Combinatorics, Probability and Computing (1997) 6, 45–48
Copyright © 1997 Cambridge University Press

Autocorrelation and the Enumeration of Strings Avoiding a Fixed String

KIMMO ERIKSSON†

Stockholm University, S-106 09 Stockholm, Sweden
(e-mail: kimmo@nada.kth.se)

Received 7 April 1995; revised 24 July 1995

Considering strings over a finite alphabet $\mathscr{A}$, say that a string is w-avoiding if it does not contain w as a substring. It is known that the number $a_w(n)$ of w-avoiding strings of length n depends only on the *autocorrelation* of w as defined by Guibas–Odlyzko. We give a simple criterion on the autocorrelations of w and w' for determining whether $a_w(n) > a_{w'}(n)$ for all large enough n.

In this paper we work with strings over a finite alphabet $\mathscr{A}$, of cardinality A. Say that a string is w-avoiding if it does not contain w as a substring. What is the number $a_w(n)$ of w-avoiding strings of length $n \geq 0$? This problem with various generalisations has been solved independently by several people. For some history and much more mathematics on this subject, we refer to Guibas and Odlyzko [2, 3]. One other area where this theory is relevant is the Hilbert series of the non-commutative graded algebra $k\langle\mathscr{A}\rangle/(w)$, where k is any field and (w) is the doublesided ideal generated by w in the k-module $k\langle\mathscr{A}\rangle$. The Hilbert series will be the generating function for the numbers $a_w(n)$. See, for example, Anich [1].

Guibas and Odlyzko [2] defined the *autocorrelation* of a k letter string w to be the binary sequence of bits $b_k b_{k-1} \dots b_1$, where $b_i = 1$ if the length i prefix of w equals the length i suffix. For example, `every eve` is a string of length 9 whose prefixes of length 1, 3 and 9 equal the corresponding suffixes; hence, corr(`every eve`) = 100000101. In particular, if the length of the word is k we trivially have $b_k = 1$. We will abuse notation somewhat and let corr(w) denote both the binary sequence and its value as a binary number.

The autocorrelation of w determines a linear recurrence for the number $a_w(n)$ of w-avoiding strings of length n. Suppose that one wants to know, given two strings w and

† This paper was written while the author was visiting MIT, supported by postdoc grants from NFR and Wenner-Gren Stiftelsen.

https://doi.org/10.1017/S0963548397002836 Published online by Cambridge University Press

w', whether there are more w-avoiding strings than w'-avoiding strings, at least for big lengths. Our main result is the following:

Theorem 0.1. *There exists an N such that $a_{w'}(n) > a_w(n)$ for all $n \geq N$ if and only if $\mathrm{corr}(w') > \mathrm{corr}(w)$ as binary numbers.*

Obviously, it would be nicer if we could specify what 'large enough' means more explicitly. We have the following conjecture, which all examples seem to indicate, but which we have not been able to prove:

Conjecture 0.2. *If $\mathrm{corr}(w') > \mathrm{corr}(w)$, then $a_{w'}(n) > a_w(n)$ from the first n where equality no longer holds.*

First, let us give the result on the importance of the autocorrelation for the number of w-avoiding words. This is an old result that several people have found, as mentioned above. We include a short proof.

Proposition 0.1. *Let w be a string of k letters $w_1 w_2 \dots w_k$ with autocorrelation $\mathrm{corr}(w) = b_k b_{k-1} \dots b_1$. The numbers $a_w(n)$ satisfy the linear recurrence*

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

of order k, for all $n \geq 0$, with start values $a_w(n) = A^n$ for $n = 0, 1, \dots, k-1$.

Proof. Let V be the set of w-avoiding words of length n, so $|V| = a_w(n)$. We shall partition V in sets V_i of sizes $A a_w(n+i-1) - a_w(n+i), i = 1, 2, \dots, k$, as follows. For any $v \in V$, choose i to be the smallest index such that $v w_1 \cdots w_i$ is not w-avoiding. Hence $v w_1 \cdots w_i$ must end with w, so $w_1 \cdots w_i$ is both a prefix and a suffix of w, and consequently $b_i = 1$. Place v in the set V_i. Then V_i is in natural bijection with the set of all words of length $n+i$ that are not w-avoiding but such that they have prefixes of length $n+i-1$ which are w-avoiding. This latter set can be constructed by taking all w-avoiding words of length $n+i-1$, appending any of the A letters of the alphabet, and then omitting all w-avoiding words. Thus $|V_i| = A a_w(n+i-1) - a_w(n+i)$, and the recurrence follows. The start values are obvious. □

When proving Theorem 0.1 we will make constant use of the recurrence (1) above in various ways. For a string w, let $g_w(x)$ be the characteristic polynomial of the recurrence (1), so

$$g_w(x) = 1 + (x - A) \sum b_i x^{i-1}.$$

The $a_w(n)$ are all positive numbers. Let r the largest positive root of the characteristic equation $g_w(x) = 0$. Observe that $r < A$ must hold. By standard knowledge of linear recurrences, the asymptotic behaviour of $a_w(n)$ is like cr^n, except in degenerate cases, but

https://doi.org/10.1017/S0963548397002836 Published online by Cambridge University Press

as is proved by Guibas and Odlyzko [2], the present case is generic. We shall now see qualitatively how the root r varies with $\mathrm{corr}(w)$.

Lemma 0.1. *Suppose w and w' are strings such that $\mathrm{corr}(w) = \mathrm{corr}(w') - 1$ as binary numbers. Let r and r' be, respectively, the largest positive roots of the characteristic equations. Then $r' \geq r$, and equality holds if and only if the alphabet size is $A = 2$ and $\mathrm{corr}(w')$ is a power of 2.*

Proof. As a binary number, we have $\mathrm{corr}(w') = \sum b'_i 2^{i-1}$ and $\mathrm{corr}(w) = \sum b_i 2^{i-1}$. Let j be the least index such that $b'_j = 1$. Then the relation $\mathrm{corr}(w) = \mathrm{corr}(w') - 1$ tells us that $b_i = b'_i$ for all $i > j$, while $b_j = 0$ and $b_i = 1$ for $i < j$. Thus we can write

$$\begin{aligned} g_{w'}(x) &= 1 + (x - A)\sum_{i>j} b'_i x^{i-1} + (x - A)x^{j-1} \qquad \text{and} \\ g_w(x) &= 1 + (x - A)\sum_{i>j} b_i x^{i-1} + (x - A)\sum_{i=1}^{j-1} x^{i-1}. \end{aligned}$$

By definition, r' and r are the largest zeros of $g_{w'}(x)$ and $g_w(x)$. We know that $g_{w'}(x)$ is positive for large enough x, so $g_{w'}(x)$ must have a largest root $r' \geq r$ if $g_{w'}(r) \leq 0$. Therefore, let us compute $g_{w'}(r)$:

$$\begin{aligned} g_{w'}(r) &= g_w(r) + (r - A)\left[r^{j-1} - \sum_{i=1}^{j-1} r^{i-1}\right] \\ &= 0 + (r - A)\left[1 + (r - 2)\sum_{i=1}^{j-1} r^{i-1}\right] \\ &= (r - A)\left[g_w(r) + (A - r)\sum_{i>j} b_i r^{i-1} + (A - 2)\sum_{i=1}^{j-1} r^{i-1}\right]. \end{aligned}$$

Here, $(r - A)$ is negative since we observed earlier that $r < A$. Now for the terms in the other factor: we have $g_w(r) = 0$; clearly $(A - r)\sum_{i>j} b_i r^{i-1}$ is non-negative, and equal to zero if and only if for all $i > j$ we have $b_i = 0$, in which case $\mathrm{corr}(w') = 2^{j-1}$; and $(A-2)\sum_{i=1}^{j-1} r^{i-1}$ is non-negative, and equal to zero if and only if $A = 2$ or $j = 1$. As noted above, it follows that $r' \geq r$, with equality if and only if $\mathrm{corr}(w') = 2^{j-1}$ and $A = 2$. (The case $\mathrm{corr}(w') = 2^{j-1}$ and $j = 1$ cannot happen, because then $\mathrm{corr}(w) = \mathrm{corr}(w') - 1 = 0$ but all non-empty strings have autocorrelation greater than zero.) □

As we saw in Lemma 0.1, there was a special case to consider: when $A = 2$ and $\mathrm{corr}(w')$ is a power of two. This case turns out to be particularly nice.

Proposition 0.2. *If the alphabet size is $A = 2$, and the two strings w and w' satisfy $\mathrm{corr}(w') = \mathrm{corr}(w) + 1 = 2^k$ (as binary numbers), then $a_{w'}(n) - a_w(n) = a_{w'}(n - k)$ for all $n \geq k$.*

https://doi.org/10.1017/S0963548397002836 Published online by Cambridge University Press

Proof. For convenience, let $a_n = a_w(n)$ and $a'_n = a_{w'}(n)$. The assumption that $\mathrm{corr}(w') = 2^k$ as a binary number implies that $\mathrm{corr}(w')$ starts with $b'_{k+1} = 1$ followed by k zeros, while $\mathrm{corr}(w) = 2^k - 1$ as a binary number means that $\mathrm{corr}(w)$ is a sequence of k ones. By Proposition 0.1, we have $a_n = \sum_{i=1}^{k} [2a_{n+i-1} - a_{n+i}]$ for all $n \geq 0$. Hence, $a_{n+1} - a_n = [2a_{n+k} - a_{n+k+1}] - [2a_n - a_{n+1}]$, so $a_n = 2a_{n+k} - a_{n+k+1}$. This last is the same recurrence as the one that a'_n obeys, by Proposition 0.1. But then also the difference $d_n \stackrel{\text{def}}{=} a'_n - a_n$ obeys this recurrence. We have $d_0 = d_1 = \cdots = d_{k-1} = 0$ but $d_k = 1$ since there is one string of length k, namely w, that is not w-avoiding, but all are w'-avoiding. Applying the recurrence to these start values on d_n we get the next k values to be $d_{k+1} = 2, d_{k+2} = 2^2, \ldots, d_{2k} = 2^k$. Hence, $d_n = a'_{n-k}$ for all $n \geq k$. □

We are now ready to prove the main result.

Proof of Theorem 0.1. We shall show that $a_{w'}(n) > a_w(n)$ for large enough n if and only if $\mathrm{corr}(w') > \mathrm{corr}(w)$. Observe that Proposition 0.1 rules out the case $\mathrm{corr}(w') = \mathrm{corr}(w)$. It clearly suffices to show that $a_{w'}(n) > a_w(n)$ for large enough n whenever $\mathrm{corr}(w') = \mathrm{corr}(w) + 1$. If the alphabet size is $A = 2$ and $\mathrm{corr}(w')$ is a power of two, this follows from Proposition 0.2. Otherwise, thanks to Guibas and Odlyzko we know that asymptotically $a_{w'}(n) \sim c' r'^n$ and $a_w(n) \sim c r^n$, and by Lemma 0.1 we know that $r' > r$ (and $c, c' > 0$). This implies of course that $a_{w'}(n) > a_w(n)$ for large enough n. □

References

[1] D. Anich, D. (1982) Non-commutative graded algebras and their Hilbert series. *J. Algebra* **78** 120–140.
[2] Guibas, L. and Odlyzko, A. (1981) String overlaps, pattern matching and nontransitive games. *J. Comb. Theory Ser. A* **30** 183–200.
[3] Guibas, L. and Odlyzko, A. (1981) Periods in strings. *J. Comb. Theory Ser. A* **30** 19–42.

https://doi.org/10.1017/S0963548397002836 Published online by Cambridge University Press