

Reconstructing Words from Subwords in Linear Time

Andreas W.M. Dress¹ and Péter L. Erdős^{2*}

¹Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22-26, 04103 Leipzig
Germany

dress@mis.mpg.de

²A. Rényi Institute of Mathematics, Hungarian Academy of Sciences, P.O. Box 127, 1364
Budapest, Hungary

elp@renyi.hu

Received August 8, 2004

AMS Subject Classification: 68R15, 92D20

Abstract. Almost 30 years ago, M. Schützenberger and L. Simon established that two n -words with letters drawn from a finite alphabet having identical sets of subwords of length up to $\lfloor n/2 \rfloor + 1$ are identical. In the context of coding theory, V.I. Levenshtein elaborated this result in a series of papers. And further elaborations dealing with alphabets and sequences with reverse complementation have been recently developed by P.L. Erdős, P. Ligeti, P. Sziklai, and D.C. Torney. However, the algorithmic complexity of actually (re)constructing a word from its subwords has apparently not yet explicitly been studied. This paper augments the work of M. Schützenberger and L. Simon by showing that their approach can be reworked so as to provide a linear-time solution of this reconstruction problem in the original setting studied in their work.

Keywords: reconstruction of words, subwords, algorithmic complexity

1. Definitions, Notations and Results

We consider

- for each $n \in \mathbb{N}$, the index set $[n] := \{1, 2, \dots, n\}$,
- a finite *alphabet* A , i.e., just a finite non-empty set,
- and *words* $w = a_1 a_2 \cdots a_n$ of length n with letters a_1, a_2, \dots, a_n drawn from the alphabet A , i.e., maps $w: [n] \rightarrow A$ from $[n]$ into A . For any such word, we denote its *length* n also by $\|w\|$ and, for any $j \in [n]$, its *restriction* to the subset $[j]$ of $[n]$ by $w_{[j]}$.

*The work of the second author was supported, in part, by Alexander v. Humboldt Stiftung, and by the Hungarian NSF, under contract Nos. T29255, T34702.

For any $k \in \mathbb{N}$ and $a \in A$, we write a^k for the word w of length k with $w(i) = a$ for all $i = 1, \dots, k$. And for any two words $w_1 \in A^{[n_1]}$ and $w_2 \in A^{[n_2]}$ of length n_1 and n_2 , respectively, we write $w_1 \circ w_2$ for the *concatenation* of w_1 and w_2 defined by

$$w_1 \circ w_2: [n_1 + n_2] \rightarrow A: i \mapsto \begin{cases} w_1(i), & \text{if } i \leq n_1, \\ w_2(i - n_1), & \text{if } i > n_1. \end{cases}$$

Further, for any word $w \in A^{[n]}$ of length n with entries from A and any letter $a \in A$, we put

$$w^{-1}(a) := \{i \in [n] \mid w(i) = a\} \text{ and } \|w\|_a := \#w^{-1}(a),$$

and for any w and a as above and any $k \in [\|w\|_a]$, we let $j_{w|a}(k)$ denote that index $j \in [n]$ for which $w(j) = a$ and $\|w_{[j]}\|_a = k$ holds, i.e., the index j of the k 's copy of a in w or, equivalently, the smallest index $j \in [n]$ with $\|w_{[j]}\|_a \geq k$:

$$j_{w|a}(k) = \min\{j \in [n] : \|w_{[j]}\|_a \geq k\}.$$

In particular,

$$\min(w^{-1}(a)) = j_{w|a}(1) \text{ and } \max(w^{-1}(a)) = j_{w|a}(\ell)$$

holds for every $a \in A$ with $\|w\|_a > 0$ and $\ell := \|w\|_a$.

Clearly, given any word $w \in A^{[n]}$ of length n , any number $\ell \in [n]$, and any letter $a \in A$ with $\|w\|_a = \ell$, there exists a unique sequence

$$w_0(a), w_1(a), \dots, w_\ell(a)$$

of words (possibly of length 0) with

$$w = w_0(a) \circ a \circ w_1(a) \circ a \cdots \circ a \circ w_\ell(a).$$

It is also obvious that $\|w_k(a)\|_a = 0$ holds for all $k = 0, 1, \dots, \ell$, and that

$$j_{w|a}(k) = k + \sum_{i \in [k]} \|w_{i-1}(a)\| \text{ and } n = \ell + \sum_{i=0}^{\ell} \|w_i(a)\| \quad (1.1)$$

holds for all $k \in [\ell]$ for these words $w_0(a), w_1(a), \dots, w_\ell(a)$.

Further, knowing the length $\|w\|$ of a word w as well as, for all but one $a \in A$, the number $\|w\|_a$ of copies of a occurring in w and the $\|w\|_a$ indices $j_{w|a}(1), j_{w|a}(2), \dots, j_{w|a}(\|w\|_a)$ encoding where the letter a occurs in w — or (cf. (1.1)) just as well the sums $\sum_{i \in [k]} \|w_{i-1}(a)\|$ for all $k \in [\|w\|_a]$ — is equivalent to knowing w .

Next, given any $n \in \mathbb{N}$, any alphabet A , and any word $w = a_1 a_2 \cdots a_n$ in $A^{[n]}$ as above, we define the subset $\binom{w}{m}$ of $A^{[m]}$ for every $m \in [n]$ by

$$\binom{w}{m} := \{a_{i_1} a_{i_2} \cdots a_{i_m} \mid i_1, i_2, \dots, i_m \in \mathbb{N}, 1 \leq i_1 < i_2 < \cdots < i_m \leq n\}.$$

Any word v in $\binom{w}{m}$ will be called a *subword of length m* of w .

We define *Schützenberger's Guessing Game* to be the task of correctly reconstructing, in a systematic fashion, a word $w \in A^{[n]}$ of length n from answers to queries about its subwords of length m , for some fixed $m \in \mathbb{N}$. It follows from results of Schützenberger and Simon (cf. [5]) and Levenstein (cf. [2, 3]) that one can always reconstruct w from answers to sufficiently many such queries in case $2m > n$ holds[†]. The shortest known proof can be found in the *Bible of Formal Language Theory* edited by M. Lothaire (cf. [4]). Here, we will consider queries of the following three types:

- (i) What is $\|w: m\|_a := \max(\|v\|_a : v \in \binom{w}{m})$?
- (ii) What is $\overline{j}_a(w|m|k) := \max(\min(v^{-1}(a) : v \in \binom{w}{m}, \|v\|_a \geq k)$?
- (iii) What is $\underline{j}_a(w|m|k) := \min(\max(v^{-1}(a) : v \in \binom{w}{m}, \|v\|_a \geq k)$?

Noting that $\|w\|_a < m$ must hold for all but at most one letter $a \in A$ in case $2m > n$, we will show that the following holds:

Theorem 1.1. *Given an alphabet A and two integers $n, m \in \mathbb{N}$ with $2m > n$, any word $w \in A^{[n]}$ can be reconstructed from answers to $\#A$ queries of the type (i), and $\lfloor n(1 - \frac{1}{\#A}) \rfloor$ queries of type (ii) and (iii), each.*

Clearly, this theorem follows immediately from the following two results:

Proposition 1.2. *With A, n, m, w as in Theorem 1.1 and a any letter in A , one has*

$$\min(\|w\|_a, m) = \|w: m\|_a,$$

and, therefore,

$$\|w\|_a = \|w: m\|_a$$

in case $\|w\|_a < m$, while

$$\|w\|_a = n - \sum_{b \in A - \{a\}} \|w: m\|_b$$

holds in case $\|w\|_a > m$. In particular, $\|w\|_a$ can be determined for all $a \in A$ from answers to exactly $\#A$ queries of type (i).

Proposition 1.3. *With A, n, m, w, a as above, $\ell := \|w\|_a < m$, and $k \in [\ell]$, one has either*

$$\underline{j}_a(w|m|k) > k \text{ and } \sum_{i \in [k]} \|w_{i-1}(a)\| = \underline{j}_a(w|m|k) + n - \ell - m$$

or

$$\overline{j}_a(w|m|\ell - k + 1) < m + k - \ell \text{ and } \sum_{i \in [k]} \|w_{i-1}(a)\| = \overline{j}_a(w|m|\ell - k + 1) - 1.$$

[†] Yet not always in case $2m \leq n$: E.g., if $A = \{a_1, a_2\}$, one has $\binom{w_1}{m} = \binom{w_2}{m} = A^{[m]}$ for the two words $w_1 := a_1 a_2 a_1 a_2 \cdots a_1 a_2$ and $w_2 := a_2 a_1 a_2 a_1 \cdots a_2 a_1$ of length $2m$.

2. Proofs

The first proposition is fairly obvious. To establish the second one, we proceed as follows: To simplify notation, we put $w_i := w_i(a)$ for all $i = 0, \dots, \ell$ and, following [4], we note that any subword v of w of length m with $\|v\|_a \geq k$ for which $\min(v^{-1}(a))$ is as large as possible, must be a subword of the word

$$w' = (w_0 \circ w_1 \cdots \circ w_{\ell-k}) \circ (a \circ w_{\ell-k+1} \circ a \circ w_{\ell-k+2} \circ \cdots \circ a \circ w_{\ell}),$$

one gets from w by dropping all but the last k copies of a in w . Thus, if

$$\sum_{i \in [\ell-k+1]} \|w_{i-1}\| \leq m - k \quad (2.1)$$

holds, the required subwords v of w are necessarily of the form

$$v = w_0 \circ w_1 \cdots \circ w_{\ell-k} \circ a \circ v'$$

for some subword v' of length $m - (1 + \sum_{i \in [\ell-k+1]} \|w_{i-1}\|)$ of the word

$$w_{\ell-k+1} \circ a \circ w_{\ell-k+2} \circ \cdots \circ a \circ w_{\ell},$$

with $\|v'\|_a = k - 1$, implying that

$$\bar{j}_a(w|m|k) = 1 + \sum_{i \in [\ell-k+1]} \|w_{i-1}\|$$

must hold in this case. Otherwise, $\sum_{i \in [\ell-k+1]} \|w_{i-1}\| > m - k$ holds and the required subwords v of w are necessarily of the form $v = v' \circ a^k$ for some subword v' of length $m - k$ of $w_0 \circ w_1 \cdots \circ w_{\ell-k}$, implying that

$$\bar{j}_a(w|m|k) = m - k + 1$$

holds in this case. So, we always have

$$\bar{j}_a(w|m|k) = 1 + \min\left(\sum_{i \in [\ell-k+1]} \|w_{i-1}\|, m - k\right),$$

or, equivalently,

$$\sum_{i \in [\ell-k+1]} \|w_{i-1}\| \geq \bar{j}_a(w|m|k) - 1, \quad (2.2)$$

with equality holding unless $\sum_{i \in [\ell-k+1]} \|w_{i-1}\| > m - k$ and $\bar{j}_a(w|m|k) - 1 \geq m - k$ hold, in which case,

$$\bar{j}_a(w|m|k) - 1 = m - k < \sum_{i \in [\ell-k+1]} \|w_{i-1}\| \quad (2.3)$$

must hold. Replacing k by $\ell - k + 1$ in (2.2) we see that

$$\sum_{i \in [k]} \|w_{i-1}\| = \underline{j}_a(w|m|k) + n - \ell - m$$

holds unless we have $m + k - \ell \leq \sum_{i \in [k]} \|w_{i-1}\|$ and $\overline{j}_a(w|m|\ell - k + 1) \geq m + k - \ell$.

Similarly, any subword v of w of length m with $\|v\|_a \geq k$ for which $\max(v^{-1}(a))$ is as small as possible, must be a subword of the word

$$w'' = (w_0 \circ a \circ w_1 \circ \cdots \circ w_{k-1} \circ a) \circ (w_k \circ w_{k+1} \cdots w_\ell),$$

one gets from w by dropping all but the first k copies of a in w . Thus, if

$$m - k \geq \sum_{i=k}^{\ell} \|w_i\| = n - \ell - \sum_{i \in [k]} \|w_{i-1}\|,$$

or, equivalently,

$$\sum_{i \in [k]} \|w_{i-1}\| > n - m - \ell + k \quad (2.4)$$

holds, the required subwords v of w are necessarily of the form

$$v = v'' \circ a \circ w_k \circ w_{k+1} \cdots w_\ell$$

for some subword v'' of $w_0 \circ a \circ w_1 \circ \cdots \circ a \circ w_{k-1}$ of length $m - 1 - \sum_{i=k}^{\ell} \|w_i\|$ with $\|v''\|_a = k - 1$. So, we must have

$$\underline{j}_a(w|m|k) = m - \sum_{i=k}^{\ell} \|w_i\| = m - n + \ell + \sum_{i \in [k]} \|w_{i-1}\| > k$$

in this case, while $\underline{j}_a(w|m|k) = k$ must hold in case

$$m - n + \ell + \sum_{i \in [k]} \|w_{i-1}\| \leq k.$$

In other words, we have

$$\underline{j}_a(w|m|k) = m - n + \ell + \sum_{i \in [k]} \|w_{i-1}\|$$

unless $\underline{j}_a(w|m|k) = k$ and

$$\sum_{i=k}^{\ell} \|w_i\| = n - \ell - \sum_{i \in [k]} \|w_{i-1}\| > m - k \quad (2.5)$$

holds.

However, either $\underline{j}_a(w|m|k) > k$ or $\overline{j}_a(w|m|\ell - k + 1) - 1 < m + k - \ell - 1$ must always hold because, otherwise, our analysis shows that $m - \ell + k \leq \sum_{i=k}^{\ell} \|w_i\| < n - \ell - m + k$ would hold, in contradiction to $2m > n$. ■

It seems fairly obvious that, provided all one can ask for are the numbers

$$\|w: m\|_a, \overline{j}_a(w|m|k), \text{ and } \underline{j}_a(w|m|k),$$

one can not do much better. However, if one can also ask for specific words v_1 in $\binom{w}{m}$ with $\|v_1\|_a \geq k$ and $\min(v_1^{-1}(a)) = \overline{j}_a(w|m|k)$ and/or words v_2 with $\|v_2\|_a \geq k$ and $\max(v_2^{-1}(a)) = \underline{j}_a(w|m|k)$, a much smaller number of queries (yet depending on the specific word w that is to be reconstructed) might do. E.g., if there exists some $a \in A$ with $\ell := \|w: m\|_a < m$, some $k \in [\ell]$, some word $v_1 \in \binom{w}{m}$ with $\|v_1\|_a \geq \ell - k + 1$ and $\min(v_1^{-1}(a)) = \overline{j}_a(w|m|\ell - k + 1)$ of the form $v'_1 \circ a^{\ell - k + 1}$, and some word $v_2 \in \binom{w}{m}$ with $\|v_2\|_a \geq k$ and $\max(v_2^{-1}(a)) = \underline{j}_a(w|m|k)$ of the form $a^k \circ v''_2$, then $w = v'_1 \circ a \circ v''_2$ must hold. It might be of some interest to investigate further such cases as well as, e.g., the average number of queries of this more specific type needed to reconstruct the words in $A^{[n]}$.

A closely related problem is to identify words of length n over a finite alphabet with reverse complementation. This problem arises in the context of molecular genetics where it is, more specifically, related to the problem of constructing efficient microarray assays. It has recently been shown by P.L. Erdős *et al.* (cf. [1]) that, in this case, subwords of length up to roughly $2n/3$ are necessary to identify the words. The complexity issue of the corresponding version of M. Schützenberger's guessing game seems to be more involved and will be addressed in subsequent papers.

References

1. P.L. Erdős, P. Ligeti, P. Sziklai, and D.C. Torney, Subwords in reverse-complement order, preprint, 2004.
2. V.I. Levenshtein, On perfect codes in deletion and insertion metric, *Discrete Math. Appl.* **2** (1992) 241–258.
3. V.I. Levenshtein, Efficient reconstruction of sequences from their subsequences or supersequences, *J. Combin. Theory Ser. A* **93** (2001) 310–332.
4. M. Lothaire, *Combinatorics on Words*, Addison-Wesley, Chapter 6, 1983, pp. 119–120.
5. I. Simon, Piecewise testable events, In: *Automata Theory and Formal Languages*, H. Brakhage ed., LNCS. **33** Springer Verlag, 1975, pp. 214–222.