6th International Conference on Applied Informatics
Eger, Hungary, January 27–31, 2004.

Reconstruction from subwords*

Péter Ligeti[a], Péter Sziklai[b]

[a]Department of Computer Science, Eötvös Loránd University
e-mail: turul@cs.elte.hu

[b]Department of Computer Science, Eötvös Loránd University
e-mail: sziklai@cs.elte.hu

Abstract

In the paper two variants of a combinatorial problem for the set F_q^n of sequences of length n over the alphabet $F_q = \{0, 1, .., q-1\}$ are considered, with some applications. The original problem was the following: for a given word $w \in F_q^n$, what is the smallest integer k such that we can reconstruct w if we know all of its subwords of length k. This problem was solved by Lothaire [8] .

We consider the following variant of this problem: the n-letter word $w = w_1...w_n$ (which is called a *DNA-word*) is composed over an alphabet consisting of q *complement pairs*:$\{i, \bar{i} : i = 0, .., q-1\}$; and denote by w^* its *reverse complement*, i.e. $w^* = \bar{w}_n...\bar{w}_1$. A DNA-word u is called a subword of w if it is a subword of either w or w^*. (Another formulation is that we identify w and w^*.) We want to reconstruct w from its subwords of length k. We give a simple proof for $k = n-1$, and apply this result for determining the automorphism group of the poset of DNA-words of length at most n, partially ordered by the above subword relation.

Key Words and Phrases: poset, reconstruction, subword

1. Introduction

Consider the q-element alphabet $F_q = \{0, 1, ..., q-1\}$. In the paper we examine the elements of the set F_q^n of sequences of length n called *words*. A subsequence u of a given word w called *subword*, with notation $u \subseteq w$. Consider a given word $w \in F_q^n$.

Definition 1.1. Let $s_k(w) = u \in F_q^k : u \subseteq w$, the multiset of all of the $\binom{n}{k}$ subwords of w of length k.

*Supported by the ETIK Grant, OTKA F-043772, T-043758 and Magyary Z. Grant

Definition 1.2. Let $s_k^*(w) = u \in F_q^k : u \subseteq w$, the set of all of the different subwords of w, of length k.

In other words the set $s_k^*(w)$ is the set $s_k(w)$ without multiplicities. Here is a simple example to show the difference between s_k and s_k^*.

Example 1.1. Let $w = 00011$. Then $s_4(w) = \{0011, 0011, 0011, 0001, 0001\}$ and $s_4^*(w) = \{0011, 0001\}$

There is two type of the *reconstruction problem*: for a given word w of length n what is the smallest k, such that we can reconstruct w from the set $s_k(w)$ or from the set $s_k^*(w)$. In Section 2 we give a short overview concerning the known results of this problem.

It is relatively easy to prove (see [3]), that $s_{n-1}^*(w)$ is enough for the reconstruction. Using this result, Erdős, Sziklai and Torney [3] determined the automorphism group of the partially ordered set (or poset) containing all words of length at most n over a q-letter alphabet. Similarly, we consider two other posets and determine their automorphism groups.

Let $u_{m,n}$ denote the word $a_1...a_n$ where $m \geq 2$, $a_1 = 0$ and $a_{i+1} \equiv a_i + 1 \bmod (m)$, ie. for $n \equiv l \bmod (m)$

$$u_{m,n} = 012...(m-1)012...(m-1)...012...(l-1)$$

furthermore, let $B^{m,n}$ denote the set of all subsequences of $u_{m,n}$ partially ordered by the subsequence relation. This is a notable word: among the n-long words over the m-element alphabet, $u_{m,n}$ has maximum number of subwords. Burosch et al. [2] determined $\mathrm{Aut}(B^{m,n})$ by algebraic way, in Section 3 using the result of [3] we give a significantly shorter proof of the theorem of Burosch et al.

In Section 4 we consider the well known *DNA-words* and we define a new type of the reconstruction problem. Let $\Gamma = \{i, \bar{i} : i = 0, 1, .., q-1\}$ be an alphabet of q pairs of symbols (called *complement pairs*); and denote by Γ^n the set of all sequences of length n over the alphabet Γ. The elements of Γ^n are called *DNA-words*. Define $\bar{\bar{i}} = i$ for all i and for a word $w = w_1w_2...w_n \in \Gamma^n$ let $w^* = \bar{w}_n...\bar{w}_1$ be the *reverse complement* of w. Note that $(w^*)^* = w$.

Denote $g \prec f$ if g is a subword of either f or $\widetilde{f}$ (we will see that this is the good formalization because of the reverse complementarity). Let $d_m^*(f)$ denote the set of all words g of length at most m which $g \prec f$. The *DNA reconstruction* problem is the following: for a given DNA-word f of length n what is the smallest m such that we can reconstruct f from the set $d_m^*(f)$? We prove in a simple way that $d_{n-1}^*(f)$ is enough to the reconstruction.

Let $D^{q,n}$ denote the poset of all DNA-words of length at most n over an alphabet of q complement pairs, partially ordered by the $\prec$ relation. As an application of the previous results we determine $\mathrm{Aut}(D^{q,n})$.

2. Known results

2.1. Reconstruction from the multiset $s_k(w)$

The original problem was first considered by Kalashnik in 1973: for an arbitrary word w of length n, what is the smallest k such that we can reconstruct w from $s_k(w)$, i.e. from the multiset of its $\binom{n}{k}$ subwords of length k?

An upper bound for k was find independently by Leon'tev and Smetanin [5] and Manvel et al [9]. Furthermore in [9] the authors find a lower bound too:

Theorem 2.1. *We can reconstruct w from $s_k(w)$ for $k \geq \frac{n}{2}$ and for $k < \log_2 n$ we can not.*

In these papers the authors use some simple combinatorial ideas and show lot of examples.

Later Krasikov and Roditty [4] found an essentially better upper bound in the following way: they proved that if for $w \neq v$ $s_k(w) = s_k(v)$, then for some s the system

$$a_1^h + a_2^h + .. + a_s^h = b_1^h + b_2^h + .. + b_s^h, h = 1, .., k-1$$

$$a_1 < a_2 < .. < a_s, b_1 < b_2 < .. < b_s$$

has a nontrivial solution with $a_i, b_i \in [0, n-1]$.

This is the Prouchet-Tarry-Escott problem of classical Diophantine analisys. Recently Borwein, Erdélyi and Kós [1] proved that this system has only trivial solutions whenever $k \geq \lfloor \frac{16}{7}\sqrt{n} \rfloor$. This yields the following:

Theorem 2.2. *We can reconstruct w from $s_k(w)$ for $k \geq \lfloor \frac{16}{7}\sqrt{n} \rfloor$.*

However the precise upper bound for k is still an open problem.

2.2. Reconstruction from the multiset $s_k^*(w)$

The second type of the reconstruction problem is the following: for a given word w of length n what is the smallest k such that we can reconstruct w from $s_k^*(w)$, i.e. from the set of its *different* subwords of length k. The following result was proved independently by Levenshtein [6] and Lothaire [8]:

Theorem 2.3. *We can reconstruct w from $s_k^*(w)$ for $k \geq \lfloor \frac{n-1}{2} \rfloor$.*

Contrast with the previous problem this result is sharp:

Example 2.1. Consider the periodic words $u = 0101..01$ and $v = 1010..10$ of length $2n$. It is easy to see, that

$$s_n^*(u) = s_n^*(v) = F_2^n$$

ie. all the binary words of length n.

In our proofs we use the following very weak version of Theorem 2.3 proved by Erdős, Sziklai and Torney [3] by constructive way:

Lemma 2.1. *If $3 \leq n$ then every word w of length n is uniquely determined by $s^*_{n-1}(w)$ i.e. its $(n-1)$-subwords.*

3. A short proof of the theorem of Burosch et al.

Before the theorem let's see some remarks. It is clear, that the levels of the poset are invariant under an automorphism. Also homogeneity (i.e. all letters of the word are the same) and total inhomogeneity (i.e. all the letters of the word are different) are kept by every automorphism.

The basic idea of our proof is the following: we consider the action of an arbitrary automorphism on the first two levels of the poset. If an automorphism fixes these levels, then inductively, because of Lemma 2.1, it is the identity on the whole poset. Then it is enough to examine the automorphisms on the letters and on the two-letter subwords. The theorem was first proved by Burosch et al [2]:

Theorem 3.1. *(i) if $1 \leq n \leq m$, then $\mathrm{Aut}(B^{m,n}) = Sym_n$;*
(ii) if $m+1 \leq n \leq 2m-1$, then $\mathrm{Aut}(B^{m,n}) = Z_2 \otimes Sym_{2m-n}$;
(iii) if $2m \leq n$, then $\mathrm{Aut}(B^{m,n}) = Z_2$.

We give here a short presentation of our train of thought, for proofs and more see Ligeti and Sziklai [7].
(i) Now $u_{m,n} = 012..(n-1)$, i.e. it is a total inhomogeneous word. Take an arbitrary automorphism $\sigma_0 \in \mathrm{Aut}(B^{m,n})$, and consider its action on the first level of the poset. Thus, this is a permutation π on $\{0,1,2,..,n-1\}$, take its inverse π^{-1}. This permutation induces an automorphism $\sigma_{\pi^{-1}}$ on the poset. Let $\sigma_1 = \sigma_0\sigma_{\pi^{-1}}$. Then σ_1 fixes all of the letters. Furthermore, σ_1 fixes all sequences of form ij where $i<j$ because $\sigma_1(ij) \neq (ji)$ as ji is not a subword of $u_{m,n}$. Then σ_1 is the identity on the two lowest levels of the poset and, by Lemma 2.1, on the whole poset.
(ii) In this case $u_{m,n} = 01...(m-1)01...(k-1)$ where $n = m+k$, $1 \leq k \leq m-1$ and let σ_0 be an arbitrary automorphism. We prove that we have strong restrictions for the images of the letters $0,1,...,k-1$, but we are free to choose the images of the remaining $2m-n$ letters (this yields the factor Sym_{2m-n}).

Remark 3.1. Let e be an element of the third level of the poset such that e contains the letters i,j only and suppose that ii is a subword of e. Then we can read from the poset whether j is the middle letter or not.

In that case $e = iij, jii$, or iji. The first two words have two subwords of length two, but the third word has three.

Remark 3.2. Let $j_1 < j_2 \leq k-1$ and $i \leq k-1$, $i \neq j_1, j_2$, then we can tell the difference between the j_1iij_2-type subwords and the j_1j_2ii-type or iij_1j_2-type subwords in the poset.

From these remarks we get the following:

Lemma 3.1. *For $i = 0, 1, 2, .., k-1$ the image of the letter i is i or $(k-i-1)$ by any automorphism.*

Now we define a mapping ρ: given a word $w = x_1x_2...x_sy_1y_2...y_tz_1z_2...z_u$, where $0 \leq x_i, z_i \leq k-1; k \leq y_i \leq m-1$; let

$$\rho(w) = z_uz_{u-1}...z_1y_1y_2...y_tx_sx_{s-1}...x_1.$$

Let ν be the mapping that changes all the letters i $(0 \leq i \leq k-1)$ for $k-1-i$ in each word (and does not changes the letters j for $k \leq j \leq m-1$). Clearly neither ρ nor ν is an automorphism but $\rho\nu$ is an involution in $\mathrm{Aut}(B^{m,n})$.

Now let σ_0 be an arbitrary automorphism, and consider its action on the letters $k, ..., m-1$, this induces a permutation π on these letters (still on the first level), take its inverse π^{-1}. This permutation induces an automorphism $\sigma_{\pi^{-1}}$ on the poset. Let $\sigma_1 = \sigma_0\sigma_{\pi^{-1}}$. Then σ_1 is the identity on the letters $k, ..., m-1$ and, as above, σ_1 fixes all sequences of form ij where $k \leq i < j$. Finally, if $\sigma_1(0) = (k-1)$ then let $\sigma = \rho\nu\sigma_1$ where and if $\sigma_1(0) = 0$ then let $\sigma = \sigma_1$. Hence $\sigma(0) = 0$.

Lemma 3.2. *If an automorphism fixes 0 then it fixes the two lowest levels of the poset.*

Now by Lemma 2.1 and Lemma 3.2 we get the part **(ii)** of Theorem 3.1.

(iii) Now the word is of the following

$$u_{m,n} = 012..(m-1)012..(m-1)..012..(l-1)$$

for $n \equiv l \bmod (m)$. The Lemma 3.1 is clearly true here, furthermore

Lemma 3.3. *For the letters $k \leq j \leq m-1$ the image of the letter j is j or (k+m-j-1).*

Now let's describe the involutory automorphism of $B^{m,n}$. Let σ^* be the mapping that reverses all the words, and let $\nu_{k,m}$ be the mapping that changes the letters in the words in the following way: for $0 \leq i \leq k-1$ the letter i is changed for $k-1-i$, and for $k \leq j \leq m-1$ the letter j is changed for $(m+k-1-j)$. Clearly neither σ^* nor $\nu_{k,m}$ is an automorphism of $B^{m,n}$, but $\sigma^*\nu_{k,m} \in \mathrm{Aut}(B^{m,n})$.

Now let σ_0 be an arbitrary automorphism, furthermore let σ be $\sigma^*\nu_{k,m}\sigma_0$ if $\sigma_0(0) = (k-1)$ and let σ be σ_0 if $\sigma_0(0) = 0$. Now $\sigma(0) = 0$. Similarly to part **(ii)**, Lemma 3.2 is true which proves the Theorem 3.1.

4. The DNA reconstruction

The motivation of this analysis is coming from the biology: based on some basic properties of DNA strands we can build a mathematical model, which is easy to handle. DNA is composed of units called *nucleotides* : A,C,G and T, these

letters are the elements of the alphabet. The letters form two complement pairs: A-T and C-G. Furthermore, DNA is double-stranded, i.e. each sequence occurs together with its reverse complement (we get the reverse complement in two steps: replacing each letter by its complement and reverse this sequence). For example the reverse complement of AACCGT is ACGGTT.

We can generalize the above properties for q complement pairs, and consider a reconstruction problem (see Section 1). It is easy to see that it makes no difference how many complement pairs build up the DNA-word.

Lemma 4.1. *We can solve a reconstruction problem of all DNA strands over an alphabet with q complement pairs iff we can do it for the similar problem for $q = 2$, i.e. iff we can reconstruct all DNA strands over the alphabet $\{\{A, T\}, \{C, G\}\}$.*

It is clear that if we can reconstruct all strands over an alphabet with k complement pairs, then we can reconstruct them over ACGT. Conversely, suppose that we can reconstruct all strands over ACGT. Then replace the first complement pair with A-T, and all the others with C-G. Now we can reconstruct the strand, and so we find the places of letters from the first complement pair in the original strand (now A-T-s are there); then we can repeat the procedure in order to find the other complement pairs.

Using this, similar to Lemma 2.1 we proved the following :

Lemma 4.2. *If $3 \leq n$ then every DNA-word f of length n is uniquely determined by $d^*_{n-1}(f)$.*

Now we can determine $\mathrm{Aut}(D^{q,n})$. We can see easily two types of automorphisms without proof: a permutation $\pi \in Sym_q$ on the complement pairs induces an automorphism σ_π on $D^{q,n}$. Denote also by Sym_q the automorphism group generated by these σ_π-s. Furthermore, consider a map which interchanges the elements of the i-th complement pair. This induces an automorphism σ^*_i on $D^{q,n}$. Denote by Z_2 the automorphism group generated by σ^*_i. Surprisingly in most cases there are no more automorphisms (note that the automorphism that reverse the order of the letters, which is a natural one, is $\sigma^*_1\sigma^*_2...\sigma^*_k$; e.g. $\sigma^*_1\sigma^*_2(ab) = \bar{a}\bar{b}$, which is identified to its reverse complement, i.e. ba).

Theorem 4.1. *(i) if $n = 1$, then $\mathrm{Aut}(D^{q,n}) = Sym_q$;*
(ii) if $n = 2$, then $\mathrm{Aut}(D^{q,n}) = Sym_q \otimes Sym_3^q \otimes Sym_4^{\binom{q}{2}}$;
(iii) if $n \geq 3$, then $\mathrm{Aut}(D^{q,n}) = Sym_q \otimes Z_2^q$.

The proof of the theorem is similar to Theorem 3.1: if an automorphism fixes the two lowest levels of the poset, then because of Lemma 4.2 it is the identity on the whole poset (we can apply Lemma 4.2 only for $n \geq 3$). The case $n = 1$ is considered only for the sake of completeness. In **(ii)** the poset has only two levels. It is clear that an automorphism transfers complement pairs to complement pairs. Take an arbitrary automorphism $\sigma_0 \in \mathrm{Aut}(D^{q,n})$ and consider its action on the set of complement pairs. Thus, this is a permutation on q elements, take its inverse π^{-1}. This permutation induces an automorphism $\sigma_{\pi^{-1}}$ on the poset $D^{q,n}$. Let

$\sigma_1 = \sigma_0\sigma_{\pi^{-1}}$. Then σ_1 fixes all of the complement pairs. Now one can partition the second level into $q + \binom{q}{2}$ blocks: we have q blocks of size 3 with elements $\{ii \equiv \overline{i}\overline{i}, i\overline{i}, \overline{i}i\}$; and $\binom{q}{2}$ blocks of size 4, with elements $\{ij \equiv \overline{j}\overline{i}, i\overline{j} \equiv j\overline{i}, \overline{i}j \equiv \overline{j}i, \overline{i}\overline{j} \equiv ji\}$ for all $i \neq j$, each block is fixed by σ_1 (setwise). This means q copies of Sym_3 and $\binom{k}{2}$ copies of Sym_4, and these automorphisms differ and commute, which proves the second part of the theorem.

In **(iii)** we can prove easily the following.

Remark 4.1. The automorphism σ_1 fixes all sequences in form of ii

To the contrary suppose that $\sigma_1(ii) = i\overline{i}$ (or equivalently $\overline{i}i$). Then we can not define $\sigma_1(iii)$.

Let σ_i^* be the automorphism which interchanges the elements of the i-th complement pair. Denote σ_2 the product of σ_1 and those σ_i^*'s for which $\sigma_1(i\overline{i}) = \overline{i}i$. Then σ_2 fixes all elements in the 3-blocks. Furthermore:

Remark 4.2. The automorphism σ_2 fixes all sequences in form of ij for all $i \neq j$.

Now by Remark 4.1 and Remark 4.2 we have that the two lowest levels of the poset are fix, which completes the proof.

References

[1] P. Borwein, T. Erdélyi, G. Kós, Littlewood-type problems on $[0, 1]$, *Proc. London Math. Soc. (3)*, 1999 **79** No. 1, 22-46.

[2] G. Burosch, H-D. O.F. Gronau, J-M. Laborde, The automorphism group of the subsequence poset $B_{m,n}$, *Order*, 1999 **(16)** No. 2, 179-194 .

[3] P.L. Erdős, P. Sziklai and D. Torney, The word poset and insertion-deletion codes, *Electronic Journal of Combinatorics*, 2001 **(8)** No. 2, Research Paper 8, 10pp. (electronic).

[4] I. Krasikov, Y. Roditty, On a reconstruction problem for sequences (English. English summary) *J. Combin. Theory Ser. A*, 1997 **77** No. 2, 344-348.

[5] V.K. Leont'ev, Yu G. Smetanin, Problems of information on the set of words, (English. English summary) *J. Math. Sci. (New York)*, 2002 **108** No. 1, 49-70.

[6] V. I. Levenshtein, Efficient reconstruction of sequences from their subsequences or supersequences (English. English summary) *J. Combin. Theory Ser. A*, 2001 **93** No. 2, 310-332.

[7] P. Ligeti, P. Sziklai, Automorphisms of subword-posets, submitted to *Discrete Math.*

[8] M. Lothaire, Combinatorics on words. *Encyclopedia of Mathematics and its Applications*, **17**. Addison-Wesley Publishing Co., Reading, Mass., 1983.

[9] B. Manvel, A. Meyerowitz, A. Schwenk, K. Smith. P. Stockmeyer, Reconstruction of sequences, *Discrete Math*, 1994 **94** No. 3, 209-219

Postal address

Péter Ligeti, Péter Sziklai
Department of Computer Science
Eötvös Loránd University 1/D,
Pázmány P. s.
1117 Budapest,
Hungary