \sum_{r=1}^{q} a_r r^{k'} + l \sum_{r=qn+1}^{qm} a_r r^{k'} + f_k^* (s_{k-1}, \ldots, s_0),
\]
where \( f_k^* \) denotes a linear function of \( s_{k-1}, \ldots, s_0 \) that is the sum of \( g_k^* \) and \( h_k^* \).

Since \( \sum_{k=1}^{k'} n_k^{p+1} \) \( \approx \frac{n^{p+1}}{p+1} \), it follows that, for \( m \cdot q \cdot k' < l \), all the equations split, and we obtain two systems. Therefore, according to what was proved earlier for the case of a periodic word (Theorem 2), for \( l \geq \frac{16}{7} \sqrt{P} + 5 \), where \( P \equiv \max (p, q) \), the subword length has to satisfy \( k \geq \frac{16}{7} \sqrt{P} + 5 \).

**Corollary 1.** For \( m = 1 \), \( \alpha \) has the form of a word with an aperiodic prefix and a periodic suffix:
\[
\alpha = a_1 a_2 \ldots a_q a_{q+1} a_{q+2} \ldots a_{q+m}.
\]
If the suffix is longer than the prefix, namely, \( l \geq q \frac{16}{7} \sqrt{P} + 5 \), then the word \( \alpha \) can be uniquely reconstructed from subwords of length \( k \geq \frac{16}{7} \sqrt{P} + 5 \) where \( P \equiv \max (p, q) \).

**Theorem 4.** Suppose that a word \( \alpha \) consists of three subwords: a periodic prefix \( \overline{\alpha} \), a periodic suffix \( \hat{\alpha} \), and an aperiodic root \( \alpha \):
\[
\alpha = \overline{\alpha} \hat{\alpha} \alpha \equiv (a_1 a_2 \ldots a_q m a_{q+1} a_{q+2} \ldots a_{q+m})^m 
\times a_{q+m} (a_{q+m+1} a_{q+m+2} \ldots a_{q+m+p})^p.
\]

Then, for \( m \geq s \cdot k' \) and \( l \geq (m-1) q \cdot k' + (q+s)^{k'} \), where \( k' = \frac{16}{7} \sqrt{P} + 5 \) and \( P \equiv \max (p, q, s) \), for word (6) to be uniquely reconstructed, it is sufficient that \( k \geq k^* \).

**Proof.** The proof is based on Theorems 2 and 3.

Consider \( s_0 (\alpha) \):
\[
s_0 (\alpha) = \sum_{r=1}^{qm} a_r + m \sum_{r=1}^{q} a_r + \sum_{r=qn+1}^{qm} a_r + l \sum_{r=qn+1}^{qm} a_r.
\]

Then, for \( mq + s < l \), we can split the equations for two subwords: \( \overline{\alpha} \hat{\alpha} \) and \( \hat{\alpha} \).

Consider \( k' \) such that \( 1 \leq k' \leq k-1 \). Omitting the detailed calculations, which are similar to those presented in the proof of Theorem 3, we obtain
\[
s_k (\alpha) = m \sum_{r=1}^{q} a_r r^{k'} + q s \sum_{r=qn+1}^{qm} a_r r^{k'}
\]
\[
+ q s \sum_{r=qn+1}^{qm} a_r r^{k'} + f_k^* (s_{k-1}, \ldots, s_0),
\]
where \( f_k^* \) is a linear function. Thus, under the condition \( l \geq mq^{k'} + (q+s)^{k'} - q^{k'} \), where \( k' = \frac{16}{7} \sqrt{P} + 5 \) and \( P = \max (m, q) \), the equations for the words \( \overline{\alpha} \hat{\alpha} \) and \( \hat{\alpha} \) can be split, so the whole word can be reconstructed under the subword length constraint \( k \geq k^* \).

However, the problem of reconstructing the word \( \overline{\alpha} \hat{\alpha} \) can be considered separately. Indeed, by Corollary 1, to reconstruct the word \( \overline{\alpha} \hat{\alpha} \) from a set of its subwords, it is sufficient that \( m \geq s \cdot k^* \), where \( k^* = \frac{16}{7} \sqrt{P} + 5 \) and \( P_1 = \max (q, s) \). Then the length of the subwords is constrained as \( k \geq k^* \).

Combining the conditions with \( k_1^* \) and \( k_2^* \), we conclude that the word \( \alpha \) can be reconstructed from subwords for \( m \geq s \cdot k^* \) and \( l \geq (m-1) q \cdot k^* + (q+s)^{k^*} \), where \( k^* = \frac{16}{7} \sqrt{P} + 5 \) and \( P \equiv \max (p, q, s) \). Here, the subword length satisfies \( k \geq k^* \).
Remark. The root length \( s = 0 \) gives the case considered in Theorem 3.

Corollary 2. The root length \( s = 1 \) is equivalent to a single symbol placed between two periodic subwords in the word \( \alpha \). In this case, \( \alpha \) can be reconstructed from subwords if \( l \geq (m - 1)q^k + (q + 1)^k \), where \( k^* = \left\lfloor \frac{16}{7} \sqrt{P} \right\rfloor + 5 \) and \( P \equiv \max(p,q) \). The subword length \( k \) is constrained by the relation \( k \leq k^* \).

Corollary 3. Suppose that a word \( \alpha \) of the form (6) is such that the generating subwords in the prefix and the suffix are identical:

\[
\alpha = (a_1a_2 \ldots a_q)^m(a_{q+1} \ldots a_{q+s}(a_{q+s+1} \ldots a_{q+s+p}))',
\]

\( q = p, \ a_t = a_j, \quad (i, j) \in \{(1, q + s + 1), \ldots, (q, q + s + p)\} \).

Assume that \( k^* = \left\lfloor \frac{16}{7} \sqrt{P} \right\rfloor + 5 \) and \( P \equiv \max(q,s) \). Then, for \( l + m \geq s^k - q^k \), the word \( \alpha \) can be uniquely reconstructed if \( k \geq k^* \).

Proof. For \( \alpha \) having the indicated structure, Eq. (7) becomes

\[
s_k(\alpha) = (m + l) \sum_{r=1}^q a_r r^k + \sum_{r=q+1}^{q+p} a_r r^k + f_k(s_{k-1}, \ldots, s_0).
\]

Therefore, for \( m + l \geq (q + s) k^*-q^k \), the equations for the generating subword of the prefix (suffix) and the root can be split.

Theorem 5. Suppose that a word \( \alpha \) has the form

\[
\alpha = ((a_1a_2 \ldots a_q)^m(a_{q+1} \ldots a_{q+s}(a_{q+s+1} \ldots a_{q+s+p}))')^N,
\]

i.e., it consists of repeats of a subword, which, in turn, consists of a periodic prefix and a suffix. Then, for \( l \geq mq^k \), where \( k^* = \left\lfloor \frac{16}{7} \sqrt{P} \right\rfloor + 5 \) and \( P \equiv \max(p,q) \), word (8) can be uniquely reconstructed if \( k \geq k^* \).

Proof. We begin with computing \( s_0(\alpha) \):

\[
s_0(\alpha) = \sum_{r=1}^{N(qm+pl)} a_r = Nm \sum_{r=1}^q a_r + Ni \sum_{r=q+1}^{q+p} a_r.
\]

For \( m < l \), the equations split:

\[
\begin{align*}
\sum_{r=q+1}^{q+p} a_r &= \left\lfloor \frac{s_0(\alpha)}{Nl} \right\rfloor, \\
\sum_{r=1}^q a_r &= \left\lfloor \frac{s_0(\alpha)}{Nl} \right\rfloor \cdot \frac{l}{m}.
\end{align*}
\]

Introducing the notation \( \bar{n} \equiv qm + pl \), we express \( s_k(\alpha) \) for \( k^* \) such that \( 1 \leq k^* \leq k - 1 \):

\[
s_k(\alpha) = \sum_{r=1}^{N\bar{n}} a_r r^k = \sum_{j=0}^{N-1} \left( \sum_{i=0}^\infty \sum_{r=iq+j} a_{r-nj-q} e^{kr} \right) + \sum_{i=0}^{l-1} \left( \sum_{r=iq+j+pi} a_{r-nj-qi+p} r^k \right)
\]

Thus, for \( m q^k < 1 \), all equations split and we obtain two systems. By Theorem 2 for a periodic word, for \( l \geq mq^k \), where \( k^* = \left\lfloor \frac{16}{7} \sqrt{P} \right\rfloor + 5 \) and \( P \equiv \max(p,q) \), the word \( \alpha \) can be reconstructed if \( k \geq k^* \).

Remark. If \( N = 1 \), we obtain the result proved in Theorem 3 for a word consisting, in fact, of a single repeat of a subword with a periodic suffix and prefix.

4. CONCLUSIONS

It was found that a periodic word of length \( n \) with period \( p \) can be uniquely reconstructed from a multiset of its subwords of fixed length \( k \) if \( k \geq \left\lfloor \frac{16}{7} \sqrt{P} \right\rfloor + 5 \). The subword length \( k \) was also estimated for words that are not periodic, but contain periodic subwords. Specifically, for words consisting of a periodic prefix and a periodic suffix, the possibility of unique reconstruction was proved for a sufficient suffix length. It was shown that words consisting of a periodic prefix, a periodic suffix, and an aperiodic root can be uniquely reconstructed from a multiset of their fragments of length \( k \) under two conditions: on the one hand, if the root is small as compared with the prefix and, on the other hand, if the prefix is small as compared with the suffix. An estimate for \( k \) was proved in the case of a periodic word generated by a subword consisting of two parts: a periodic suffix and a prefix.

FUNDING

This work was supported by the Russian Foundation for Basic Research, project no. 19-07-00150.
ON THE WORD FRAGMENT LENGTH FOR UNAMBIGUOUS RECONSTRUCTION

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

1. Yu. I. Zhuravlev, “On the algebraic approach to recognition and classification problems,” Probl. Kibern. 33, 5–68 (1978).
2. V. K. Leont’ev and Yu. G. Smetanin, “On the reconstruction of vectors from a set of their fragments,” Dokl. Akad. Nauk SSSR 302 (6), 1319–1322 (1988).
3. B. Manvel et al., “Reconstruction of sequences,” Discrete Math. 94 (3), 209–219 (1991).
4. V. K. Leont’ev, “Reconstruction of binary words from their fragments,” Dokl. Akad. Nauk 330 (4), 434–436 (1993).
5. M. Lothaire, Combinatorics on Words (Cambridge Univ. Press, Cambridge, 1997), Vol. 17.
6. V. I. Levenshtein, “Efficient reconstruction of sequences from their subsequences or supersequences,” J. Combinatorial Theory Ser. A 93 (2), 310–332 (2001).
7. J. Lin et al., “Experiencing SAX: A novel symbolic representation of time series,” Data Mining Knowl. Discovery 15 (2), 107–144 (2007).
8. T. Batu et al., “Reconstructing strings from random traces,” Departmental Papers (CIS) (2004), p. 173.
9. S. Kannan and A. McGregor, “More on reconstructing strings from random traces: Insertions and deletions,” Proceedings of International Symposium on Information Theory 2005 (ISIT 2005) (IEEE, 2005), pp. 297–301.
10. Yu. Smetanin and M. Ul’yanov, “Determining the characteristics of Kolmogorov complexity of time series: An approach based on symbolic descriptions” (2013).
11. D. Gusfield, “Algorithms on stings, trees, and sequences: Computer science and computational biology,” ACM SIGACT News 28 (4), 41–60 (1997).
12. I. Krasikov and Ye. Roditty, “On a reconstruction problem for sequences,” J. Combinatorial Theory Ser. A 77 (2), 344–348 (1997).
13. A. D. Scott, “Reconstructing sequences,” Discrete Math. 175 (1–3), 231–238 (1997).
14. M. Dudik and L. J. Schulman, “Reconstruction from subsequences,” J. Combinatorial Theory Ser. A 103 (2), 337–348 (2003).
15. V. K. Leont’ev, “Problems of word reconstruction from fragments and their applications,” Diskret. Anal. Issled. Oper. 2 (2), 26–48 (1995).
16. P. Borwein, T. Erdélyi, and G. Kós, “Littlewood-type problems on [0, 1],” Proc. London Math. Soc. 79 (1), 22–46 (1999).
17. V. A. Alekseev and Yu. G. Smetanin, “On the possibility of periodic word reconstruction from subwords of fixed length,” Chebyshev. Sb. 22 (1), 57–66 (2021).

Translated by I. Ruzanova