Subwords in reverse-complement order - Extended abstract *

Péter L. Erdős
A. Rényi Institute of Mathematics
Hungarian Academy of Sciences,
Budapest, P.O. Box 127, H-1364 Hungary
elp@renyi.hu

Péter Ligeti and Péter Sziklai
Dept. of Comp. Science, Eötvös University, Budapest
turul@cs.elte.hu sziklai@cs.elte.hu

David C. Torney
Theoretical Biology and Biophysics,
Mailstop K710, Los Alamos National Laboratory,
Los Alamos, New Mexico, 87545, USA;
dct@lanl.gov

May 23, 2006

Abstract

We examine finite words over an alphabet $\Gamma = \{a, \bar{a}; b, \bar{b}\}$ of pairs of letters, where each word $w_1 w_2 ... w_t$ is identified with its *reverse complement* $\bar{w}_t ... \bar{w}_2 \bar{w}_1$ (where $\bar{\bar{a}} = a, \bar{\bar{b}} = b$). We seek the smallest k such that every word of length n, composed from Γ, is uniquely

*This work was supported, in part, by Hungarian NSF, under contract Nos. AT48826, NK62321, F043772, N34040, T34702, T37846, T43758, ETIK, Magyary Z. grant and by the U.S.D.O.E.

Dagstuhl Seminar Proceedings 06201
Combinatorial and Algorithmic Foundations of Pattern and Association Discovery
http://drops.dagstuhl.de/opus/volltexte/2006/785

determined by the set of its subwords of length up to k. Our almost sharp result ($k \sim 2n/3$) is an analogue of a classical result for "normal" words. This problem has its roots in bioinformatics.

AMS Classification: 05D05, 68R15.

1 Introduction

Let Δ be a finite alphabet and let Δ^* denote the set of all finite sequences over Δ, called *words.* For $s, w \in \Delta^*$ we say that s is a *subword* of w ($s \leq w$) if s is a (not necessarily consecutive) subsequence of w. (Note, there some authors have called these constructs "subsequences", reserving "subword" for consecutive subsequences.) The length of w is denoted $|w|$. The following result was independently rediscovered repeatedly; as far as we are aware the problem originally was posed by M.P. Schützenberger and I. Simon.

Theorem 1 (Simon [8]) *Every word $w \in \Delta^*$ with at most $2m-1$ letters is completely determined by its length and by the set of all its subwords of length at most m.*

The pair of words *abababa* and *bababab* shows clearly that this result is sharp. In Simon's paper it was noted that it suffices to prove the theorem for the two-letter case: $\Delta = \{a, b\}$. Perhaps the shortest proof of Theorem 1 is due to Jacques Sakarovitch and Imre Simon (see [6], pp. 119–120); we were influenced by this nice proof.

V. L. Levenshtein in his papers [3, 4, 5] considers more generalizations of the reconstruction problem. In [3] the author examines which other sets of subwords or super-words determine uniquely the original word, in [4] the maximum size of the set of common subwords (or super-words) of two different words of a given length is given in a recursive way. In [5] every unknown sequence is reconstructed from its versions distorted by errors of a certain type, which are considered as outputs of repeated transmissions over a channel, and a minimal number of transmissions sufficient to reconstruct the original word (either exactly or with a given probability) is given. In both of the latter papers simple reconstruction algorithms are given.

In this paper we study another version of the Schützenberger-Simon's problem. Let $\Gamma = \{a, \bar{a}; b, \bar{b}\}$ be an alphabet where the letters come in pairs (called *complement pairs*); and let Γ^* denote the set of all finite sequences,

called *words*, composed from Γ. Define $\bar{\bar{a}} = a$, $\bar{\bar{b}} = b$ and for a word $w = w_1 w_2 ... w_t \in \Gamma^*$ let $\widetilde{w} = \overline{w_t}\,\overline{w_{t-1}}\ ...\ \overline{w_1}$, the *reverse complement* of w. Note that $\widetilde{(\widetilde{w})} = w$. Now we want to keep the essence of the previous partial ordering, while, in our poset, **each word is identified with its reverse complement.**

As in the foregoing Theorem, we do not address effective *reconstruction* per se; our concern is the prefatory problem of determining the minimal m such that the subwords of length up to m determine each word of length n. In the "classical case" the reconstruction problem was recently addressed (see: A. Dress and P.L. Erdős, [1]). In the reverse complement case the problem seems to be more complicated, and no results are presently available.

Our problem and definitions have biological motivations (for details see [2]). DNA typically exists as paired, reverse complementary words or *strands*: the Watson-Crick double helix, with its four letters, A, C, G and T paired via $\bar{A} = T$ and $\bar{C} = G$. Corresponding DNA codes could involve the insertion-deletion metric — with bounded *similarity* between two strands: the length of the longest subword common either to the strands or common to one strand and the reverse complement of the other.

Another common task is to decide rapidly and efficiently whether a given DNA double-strand (for example an erroneous gene, which is associated with illness) is present in a sample. This setting typically invokes microarrays: ten thousand or so of relatively short DNA words (called *probes*) are fixed on a glass slide. The sample reacts with the probes, and the probes which bind material from the sample are determined. We may model this process with our definition, i.e., to say that binding occurs if the probe is a subword of either strand. One may argue that the physicochemical laws don't allow each subword of the long DNA word to bind effectively because, for instance, "blocks" of consecutive matches may be required for binding. Although this is a perfectly legitimate objection, our aim is to provide additional background for such applications.

2 Main results

In this section we formulate our main results. Let's recall that in our partial order every word is identified with its reverse complement. Therefore, if in this partial order the word g is smaller than the word f, then it can happen that g is a subword of f or it is a subword of its reverse complement $\widetilde{f}$. For

convenience, if we do not know (or do not care) which is the case, then we will say that the word g *precedes* the word f ($g \prec f$). Let $S(m, f)$ denote the set of all words of length $\leq m$, which precede f. We seek to determine when $S(m, f)$ uniquely defines f.

One may note essential differences between this and the original problem; here, for instance, we may have more subwords but we do not distinguish between individual subwords belonging to a word or to its reverse complement. This difference is evident when the alphabet consists of a letter and its complement.

Let's consider the following example:

$$\mathcal{F}' = \bar{a}^{2k+\varepsilon}\, a^k \quad \text{and} \quad \mathcal{G}' = \bar{a}^{2k+\varepsilon-1}\, a^{k+1}, \tag{1}$$

where $\varepsilon \in \{0, 1, 2\}$ and $k \geq 1$ and $(k, \varepsilon) \neq (1, 0)$. The length of both words is $3k + \varepsilon$. On the one hand the subword $\bar{a}^{2k+\varepsilon}$ of $\mathcal{F}'$ satisfies $\bar{a}^{2k+\varepsilon} \not\prec \mathcal{G}'$. On the other hand it is easy to check that

$$S(2k + \varepsilon - 1, \mathcal{F}') = S(2k + \varepsilon - 1, \mathcal{G}').$$

In this paper we prove the following result:

Theorem 2 *Every word $f \in \{a, \bar{a}\}^*$ of length at most $3m - 1$ is uniquely determined by its length and by the set*

$$D'(f) := S(2m, f).$$

The next example illustrates that if our words contain letters from more than one complement pair, then they are "easier to distinguish". Consider the following words:

$$\mathcal{F} = \bar{a}^{2k+\varepsilon}\, \bar{b}\, b\, a^k \quad \text{and} \quad \mathcal{G} = \bar{a}^{2k+\varepsilon-1}\, \bar{b}\, b\, a^{k+1}, \tag{2}$$

where $\varepsilon \in \{0, 1, 2\}$ and $k \geq 1$ and $(k, \varepsilon) \neq (1, 0)$. The length of both words is $3k + 2 + \varepsilon$. On the one hand the subword $\bar{a}^{2k+\varepsilon}$ of $\mathcal{F}$ satisfies $\bar{a}^{2k+\varepsilon} \not\prec \mathcal{G}$. On the other hand it is easy to verify that

$$S(2k + \varepsilon - 1, \mathcal{F}) = S(2k + \varepsilon - 1, \mathcal{G}).$$

We have the following statement:

Theorem 3 *Every word $f \in \Gamma^*$ of length at most $3m+1$ $(m > 1)$ containing both $(a$ or $\bar{a})$ and $(b$ or $\bar{b})$ is uniquely determined by its length and by the set*

$$D(f) := S(2m, f).$$

The examples *abab* and *abba* show that in case of $m = 1$ the statement is not true.

Please recognize that due to our definitions, the expression "uniquely determined" means "uniquely determined, up to reverse complementation". The statement pertains to the case of $\varepsilon = 2$ in the example.

3 Easy consequences

There are some immediate consequences of the results of Section 2. Here we give just some examples: in the case when our words contain letters from one complement pair only, the following result holds:

Corollary 4 *Every word $f \in \{a, \bar{a}\}^*$ of length at most n is uniquely determined by its length and by the set $S\left(\left\lceil \frac{2(n+2)}{3} \right\rceil, f\right)$.*

Proof: Let m be the smallest integer such that $n \leq 3m-1$. Then $\left\lceil \frac{2(n+2)}{3} \right\rceil \geq 2m$ and Theorem 2 applies. ■

And, correspondingly, for the case of words containing letters from two complement pairs:

Corollary 5 *Every word $f \in \Gamma^*$ of length at most n containing both $(a$ or $\bar{a})$ and $(b$ or $\bar{b})$ is uniquely determined by its length and by the set $S\left(\left\lfloor \frac{2(n+1)}{3} \right\rfloor, f\right)$.*

Proof: The statement is straightforward: Let m be the smallest integer such that $n \leq 3m + 1$. Then $\left\lfloor \frac{2(n+1)}{3} \right\rfloor \geq 2m$, therefore Theorem 3 applies. ■

Our instinct says that Corollaries 4 and 5 are not sharp. We suspect that the truth is the following:

Conjecture 6 *Each word of length at most $3m+2+\varepsilon$ containing both $(a$ or $\bar{a})$ and $(b$ or $\bar{b})$ is uniquely determined by its length and by the set $S(2m+\varepsilon, f)$. Furthermore, each word of length at most $3m + \varepsilon$ containing only a or $\bar{a}$ is uniquely determined by its length and by the set $S(2m + \varepsilon, f)$.*

If our words are self-reverse complementary, then we are back to the original problem:

Remark 7 *Let the words f and $g \in \Gamma^*$ (of length at most n) be self-reverse complementary, that is $f = \tilde{f}$ and $g = \tilde{g}$. Now if $S(\lceil (n+1)/2 \rceil, f) = S(\lceil (n+1)/2 \rceil, g)$ then $f = g$.*

Proof: If for the word w we have $w \prec f$ and $f = \tilde{f}$, then w is a subword of f as well as of $\tilde{f}$. Therefore Theorem 1 applies. ■

For the original problem it was almost trivial that from the result for the case of 2-letter alphabet one derives an (approximate) result for the case of k-element alphabets as well. The situation here is similar but the proof requires some work:

Theorem 8 *Theorem 3 remains valid if the word f contains letters from $k \geq 2$ different complement pairs.*

Proof: We use induction on the number k of different complement pairs present. The case of two pairs present is Theorem 3. Assume that the statement is valid for the case of $k-1$ different pairs present. Let f and g be words with length $|f| = |g| \leq 3m+1$, and in both words let there be k different complement pairs present. The alphabet is $\{a_1, \bar{a}_1, ..., a_k, \bar{a}_k\}$. Let $A_{1,2}, \bar{A}_{1,2}$ be a new pair of complementary letters, and $f_{1,2}$ be the word derived from f by identifying all occurrences of a_1 and a_2 with $A_{1,2}$ and all occurrences of $\bar{a}_1$ and $\bar{a}_2$ with $\bar{A}_{1,2}$. The word $g_{1,2}$ is derived similarly. The new words contain letters from $k-1$ different pairs and $D(f_{1,2}) = D(g_{1,2})$. The inductive hypothesis gives that $f_{1,2} = g_{1,2}$ (one might need to exchange the names of $g_{1,2}$ and $\widetilde{g}_{1,2}$). Furthermore, for the subwords $f^*_{1,2}$ and $g^*_{1,2}$ consisting of all occurrences of the letters $\{a_1, \bar{a}_1, a_2, \bar{a}_2\}$ we have $D(f^*_{1,2}) = D(g^*_{1,2})$; therefore, we can apply Theorem 3. Whence $f^*_{1,2} = g^*_{1,2}$ or $f^*_{1,2} = \widetilde{g}^*_{1,2}$.

In the case of $f^*_{1,2} = g^*_{1,2}$ interleaving $f_{1,2}$ and $f^*_{1,2}$ we can determine f which is identical to g. In case of ($f_{1,2} = \widetilde{g}_{1,2}$ and $f^*_{1,2} = \widetilde{g}^*_{1,2}$) we can proceed similarly. However, it can happen that

$$f_{1,2} = g_{1,2} \quad \text{but} \quad f_{1,2} \neq \widetilde{g}_{1,2} \quad \text{while} \tag{3}$$

$$f^*_{1,2} \neq g^*_{1,2} \quad \text{but} \quad f^*_{1,2} = \widetilde{g}^*_{1,2}. \tag{4}$$

The value $|f^*_{1,2}|$ cannot be odd, since otherwise $f_{1,2}\left(\frac{|f^*_{1,2}|+1}{2}\right) = g_{1,2}\left(\frac{|g^*_{1,2}|+1}{2}\right)$, therefore $f^*_{1,2} = \widetilde{g}^*_{1,2}$ cannot occur. So let $|f^*_{1,2}| = \ell$ be even. From Condition

(4) it follows that there is an index $j \leq \ell/2$ such that, say, $f^*_{1,2}(j) = a_1$, $g^*_{1,2}(j) = a_2$, while $f^*_{1,2}(\ell+1-j) = \bar{a}_2$ and $g^*_{1,2}(\ell+1-j) = \bar{a}_1$. From Condition (3) it follows that there is a subscript $i \leq (3m+1)/2$ such that, say, $f_{1,2}(i) = a_3$ (therefore $g_{1,2}(i) = a_3$ also holds) while $g_{1,2}(3m+2-i) = b$ where $b \neq \bar{a}_3$. If $b \in \{a_1, ..., a_k\}$, then introducing the new letters $B_1, \bar{B}_1, B_2, \bar{B}_2$, substitute all occurrences of a_1 and a_3 with B_1, all occurrences of $\bar{a}_1, \bar{a}_3$ with $\bar{B}_1$, all occurrences of the letters $a_2, a_4, ..., a_k$ with B_2, and, finally, all occurrences of the letters $\bar{a}_2, \bar{a}_4, ..., \bar{a}_k$ with $\bar{B}_2$ in the original words. The result is the words f^B and g^B which satisfy the conditions of Theorem 3 while clearly $f^B \neq g^B$ and $f^B \neq \widetilde{g^B}$, a contradiction.

If, however, $b \in \{\bar{a}_1, \bar{a}_2, \bar{a}_4, ..., \bar{a}_k\}$, then we may define a bipartition of the alphabet, where letters b and a_3 belong to different classes, and letters a_1 and a_2 also belong to different classes. Then substitute all occurrences of the letters from the first class of the bipartition with $C_1, \bar{C}_1$ and the letters from the second class with $C_2, \bar{C}_2$, respectively. The new words clearly satisfy the conditions of Theorem 3; however, the consequence of Theorem 3 does not hold. ∎

This proof suggests that the existence of letters from more complement pairs decreases the necessary subword length in the result.

Because our approach does not work for very short words, we use the following enumerative result:

Remark 9 *Theorem 2 and 3 were tested by a computer program for short words (for $|f| \leq 13$ and for selected words with $|f| \leq 18$) and were found valid. Therefore our proofs need only address sufficiently long words, allowing reasoning which is effective above a (usually very small) length.*

The general approach to prove our main results is similar to the one in the proof of Theorem 8: identify a subword of the word under investigation which distinguishes the word and its reverse complement from each other. Such a subword can identify the word itself. The greater the similarity between the word and its reverse complement, the harder to find such a subword but, compensating for this difficulty, more is known about the structure of such words.

References

[1] A.W.M. Dress - P.L. Erdős: Reconstructing Words from Subwords in Linear Time, *Annals of Combinatorics*, **8** (4) (2004), 457–462.

[2] A.G. D'yachkov - P.L. Erdős - A.J. Macula - V.V. Rykov - D.C. Torney - C.-S. Tung - P.A. Vilenkin - P. S. White: Exordium for DNA Codes *J. Comb. Opt.* **7** (2003), 369–379.

[3] V.I. Levenshtein: On perfect codes in deletion and insertion metric, *Discrete Math. Appl.* **2** (1992), 241–258.

[4] V.I. Levenshtein: Efficient reconstruction of sequences from their subsequences or supersequences, *J. Comb. Theory* (A) **93** (2001), 310–332.

[5] V.I. Levenshtein: Efficient reconstruction of sequences, *IEEE Tr. Inf. Theory* **47** (1) (2001), 2–22.

[6] M. Lothaire: Combinatorics on words, Addison-Wesley, (1983). Chapter 6, pp. 119–120.

[7] J. Manuch: Characterization of a word by its subwords. (Rozenberg, Grzegorz (ed.) *et al.*), Developments in language theory. Foundations, applications, and perspectives. Proceedings of the 4th international conference, Aachen, Germany, July 6-9, 1999. Singapore: World Scientific. (2000), 210–219.

[8] I. Simon: Piecewise testable events, (H. Brakhage ed.), Automata Theory and Formal Languages, LNCS. **33**, Springer Verlag, (1975), 214–222.