Combinatorics, Probability and Computing (2002) 11, 393–402. © 2002 Cambridge University Press
DOI: 10.1017/S0963548302005187 Printed in the United Kingdom

Pattern Avoidance and Overlap in Strings

MARIANNE MÅNSSON

Department of Mathematics, Chalmers University of Technology,SE 412 96 Göteborg, Sweden
(e-mail: marianne@math.chalmers.se)

Received 1 May 2001; revised 12 August 2001

Consider a finite alphabet Ω and patterns which consist of characters from Ω. For a given pattern w, let $\mathrm{cor}(w)$ denote its autocorrelation, which can be seen as a measure of the amount of overlap in w. Letting $a_w(n)$ denote the number of strings over Ω of length n which do not contain w as a substring, the main result of this paper is: If $\mathrm{cor}(w) > \mathrm{cor}(w')$ then $a_w(n) - a_{w'}(n) > (|\Omega| - 1)(a_w(n-1) - a_{w'}(n-1))$ for $n \geqslant N$, and the value of N is given. This result confirms a conjecture by Eriksson [2], which was previously proved to be true by Cakir, Chryssaphinou and Månsson [1] when $|\Omega| \geqslant 3$.

1. Introduction and main result

Let Ω denote a finite alphabet of size $q \geqslant 2$, and call any finite sequence composed of characters from the alphabet Ω a string. For a given string $w = w_1 \cdots w_k$, $w_i \in \Omega$, which will be referred to as a pattern, let $a_w(n)$ denote the number of strings of length n, which do not contain w as a substring of consecutive characters. We say that these strings avoid w. In this paper we will consider how the structure of the pattern w affects $a_w(n)$.

Guibas and Odlyzko [3] introduced the notion of autocorrelation of a pattern w. If $|w| = k$, where $|w|$ denotes the length of w, it is defined to be the binary sequence $b_k b_{k-1} \cdots b_1$, where $b_i = 1$ if $w_j = w_{k-i+j}$, $j = 1, \ldots, i$, that is, if there is an overlap of size i. Sometimes it is convenient to view this sequence as a number in some base, and with some abuse of notation we let $\mathrm{cor}(w)$ denote both the sequence and its numerical value in, say, base 2. For example, if $\Omega = \{H, T\}$, and $w = TTHTHTTHT$, then $\mathrm{cor}(w) = 100001001$, and if it is viewed as a binary number, then $\mathrm{cor}(w) = 529$. Autocorrelation can be seen as a measure of the amount of overlap in w.

By means of generating functions, Guibas and Odlyzko [3] derived the following recurrence equation for $a_w(n)$, where the convention $a_w(0) = 1$ is used:

If $|w| = k$ *and* $\mathrm{cor}(w) = b_k b_{k-1} \cdots b_1$, *then, for* $n \geqslant 0$,

$$a_w(n) = \sum_{i=1}^{k} b_i [q\, a_w(n+i-1) - a_w(n+i)]. \tag{1.1}$$

https://doi.org/10.1017/S0963548302005187 Published online by Cambridge University Press

Furthermore, Guibas and Odlyzko [3] showed 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 [2], who gave another, combinatorial, proof of (1.1), also proved that there exists an N such that $a_w(n) > a_{w'}(n)$, for $n \geqslant N$, if and only if $\mathrm{cor}(w) > \mathrm{cor}(w')$. Furthermore, Eriksson [2] states the following conjecture concerning the value of N:

If $\mathrm{cor}(w) > \mathrm{cor}(w')$, *then* $a_w(n) > a_{w'}(n)$ *from the first* n *where equality no longer holds.*

Cakir, Chryssaphinou and Månsson [1] proved that Eriksson's conjecture is true, by giving a lower bound on $a_w(n) - a_{w'}(n)$, however, only under the annoying assumption that $q \geqslant 3$:

If w *and* w' *are patterns of length* k *with* $\mathrm{cor}(w) > \mathrm{cor}(w')$, *where* $\mathrm{cor}(w) = b_k \cdots b_1$ *and* $\mathrm{cor}(w') = b'_k \cdots b'_1$, *and* $r = \max\{i : b_i \neq b'_i\}$, *then*

$$a_w(n) - a_{w'}(n) = \begin{cases} 0, & \text{if } n < 2k - r, \\ 1, & \text{if } n = 2k - r, \end{cases} \tag{1.2}$$

and if $q \geqslant 3$, *then, for* $n \geqslant 2k - r$,

$$a_w(n) - a_{w'}(n) > (q-2) \sum_{i=1}^{n-1} [a_w(i) - a_{w'}(i)]. \tag{1.3}$$

If the patterns are of different lengths, $|w| = k$ *and* $|w'| = j$, $j < k$, *then the formulae above hold with* $2k - r$ *replaced by* j *throughout.*

The proof of (1.2) in Cakir, Chryssaphinou and Månsson [1] is a simple use of (1.1), while the proof of (1.3) is more involved. The case where $q = 2$ was left open except for some special cases. In the present paper, we show that the conjecture of Eriksson is, as expected, true also when $q = 2$.

Theorem 1.1. *If* w *and* w' *are patterns of length* k *with* $\mathrm{cor}(w) > \mathrm{cor}(w')$, *and* $r = \max\{i : b_i \neq b'_i\}$, *then, for* $n \geqslant 2k - r$,

$$a_w(n) - a_{w'}(n) > (q-1)[a_w(n-1) - a_{w'}(n-1)]. \tag{1.4}$$

If the patterns are of different lengths, $|w| = k$ *and* $|w'| = j$, $j < k$, *then* (1.4) *holds for* $n \geqslant j$.

Before proving this theorem, we need some results and observations on the avoidance of a pattern. An extensive list of references on results on the occurrence of patterns can be found in Régnier and Szpankowski [4].

2. Results on the avoidance of a pattern

Strings which 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

$$a_w(n) = q^n, n = 1, \ldots, k-1, \quad \text{and} \quad a_w(k) = q^k - 1. \tag{2.1}$$

The number of strings of length n for which w does not occur in the first $n-1$ positions

https://doi.org/10.1017/S0963548302005187 Published online by Cambridge University Press

is equal to $q\,a_w(n-1)$. 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 $a_w(n)$. Thus $q\,a_w(n-1)-a_w(n)$ is the number of strings of length n ending with w, that is, in positions $n-k+1$ to n, and avoiding w in its first $n-1$ positions. Hence it is true that

$$q\,a_w(n-1)-a_w(n)\quad\begin{cases}=0, & \text{if } n\leqslant k-1,\\ =1, & \text{if } n=k,\\ >0, & \text{if } n\geqslant k.\end{cases}\tag{2.2}$$

Now, let us introduce the notation $h_w(n)=q\,a_w(n-1)-a_w(n)$. Then, by (2.2), $qh_w(n-1)-h_w(n)=0$ for $n<k$, and $qh_w(k-1)-h_w(k)=-1$. Above we considered what happened when adding a character at the end of the strings which avoid w. We now repeat this argument for the strings in which w occurs for the first time at the very end, by considering what happens when a character is added in the beginning. For $n>k$, $h_w(n-1)>0$, and then $qh_w(n-1)$ can be interpreted as the number of strings of length n ending with w and avoiding w in positions $2,\ldots,n-1$. As above, these strings can be divided into two groups: those which start with w and those which do not start with w. The number of strings in the latter group is $h_w(n)$. Hence $qh_w(n-1)-h_w(n)$, $n>k$, is the number of strings of length n which both begin and end with w, but avoid w in all other places, which is certainly a nonnegative number.

Using the convention that $b_0=1$, let

$$s=\max_{0\leqslant j\leqslant k-1}\{j : b_j=1\}.\tag{2.3}$$

The shortest possible string of length $>k$ which both begins and ends with w is of length $2k-s$, and there obviously exists only one such string. To summarize,

$$qh_w(n-1)-h_w(n)\quad\begin{cases}=0, & \text{if } n<k,\\ =-1, & \text{if } n=k,\\ =0, & \text{if } k<n<2k-s,\\ =1, & \text{if } n=2k-s,\\ \geqslant 0, & \text{if } n>2k-s.\end{cases}\tag{2.4}$$

Note that (2.4) can be verified formally by using (1.1).

In Cakir, Chryssaphinou and Månsson [1], the proof of (1.3) was divided into the cases where $b_{k-1}=0$ and $b_{k-1}=1$. In the case where $b_{k-1}=1$, the proof relied on the fact that if $b_{k-1}=1$ then $b_1=\cdots=b_k=1$; the implication follows since, by definition of b_{k-1}, $w_1=w_2$, $w_2=w_3,\ldots,w_{k-1}=w_k$, and hence $w_1=w_2=\cdots=w_k$. In the present paper the autocorrelation structure plays a greater role. We use the following observation, where $\lfloor\cdot\rfloor$ denotes the integer part:

$$\begin{aligned}&\text{if}\quad b_j=1\quad\text{for some}\quad j\in\{0,1,\ldots,k-1\},\\ &\text{then}\quad b_{k-(k-j)t}=1\quad\text{for all}\quad t\in\{1,\ldots,\lfloor\tfrac{k}{k-j}\rfloor\},\end{aligned}\tag{2.5}$$

which follows by the definition of autocorrelation, as in the case where $b_{k-1}=1$.

A key tool in both Cakir, Chryssaphinou and Månsson [1] and the present paper is the recurrence equation for $a_w(n)$ given in (1.1), where $a_w(n)$ is expressed in terms of 'future'

https://doi.org/10.1017/S0963548302005187 Published online by Cambridge University Press

values of $a_w(i)$, that is, $i > n$. Here, however, it is more convenient to express $a_w(n)$ in terms of $a_w(i)$, $i < n$, as in the following equation, which follows directly from (1.1). For $n \geqslant k$,

$$a_w(n) = q\,a_w(n-1) - a_w(n-k) + \sum_{i=1}^{k-1} b_i[q\,a_w(n-k+i-1) - a_w(n-k+i)]. \tag{2.6}$$

Note that it follows immediately from this recurrence equation and (2.1) that $a_w(n) = a_{w'}(n)$, for all n, if $\mathrm{cor}(w) = \mathrm{cor}(w')$.

We shall use the convention that $h_w(n) = 0$, $n = 0, -1, -2, \ldots$, and, to simplify the notation, we drop the subscript in h_w when there is no risk of confusion.

Lemma 2.1. *Assume that w is a pattern of length k with autocorrelation $\mathrm{cor}(w) = b_k b_{k-1} \cdots b_1$, and let s be defined by* (2.3). *Then, for $n \geqslant 1$,*

$$q\,a_w(n-1) - a_w(n) \geqslant \begin{cases} (q-1)\displaystyle\sum_{i=1}^{k-s}[q\,a_w(n-1-i) - a_w(n-i)], & \text{if } 0 < s < k, \qquad (2.7) \\ (q-1)\displaystyle\sum_{i=1}^{k-1}[q\,a_w(n-1-i) - a_w(n-i)], & \text{if } s = 0. \qquad (2.8) \end{cases}$$

Proof. The proof is divided into six parts:

(i) $n \leqslant k$,	(ii) $k < n < 2k - s$,
(iii) $n = 2k - s$,	(iv) $2k - s < n < 2k$, $1 \leqslant s \leqslant k - 1$,
(v) $n \geqslant 2k$, $1 \leqslant s \leqslant k - 1$,	(vi) $n > 2k$, $s = 0$.

(i) By (2.2), $h(k) = 1$ and $h(n) = 0$ for $1 \leqslant n \leqslant k - 1$, so the lemma is obviously true for $n \leqslant k$.

(ii) Now we consider n such that $k < n < 2k - s$, and hence we can assume that $0 \leqslant s \leqslant k - 2$. For these n it follows by (2.4) that

$$h(n) = qh(n-1) = (q-1)h(n-1) + h(n-1).$$

Furthermore, if $k < n - 1$, then $h(n-1) = (q-1)h(n-2) + h(n-2)$, so that

$$h(n) = (q-1)(h(n-1) + h(n-2)) + h(n-2).$$

Repeating this argument, and using that $h(k) = 1$, we get

$$h(n) = 1 + (q-1)\sum_{i=1}^{n-k} h(n-i). \tag{2.9}$$

Since $h(j) = 0$ for $j < k$, the lemma is true for $n < 2k - s$.

https://doi.org/10.1017/S0963548302005187 Published online by Cambridge University Press

(iii) By (2.4) and (2.9)

$$\begin{aligned} h(2k-s) &= qh(2k-s-1)-1 \\ &= (q-1)h(2k-s-1)+h(2k-s-1)-1 \\ &= (q-1)\sum_{i=1}^{k-s} h(2k-s-i), \end{aligned}$$

so the lemma holds also for $n = 2k-s$.

(iv) In this step we assume that $s > 0$, and consider n such that $2k-s < n < 2k$. Fix n and make the induction hypothesis that (2.7) is true for all $m < n$. By (2.6), and since $b_i = 0$ for $s < i < k$, we get

$$\begin{aligned} h(n) &= qa(n-1)-a(n) \\ &= q\left(qa(n-2)-a(n-k-1)+\sum_{i=1}^{k-1} b_i[qa(n-k+i-2)-a(n-k+i-1)]\right) \\ &\quad -\left(qa(n-1)-a(n-k)+\sum_{i=1}^{k-1} b_i[qa(n-k+i-1)-a(n-k+i)]\right) \\ &= qh(n-1)-h(n-k)+\sum_{i=1}^{s} b_i[qh(n-k+i-1)-h(n-k+i)]. \end{aligned} \tag{2.10}$$

Recall from (2.4) that

$$q\,h(j-1)-h(j) \quad \begin{cases} =0, & \text{if } j<k, \\ =-1, & \text{if } j=k, \\ \geqslant 0, & \text{if } j>k, \end{cases}$$

and note that, if $i = 2k-n$, then $n-k+i = k$ and $0 < i < s$, so that

$$\sum_{i=1}^{s} b_i[qh(n-k+i-1)-h(n-k+i)] \geqslant -b_{2k-n} \geqslant -1. \tag{2.11}$$

Moreover, $h(n-k) = 0$ if $n < 2k$, and it follows by (2.10), (2.11) and the induction hypothesis that

$$\begin{aligned} h(n) &\geqslant qh(n-1)-1 \\ &= (q-1)h(n-1)+h(n-1)-1 \\ &\geqslant (q-1)[h(n-1)+\cdots+h(n-(k-s)-1)]-1 \\ &\geqslant (q-1)[h(n-1)+\cdots+h(n-(k-s))], \end{aligned}$$

where the last inequality follows from $n-(k-s)-1 \geqslant k$ and $h(n-(k-s)-1) \geqslant h(k) = 1$, by the induction hypothesis.

(v) In the final step in the case $s > 0$, we take $n \geqslant 2k$, and make the assumption that (2.7) is true for all $m < n$. In this case $qh(n-k+i-1)-h(n-k+i) \geqslant 0$, for all $i = 1,\ldots,k-1$,

https://doi.org/10.1017/S0963548302005187 Published online by Cambridge University Press

and it follows from (2.10) that

$$\begin{aligned} h(n) &= qh(n-1) - h(n-k) + \sum_{i=1}^{s} b_i[qh(n-k+i-1) - h(n-k+i)] \\ &\geqslant qh(n-1) - h(n-k) \\ &\geqslant (q-1)h(n-1) - h(n-k) + (q-1)\sum_{i=1}^{k-s} h(n-1-i) \\ &\geqslant (q-1)\sum_{i=1}^{k-s} h(n-i), \end{aligned}$$

where the last two inequalities follow by the induction hypothesis.

(vi) What remains to be proved is the case where $s = 0$ and $n > 2k$. Fix $n > 2k$ and assume that (2.8) holds for all $m < n$. Then, by (2.10),

$$\begin{aligned} h(n) &= qh(n-1) - h(n-k) \\ &\geqslant (q-1)h(n-1) - h(n-k) + (q-1)\sum_{i=1}^{k-1} h(n-1-i) \\ &\geqslant (q-1)\sum_{i=1}^{k-1} h(n-i). \end{aligned}$$

□

3. Proof of Theorem 1.1

To simplify the notation, we will let

$$a(n) = a_w(n), \quad a'(n) = a_{w'}(n), \quad \Delta(n) = a_w(n) - a_{w'}(n),$$
$$h(n) = q\,a_w(n-1) - a_w(n) \quad \text{and} \quad h'(n) = q\,a_{w'}(n-1) - a_{w'}(n),$$

when this is more convenient. The proof of Theorem 1.1 proceeds by induction, and the cases of equal and unequal lengths of the patterns are handled separately.

3.1. w and w' of equal length

Assume that $|w| = |w'| = k$. By (1.2) we have $\Delta(n) = 0$, $n < 2k - r$, and $\Delta(2k - r) = 1$. Hence the statement of the theorem is true for $n = 2k - r$. Fix $n > 2k - r$ and make the induction hypothesis that (1.4) is true for all m such that $2k - r \leqslant m < n$. Using (2.6) and the fact that $b_i = b'_i$, $i = r+1, \ldots, k-1$, $b'_r = 0$ yields

$$\begin{aligned} \Delta(n) = q\Delta(n-1) &+ \sum_{i=r+1}^{k-1} b_i[q\Delta(n-k+i-1) - \Delta(n-k+i)] \\ &+ \sum_{i=1}^{r} b_i h(n-k+i) - \sum_{i=1}^{r-1} b'_i h'(n-k+i) - \Delta(n-k), \end{aligned} \tag{3.1}$$

and we need to show that the right-hand side of (3.1) is strictly larger than $(q-1)\Delta(n-1)$. The first line of (3.1) will be bounded by repeated application of the induction hypothesis,

https://doi.org/10.1017/S0963548302005187 Published online by Cambridge University Press

while the second line will be bounded by means of Lemma 3.1 below, in which proof the autocorrelation structure plays a crucial role, and by the induction hypothesis. These two bounds together prove (1.4).

Lemma 3.1. *Suppose* $|w| = |w'| = k$ *and* $\mathrm{cor}(w) > \mathrm{cor}(w')$. *Then, for* $n \geqslant 2k - r$,

$$\sum_{i=1}^{r} b_i h(n-k+i) - \sum_{i=1}^{r-1} b'_i h'(n-k+i) \geqslant \sum_{i=1}^{r-1} [q\Delta(n-k+i-1) - \Delta(n-k+i)]. \tag{3.2}$$

Proof. If $r = 1$ then the left-hand side of (3.2) is $b_1 h(n-k+1)$, which is nonnegative, while the right-hand side vanishes and the lemma is obviously true. In the rest of the proof we assume that $r = 2, \ldots, k-1$.

Let s be the number defined in (2.3) pertaining to the word w, and set $\gamma = k - s$. Note that, since $b_r = 1$, it is obvious that $s \geqslant r$, and that for some $t \in \{1, 2, \ldots\}$ either $r = k - \gamma t$, or $k - \gamma t > r > k - \gamma(t+1)$. In the first case $b_r = b_{r-\gamma} = b_{r-2\gamma} = \cdots = 1$, by (2.5), and, letting $R = k - \gamma(t+1)$, it follows in the latter case that $b_r = b_R = b_{R-\gamma} = b_{R-2\gamma} = \cdots = 1$. Let Γ denote the set $\{k-\gamma, k-2\gamma, \ldots\} \cap \{1, 2, \ldots, r\} \cup \{r\}$. Then $b_i = 1$ for $i \in \Gamma$. Using the fact that $b_j h(j) \geqslant 0$ for all j, it follows that

$$\sum_{i=1}^{r} b_i h(n-k+i) \geqslant \sum_{i \in \Gamma} h(n-k+i). \tag{3.3}$$

Furthermore $s \geqslant r > 1$, so by Lemma 2.1

$$h(n-k+i) \geqslant (q-1) \sum_{j=1}^{\gamma} h(n-k+i-j),$$

for all i. If the elements of Γ are ordered by size, the smallest element and the distance between two consecutive elements are at most γ, and we get

$$\sum_{i \in \Gamma} h(n-k+i) \geqslant (q-1) \sum_{i=1}^{r-1} h(n-k+i).$$

This inequality, together with (3.3) and $b'_j h'(j) \leqslant h'(j)$, yields

$$\begin{aligned} \sum_{i=1}^{r} b_i h(n-k+i) - \sum_{i=1}^{r-1} b'_i h'(n-k+i) & \\ \geqslant \sum_{i=1}^{r-1} h(n-k+i) - \sum_{i=1}^{r-1} h'(n-k+i) & \\ = \sum_{i=1}^{r-1} [q\Delta(n-k+i-1) - \Delta(n-k+i)], & \end{aligned}$$

and the lemma is proved. □

https://doi.org/10.1017/S0963548302005187 Published online by Cambridge University Press

By (3.1), the above lemma and the induction hypothesis, it follows that

$$\begin{aligned}\Delta(n) &\geqslant q\Delta(n-1) + \sum_{i=r+1}^{k-1} b_i[q\Delta(n-k+i-1) - \Delta(n-k+i)] \\ &\quad + \sum_{i=1}^{r-1}[q\Delta(n-k+i-1) - \Delta(n-k+i)] - \Delta(n-k) \\ &\geqslant q\Delta(n-1) + \sum_{i=r+1}^{k-1} b_i[q\Delta(n-k+i-1) - \Delta(n-k+i)] - \Delta(n-k+r-1).\end{aligned}$$

It remains to prove that

$$\Delta(n-1) + \sum_{i=r+1}^{k-1} b_i[q\Delta(n-k+i-1) - \Delta(n-k+i)] > \Delta(n-k+r-1). \tag{3.4}$$

If $r = k-1$, (3.4) is clearly true, since $\Delta(n-1) > \Delta(n-2)$ by the induction hypothesis, and the theorem is proved in this case. Now, assume that $1 \leqslant r < k-1$, and note that

$$\Delta(j) + b_i[q\Delta(j-1) - \Delta(j)] = \begin{cases} \Delta(j), & \text{if } b_i = 0, \\ q\Delta(j-1), & \text{if } b_i = 1, \end{cases}$$

and, by the induction hypothesis, $\Delta(j) \geqslant (q-1)\Delta(j-1)$, for $j < n$, so that

$$\Delta(j) + b_i[q\Delta(j-1) - \Delta(j)] \geqslant (q-1)\Delta(j-1), \tag{3.5}$$

with strict inequalities for $j \geqslant 2k-r$. Using (3.5) repeatedly for $j = n-1$ down to $j = n-k+r+1$, we obtain

$$\Delta(n-1) + \sum_{i=r+1}^{k-1} b_i[q\Delta(n-k+i-1) - \Delta(n-k+i)] > (q-1)\Delta(n-k+r).$$

Since $\Delta(n-k+r) \geqslant \Delta(n-k+r-1)$, by another use of the induction hypothesis, (3.4) is proved, which concludes the proof of the theorem when the patterns are of equal length.

3.2. w and w' of unequal length

We must prove the result corresponding to (1.4) in the case of different lengths of the patterns: $|w| = k$ and $|w'| = j < k$. The basic step in the induction follows by (2.1): $a(n) = a'(n)$ for $n < j$, and for $n = j$ we have $a(j) = q^j$, while $a'(j) = q^j - 1$. Hence

$$1 = \Delta(j) > (q-1)\Delta(j-1) = 0.$$

Fix $n > j$ and assume that (1.4) is true for all m such that $j \leqslant m < n$. First we consider the case where $|w| = k$, $|w'| = k-1$,

$$\operatorname{cor}(w) = 1\underbrace{00\ldots00}_{k-1} \quad \text{and} \quad \operatorname{cor}(w') = \underbrace{11\ldots11}_{k-1}.$$

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

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

https://doi.org/10.1017/S0963548302005187 Published online by Cambridge University Press

by (2.6). Hence

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

and

$$a'(n) = q\,a'(n-1) - (q-1)\,a'(n-k). \tag{3.6}$$

Furthermore $b_1 = \cdots = b_{k-1} = 0$, so that, by (2.6),

$$a(n) = q\,a(n-1) - a(n-k). \tag{3.7}$$

Using (3.6), (3.7), the induction hypothesis and the fact that $a'(n-k) \geqslant 0$ yields

$$\begin{aligned}\Delta(n) &= (q-1)\Delta(n-1) + \Delta(n-1) - \Delta(n-k) + (q-2)a'(n-k)\\ &> (q-1)\Delta(n-1).\end{aligned} \tag{3.8}$$

If w and w' have arbitrary autocorrelations, and $|w| = k$ and $|w'| = j < k$, we choose patterns v_i and v_i', $i = 1,\ldots,k-j$ with autocorrelations

$$\mathrm{cor}(v_i) = 1\underbrace{00\ldots00}_{k-i} \quad \text{and} \quad \mathrm{cor}(v_i') = \underbrace{11\ldots11}_{k-i}.$$

Note that such patterns always exist. Then $a(n) - a'(n)$ can be written as a telescoping sum as

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

By (3.8)

$$a_{v_i}(n) - a_{v_i'}(n) > (q-1)[a_{v_i}(n-1) - a_{v_i'}(n-1)],$$

$i = 1,\ldots,k-j$. Furthermore w and v_1 are of the same length, which holds also for v_i' and v_{i+1}, $i = 1,\ldots,k-j-1$, and for v_{k-j}' and w'. Thus the other summands are handled by the first part of this theorem, and we finally get

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

□

4. Concluding remarks

The main result in this paper, Theorem 1.1, states that, if the pattern w has a larger autocorrelation than w', then the number of strings of length n which avoid w, $a_w(n)$, is larger than $a_{w'}(n)$. The occurrence of patterns in strings is important in various areas: for

https://doi.org/10.1017/S0963548302005187 Published online by Cambridge University Press

instance, in computer science, where autocorrelation can be used to study the behaviour of pattern matching algorithms, and in molecular biology, where the frequency of certain patterns in DNA sequences is important in gene recognition. A simple model for DNA sequences is that of sequences of independent random variables distributed over the alphabet $\{A, C, G, T\}$. Hence, in the latter example, and in many other applications, the framework is probabilistic.

Our results can easily be reformulated in probabilistic terms: Assume that $X_1, X_2, \ldots$ is a sequence of random variables which are independent and uniformly distributed over Ω. Let $s_w(n)$ be the probability that w does not occur as a substring in $X_1, \ldots, X_n$. Since the total number of outcomes of $X_1, \ldots, X_n$ is q^n, and $a_w(n)$ of them do not contain w, it follows that $s_w(n) = a_w(n)/q^n$, and Theorem 1.1 can be stated in terms of $s_w(n)$. A natural generalization of the problem in this probabilistic setting is to allow a non-uniform distribution over the alphabet. If the distribution is too skewed, the autocorrelation is of minor importance for $s_w(n)$ – it is, rather, the frequency of different characters that is crucial. It might be interesting to study under which conditions on the distribution the results in this paper still hold.

Acknowledgement

The author is grateful to the referee for useful comments.

References

[1] Cakir, I., Chryssaphinou, O. and Månsson, M. (1999) On a conjecture by Eriksson concerning overlap in strings. *Combin. Probab. Comput.* **8** 429–440.
[2] Eriksson, K. (1997) Autocorrelation and the enumeration of strings avoiding a fixed string. *Combin. Probab. Comput.* **6** 45–48.
[3] Guibas, L. J. and Odlyzko, A. M. (1981) String overlaps, pattern matching and nontransitive games. *J. Combin. Theory Ser. A* **30** 183–200.
[4] Régnier, M. and Szpankowski, W. (1998) On pattern frequency occurrences in a Markovian sequence. *Algorithmica* **22** 631–649.

https://doi.org/10.1017/S0963548302005187 Published online by Cambridge University Press