

Subwords in Reverse-Complement Order*

Péter L. Erdős¹, Péter Ligeti², Péter Sziklai², and David C. Torney³

¹A. Rényi Institute of Mathematics, Hungarian Academy of Sciences, Budapest,
P.O. Box 127, H-1364, Hungary
elp@renyi.hu

²Department of Computer Science, Eötvös University, Pázmány Péter sétány 1/C,
H-1117 Budapest, Hungary
{turul, sziklai}@cs.elte.hu

³Theoretical Biology and Biophysics, Mailstop K710, Los Alamos National Laboratory,
Los Alamos, New Mexico, 87545, USA
dtorney@earthlink.net

Received October 19, 2005

AMS Subject Classification: 05D05, 68R15

Abstract. We examine finite words over an alphabet $\Gamma = \{a, \bar{a}; b, \bar{b}\}$ of pairs of letters, where each word $w_1 w_2 \cdots w_t$ is identified with its *reverse complement* $\bar{w}_t \cdots \bar{w}_2 \bar{w}_1$ (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 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.

Keywords: combinatorics of words, Levenshtein distance, DNA codes, reconstruction of words

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, that some authors have called these constructs “subsequences”, reserving “subword” for consecutive subsequences.) The length of w is denoted by $|w|$. The following result was independently rediscovered repeatedly; as far as we are aware the problem originally was posed by Schützenberger and Simon. (In the bibliography we try to give the original sources relevant to our problem. It is not our intention, however, to give a comprehensive bibliography.)

Theorem 1.1. (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 .*

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

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 1.1 is due to Sakarovitch and Simon (see [6, pp. 119–120]); we were influenced by this nice proof.

Levenshtein in his papers [3–5] 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 this 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 $w = w_1 w_2 \cdots w_t \in \Gamma^*$ let $\bar{w} = \bar{w}_t \bar{w}_{t-1} \cdots \bar{w}_1$, the *reverse complement* of w . Note that $(\bar{\bar{w}}) = w$. 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* essentially; 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, Dress and 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 do not 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.

Before we list our main results, let us remark that our problem is a special case of a general class of problems, in which group orbits substitute for the classes of words and their reverse complements. The group must have a well defined action on all subwords

— an induced action based, for instance, on permuting letter identities and letter positions. (The group considered herein is of order two.) A permutation may, for example, act on the positions included in subwords through the respective complete ordering. Thus, one version of the general problem is:

Given the k -spectra of the words for its orbits (the set of subwords of up to length k occurring in any of these words), find a characterization of all the (permutation) groups which yield k -spectra one-to-one correspondence with these orbits. For the general problem, the respective partial order would be inclusion when any member of the orbit occurs as a subword.

2. Main Results

In this section we formulate our main results. Let us 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 \tilde{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 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 us consider the following example:

$$\mathcal{F}' = \bar{a}^{2k+\varepsilon} a^k \quad \text{and} \quad \mathcal{G}' = \bar{a}^{2k+\varepsilon-1} a^{k+1}, \quad (2.1)$$

where $\varepsilon \in \{0, 1, 2\}$, $k \geq 1$ 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 2.1. *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 proof of this result can be found in Section 4.

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:

$$\mathcal{F} = \bar{a}^{2k+\varepsilon} \bar{b} b a^k \quad \text{and} \quad \mathcal{G} = \bar{a}^{2k+\varepsilon-1} \bar{b} b a^{k+1}, \quad (2.2)$$

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 2.2. *Every word $f \in \Gamma^*$ of length at most $3m + 1$ ($m > 1$) containing both $(a \text{ or } \bar{a})$ and $(b \text{ or } \bar{b})$ 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. The proof of this result can be found in Section 5.

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. For example in the case when our words contain letters from one complement pair only, one may formulate the following result.

Corollary 3.1. *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 2.1 applies. ■

Correspondingly, for the case of words containing letters from two complement pairs, we have

Corollary 3.2. *Every word $f \in \Gamma^*$ of length at most n containing both $(a \text{ or } \bar{a})$ and $(b \text{ or } \bar{b})$ 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 2.2 applies. ■

Our instinct says that Corollaries 3.1 and 3.2 are not sharp. We suspect that the truth is the following:

Conjecture 3.3. *Each word of length at most $3m + 2 + \varepsilon$ containing both $(a \text{ or } \bar{a})$ and $(b \text{ or } \bar{b})$ 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 3.4. 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 1.1 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 3.5. *Theorem 2.2 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 2.2. 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 $\{a_1, \bar{a}_1, \dots, a_k, \bar{a}_k\}$. 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 $\tilde{g}_{1,2}$). 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 2.2. Whence $f_{1,2}^* = g_{1,2}^*$ or $f_{1,2}^* = \tilde{g}_{1,2}^*$.

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 $f_{1,2} = \tilde{g}_{1,2}$ and $f_{1,2}^* = \tilde{g}_{1,2}^*$ we can proceed similarly. However, it can happen that

$$f_{1,2} = g_{1,2}, \text{ but } f_{1,2} \neq \tilde{g}_{1,2}, \text{ while} \tag{3.1}$$

$$f_{1,2}^* \neq g_{1,2}^*, \text{ but } f_{1,2}^* = \tilde{g}_{1,2}^*. \tag{3.2}$$

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 $f_{1,2}^* = \tilde{g}_{1,2}^*$ cannot occur. So let $|f_{1,2}^*| = \ell$ be even. From Condition (3.2) it follows that there is an index $j \leq \ell/2$ such that $f_{1,2}^*(j) = a_1$, $g_{1,2}^*(j) = a_2$, while $f_{1,2}^*(\ell + 1 - j) = \bar{a}_2$ and $g_{1,2}^*(\ell + 1 - j) = \bar{a}_1$. From Condition (3.1) it follows that there is a subscript $i \leq (3m + 1)/2$ such that $f_{1,2}(i) = a_3$ (therefore $g_{1,2}(i) = a_3$ also holds) while $g_{1,2}(3m + 2 - i) = b$ where $b \neq \bar{a}_3$. If $b \in \{a_1, \dots, a_k\}$, 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, \dots, a_k with B_2 , and, finally, all occurrences of the letters $\bar{a}_2, \bar{a}_4, \dots, \bar{a}_k$ with \bar{B}_2 in the original words. The result is the words f^B and g^B which satisfy the conditions of Theorem 2.2 while clearly $f^B \neq g^B$ and $f^B \neq \tilde{g}^B$, a contradiction.

If, however, $b \in \{\bar{a}_1, \bar{a}_2, \bar{a}_4, \dots, \bar{a}_k\}$, 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 2.2; however, the consequence of Theorem 2.2 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 3.6. Theorems 2.1 and 2.2 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.

In the next two sections we prove our main results. The general approach used is similar to the one in the proof of Theorem 3.5: 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, the more is known about the structure of such words.

4. The Proof of Theorem 2.1

Assume that f and g are words in $\{a, \bar{a}\}^*$ of the same length such that

$$|f| = |g| \leq 3m - 1 \quad \text{and} \quad D'(f) = D'(g) = D'.$$

Due to Remark 3.4, we may assume that f is not self-reverse complementary. Denote by $A(w)$ the number of a 's in the word w , and define $\bar{A}(w)$ analogously. Without loss of generality we may assume that both words f and g are written in the form where $A(f) \geq \bar{A}(f)$ and $A(g) \geq \bar{A}(g)$. At first assume that $A(f) > A(g)$, which also means that $\bar{A}(f) < \bar{A}(g)$. If $A(f) > 2m$, then take an arbitrary subword g' of g such that $A(g'), \bar{A}(g') \geq \bar{A}(f) + 1$. It is clear that $g' \not\sim f$. If, instead, $A(f) \leq 2m$, then take the subword f' of f containing $A(g) + 1$ a 's. It is also clear that $f' \not\sim g$ and that $|f'|, |g'| \leq 2m$, which constitutes a contradiction. Therefore, in this proof henceforth we assume that we have

$$A := A(f) = A(g) \quad \text{and} \quad \bar{A} := \bar{A}(f) = \bar{A}(g). \quad (4.1)$$

Before proceeding we introduce one more notion: a word contains a *run* of length k when it contains k consecutive copies of a certain letter.

4.1. The Case $\bar{A} < A$

In this case we know that $f \not\sim \tilde{f}$ and $g \not\sim \tilde{g}$, and each subword of f or g containing at least $\bar{A} + 1$ a 's obeys these inequalities. All subwords from $S(2m, f)$, containing at least $\bar{A} + 1$ a 's, are subwords of g , because they cannot be subwords of \tilde{g} — and correspondingly, the analogous statement holds for the subwords from $S(2m, g)$.

Our words f and g can be written in the following form:

$$f := a^{I_0} \bar{a} a^{I_1} \bar{a} \dots \bar{a} a^{I_s} \quad \text{and} \quad g := a^{J_0} \bar{a} a^{J_1} \bar{a} \dots \bar{a} a^{J_s},$$

where $s = \bar{A}$, and any I_l or J_l can be zero. If $f \neq g$, then the subset

$$L := \{l \in \{0, \dots, s\} \mid I_l \neq J_l\}$$

has at least two elements. Without loss of generality we may assume that $I_\ell = \min\{I_l, J_l : l \in L\}$, i.e., f contains a shortest run — of those letters indexed by L . Then consider the subword g' of g containing all its \bar{a} 's, containing at least $I_\ell + 1$ a 's in the ℓ -th run of a 's, and finally containing, as needed, other copies of a 's so that altogether there are at least $\bar{A} + 1$ a 's. Then, due to the definition, g' is not a subword of f , furthermore, by the number of a 's, it is also clear that \tilde{g}' is also not a subword of f . We know that

$$|g'| \leq \max \left\{ \left(\left\lceil \frac{\bar{A}}{2} \right\rceil - 1 \right) + 1 + \bar{A}, 2\bar{A} + 1 \right\},$$

since the left argument of the maximum includes, within its parentheses, the largest possible value for I_k . If $|g'| \leq 2\bar{A} + 1 \leq 2m$ holds, then there is a contradiction. Therefore this method shows that $D'(f)$ and $D'(g)$ must be different while $\bar{A} + 1 \leq m$. Continuing the proof from now on (in this section) we assume that

$$\bar{A} > m - 1. \tag{4.2}$$

Hence, in this case

$$A = 3m - 1 - \bar{A} \leq 2m - 1. \tag{4.3}$$

Denote by $\tilde{f}(a, \ell)$ the subword of f containing all a 's and the ℓ -th run of \bar{a} 's. By our assumptions these are subwords of g , but, as we have just seen, not subwords of \tilde{g} . Therefore both f and g can be written in the following forms:

$$f = a^{r_0} \bar{a}^{s_1} a^{r_1} \bar{a}^{s_2} \dots \bar{a}^{s_t} a^{r_t} \quad \text{and} \quad g = a^{r'_0} \bar{a}^{z_1} a^{r'_1} \bar{a}^{z_2} \dots \bar{a}^{z_t} a^{r'_t}, \tag{4.4}$$

where r_0 or r_t can be zero, while r_1, \dots, r_{t-1} and all s_i and z_i are non-zero.

Now we are going to show that for all i we also have $s_i = z_i$ (which, of course, implies that $f = g$).

Let $F \in \{x, y\}^*$ be an arbitrary word and assume it is written in the form

$$F = x^{r_0} y^{s_1} x^{r_1} y^{s_2} \dots y^{s_t} x^{r_t}, \tag{4.5}$$

where the runs are not empty (except, possibly, the very first and last). That is $r_0, r_t \geq 0$ and all other superscripts > 0 . A subword W of F is *well recognizable for the pair* x, y if one can reconstruct exactly which letter of W comes from which x - or y -runs of F . (Reverse complementation is not taken into consideration here. Generally we will ensure separately that the well recognizable subword's reverse complement is not a subword of the original.) It is clear that if the subword W' of F contains W as a subword, then W' is also well recognizable. The subword F_1 containing one letter from each run is clearly well recognizable. Even better, if r_0 and r_t are both non-zero (or, oppositely, both zero), then the reverse complement of this subword is automatically not a subword of F . But when F has a large number of runs (say each run consists of one letter), then one can find much shorter well recognizable subwords.

Proposition 4.1. *Let $W(F)$ be the subword of F defined as follows:*

- (I) $W(F)$ retains at least one x from each x -run.
- (II) If r_0 or $r_t > 1$, then $W(F)$ contains one x from the respective run and one y from the neighboring y -run.

- (III) From all other x -runs with precisely two letters, let $W(F)$ contain both.
- (IV) From all other x -runs with at least three letters, $W(F)$ contains one x from the run and one y from both adjacent runs.
- (En1) If between two previously chosen y 's there are only two-letter x -runs, then keep one x from each of these runs and take one element from each y -run in-between.
- (En2) From every run of y 's, remove all but one.

Then the resulting $W(F)$ is a well recognizable subword of F for the pair x, y .

(The two last procedures enhance the previously constructed well recognizable words, that give their different kinds of names.) Proposition 4.1 may be thought of as an algorithm, whose six steps are applied sequentially in a single pass. Thus, its validity is evident. Let us remark that without operation (En1) the subword $W(F)$ would be still a well recognizable subword, but this operation decreases the number of letters by one with each application. Note that $W(F)$ never has more letters than the total number of runs in f and neither is it ever shorter than the number of x -runs. However, this construction is sensible for one-letter runs and in their presence it produces well recognizable words with fewer letters than the total number of runs.

Note also that any well recognizable subword of f in Condition (4.4) is also a well recognizable subword of g .

Assume now that $f \neq g$, that is the series s_1, \dots, s_t and z_1, \dots, z_t are different. Then the set

$$L := \{l \in \{1, \dots, t\} \mid s_l \neq z_l\}$$

has at least two elements, since the total number of \bar{a} 's are the same in both of our words. Without loss of generality we may assume that $z_\ell = \min\{s_l, z_l : l \in L\}$. At first take the subword f_1 of f containing all its a 's and $z_\ell + 1$ \bar{a} 's from the ℓ -th \bar{a} -run. This word is clearly a well recognizable one, and, due to $A > \bar{A}$, its reverse complement is not a subword of f or g . Therefore, if $A + z_\ell + 1 \leq 2m$, then $f_1 \in D'(f)$ but $f_1 \notin D'(g)$, a contradiction.

If, however, this is not the case, then $|f_1| = 2m + \alpha$ and

$$A = 2m + \alpha - (z_\ell + 1), \tag{4.6}$$

$$\bar{A} = 3m - 1 - A = m - \alpha + z_\ell,$$

where $\alpha \geq 1$. By the minimality of z_ℓ there is another \bar{a} -run in f with at least z_ℓ elements. Therefore there are at most

$$t \leq 2 + \bar{A} - (2z_\ell + 1) = m + 1 - (z_\ell + \alpha) \tag{4.7}$$

\bar{a} -runs in the word f , and there is at most one more: that is, at most $m + 2 - (z_\ell + \alpha)$ a -runs in f .

Recall that the subword f_1 is not in $D'(f)$ because it has α extra letters and $z_\ell \geq \alpha \geq 1$ (viz. (4.6)).

Assume at first that $r_0, r_t > 0$. Then consider the subword f_2 of the word f containing one letter from each run except the ℓ -th \bar{a} -run, which contains $z_\ell + 1$ \bar{a} 's. This word is well recognizable, and \tilde{f}_2 is not a subword of f or g because they do not contain

enough \bar{a} -runs. Furthermore, f_2 is also clearly not a subword of g , since in the ℓ -th \bar{a} -run there are too many letters. Due to (4.7) we know that

$$|f_2| \leq 1 + 2t + z_\ell \leq 1 + 2[m + 1 - (z_\ell + \alpha)] + z_\ell = 2m + 3 - 2\alpha - z_\ell \leq 2m,$$

since $z_\ell \geq \alpha \geq 1$. Therefore $f_2 \in D'(f)$ but $f_2 \notin D'(g)$, a contradiction.

If $r_0 = r_t = 0$ then we can repeat the previous reasoning since \tilde{f}_2 is not a subword of f or g because there are not enough a -runs in them. If, say, $r_0 > 0$ and $r_t = 0$, then we cannot rule out that the reverse complement of f_2 is a subword of g . In this case there are precisely t ($\leq m + 1 - (z_\ell + \alpha)$) a -runs in f . Construct the subword f_3 of f as follows: it contains one letter from each run except the ℓ -th \bar{a} -run, which contains $z_\ell + 1$ \bar{a} 's. Then f_3 looks like f_2 but it has one fewer element, due to $r_t = 0$. It is a well recognizable subword of f but not a subword of g . Its length is

$$|f_3| = 2t + z_\ell < |f_2|,$$

therefore also $f_3 \in D'(f)$. In general, this would yield a contradiction, but if $r_{t-\ell} > z_\ell$, then \tilde{f}_3 could be a subword of g . But then let f_4 be constructed from f_3 by adding z_ℓ more a letters to the $(t - z_\ell)$ -th a -run. This f_4 is clearly a subword of f but not a subword of g or \tilde{g} . Finally

$$|f_4| = |f_3| + z_\ell \leq 2m + 2 - 2\alpha \leq 2m.$$

Therefore $f_4 \in D'(f)$ but $f_4 \notin D'(g)$, a contradiction. The case $\bar{A} < A$ is proved.

4.2. The Case $\bar{A} = A$

In this case we can prove a slightly stronger version of Theorem 2.1: we can suppose that $|f| \leq 3m$. Now $|f| = |g|$ is even, i.e., $m = 2k$ and the two words are of the form

$$f = a^{r_0} \bar{a}^{s_1} a^{r_1} \bar{a}^{s_2} \dots \bar{a}^{s_t} a^{r_t} \quad \text{and} \quad g = a^{R_0} \bar{a}^{z_1} a^{R_1} \bar{a}^{z_2} \dots \bar{a}^{z_T} a^{R_T}, \quad (4.8)$$

where $r_0 + \dots + r_t = s_1 + \dots + s_t = R_0 + \dots + R_T = z_1 + \dots + z_T = A = 3k$ and at least one of r_0, r_t and at least one of R_0, R_T is positive, otherwise we exchange the names of f and \tilde{f} , and similarly for g as well. Now without loss of generality we may assume that $r_0 > 0$. Then in g we have $R_0 > 0$. Otherwise the subword $a\bar{a}^A$ of f does not precede g (since there are not enough \bar{a} 's after the first a in g , and not enough a 's before the last \bar{a} in \tilde{g}).

If $r_t > 0$ also holds, then consider the subword $f_1 = \bar{a}^A a$. If $3k + 1 \leq 4k$ then $f_1 \in D'(f)$ but \tilde{f}_1 is not a subword of g , since there are not enough a 's after the first \bar{a} in g . Therefore f_1 itself is a subword of g and we have $R_T > 0$; otherwise, there are not enough \bar{a} 's before the last a in g . It also means that f_1 is a well recognizable subword of f and g as well. Therefore $r_t = 0 \Leftrightarrow R_T = 0$. (If, however, $|f| \leq 4$, then applying Remark 3.6 completes the proof.)

Assume at first that

$$r_t, R_T > 0. \quad (4.9)$$

Denote by F_i the subword of f derived from f_1 by inserting one a from the i -th a -run. If $A \geq 6$ then $F_i \in D'(f)$. These words together, for all i , describe the length of the

\bar{a} -runs in f , and all those runs are the complete union of some consecutive \bar{a} -runs in g . Repeating the process with g , yielding G_i 's, we have the similar correspondence between the \bar{a} -runs of f and g . Therefore the \bar{a} -run structures of f and g are identical: $t = T$, and $s_i = z_i; i = 1, \dots, t$. (If $A \leq 5$ then Remark 3.6 finishes the proof.) Therefore our words are of the form

$$f = a^{r_0} \bar{a}^{s_1} a^{r_1} \bar{a}^{s_2} \dots \bar{a}^{s_t} a^{r_t} \quad \text{and} \quad g = a^{R_0} \bar{a}^{s_1} a^{R_1} \bar{a}^{s_2} \dots \bar{a}^{s_t} a^{R_t}. \quad (4.10)$$

Assume now that $f \neq g$: that is, the series r_0, \dots, r_t and R_0, \dots, R_t are different. Then the set

$$L := \{l \in \{0, \dots, t\} \mid r_l \neq R_l\}$$

has at least two elements, since the total number of a 's is A in both words. Without loss of generality we may assume that $R_\ell = \min\{r_l, R_l : l \in L\}$. Consider the f -subword $f_2 = \bar{a}^{s_1 + \dots + s_\ell} a^{R_\ell + 1} \bar{a}^{s_{\ell+1} + \dots + s_t} a$. This is clearly neither a subword of g nor of \tilde{g} . Therefore $A + R_\ell + 2 > 4k$, implying that $R_\ell \geq k - 1$. Due to the selection procedure for R_ℓ there is another a -run in f of length at least R_ℓ . Then all the other a -runs in f altogether contain $\leq 3k - (2R_\ell + 1)$ letters; hence the numbers of \bar{a} -runs are limited: $t \leq 3k - 2R_\ell$. Let the subword f_3 contain one letter from each different run in f , and contain R_ℓ more letters from the ℓ -th a -run. This word has at most $2(3k - 2R_\ell) + 1 + R_\ell = 6k - 3R_\ell + 1 \leq 3k + 4$ letters (here we used $R_\ell \geq k - 1$). Since f_3 is a subword of f but does not precede g and this is a contradiction (unless $k \leq 2$, when $|f| \leq 12$ and Remark 3.6 applies; or $k = 3$ and the length of word f 's a -runs are 3, 2, 1, 1, 1, 1 which allows again the use of Remark 3.6), Theorem 2.1 is established for this case.

From now on we assume that (4.9) does not hold: that is we have

$$r_t = R_T = 0. \quad (4.11)$$

(Let us recall that at that point we do not know whether the number of runs in f and g are equal or different.) Let $f(a; i)$ denote the subword of f containing all its a 's, furthermore one \bar{a} from the i -th \bar{a} -run of $f; i = 1, \dots, t$.

Claim: Every $f(a; i)$ is a subsequence of g or every $f(a; i)$ is a subsequence of \tilde{g} or both hold.

Indeed, if every $f(a; i)$ is a subsequence of both words then there is nothing to prove. Therefore assume that there is an index i such that $f(a; i)$ is a subsequence of g but not of \tilde{g} . Then for all indices $l \neq i$ the subword $f(a; l)$ is also a subword of g . Indeed, if there is an index l , such that the subword $f(a; l)$ was a subword of \tilde{g} but not of g , then consider the analogous subword $f(a; i, l)$ of f , containing altogether $A + 2$ letters (all a 's and one letter from the i -th and one from the l -th \bar{a} -run). This would not be a subword either of g or \tilde{g} , a contradiction, if $A \geq 6$ (if $A < 6$ then Remark 3.6 applies). The Claim is proved.

Therefore we may assume that all $f(a; i)$ are subwords of g ; therefore $t \leq T$, and one can make t groups g_1^*, \dots, g_t^* of consecutive a -runs in g such that the total length of a -runs within g_j^* is equal to s_j . Repeat the whole process for the subwords $g(a; i)$. It still might be necessary to substitute \tilde{f} for f , but due to (4.11) this already implies that

$$t = T. \quad (4.12)$$

But from this equation it also follows that each $g(a; i)$ is a subword of f , since they are just the image in g of the subwords $f(a; i)$. Therefore we also have $r_i = R_i$ for all i .

Now repeat the whole process for the analogous subwords $f(\bar{a}; i)$ of f . This yields

$$(s_i = z_i, \text{ for all } i) \quad \text{or} \quad (s_i = R_{t-i}, \text{ for all } i).$$

In the first case the proof is complete. Assume that this is not the case. Then the second relation series holds. But repeating the whole process again for the analogous subwords $g(\bar{a}, i)$ then we get that $z_i = r_{t-i}$, for all i . Since we have $r_i = R_i$ it follows that $s_i = z_i$ for all i , which contradicts our assumption, and Theorem 2.1 is proved.

5. Proof of Theorem 2.2

In this section, for conciseness, we will use the notation \hat{a} for both a and \bar{a} and \hat{b} for both b and \bar{b} , when the actual value of \hat{a} or \hat{b} is immaterial. With this notation every word of Γ^* can be considered as a word from $\{\hat{a}, \hat{b}\}^*$. Assume that f and g are words in Γ^* of the same length such that

$$|f| = |g| \leq 3m + 1 \quad \text{and} \quad D(f) = D(g) = D. \quad (5.1)$$

Without loss of generality we may also assume, due to Remark 3.4, that at least one of the two words, say g , is not self-reverse complementary. Furthermore let

$$p = \max\{|s| : s \in D \cap \hat{a}^*\} \quad \text{and} \quad q = \max\{|s| : s \in D \cap \hat{b}^*\}.$$

Without loss of generality we can assume that $q \leq p$. Let $f(a)$ denote the subword of f consisting of all \hat{a} 's. The notation $f(b)$, $g(a)$, and $g(b)$ are analogous. Then, by definition, $|f(a)| \geq p$ and $|f(b)| \geq q$; hence

$$2q \leq p + q \leq |f(a)| + |f(b)| = |f| \leq 3m + 1,$$

and consequently $q \leq \frac{3m+1}{2} < 2m$ if $1 < m$. This implies that $|f(b)| = |g(b)| = q$. It also implies that $|f(a)| = |g(a)|$ holds. We remark that $|f(a)|$ may exceed p . (Note that if q is odd, then the subwords containing all \hat{b} 's are different from their reverse complements.)

Due to these properties there exist non-negative integers $t, T; i_0, \dots, i_t; r_1, \dots, r_t; j_0, \dots, j_T$; and R_1, \dots, R_T such that

$$f = \hat{a}^{i_0} \hat{b}^{r_1} \hat{a}^{i_1} \dots \hat{b}^{r_t} \hat{a}^{i_t} \quad \text{and} \quad g = \hat{a}^{j_0} \hat{b}^{R_1} \hat{a}^{j_1} \dots \hat{b}^{R_T} \hat{a}^{j_T}, \quad (5.2)$$

where t can be equal to T , and i_0, i_t, j_0, j_T can be zero, while all other superscripts are nonnegative integers, and, furthermore, where $i_0 + \dots + i_t = j_0 + \dots + j_T = |f(a)|$ and $r_1 + \dots + r_t = R_1 + \dots + R_T = |f(b)|$. Since $q \leq 2m$, the subwords $f(b)$ and $g(b)$ belong to $S(2m, f) = D$; therefore $f(b) = g(b)$ or $f(b) = \tilde{g}(b)$, or both. Let us remark that we have our general form (4.5) with letters \hat{a} and \hat{b} ; therefore Proposition 4.1 applies to these words.

For two words w and u denote by $w \simeq u$ if both of $w \prec u$ and $u \prec w$ hold. The following observation will be useful later.

Proposition 5.1. *Assume that $T = t$, $i_k = j_k$ for $k = 0, \dots, t$ and $r_l = R_l$ for $l = 1, \dots, t$, and furthermore $f(a) \simeq g(a)$ and $f(b) \simeq g(b)$. Then $f \simeq g$.*

Proof. Suppose instead that $f \neq g$ and $f \neq \tilde{g}$. We can obtain f by interleaving the runs of $f(a)$ and $f(b)$. Since $f \neq g$ it is easy to see that we must get g from the runs of $\widetilde{f(a)}$ and $f(b)$. If at least one of $f(a)$ and $f(b)$ is self-reverse complementary, then we get $f = \tilde{g}$ or $f = g$, a contradiction. Suppose now that $f(a) \neq \widetilde{f(a)}$ and $f(b) \neq \widetilde{f(b)}$. Then due to Theorem 1.1 there exists a subword a_* of length at most $\lceil (|f(a)| + 1)/2 \rceil$, such that, say, $a_* \leq f(a)$, but $a_* \not\leq \widetilde{f(a)}$. We get b_* of length at most $\lceil (|f(b)| + 1)/2 \rceil$ similarly. Now let f_* be the word obtained from interleaving a_* and b_* . Clearly $f_* \prec f$ but $f_* \not\prec g$. Hence if $|f| > 7$, then $|f_*| \leq \lceil (|f(a)| + 1)/2 \rceil + \lceil (|f(b)| + 1)/2 \rceil = \lceil (f + 2)/2 \rceil = \lceil (3m + 3)/2 \rceil \leq 2m$, a contradiction. (The cases $|f| \leq 7$ are covered by Remark 3.6.) ■

Next we are going to show that the conditions of Proposition 5.1 hold.

At first we show that the run structures in $f(b)$ and in at least one of $g(b)$ and $\tilde{g}(b)$ are identical. Denote by $f(b; \ell)$ the subword consisting of all its \hat{b} 's and one letter from the ℓ -th \hat{a} -run. Since $|f(b; \ell)| \leq 2m$, $m > 1$, this belongs to $D(f) = D(g)$.

Claim: Every $f(b; \ell)$ is a subsequence of g or a subsequence of \tilde{g} or both hold.

Indeed, if every $f(b; \ell)$ is a subsequence of both words then there is nothing to prove. Therefore assume that for a particular k the word $f(b; k)$ is a subword of, say, g but not of \tilde{g} . Then for all ℓ the words $f(b; \ell)$ are subwords of g as well. Indeed, if there is a $j \neq k$ such that $f(b; j)$ is a subword of \tilde{g} but not of g , then the f -subword $f(b; k, j)$, defined analogously, is not a subword of either g or \tilde{g} . Because $|f(b; k, j)| \leq (3m + 1)/2 + 2$, this yields a contradiction for $m \leq 5$. (The cases $m \leq 4$ are covered by Remark 3.6.) The Claim is proved.

So we can assume that every $f(b; \ell)$ is a subsequence of, say, g . Therefore $t \leq T$, and one can construct t groups g_1^*, \dots, g_t^* of consecutive \hat{b} -runs in g such that the total length of the \hat{b} -runs within g_j^* is equal to r_j . Repeat the whole process for the subwords $g(a; i)$. It is possible that we had to substitute \tilde{f} for f , but this already implies that $t = T$. But from this equation it also follows that each $g(a; \ell)$ can be chosen to be a subword of f since, as we know, the subwords $f(a; i)$ can be found in g . Therefore we also have $r_i = R_i$ for all i and

$$f = \hat{a}^{i_0} \hat{b}^{r_1} \hat{a}^{i_1} \dots \hat{b}^{r_t} \hat{a}^{i_t} \quad \text{and} \quad g = \hat{a}^{j_0} \hat{b}^{r_1} \hat{a}^{j_1} \dots \hat{b}^{r_t} \hat{a}^{j_t}, \tag{5.3}$$

where the \hat{b} -runs with the same superscripts are identical. Furthermore, we also know that the number of non-empty \hat{a} -runs in f and g are equal as well. Indeed, if the multiset $\{i_0, i_r\}$ has no fewer non-zero elements than the multiset $\{j_0, j_r\}$, then the word containing one \hat{a} from the nonempty runs indexed by the first multiset and $f(b)$ establishes this relation. Therefore the number of non-empty \hat{a} -runs in f and g is the same, say r' : equal to $t - 1$, t or $t + 1$.

It remains to prove that $f(a) \simeq g(a)$ and that g can be written in a form such that $i_k = j_k$ for all possible k . (Note that if one must interchange g and \tilde{g} then we will show that in that case $f(b) = \tilde{f}(b)$.)

5.1. The Case $q = 1$

Let us start with the special case $q = 1$. Now without loss of generality we may assume that both words are written in the form where $\hat{b} = b$ (otherwise we can take the reverse complement form of the word). Now any subword of f containing the letter b should be contained in g in its original form because changing the subword into its reverse complement would change b into \bar{b} . Since $|f(a)| = |g(a)|$, $i_0 + i_1 = j_0 + j_1$.

If the multisets $\{i_0, i_1\}$ and $\{j_0, j_1\}$ were different, then there would exist a unique smallest element within them, say, the i_1 : we have $i_0 > j_0$, $j_1 > i_1$. Take a subword u of g of the form

$$u = b\hat{a}^{i_1+1}.$$

This subword clearly does not precede f (there are not enough \hat{a} 's after b in the word f). Since $|u| \leq (3m+1)/2 \leq 2m$, $m > 1$, therefore $D(f) \neq D(g)$, a contradiction. The ordered pairs (i_0, i_1) and (j_0, j_1) coincide. Denote by f_0 the longest simple subword of f ending with b and by f_1 the longest subword of f starting with b . The definitions of g_0 and g_1 are similar. Now f_0 and g_0 are words of the same length, and all their subwords of length $\leq 2m$, ending with b coincide as well. Denote by f_0^* and g_0^* the same words without their b terminuses. Then we know that all subwords of length $\lceil (|f_0^*| + 1)/2 \rceil$ of f_0^* and g_0^* are the same over the alphabet a, \bar{a} , in the simple subword relation. Application of Theorem 1.1 gives that $f_0^* = g_0^*$ in the original ordering. Furthermore, the same applies to f_1^* and g_1^* ; therefore we have proved that $f = g$.

From now on we assume that $1 < q \leq (3m+1)/2$. Therefore $|f(a)| = 3m+1-q \leq 3m-1$. Now considering the elements $\hat{a}^k \in D$ and applying Theorem 2.1 we get that

$$f(a) \simeq g(a).$$

The only remaining goal is to prove that the \hat{a} -structure of the words are the same, i.e., $i_k = j_k$ for all k .

5.2. The Case $1 < q \leq m+1$

Proposition 5.2. *If $1 < q \leq m+1$ and there are two indices $\ell \in \{0, \dots, t\}$ for which*

$$q + i_\ell > 2m, \tag{5.4}$$

then we have $t = 2$, $q = m+1$, $i_0 = i_1 = j_0 = j_1 = m$.

Proof. Indeed, if $q \leq m$ and if there are two distinct indices $k \neq l$ satisfying (5.4) then

$$q + i_l + q + i_k \geq 2m + 1 + 2m + 1;$$

therefore

$$q + i_l + i_k \geq 4m + 2 - q \geq 3m + 2 > |f|,$$

a contradiction.

If, however, $q = m+1$ and $i_0 = i_1 = m$, then $j_0 = j_1$ as well. Otherwise we would have, say, $j_0 < i_1 < j_1$. Then a g -subword consisting of one letter from the middle \hat{b} -run and $i_1 + 1$ letters from the j_1 -run is clearly shorter than $2m$ but does not precede f , a contradiction. Let us remark that in this case Proposition 5.1 is applicable directly, and Theorem 2.2 is proved. ■

If there is precisely one index ℓ satisfying (5.4), then the corresponding run will be called a *long* run, while the other runs are called *short*. Denote by $f^*(b; k)$ the f -subword consisting of all its \hat{b} 's and the complete k -th \hat{a} -run. For short runs the length of these words is at most $2m$; therefore these belong to $D(f) = D(g)$. Assume for a moment that $f(b) = g(b) \neq \widetilde{g(b)}$. Then $f^*(b; k)$ is not a subword of \widetilde{g} for any short run, and therefore we can find equality of the lengths of the short runs, i.e., $i_k = j_k$ for short runs. Furthermore, because of Proposition 5.2 (i) there is only one \hat{a} -run (the ℓ -th), whose length can not be ascertained from the subwords, but then $|i_\ell| = (3m + 1 - q) - \sum_{k \neq \ell} |i_k| = (3m + 1 - q) - \sum_{k \neq \ell} |j_k| = |j_\ell|$, which completes the proof in this case. Therefore from now on we assume that

$$f(b) = g(b) = \widetilde{g(b)}$$

holds as well. (We also know that $q = |f(b)|$ is even, but this is not important.)

Case 1. Assume at first that there is a long run in the word f and this is the ℓ -th one. Then g also has at least one long run. Indeed, let u_1 denote an $(2m - q)$ -letter subword of the long run. Then the f -subword $f(b) \cup u_1$ belongs to $D(g)$, and the image of u_1 is contained in a long \hat{a} -run of g . However, g cannot contain two long runs, otherwise Proposition 5.2 would apply, a contradiction. Therefore g contains exactly one long run and we may assume that f and g contain their respective long runs at the same index ℓ . Let us assume now that $\ell \neq t - \ell$. Then denote by f_ℓ^* the subword containing everything except the ℓ -th and $(t - \ell)$ -th \hat{a} -runs. This has at most $2m$ letters, and therefore belongs to $D(f)$: that is, it precedes the analogously defined g -subword g_ℓ^* . Similarly g_ℓ^* precedes f_ℓ^* . Consequently we know that $f_\ell^* \simeq g_\ell^*$. This means that

- (a) $f_\ell^* = g_\ell^*$, or
- (b) $f_\ell^* = \widetilde{g_\ell^*}$,

or both. But all the three possibilities imply that $i_\ell + i_{t-\ell} = j_\ell + j_{t-\ell}$. If (b) does not hold then there is a $k \neq \ell, t - \ell$ such that $f(b; k)$ is not a subword of $g(b; t - k)$. But since $i_{t-k} \neq 0$, the subword $f(b; k, t - \ell)$ (consisting of all \hat{b} 's and one element of the k -th and one element of the $(t - \ell)$ -th \hat{a} -runs each) which is not longer than $2m$, is therefore a subword of $g(b; k, t - \ell)$, and vice versa, which shows that Proposition 5.1 is applicable. If, however, (b) holds but (a) does not, then there is a k such that $f(b; k)$ is not a subword of $g(b; k)$. Then let u denote an $2m - q - i_k$ element subword of the long run in f . Let f^l be the word consisting of u and $f(b; k)$. This is not a subword of g but also not a subword of $\widetilde{g(b; t - k, t - \ell)}$ unless q is very close to m and $j_{t-\ell}$ is also close to m . But then we have a small run-number r and then there is a well recognizable subword of f with at most $2r + 1$ letters and repeating the previous reasoning we get the contradiction.

We now come to the case when $\ell = t - \ell$ and t is odd. But then if f_ℓ^* has at most $2m$ letters, which allows us to show as before that $f_\ell^* \simeq g_\ell^*$, and then we can apply Proposition 5.1 again. If this is not the case then we have $q = m + 1$ and $i_\ell = m$. If we have at least four non-empty \hat{a} -runs then for all $k \neq \ell$ we have $f(b; k, t - k) \simeq g(b; k, t - k)$, showing that $i_\ell = j_\ell$. Furthermore, it is impossible, as usual, that for k_1, k_2 we have $f(b; k_1, t - k_1) = g(b; k_1, t - k_1)$ while $f(b; k_2, t - k_2) = \widetilde{g(b; k_2, t - k_2)}$. (We can use the previous technique again.) So Proposition 5.1 is applicable again.

Case 2. Next suppose that there is no long run. Then all $f(b; k) \in D(f) = D(g)$. Assume that for all k the subword $f(b; k, t - k)$ has length $\leq 2m$. Then for all k we have $f(b; k, t - k) \simeq g(b; k, t - k)$. Moreover, as usual, we can show that if there is a k such that $f(b; k, t - k)$ is equal to $g(b; k, t - k)$ but not to its reverse complement; then for all other $l \neq k$ we also have $f(b; l, t - l) = g(b; l, t - l)$. Indeed, if this is not the case then there is a subword f_1 of $f(b; k, t - k)$ with at most $\lceil (i_k + i_{t-k})/2 \rceil$ letters from its \hat{a} -runs showing that $f(b; l, t - l) \neq \tilde{g}(b; l, t - l)$. Similarly, there is a subword f_2 of $f(b; l, t - l)$ with at most $\lceil (i_l + i_{t-l})/2 \rceil$ letters from its \hat{a} -runs showing that $f(b; l, t - l) \neq g(b; l, t - l)$. Putting together these two subwords we get a word from $D(f)$ which does not belong to $D(g)$, a contradiction, except that $q = m + 1$ and both \hat{a} -run pairs contain exactly $m - 1$ letters, where m is odd. But again, we can find a well recognizable word with ten letters, and repeating the whole process we are done.

So what remains is that we have an ℓ such that $q + i_\ell + i_{t-\ell} > 2m$. Then for all other $k \neq \ell, t - \ell$ we have $f(b; k, t - k) \simeq g(b; k, t - k)$. (Otherwise we have four non-empty \hat{a} -runs, and finding a well recognizable word with eight letters finishes the proof.) Again we can show that, say, $f(b; k, t - k)$ is equal to $g(b; k, t - k)$. Of course, we get that $i_\ell + i_{t-\ell} = j_\ell + j_{t-\ell}$. Then the multisets $\{i_\ell, i_{t-\ell}\}$ and $\{j_\ell, j_{t-\ell}\}$ are the same. Otherwise there would be a clear maximum, say i_ℓ and then $f(b; i_\ell)$ does not precede g , a contradiction. So we are done except that $i_\ell = j_{t-\ell} \neq j_\ell = i_{t-\ell}$. If for all $k \neq \ell, t - \ell$ we have $f(b; k, t - k) = \tilde{g}(b; k, t - k)$, then we can apply Proposition 5.1 to obtain $f = \tilde{g}$, or there is a k which does not satisfy this. As usual, we can construct a subword of f with $\lceil (i_k + i_{t-k})/2 \rceil + \lceil (i_\ell + i_{t-\ell})/2 \rceil$ letters from the respective \hat{a} -runs which does not precede g : a contradiction, except that again those four runs contain all the \hat{a} 's. Repeating the reasoning, we can construct a well recognizable word of length at most, say, 10. So the case $1 < q \leq m + 1$ is solved.

5.3. The Case $q > m + 1$

In this case we have $p = |f(a)| \leq 2m - 1$. Therefore any subword f_k consisting of $f(a)$ and an arbitrary letter from the k -th \hat{b} -run belongs to $D(f)$. If $f(a) \neq \tilde{f}(a)$ then it also means that for all k the subword f_k is a subword of g , and therefore for all k we have $i_k = j_k$. Proposition 5.1 completes the proof.

So we may assume that $f(a) = \tilde{f}(a)$. Suppose that there is a k such that f_k is a subword of g but not of \tilde{g} . Assume furthermore that there is an ℓ such that f_ℓ is a subword of \tilde{g} but not of g . (If this second subword does not exist then we already have that the lengths of the \hat{a} -runs in f and g are identical.) Let $f_{k,\ell}$ denote the “union” of the former two subwords, then it is a subword of f but not a subword either of g or of \tilde{g} . If $q > m + 2$ then $f_{k,\ell} \in D(f)$ therefore it is a contradiction and we are done. But $q = m + 2$ can not be true, otherwise $p = 2m - 1$ would hold, and therefore $f(a) \neq \tilde{f}(a)$, a contradiction. Theorem 2.2 is fully proved. ■

References

1. A.W.M. Dress and P.L. Erdős, Reconstructing words from subwords in linear time, *Ann. Combin.* **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, and P.S. White, Exordium for DNA Codes, *J. Comb. Optim.* **7** (4) (2003) 369–379.
3. V.I. Levenshtein, On perfect codes in deletion and insertion metric, *Discrete Math.* **3** (1) (1991) 3–20; Translation in *Discrete Math. Appl.* **2** (1992) 241–258.
4. V.I. Levenshtein, Efficient reconstruction of sequences from their subsequences or supersequences, *J. Combin. Theory Ser. A* **93** (2001) 310–332.
5. V.I. Levenshtein, Efficient reconstruction of sequences, *IEEE Trans. Inform. Theory* **47** (1) (2001) 2–22.
6. M. Lothaire, *Combinatorics on Words*, *Encyclopedia of Mathematics and its Applications* **17**, Addison-Wesley, Reading, Mass., 1983.
7. J. Manuch, Characterization of a word by its subwords, In: *Developments in Language Theory*, G. Rozenberg, et al. Ed., World Scientific Publ. Co., Singapore, (2000) pp. 210–219.
8. I. Simon, Piecewise testable events, *Lecture Notes in Comput. Sci.* **33** (1975) 214–222.