On a conjecture by Eriksson concerning overlap in strings

Cakir, I ; Chryssaphinou, O ; Månsson, M

Abstract: Consider a finite alphabet Ω and strings consisting of elements from Ω. For a given string w, let cor(w) denote the autocorrelation, which can be seen as a measure of the amount of overlap in w. Furthermore, let aw(n) be the number of strings of length n that do not contain w as a substring. Eriksson [4] stated the following conjecture: if cor(w)>cor(w), then aw(n)>aw (n) from the first n where equality no longer holds. We prove that this is true if [mid R:]Ω[mid R:][gt-or-equal, slanted]3, by giving a lower bound for aw(n)−aw (n).

DOI: https://doi.org/10.1017/S0963548399003806

Combinatorics, Probability and Computing (1999) 8, 429–440. Printed in the United Kingdom

On a Conjecture by Eriksson Concerning Overlap in Strings

ISA CAKIR[1], OURANIA CHRYSSAPHINOU[2]† and MARIANNE MÅNSSON[3]‡

[1] Abteilung für Angewandte Mathematik, Universität Zürich-Irchel, Winterthurerstrasse 190, CH 8057 Zürich, Switzerland
(e-mail: isa.cakir@winterthur.ch)

[2] Department of Mathematics, University of Athens, Panepistemiopolis, GR 157 84 Athens, Greece
(e-mail ocrysaf@atlas.uoa.gr)

[3] Department of Mathematics, Chalmers University of Technology, S 412 96 Göteborg, Sweden
(e-mail marianne@math.chalmers.se)

Received 5 August 1997; revised 30 June 1998

Consider a finite alphabet Ω and strings consisting of elements from Ω. For a given string w, let $\mathrm{cor}(w)$ denote the autocorrelation, which can be seen as a measure of the amount of overlap in w. Furthermore, let $a_w(n)$ be the number of strings of length n that do not contain w as a substring. Eriksson [4] stated the following conjecture: *if* $\mathrm{cor}(w) > \mathrm{cor}(w')$, *then* $a_w(n) > a_{w'}(n)$ *from the first* n *where equality no longer holds.* We prove that this is true if $|\Omega| \geqslant 3$, by giving a lower bound for $a_w(n) - a_{w'}(n)$.

1. Introduction

In this paper we consider a finite alphabet $\Omega = \{\omega_1, \ldots, \omega_q\}$ of size $q \geqslant 2$. A finite sequence of elements from the alphabet Ω is called a string. For a given string $w = (w_1, \ldots, w_k)$, $w_i \in \Omega$, $k \geqslant 2$, which we refer to as a pattern, we consider the number of strings of length n that do not contain w as a substring. Following the notation of Eriksson [4], we denote this number by $a_w(n)$ and say that these strings avoid w. Furthermore, we write $|w|$ for the length of the pattern w.

Guibas and Odlyzko [6] introduced the notion of autocorrelation of a pattern w. If $|w| = k$, it is defined to be the binary sequence $(b_k b_{k-1} \ldots b_1)$, where $b_i = 1$ if $w_j =$

† This work was done while the author was visiting Universität Zürich.

‡ This work was done while the author was visiting Universität Zürich, supported by the Swedish Natural Science Research Council (NFR).

$w_{k-i+j}, j = 1, \ldots, i$, that is, if there is an overlap of size i. This sequence can be viewed as a binary number, and with some abuse of notation we let $\mathrm{cor}(w)$ denote both the sequence and its value as a binary number. For example, if $\Omega = \{A, C, G, T\}$, that is, the DNA alphabet, and $w = AACATTAACA$, then $\mathrm{cor}(w) = (1000001001)$ and if it is viewed as a binary number, then $\mathrm{cor}(w) = 2^9 + 2^3 + 2^0$.

In Guibas and Odlyzko [6, 7] several results are derived in terms of autocorrelations and correlations (which concern the overlap between different patterns). One of their results is that asymptotically $a_w(n) \sim c_w \theta_w^n$, where $c_w, \theta_w > 0$ are constants that depend on the autocorrelation of w. Eriksson [4] shows that if $\mathrm{cor}(w) > \mathrm{cor}(w')$ then $\theta_w > \theta_{w'}$, and from these facts his main theorem follows: there exists an N such that $a_w(n) > a_{w'}(n)$, $n \geqslant N$, if and only if $\mathrm{cor}(w) > \mathrm{cor}(w')$. Furthermore, Eriksson [4] states a conjecture concerning the value of N which reads as follows: *if* $\mathrm{cor}(w) > \mathrm{cor}(w')$, *then* $a_w(n) > a_{w'}(n)$ *from the first n where equality no longer holds.*

In this paper we prove that this conjecture is true for $q \geqslant 3$. In fact, we prove more than the conjecture: we give a lower bound for $a_w(n) - a_{w'}(n)$. We also give the precise value of n for which $a_w(n) \neq a_{w'}(n)$ for the first time, if $q \geqslant 2$. In the case of $q = 2$, Eriksson [4] proved the conjecture in one special case and we have succeeded in another special case. The general case for $q = 2$ remains open.

Considering a sequence of independent random variables with a uniform distribution over Ω, we state the results in terms of probabilities. Finally, we discuss the connection between the autocorrelation of a pattern and the expected waiting time for its first occurrence in a random sequence.

2. Results

In this section the results are presented, while the proofs are deferred to Section 4. The main result of this paper is as follows.

Theorem 2.1. *Assume w and w' are patterns of length k with* $\mathrm{cor}(w) > \mathrm{cor}(w')$, *where* $\mathrm{cor}(w) = (b_k \ldots b_1)$ *and* $\mathrm{cor}(w') = (b'_k \ldots b'_1)$, *and let* $r = \max\{i : b_i \neq b'_i\}$.

(i) *Then*

$$a_w(n) = a_{w'}(n), \qquad n < 2k - r, \tag{2.1}$$

and

$$a_w(2k - r) = a_{w'}(2k - r) + 1. \tag{2.2}$$

(ii) *If $q \geqslant 3$, then, for $n \geqslant 2k - r$,*

$$a_w(n) - a_{w'}(n) > (q - 2) \sum_{i=1}^{n-1} a_w(i) - a_{w'}(i). \tag{2.3}$$

(iii) *If the patterns are of different lengths, $|w| = k$ and $|w'| = j$, $j < k$, then $2k - r$ in the above formulae should be replaced by j.*

Observe the assumption of Theorem 2.1(ii): $q \geqslant 3$. In the case of $q = 2$, Eriksson [4] proved the following proposition, which handles one special case.

Proposition 2.2 (Proposition 0.2, Eriksson [4]). *Assume $q = 2$ and that the autocorrelations of w and w' satisfy* $\mathrm{cor}(w) = 2^k$ *and* $\mathrm{cor}(w') = 2^k - 1$, *that is,* $|w| = k + 1$, $|w'| = k$, $\mathrm{cor}(w) = (1\underbrace{00\ldots00}_{k})$ *and* $\mathrm{cor}(w') = (\underbrace{11\ldots11}_{k})$. *Then* $a_w(n) - a_{w'}(n) = a_w(n-k)$ *for all* $n \geqslant k$.

We have succeeded in another special case.

Proposition 2.3. *If* $|w| = |w'| = 2$ *with* $\mathrm{cor}(w) = (11)$ *and* $\mathrm{cor}(w') = (10)$, *then, for* $q = 2$ *and* $n \geqslant 2$,

$$a_w(n) - a_{w'}(n) > a_w(n-1) - a_{w'}(n-1) + a_w(n-2) - a_{w'}(n-2).$$

We will now consider independent random variables $X_1, X_2, \ldots$, taking values in Ω with the probabilities $P(X_i = \omega_j) = 1/q$, for all i and $\omega_j \in \Omega$. Let $s_w(n)$ be the probability that w does not occur in the first n trials. Since the total number of outcomes of $X_1, \ldots, X_n$ is q^n, and among them $a_w(n)$ do not contain w, it follows that

$$s_w(n) = a_w(n)/q^n,$$

and we can state Theorem 2.1 in terms of probabilities.

Corollary 2.4. *Let the assumptions of Theorem 2.1 be in force.*

(i) *Then* $s_w(n) = s_{w'}(n)$, $n < 2k - r$, *and* $s_w(2k-r) = s_{w'}(2k-r) + 1/q^{2k-r}$.

(ii) *If* $q \geqslant 3$, *then, for* $n \geqslant 2k - r$,

$$s_w(n) - s_{w'}(n) > (q-2)\sum_{i=1}^{n-1}[s_w(i) - s_{w'}(i)]\, q^{-n+i}.$$

(iii) *If the patterns are of different lengths,* $|w| = k$ *and* $|w'| = j$, $j < k$, *then* $2k - r$ *in the above formulae should be replaced by* j.

Of course, Proposition 2.2 and Proposition 2.3 can also be formulated in probabilistic terms.

Let T_w denote the waiting time until the first occurrence of w, that is, if $|w| = k$ then T_w is the smallest i such that $X_{i-k+j} = w_j$, $j = 1, \ldots, k$. From Chryssaphinou, Papastavridis and Tsapelas [3], for example, we get the following formula for the expectation of T_w:

$$E[T_w] = \sum_{i=1}^{|w|} b_i q^i.$$

Hence $E[T_w] > E[T_{w'}]$ if and only if $\mathrm{cor}(w) > \mathrm{cor}(w')$ and the corollary below follows immediately.

Corollary 2.5. *The results of Theorem 2.1 and Corollary 2.4 hold if* $E[T_w] > E[T_{w'}]$.

For more literature concerning waiting times, we refer to Blom and Thorburn [1] and Li [8], for example, and, for the case of dependent X_is, to Chryssaphinou and Papastavridis [2], Gerber and Li [5] and Rudander [9].

3. Preparatory results and remarks

In this section we present some results, remarks and conventions that will be useful below.

Strings that are shorter than the pattern w can, of course, not include w as a substring. Also, all strings of the same length as w, except the pattern itself, avoid w. Hence, if $|w| = k$ then

$$a_w(n) = q^n, n = 1, \ldots, k-1, \quad \text{and} \quad a_w(k) = q^k - 1. \tag{3.1}$$

The following proposition, which is the key tool in this paper, gives a recurrence equation for $a_w(n)$ for all $n \geqslant 0$, when the convention that $a_w(0) = 1$ is used.

Proposition 3.1. *Assume that w is a pattern of length k with autocorrelation* $\text{cor}(w) = (b_k b_{k-1} \ldots b_1)$. *Then, for $n \geqslant 0$,*

$$a_w(n) = \sum_{i=1}^{k} b_i [q\, a_w(n+i-1) - a_w(n+i)]. \tag{3.2}$$

A proof of the above proposition can be found, for example, in Eriksson [4]. In (3.2) $a_w(n)$ is expressed in terms of 'future' values of $a_w(i)$, that is, $i > n$. For us, however, it is more convenient to express $a_w(n)$ in terms of $a_w(i)$, $i < n$, as in the following two equations, which follow immediately from (3.2). In the second one, we use the convention that $b_0 = 1$. This convention will be used repeatedly throughout the paper. For $n \geqslant k$,

$$a_w(n) = q\, a_w(n-1) - a_w(n-k) + \sum_{i=1}^{k-1} b_i [q\, a_w(n-k+i-1) - a_w(n-k+i)] \tag{3.3}$$

and

$$a_w(n) = \sum_{i=0}^{k-1} [q\, b_{i+1} - b_i] a_w(n-k+i). \tag{3.4}$$

Remark 3.2. Note that it follows immediately from the recurrence equations for $a_w(n)$ that $a_w(n) = a_{w'}(n)$ for all n if $\text{cor}(w) = \text{cor}(w')$. That the reverse implication holds follows from Theorem 2.1(i), which in turn is a consequence of the recurrence relation, as can be seen in its proof.

The number of strings of length $n+1$ for which w does not occur in the first n positions is equal to $q\, a_w(n)$. These strings can be divided into two groups: those that end with w and those that do not end with w. The number of strings in the latter of these groups is $a_w(n+1)$. Thus $q\, a_w(n) - a_w(n+1)$ is the number of strings of length $n+1$ ending with w and avoiding w in its first n positions. It is hence true that, for $n < k-1$,

$$q\, a_w(n) - a_w(n+1) = 0, \tag{3.5}$$

and, for $n \geqslant k-1$,

$$q\, a_w(n) - a_w(n+1) > 0. \tag{3.6}$$

Furthermore, we have the following lemma valid for $n \geqslant k-1$ and all $q \geqslant 2$.

Lemma 3.3. *If $q \geqslant 2$, then, for $n \geqslant k-1$,*

$$q\, a_w(n) - a_w(n+1) > (q-2) \sum_{i=1}^{n-1} q\, a_w(i) - a_w(i+1). \tag{3.7}$$

A proof of this lemma is given in the next section.

Remark 3.4. By the definition of autocorrelation it follows that, if a pattern $w = (w_1, \ldots, w_k)$ is of length k, then $b_k = 1$. Furthermore, if $b_{k-1} = 1$, then

$$\mathrm{cor}(w) = (\overbrace{11\ldots11}^{k}),$$

since, if $b_{k-1} = 1$, then $w_2 = w_1$, $w_3 = w_2$, …, $w_k = w_{k-1}$, by the definition of b_{k-1}, and hence $w_1 = w_2 = \cdots = w_k$, which implies $b_1 = \cdots = b_k = 1$.

4. Proofs

In this section proofs of the results in the previous sections are given. To simplify the notation, we will let

$$a(n) = a_w(n), \quad a'(n) = a_{w'}(n), \quad \Delta(n) = a_w(n) - a_{w'}(n),$$

$$f(n) = q\, a_w(n) - a_w(n+1) \quad \text{and} \quad f'(n) = q\, a_{w'}(n) - a_{w'}(n+1), \tag{4.1}$$

when this is more convenient.

Proof of Lemma 3.3. By (3.6) the inequality (3.7) is true if $q = 2$. In the rest of the proof we assume that $q \geqslant 3$, and prove the inequality (3.7) by induction on n. By (3.1) we first note that (3.7) obviously holds for $n = k-1$:

$$1 = q\, a(k-1) - a(k) > (q-2) \sum_{i=0}^{k-2} q\, a(i) - a(i+1) = 0.$$

Assume that (3.7) is true for all m such that $k-1 \leqslant m < n$. By (3.4) the difference $f(n)$ can be written as

$$\begin{aligned} f(n) &= q \sum_{i=0}^{k-1} [q\, b_{i+1} - b_i] a(n-k+i) - \sum_{i=0}^{k-1} [q\, b_{i+1} - b_i] a(n-k+i+1) \\ &= \sum_{i=0}^{k-1} [q\, b_{i+1} - b_i] f(n-k+i). \end{aligned} \tag{4.2}$$

We will distinguish between the cases where $b_{k-1} = 0$ and $b_{k-1} = 1$. Assume first that $b_{k-1} = 0$ and recall that $b_k = 1$ (Remark 3.4), in which case $q\,b_k - b_{k-1} = q$. It is clear that

$$\begin{aligned}[q\,b_k - b_{k-1}]f(n-1) &= [q-2+2]f(n-1)\\ &= (q-2)f(n-1) + 2f(n-1).\end{aligned}$$

Apply the induction hypothesis on the rightmost part of the above equation to get

$$[q\,b_k - b_{k-1}]f(n-1) > (q-2)f(n-1) + 2(q-2)\sum_{i=1}^{n-2} f(i),$$

which inserted in (4.2) yields

$$f(n) > (q-2)f(n-1) + 2(q-2)\sum_{i=1}^{n-2} f(i) + \sum_{i=0}^{k-2}[q\,b_{i+1} - b_i]f(n-k+i). \tag{4.3}$$

The troublesome part of (4.3) is $[q\,b_{i+1} - b_i]f(n-k+i)$, since $q\,b_{i+1} - b_i$ can be negative, while $f(n-k+i) \geqslant 0$ by (3.5) and (3.6). To handle this, each such term will be considered together with the corresponding term in the other sum in (4.3). By the assumption that $q \geqslant 3$, we have that $2(q-2) + q\,b_{i+1} - b_i \geqslant 2q - 5 = (q-2) + (q-3) \geqslant q-2$. Thus

$$\begin{aligned}f(n) &> (q-2)f(n-1) + 2(q-2)\sum_{i=1}^{n-k-1} f(i) + \sum_{i=0}^{k-2}[2(q-2) + q\,b_{i+1} - b_i]f(n-k+i)\\ &\geqslant (q-2)\sum_{i=1}^{n-1} f(i),\end{aligned}$$

and the lemma is proved for the case of $b_{k-1} = 0$.

If $b_{k-1} = 1$, then $\mathrm{cor}(w) = (\overbrace{11\ldots11}^{k})$ by Remark 3.4, and minor changes in the argument are needed. Inserting $q\,b_{i+1} - b_i = q-1$, for all i, in (4.2) yields

$$f(n) = \sum_{i=0}^{k-1}(q-1)f(n-k+i). \tag{4.4}$$

Now we apply the induction hypothesis on the summand $(q-1)f(n-k)$:

$$\begin{aligned}(q-1)f(n-k) &= (q-2)f(n-k) + f(n-k)\\ &> (q-2)f(n-k) + (q-2)\sum_{i=1}^{n-k-1} f(i).\end{aligned}$$

Inserting this bound in (4.4) gives the result for this case. □

Now the main theorem, Theorem 2.1, will be proved. This is done mainly by means of Lemma 3.3 and ideas similar to those used in its proof.

Proof of Theorem 2.1.

(i) From (3.1) we know that $a(n) = a'(n), n = 1, \ldots, k$. We will now show that equality also holds for $n = k+1, \ldots, 2k-r-1$. Pick an $n \in \{k+1, \ldots, 2k-r-1\}$ and make the induction hypothesis that $a(m) = a'(m)$ for all $m < n$.

Using (3.3) and that $f(i) = 0$ for $i < k-1$ by (3.5), we get

$$\begin{aligned} a(n) &= q\,a(n-1) - a(n-k) + \sum_{i=1}^{k-1} b_i f(n-k+i-1) \\ &= q\,a(n-1) - a(n-k) + \sum_{i=2k-n}^{k-1} b_i f(n-k+i-1). \end{aligned}$$

This sum involves only $b_i, i \geqslant 2k-n > r$, and $a(i)$, $i < n$. These b_is satisfy $b_i = b'_i$ and $a(i) = a'(i)$, $i < n$, by hypothesis. Hence

$$a(n) = q\,a'(n-1) - a'(n-k) + \sum_{i=2k-n}^{k-1} b'_i f'(n-k+i-1).$$

Furthermore, since $f'(i) = 0$, $i < k-1$, by (3.5), we get that $a(n) = a'(n)$.

The next step is to show that $a(2k-r) = a'(2k-r)+1$. Using (3.3) and, as above, that $f(i) = f'(i) = 0$, $i < k-1$, we get

$$\begin{aligned} a(2k-r) &= q\,a(2k-r-1) - a(k-r) + \sum_{i=1}^{k-1} b_i f(k-r+i-1) \\ &= q\,a(2k-r-1) - a(k-r) + b_r f(k-1) + \sum_{i=r+1}^{k-1} b_i f(k-r+i-1). \end{aligned}$$

Since $f(k-1) = 1$ by (3.1), $b_r = 1$, $b'_r = 0$ and $a(n) = a'(n)$, $n < 2k-r$, we have

$$\begin{aligned} a(2k-r) &= f(k-1) + a'(2k-r) \\ &= a'(2k-r) + 1, \end{aligned}$$

which completes the proof of (i).

(ii) From (i) it follows that

$$1 = a(2k-r) - a'(2k-r) > (q-2) \sum_{i=1}^{2k-r-1} a(i) - a'(i) = 0,$$

so (2.3) is true for $n = 2k-r$. To prove the case of an arbitrary n, we will use induction and make the assumption that (2.3) is true for all m such that $2k-r \leqslant m < n$.

First we will consider the case where $b_{k-1} = b'_{k-1} = 0$. Note that this assumption implies that $k \geqslant 3$. We use (3.4) and the fact that $b_i = b'_i$, $i = r+1, \ldots, k$, to express the difference as

$$\begin{aligned} a(n) - a'(n) &= \sum_{i=0}^{k-1} [q\,b_{i+1} - b_i] a(n-k+i) - \sum_{i=0}^{k-1} [q\,b'_{i+1} - b'_i] a'(n-k+i) \\ &= \sum_{i=r+1}^{k-1} [q\,b_{i+1} - b_i] \left\{ a(n-k+i) - a'(n-k+i) \right\} \end{aligned} \tag{4.5}$$

$$+\sum_{i=0}^{r}[q\,b_{i+1}-b_i]a(n-k+i)-\sum_{i=0}^{r}[q\,b'_{i+1}-b'_i]a'(n-k+i). \tag{4.6}$$

The lines (4.5) and (4.6) in the equation above will be considered separately, and we denote them by (A) and (B), respectively. Recall that $\Delta(i)=a(i)-a'(i)$. Since $b_k=1$ and $b_{k-1}=0$, it follows that

$$\text{(A)} \quad = \quad q\,\Delta(n-1)+\sum_{i=r+1}^{k-2}[q\,b_{i+1}-b_i]\Delta(n-k+i), \tag{4.7}$$

and, by the induction hypothesis,

$$\begin{aligned} q\,\Delta(n-1) &= (q-2)\Delta(n-1)+2\Delta(n-1) \\ &> (q-2)\Delta(n-1)+2(q-2)\sum_{i=1}^{n-2}\Delta(i). \end{aligned} \tag{4.8}$$

Combining relations (4.7) and (4.8) and arguing as in the proof of Lemma 3.3, we get

$$\begin{aligned} \text{(A)} &> (q-2)\Delta(n-1)+2(q-2)\sum_{i=1}^{n-2}\Delta(i)+\sum_{i=r+1}^{k-2}[q\,b_{i+1}-b_i]\Delta(n-k+i) \\ &= (q-2)\Delta(n-1)+2(q-2)\sum_{i=1}^{n-k+r}\Delta(i)+\sum_{i=r+1}^{k-2}[2(q-2)+q\,b_{i+1}-b_i]\Delta(n-k+i) \\ &\geqslant (q-2)\Delta(n-1)+(q-2)\sum_{i=1}^{n-k+r}\Delta(i)+(q-2)\sum_{i=1}^{n-k+r}\Delta(i)+(q-2)\sum_{i=n-k+r+1}^{n-2}\Delta(i) \\ &= (q-2)\sum_{i=1}^{n-1}\Delta(i)+(q-2)\sum_{i=1}^{n-k+r}\Delta(i). \end{aligned} \tag{4.9}$$

Note that the goal is to show that (A) + (B) $> (q-2)\sum_{i=1}^{n-1}\Delta(i)$. The latter sum on the right-hand side of (4.9), which we know is positive by the induction hypothesis, will be used to handle (B), which might be negative, as we shall see.

The next step is to rewrite (B) in a more tractable way, as follows.

$$\begin{aligned} \text{(B)} &= q\,b_{r+1}a(n-k+r)-b_0a(n-k)+\sum_{i=1}^{r}b_if(n-k+i-1) \\ &\quad -\left(q\,b'_{r+1}a'(n-k+r)-b'_0a(n-k)+\sum_{i=1}^{r}b'_if'(n-k+i-1)\right). \end{aligned}$$

Using $b_0=b'_0=1$, $b_r=1$, $b'_r=0$, $b_{r+1}=b'_{r+1}$, $b_1,\ldots,b_{r-1}\geqslant 0$, $b'_1,\ldots,b'_{r-1}\leqslant 1$, it follows that

$$\text{(B)} \geqslant q\,b_{r+1}\Delta(n-k+r)-\Delta(n-k)+f(n-k+r-1)-\sum_{i=1}^{r-1}f'(n-k+i-1). \tag{4.10}$$

By Lemma 3.3,

$$\begin{aligned}\sum_{i=1}^{r-1} f'(n-k+i-1) &< \frac{1}{q-2} f'(n-k+r-1) \\ &\leqslant f'(n-k+r-1),\end{aligned}$$

which together with (4.10) yields

$$\begin{aligned}\text{(B)} &> q\,b_{r+1}\Delta(n-k+r) - \Delta(n-k) + f(n-k+r-1) - f'(n-k+r-1) \\ &= [q\,b_{r+1}-1]\Delta(n-k+r) - \Delta(n-k) + q\,\Delta(n-k+r-1). \qquad (4.11)\end{aligned}$$

Furthermore, $q\,b_{r+1}-1 \geqslant -1$ and $\Delta(n-k+r-1) \geqslant 0$ by the induction hypothesis, so that

$$\text{(B)} \geqslant -\Delta(n-k+r) - \Delta(n-k). \qquad (4.12)$$

This expression is clearly negative, but recall that we have an 'extra' contribution from (4.9), which is positive. Summing (4.9) and (4.12) concludes the proof of (ii) in the case where $b_{k-1} = b'_{k-1} = 0$:

$$\begin{aligned}\Delta(n) &> (q-2)\sum_{i=1}^{n-1}\Delta(i) + (q-2)\sum_{i=1}^{n-k+r}\Delta(i) - \Delta(n-k+r) - \Delta(n-k) \\ &= (q-2)\sum_{i=1}^{n-1}\Delta(i) + (q-2)\sum_{i=1,i\neq n-k}^{n-k+r-1}\Delta(i) + (q-3)[\Delta(n-k+r) + \Delta(n-k)] \\ &> (q-2)\sum_{i=1}^{n-1}\Delta(i). \qquad (4.13)\end{aligned}$$

What remains of the proof of (ii) is to examine the special case where $b_{k-1} = 1$. This means that $b_i = 1$ for all $i = 1,\ldots,k$ by Remark 3.4. Furthermore, $b'_{k-1} = 0$; otherwise the autocorrelations would be the same.

What we need to show is that $\Delta(n) > (q-2)\sum_{i=1}^{n-1}\Delta(i)$ when the hypothesis that this inequality holds for all m such that $2k - r \leqslant m < n$ is made. We use (3.3) and the fact that $b_1 = \ldots = b_{k-1} = 1$ and $b'_{k-1} = 0$ to write

$$\begin{aligned}\Delta(n) &= q\,a(n-1) - a(n-k) + \sum_{i=1}^{k-1} f(n-k+i-1) \\ &\quad - \left(q\,a'(n-1) - a'(n-k) + \sum_{i=1}^{k-2} b'_i f'(n-k+i-1)\right).\end{aligned}$$

Using $b'_i \leqslant 1$, that $(q-2)\sum_{i=1}^{k-2} f'(n-k+i-1) < f'(n-2)$ by Lemma 3.3, and that $\sum_{i=1}^{k-1} f(n-k+i-1) \geqslant f(n-2)$, since $f(i) \geqslant 0$ for all i by (3.5) and (3.6), yields

$$\begin{aligned}\Delta(n) &> q\,\Delta(n-1) - \Delta(n-k) + f(n-2) - f'(n-2) \\ &= (q-1)\Delta(n-1) - \Delta(n-k) + q\,\Delta(n-2). \qquad (4.14)\end{aligned}$$

Now we need to distinguish between the cases $k = 2$ and $k \geqslant 3$. When $k \geqslant 3$, we apply the induction hypothesis on $2\Delta(n-2)$ to conclude the proof:

$$\begin{aligned}\Delta(n) &> (q-1)\Delta(n-1) - \Delta(n-k) + (q-2)\Delta(n-2) + 2(q-2)\sum_{i=1}^{n-3}\Delta(i)\\ &\geqslant (q-2)\sum_{i=1}^{n-1}\Delta(i) + (q-3)\Delta(n-k) + (q-2)\sum_{i=1, i\neq n-k}^{n-3}\Delta(i)\\ &\geqslant (q-2)\sum_{i=1}^{n-1}\Delta(i).\end{aligned}$$

If $k = 2$ it follows from (4.14) that

$$\Delta(n) > (q-1)\Delta(n-1) + (q-1)\Delta(n-2). \tag{4.15}$$

Observe that this inequality also holds for $q = 2$. Now the induction hypothesis can be applied to either of the terms in (4.15); we choose the latter one to get

$$\begin{aligned}\Delta(n) &> (q-1)\Delta(n-1) + (q-2)\Delta(n-2) + (q-2)\sum_{i=1}^{n-3}\Delta(i)\\ &\geqslant (q-2)\sum_{i=1}^{n-1}\Delta(i),\end{aligned}$$

which ends the proof of Theorem 2.1(ii).

(iii) What remains of the proof of Theorem 2.1 is to show the results corresponding to (i) and (ii) in the case of different lengths of the patterns: $|w| = k$ and $|w'| = j < k$. This case requires a somewhat different technique, but the first step is as before: by (3.1), $a(n) = a'(n)$ for $n < j$, and for $n = j$ we have $a(j) = q^j$, while $a'(j) = q^j - 1$. Hence the equalities in (2.1) and (2.2) hold if $2k - r$ is replaced by j.

As usual the proof proceeds with induction, and by the above the basic step follows:

$$1 = a(j) - a'(j) > (q-2)\sum_{i=0}^{j-1} a(i) - a'(i) = 0.$$

We assume that (2.3) is true for all m, $j \leqslant m < n$, and will show that it then holds for n. First we will consider the case where $|w| = k$, $|w'| = k-1$, $\mathrm{cor}(w) = (1\underbrace{00\ldots00}_{k-1})$ and $\mathrm{cor}(w') = (\underbrace{11\ldots11}_{k-1})$.

Since $b'_1 = \cdots = b'_{k-1} = 1$,

$$a'(n) = \sum_{i=0}^{k-2}(q-1)\,a'(n-k+1+i),$$

by (3.4). Hence

$$a'(n) - a'(n-1) = (q-1)\,a'(n-1) - (q-1)\,a'(n-k),$$

which yields

$$a'(n) \quad = \quad q\,a'(n-1) - (q-1)\,a'(n-k). \tag{4.16}$$

Furthermore, $b_1 = \cdots = b_{k-1} = 0$ so that, by (3.3),

$$a(n) \quad = \quad q\,a(n-1) - a(n-k). \tag{4.17}$$

Using (4.16), (4.17) and applying the induction hypothesis on $2\Delta(n-1)$ yields

$$\begin{aligned}
\Delta(n) &= (q-2)\Delta(n-1) + 2\Delta(n-1) - \Delta(n-k) + (q-2)a'(n-k) \\
&> (q-2)\Delta(n-1) - \Delta(n-k) + (q-2)a'(n-k) + 2(q-2)\sum_{i=1}^{n-2}\Delta(i) \\
&= (q-2)\sum_{i=1}^{n-1}\Delta(i) + (q-3)\Delta(n-k) + (q-2)a'(n-k) + (q-2)\sum_{i=1, i\neq n-k}^{n-2}\Delta(i) \\
&\geqslant (q-2)\sum_{i=1}^{n-1}\Delta(i).
\end{aligned} \tag{4.18}$$

In the general case we assume $|w| = k$ and $|w'| = j < k$. Choose patterns v_i and v_i', $i = 0, \ldots, k-j$ with autocorrelations $\mathrm{cor}(v_i) = (1\underbrace{00\ldots00}_{k-i})$ and $\mathrm{cor}(v_i') = (\underbrace{11\ldots11}_{k-i})$. Note that such patterns always exist. Then $a(n) - a'(n)$ can be written in the form of a telescoping sum, as follows:

$$\begin{aligned}
a(n) - a'(n) &= a(n) - a_{v_1}(n) + \sum_{i=1}^{k-j}(a_{v_i}(n) - a_{v_i'}(n)) \\
&\quad + \sum_{i=1}^{k-j-1}(a_{v_i'}(n) - a_{v_{i+1}}(n)) + a_{v_{k-j}'}(n) - a'(n).
\end{aligned}$$

By (4.18),

$$a_{v_i}(n) - a_{v_i'}(n) > (q-2)\sum_{j=1}^{n-1} a_{v_i}(i) - a_{v_i'}(i),$$

$i = 1, \ldots, k-j$. Furthermore, w and v_1 are of the same lengths, which also holds for v_i' and v_{i+1}, $i = 1, \ldots, k-j-1$, and for v_{k-j}' and w', so the other summands can be handled by Theorem 2.1(ii) and we finally get

$$\begin{aligned}
a(n) - a'(n) &> (q-2)\sum_{j=1}^{n-1}\left\{a(j) - a_{v_1}(j) + \sum_{i=1}^{k-j}\left(a_{v_i}(j) - a_{v_i'}(j)\right)\right. \\
&\quad \left. + \sum_{i=1}^{k-j-1}(a_{v_i'}(j) - a_{v_{i+1}}(j)) + a_{v_{k-j}'}(j) - a'(j)\right\} \\
&= (q-2)\sum_{j=1}^{n-1} a(j) - a'(j).
\end{aligned}$$

□

Proof of Proposition 2.3. This follows directly from (4.15) and the observation following it. □

Acknowledgement

The authors would like to thank Niklas Engsner for valuable comments on the manuscript.

References

[1] Blom, G. and Thorburn, D. (1982) How many random digits are required until given sequences are obtained? *J. Appl. Probab.* **19** 518–531.
[2] Chryssaphinou, O. and Papastavridis, S. (1990) The occurrence of sequence patterns in repeated dependent experiments. *Th. Probab. Appl.* **35** 145–152.
[3] Chryssaphinou, O., Papastavridis, S. and Tsapelas, T. (1994) On the waiting time of appearance of given patterns. In *Runs and Patterns in Probability* (A. Godbole and S. Papastavridis, eds), Kluwer, pp. 231–241.
[4] Eriksson, K. (1997) Autocorrelation and the enumeration of strings avoiding a fixed string. *Combinatorics, Probability and Computing* **6** 45–48.
[5] Gerber, H. U. and Li, S. R. (1981) The occurrence of sequence patterns in repeated experiments and hitting times in a Markov chain. *Stoch. Proc. Appl.* **11** 101–108.
[6] Guibas, L. J. and Odlyzko, A. M. (1981) String overlaps, pattern matching and nontransitive games. *J. Combin. Theory Ser. A* **30** 183–200.
[7] Guibas, L. J. and Odlyzko, A. M. (1981) Periods in strings. *J. Combin. Theory Ser. A* **30** 19–42.
[8] Li, S. R. (1980) A martingale approach to the study of occurrence of sequence patterns in repeated experiments. *Ann. Probab.* **8** 1171–1176.
[9] Rudander, J. (1996) On the first occurrence of a given pattern in a semi-Markov process. Vol. 2 of *Uppsala Dissertations in Mathematics*, ISSN 1401-2049, Uppsala University.