JOURNAL OF COMBINATORIAL THEORY, Series A **30**, 183–208 (1981)

String Overlaps, Pattern Matching, and Nontransitive Games

L. J. GUIBAS

Xerox Palo Alto Research Center, Palo Alto, California 94304

AND

A. M. ODLYZKO

Bell Laboratories, Murray Hill, New Jersey 07974

Communicated by the Managing Editors

Received October 2, 1978; revised July 12, 1979

This paper studies several topics concerning the way strings can overlap. The key notion of the *correlation* of two strings is introduced, which is a representation of how the second string can overlap into the first. This notion is then used to state and prove a formula for the generating function that enumerates the q-ary strings of length n which contain none of a given finite set of patterns. Various generalizations of this basic result are also discussed. This formula is next used to study a wide variety of seemingly unrelated problems. The first application is to the nontransitive dominance relations arising out of a probabilistic coin-tossing game. Another application shows that no algorithm can check for the presence of a given pattern in a text without examining essentially all characters of the text in the worst case. Finally, a class of polynomials arising in connection with the main result are shown to be irreducible.

1. INTRODUCTION

We are interested in the combinatorial structure of the occurrence of patterns in strings. Let us fix an alphabet Ω of size $q \geqslant 2$, so that all strings to be considered will be composed of characters from Ω. Our basic results, from which most of the others are derived, deal with the enumeration of strings of a given length which do not contain any one of a given set of other srings, which we will refer to as patterns. A crucial quantity for our investigations will be the correlation of two patterns X and Y. The *correlation* of X and Y, to be denoted by XY, is a string over $\{0, 1\}$ with the same length as X. The ith bit (from the left) of XY is determined as follows: place Y under X so that its leftmost character is under the ith character of X

0097-3165/81/020183–26$02.00/0

(from the left). Then, if all the pairs of characters in the overlapping segment are identical, the ith bit of XY is 1, else it is 0. For example, if $\Omega = \{H, T\}$, $X = HTHTTH$ and $Y = HTTHT$, then $XY = 001001$, as depicted below:

```
X: HTHTTH
Y: HTTHT      0
    HTTHT     0
     HTTHT    1
      HTTHT   0
       HTTHT  0
        HTTHT 1
```

Note that $YX = 00010$, so that in general $XY \neq YX$. The correlation XY has been previously termed the "leading number" of X and Y, but we chose to avoid this name, since it suggests symmetry. It makes sense to define the *autocorrelation* of X as XX. Thus for the Y above, $YY = 10010$. XX is a representation of the set of *periods* of X, i.e., those shifts that cause X to overlap itself. The question of characterizing those binary patterns that are correlations for some patterns is dealt with in a separate paper [8]. It is shown there that there are on the order of $\exp(c(\log n)^2)$ different autocorrelations of length n, no matter what the initial alphabet Ω is (e.g., there exist exactly 116 different autocorrelations of length 20), and results are obtained on the number of patterns that have a given autocorrelation. In this paper, however, we will be concerned more with several combinatorial problems in whose solutions string correlations play central roles. Other results which involve correlations are presented in [9, 10].

We often wish to interpret the correlation XY as a number in some base t, or else a polynomial in the variable t, in which case we write XY_t. Thus, for the above example,

$$XY_t = t^3 + 1, \qquad XY_2 = 9.$$

Two more final points of terminology: we write $|X|$ for the length of X, with $|X| = 0$ if X is the empty string, and we call a set of patterns $\{A, B,..., Y\}$ *reduced*, if I is never a substring of J, for any two patterns I, J in our set.

Suppose that $\{A, B,..., T\}$ is a reduced set of patterns. Let $f(n) = f(A, B,..., T; n)$ denote the number of strings of length n over our alphabet that do not contain any of $A, B,..., T$. We denote by $F(z)$ the corresponding generating function

$$F(z) = \sum_{n=0}^{\infty} f(n) z^{-n}.$$

Let $f_H(n)$ denote the number of strings of length n that end with H and do not contain any of A, B,..., T except for that single appearance of H at the end of the string, and let $F_H(z)$ be the generating function of $f_H(z)$. Our basic result is the following system of equations:

THEOREM 1. *If $\{A,..., T\}$ is a reduced set of patterns, then the generating functions $F(z)$, $F_A(z)$,..., $F_T(z)$ satisfy the following system of linear equations*:

$$\begin{gathered}(z-q)\,F(z)+zF_A(z)+zF_B(z)+\cdots+zF_T(z)=z\\ F(z)-zAA_zF_A(z)-zBA_zF_B(z)-\cdots-zTA_zF_T(z)=0\\ \cdots\\ F(z)-zAT_zF_A(z)-zBT_zF_B(z)-\cdots-zTT_zF_T(z)=0.\end{gathered}\tag{1.2}$$

An important observation is that the above system of equations is nonsingular. That is, if we let $\phi(z)=\phi(A,B,..., T; z)$ be the determinant

$$\phi(z)=\begin{vmatrix} z-q & z & \cdots & z\\ 1 & -zAA_z & \cdots & -zTA_z\\ & \cdots & & \\ 1 & -zAT_z & \cdots & -zTT_z\end{vmatrix},\tag{1.3}$$

then the fact that the set $\{A,..., T\}$ is reduced implies that in each column the highest degree polynomial occurs on the diagonal, and the only other polynomial in that column that can have equal degree is in the first row. Hence in the expansion of $\phi(z)$, the unique highest degree monomial comes from the product of the diagonal terms, so that $\phi(z)$ is a nonzero polynomial of degree $1+|A|+|B|+\cdots+|T|$. Therefore we can solve for each one of $F(z)$, $F_A(z)$,..., $F_T(z)$ and find that each one is a rational function of z with denominator equal to $\phi(z)$. (This shows that $f(n)$ and the $f_H(n)$ satisfy linear recurrences with characteristic polynomial $\phi(z)$.) For example, suppose that we exclude only a single pattern A. Then Theorem 1.1 implies that

$$F(z)=\frac{zAA_z}{1+(z-q)AA_z},\qquad F_A(z)=\frac{1}{1+(z-q)AA_z},\tag{1.4}$$

and that $f(n)=\sum a_jf(n-j)$ for $n\geqslant |A|$, where $1+(z-q)AA_z=z^{|A|}-\sum a_jz^{|A|-j}$, and j runs from 1 to $|A|$.

The above formula (as well as the more general formulas derivable from Theorem 1.1) give an easy way to estimate the numbers $f(n)$ and $f_H(n)$ through the partial fraction decomposition of $F(z)$ and $F_H(z)$ (cf. [9, 24]). For example, when we exclude only a single pattern A, the polynomial $1+(z-q)AA_z$ has exactly one zero θ of largest absolute value, and $q-\theta$ is very small (but positive). Hence $f(n)$ is asymptotic to $c\theta^n$ as $n\to\infty$, where c is a positive constant that can be explicitly calculated in terms of the correlation (see Section 7). Furthermore, θ varies monotonically with AA_q, so that for any two patterns A and B, asymptotically there will be more strings not containing A than strings not containing B precisely when $AA_q > BB_q$ (cf.]9]). It is a rather remarkable fact that this result holds not only asymptotically, but uniformly; that is, if $AA_q > BB_q$, then for any length n, there will be at least as many strings of length n that do not contain A as there are those that do not contain B. This result will be proved in Section 7.

The above results can be extended to some extent to the case where several patterns are excluded, but some care must be exercised. For example, the set $\{TH, HH, TTT\}$ is reduced, but $f(n) = 0$ for $n \geqslant 4$. This fact is obvious, but can also be deduced from Theorem 1.1. Let us call a reduced set of patterns $\{A, \ldots, T\}$ *consistent* if there exist arbitrarily long strings not containing any of $A, \ldots, T$. Then Theorem 1.1 provides us with an algorithmic way to test whether a set is consistent; we evaluate $F(z)$, and check whether it's a polynomial in $1/z$.

Theorem 1.1 can be generalized in several ways. In Section 2, we will actually prove Theorem 2.1, which deals with the enumeration of strings of a given length which exclude $\{A, \ldots, T\}$ but which start with a given string X. (Taking X to be the empty string, we will immediately obtain Theorem 1.1. The reason for proving Theorem 2.1 is that it has applications to the nontransitive game we discuss below.) Also, at the end of Section 2 we will show how these techniques can be extended to the enumeration of appearances of $A, \ldots, T$, but these results rapidly become very complex and hard to apply. It is possible to go even further, but again with some loss of utility. Goulden and Jackson [7] have generalized our results to show that for an alphabet $\Omega = \{1, \ldots, q\}$, if ·$\{A_1, \ldots, A_m\}$ is a set of patterns, then the power series in $x_1, \ldots, x_q, y_1, \ldots, y_m$, where the coefficient of $\prod x_i^{a_i} \cdot \prod y_j^{b_j}$ is the number of strings composed of a_i i's, $1 \leqslant i \leqslant q$, which contain exactly b_j of the A_j's, is a rational function.

There is one further kind of generalization which we will discuss in Section 3. The number $f(n)\, q^{-n}$ may be regarded as the probability that none of $A, \ldots, T$ will occur in n throws with a fair q-sided die. However, we can also consider a q-sided die in which different faces have different probabilities, and ask for the probability that none of $A, \ldots, T$ will occur in n throws, or the probability that exactly k A's occur, and so on. In Section 3 we will prove Theorem 3.3, which generalizes Theorem 1.1 to this case. However, the results become much less elegant.

Before proceeding to discuss our other results, we should give references to previous work. Expressions such as (1.4) have been known for a long time for some specific patterns A (cf. [5]). Furthermore, it seems to have been widely known that functions such as $F(z)$ are always rational (cf. [12]). However, the first published account that we are aware of that contained a closed-form formula like (1.4) in the case of a single excluded pattern is that of Solev'ev [24]. The recurrence for $f(n)$, again when only a single pattern is excluded, was also given by Harborth]11]. More importantly, it has been pointed out to us that Roberts [22] has found two methods for obtaining the generating function $F(z)$ of Theorem 1.1 for any reduced set of excluded patterns. One of his methods used Markov chains, while the other is essentially the same as ours. Since [22] is apparently not going to be published, we hereby present, with Roberts' permission, our proofs of these basic results.

Results such of Theorem 1.1 or Theorem 3.3 solve various clustering problems [14, 16, 21, 23] which have applications in quality control, where it is imporant to know how unlikely a given event is. There are also other applications to prefix-synchronized codes [9]. In this paper we will aply these basic results to the study of certain nontransitive dominance relations arising in games of chance and to the worst-case behavior of pattern-matching algorithms. (For another, slightly less related work on pattern matching, see [10].) The nontransitive game we will consider is the penny-flipping game invented by Penney [17] (see also [6, 24]). In its original formulation it starts out with two players who agree on some integer $k \geqslant 2$. Player I then selects a sequence A of k heads or tails, and Player II, knowing what A is, selects another sequence B of length k. The players then flip a coin until either A or B appears as a block of k consecutive outcomes. (The game will terminate with probability one.) Player I wins if A appears before B does. The intriguing feature of this game is the fact that if $k \geqslant 3$, then no matter what A is, Player II can choose a B so that his probability of winning will be greater than 1/2. (If $k = 2$ and $A = HT$, say, then Player II cannot expect to win with probability greater than 1/2.) This is easy to see when $A = HH \cdots H$; if Player II selects $B = TH \cdots H$, then the only way for Player I to win is for heads to come up on each of the first k throws, so that the odds in favor of B (the probability that B comes up first divided by the probability that A comes up first) are $2^k - 1$ to 1. A somewhat similar situation occurs for other A's, as we will explain.

We consider a slightly more general version of the game, in which the penny is replaced by a q-sided die, and q-sequences A and B can be of different lengths, provided only that neither A occurs in B nor B in A. There are many possible ways to calculate the odds that B will occur before A (a partial list is given in [6]), most of which are very involved. By far the simplest, however, is a formula due to Conway [6]: the odds that B will win are given by

$$\frac{AA_q - AB_q}{BB_q - BA_q} \tag{1.5}$$

Conway's proof of this elegant formula was never published [4]. The first widely disseminated proof is due to Collings [3], who proved it using results on the waiting times till the apearance of a pattern (which we will discuss later). Other proofs have recently been obtained by Li [15] using martingales, and by Wendel [25] using Markov processess.

We will present here two proofs of Conway's formula (1.5). The first one follows immediately from Theorem 1.1. We let $\{A, B\}$ be the set of excluded patterns. Then $f_A(n)\,q^{-n}$ is precisely the probability that A wins on the nth throw. Hence the probability that A wins is $\sum f_A(n)\,q^{-n} = F_A(q)$. Now when we solve the system (1.2), we obtain

$$F_A(z) = \frac{BB_z - BA_z}{(z-q)(AA_z \cdot BB_z - AB_z \cdot BA_z) + AA_z + BB_z - AB_z - BA_z}$$

and therefore the probability that A wins is given by $(BB_q - BA_q)/(AA_q + BB_q - AB_q - BA_q)$. Similarly the probability that B wins is given by $(AA_q - AB_q)/(AA_q + BB_q - AB_q - BA_q)$, and so the odds in favor of B are given by (1.5).

Before giving the other proof of (1.5), we will discuss some generalizations. First of all, we can have more than two players. For example, if the chosen sequences (which again in this setting do not have to be of equal lengths, but which have to be reduced to avoid trivial situations) are $A, B,\ldots, T$, then apply Theorem 1.1 with $\{A,\ldots, T\}$ as the excluded set. The probability that A, say, wins, is again given by $F_A(q)$, and can be calculated in terms of the correlations. (Note that $F_A(q)$ is always finite, so that even if the determinant $\phi(z)$ of the system (1.2) vanishes at $z = q$, the numerator of the expansion for $\phi(z)$ has to vanish there as well, and we can obtain $F_A(q)$ by cancelling the common factors.)

The analysis of the r-person game is considerably more involved than that of the 2-person one, in part because of the problem of coalitions. Just for reference, let us state that the probability that A appears before either B or C (provided $\{A, B, C\}$ is reduced set) is given by Q_1/Q_2, where (letting HS denote HS_q)

$$Q_1 = (BB - BA)\,CC + (BA - BC)\,CB + (BC - BB)\,CA,$$

and

$$\begin{aligned} Q_2 = {} & (AC - AB + CB - CC)\,BA + (BC - CC)\,AB \\ & + (CC - CB - BC + BB)\,AA + (CC - AC)\,BB \\ & + (AC - BC)\,CB + (BC - BB - AC + AB)\,CA. \end{aligned}$$

Another generalization is to allow the die to be biased, so that different sides appear with varying probabilities. A formula for the odds that B wins similar to (1.5) can be derived in this case also (see Section 3). However, the question of nontransitivity becomes much more involved. For example, if 0 comes up with a probability very close to 1, then the first player will be practically guaranteed to win if the chooses $A = 0 \cdots 0$. Even for $q = 2$ it is an open question to determine, for a given k, those probabilities which enable Player II to obtain odds better than 1 to 1 against all possible choices for Player I.

Because of the complexity of dealing with unequal probabilities or more than two players, we will confine ourselves to the two-player, fair q-sided die situation, in which both players choose sequences of length k. The fact that for $q = 2$ and $k \geqslant 3$ player II has an advantage seems to have been noted by several people. Ramshaw [19] has even found a very simple algorithm for player II's choice if $q = 2$, which guarantees him odds of at least 6/5. It turns out that properties of correlations can be used to deduce the optimal strategy for Player II. In Section 4 we will prove the following result, which answers a question posed by Conway [4]:

THEOREM 1.6. *If $A = a_1 \cdots a_k$ is the choice of Player I, then all the choices B for Player II which maximize his probability of winning are of the form $B = ba_1 \cdots a_{k-1}$, for a suitable b.*

We should mention that the case $k = 2$ of the theorem above follows immediately from Conway's formula (1.5), and so in the proof we will only consider $k \geqslant 3$. (Note that if $k = 2$ and $q \geqslant 3$, the game is again nontransitive.)

In general there does not seem to be any simple rule for the optimal choice of the character b. However, the choice of b presented in Section 4 can be shown to give odds of almost $q/(q-1)$ (for large k). With additional work it can even be shown that for $q = 2$, Player II can always obtain odds of at least 9/5, and he can be held to these odds only for $k = 4$.

Let us introduce the expected waiting time; i.e., the expected number of throws until the desired sequence appears. Then the waiting time for A is qAA_q. This was proved first by Solov'ev [24], and then rediscovered by Nielsen [16] and Collings [3]. We can prove it very simply using Theorem 1.1; if A is the only excluded pattern, then $f(n)\,q^{-n}$ is the probability that one has to wait more than n throws for the appearance of A, and so $F(q) = \sum f(n)\,q^{-n}$ is the expected waiting time. But by (1.3) this is just qAA_q. Thus although all patterns of length k appear about equally often in random strings, the time of their first appearance depends strongly on their autocorrelations. This fact makes the nontransitivity of this game all the more surprising, since a pattern with the correlation $10 \cdots 0$ is expected to show up very early. Another, similar result, however, helps explain the situation; namely, the expected number of additional throws needed to obtain A if we start with B is $qAA_q - qBA_q$. This was first proved by Collings [3]. Another proof can be obtained from Theorem 2.1, by letting $X = B$ and taking the set of excluded patterns to consist of A. Considering these two results on waiting times as is done by Collings [3] leads immediately to another proof of Conway's formula (1.5).

We shall now leave nontransitive games and present another application of our basic theory, this time to the analysis of the worst-case behavior of string searching algorithms. We consider the problem of finding whether a given pattern A of length k occurs in a string S of length n. The cost of the search is measured by the number of characters of S that the algorithm has to look at. Several algorithms are known [13] which never need to access S more than cn times, where c is some constant greater than or equal to 1 (typically these algorithms end up accessing some letters several times). One particular example, the Boyer-Moore algorithm [2], turns out not only to be linear in its worst-case behavior [10, 13], but is also sublinear on the average; i.e., on the average it only has to look at αn characters of S, where $\alpha < 1$ is a constant depending on the pattern A and the size q of the alphabet. The question then immediately arises as to whether there is an algorithm that is sublinear in the worst case; i.e., which never looks at more than βn characters of S, where $\beta < 1$. This question was answered in the negative by Rivest [20]:

THEOREM 1.7. *Let A be any pattern of length k. Any algorithm that purports to decide for an arbitrary string S of length n whether A appears in S will need to examine at least $n-k+1$ characters of S for some string S. Furthermore, the algorithm will need to examine all n characters in the worst case for infinitely many values of n.*

We will present another proof of this result in Section 5. It differs from Rivest's in that (1.3) gives us an explicit form for the recursion satisfied by $f(n)$, the number of strings of length n that do not contain A. We should mention that the bound $n-k+1$ is best possible, as is shown in [20].

In Section 6 we will consider a considerably different result related to string autocorrelations. If $f(n)$ counts all the q-ary strings of length n that do not contain a particular pattern A of length k, then (1.3) gives us a very simple expression for the generating function $F(z) = \sum f(n)\, z^{-n}$, which shows, in particular, that $f(n)$ satisfies a k-term linear recurrence with constant coefficients. One might ask whether this is indeed the simplest possible expression of this kind. If $q = 2$ and $A = 10 \cdots 0$ is of length k, then

$$\frac{zAA_z}{1+(z-2)AA_z} = \frac{z^k}{1+(z-2)z^{k-1}} = \frac{zBB_z}{1+(z-2)BB_z} - \frac{z}{z-1},$$

where $B = 0 \cdots 0$ is of length $k-1$, which is simply a reflection of the combinatorial fact that any string S of length n that contains B also contains A, unless $S = 0 \cdots 0$. The question is thus whether other generating functions

$F(z)$ can be decomposed this way. The answer, at least for $q \geqslant 3$, is no; the polynomials $1 + (z-q)AA_z$ are irreducible, so that the $F(z)$ cannot be written as sums of rational functions with rational coefficients whose denominators have degrees smaller than k. This follows from the following more general result.

THEOREM 1.8. *If $p(z)$ is a polynomial with coefficients* 0 *and* 1, *and* $q \in Z^+$, $q \geqslant 3$, *then $(z-q)p(z)+1$ is irreducible.*

This theorem, and some generalizations, will be proved in Section 6. Here we will only mention that the case $q = 2$ is still open. We saw above that if $p(z) = z^m$, then $(z-2)p(z)+1$ is reducible. We conjecture that this is the only such case, and that if $p(1) \neq 1$, then $(z-2)p(z)+1$ is irreducible (still with the assumption that the coefficients of $p(z)$ are 0 and 1). If true, this would prove that in the binary case the expansion (1.3) is usually the simplest possible.

Finally, in the last section, we prove the result mentioned before, namely that the number of strings of any given length that do not contain A is a monotonic increasing function of AA_q.

2. BASIC GENERATING FUNCTIONS

In this section we prove a generalization of Theorem 1.1 and indicate some extensions of it. Let $\{A, B, \ldots, T\}$ be a reduced set of patterns, and let X be a pattern that contains none of $A, B, \ldots, T$ (but which may itself be contained in one or more of them). We allow X to be the empty pattern, in which case we set $|X| = 0$. We let $f(n) = f(X; A, \ldots, T; n)$ denote the number of strings of length n over our alphabet which start with X and do not contain any of $A, B, \ldots, T$. We also let $f_H(n)$ denote the number of strings of length n which start with X, end with H, and do not contain any of $A, B, \ldots, T$ except for that single appearance of H at the end of the string. We define the generating functions

$$F(z) = \sum_{n=0}^{\infty} f(n)\, z^{-n}, \qquad F_H(z) = \sum_{n=0}^{\infty} f_H(n)\, z^{-n}.$$

THEOREM 2.1. *With notation as above, $F(z), F_A(z), \ldots, F_T(z)$ satisfy the following system of linear equations*:

$$\begin{gathered}
(z-q)\,F(z) + zF_A(z) + zFB(z) + \cdots + zF_T(z) = z^{1-|X|} \\
F(z) - zAA_ZF_A(z) - zBA_zF_B(z) \cdots - zTA_zF_T(z) = z^{1-|H|}XA_z \\
\cdots \\
F(z) - zAT_zF_A(z) - zBT_zF_B(z) \cdots - zTT_zF_T(z) = z^{1-|H|}XT_z,
\end{gathered} \tag{2.2}$$

where if X is the empty string, we set $|X| = 0$ and $XY_z = 0$ for all Y.

Proof. If, for $n \geqslant |X|$, we adjoin any character from our alphabet to one of the strings counted by $f(n)$, the resulting string will be counted by exactly one of $f(n+1)$, $f_A(n+1), \ldots, f_T(n+1)$. Hence

$$qf(n) = f(n+1) + f_A(n+1) + \cdots + f_T(n+1) \qquad \text{for} \quad n \geqslant |X|.$$

If we now multiply both sides of this equation by z^{-n}, sum on $n \geqslant |X|$, and use the fact that $f(|X|) = 1$, $f_A(|X|) = 0$, we obtain

$$qF(z) = zF(z) - z^{1-|X|} + zF_A(z) + \cdots + zF_T(z),$$

which is the first equation in our system above.

Next, let $H = h_1 \cdots h_r$ be one of $A, \ldots, T$. If $n \geqslant |X|$ and $Y = y_1 \cdots y_n$ is counted by $f(n)$, consider the string $Y * H = y_1 \cdots y_n h_1 \cdots h_r = z_1 \cdots z_{n+r}$ (see Fig. 1). There is a first occurrence of one of $A, B, \ldots, T$ in $Y * H$; say $G = g_1 \cdots g_s$ occurs at position $t, g_1 \cdots g_s = z_{t-s+1} \cdots z_t$, and $z_1 \cdots z_t$ is counted by $f_G(t)$. Since G does not occur in Y, we must have $t > n$, and hence $g_{s-1+n+1} \cdots g_s = h_1 \cdots h_{t-n}$, so that the $(t-n)$th element in GH, counting from the right, is 1, which we denote by $t - n \in GH$. Conversely, if $t - n \in GH$, then any string counted by $f_G(t)$ arises from the concatenation of a string counted by $f(n)$ and H. Hence

$$\begin{aligned} f(n) = & \sum_{r \in AH} f_A(n+r) \\ & + \sum_{r \in BH} f_B(n+r) + \cdots + \sum_{r \in TH} f_T(n+r) \qquad \text{for} \quad n \geqslant |X|. \end{aligned} \tag{2.2}$$

To complete our proof we need to consider the sums on the right side above for $n < |X|$. (If we were only interested in proving Theorem 1, this would be unnecessary, as $|X| = 0$ there.) We claim that if $n < |X|$, then

$$\sum_{r \in AH} f_A(n+r) + \cdots + \sum_{r \in TH} f_T(n+r) = \begin{cases} 1 & \text{if } |X| - n \in XH, \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

Denote $X = x_1 \cdots x_{|X|}$, $H = h_1 \cdots h_{|H|}$. First of all, suppose that $f_G(n+r)$ is positive for some $r \in GH$, and let $Z = z_1 \cdots z_{n+r}$ be counted by $f_G(n+r)$. Then $z_{n+r-|G|+1} \cdots z_{n+r} = G$, $z_1 \cdots z_{|X|} = X$. But $r \in GH$ means that $z_{n+1} \cdots z_{n+r} = h_1 \cdots h_r$, so in particular $z_{n+1} \cdots z_{|X|} = h_1 \cdots h_{|X|-n}$, and $|X| - n \in XH$. (Fig. 2). Thus unless $|X| - n \in XH$, the sum in (2.3) is zero.

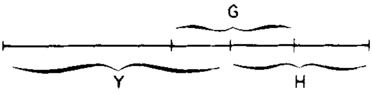

FIG. 1. If G occurs in the concatenation of Y and G, but occurs in neither Y nor H separately, then a suffix of G is a prefix of H.

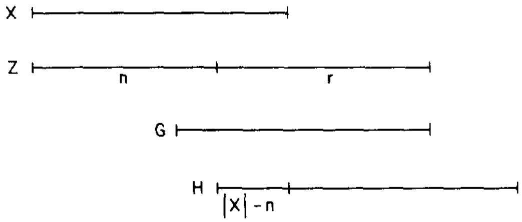

FIG. 2. If X is a prefix of Z, G a suffix of Z, $n < |X|$, and the last r characters of G equal the first r characters of H, then the first $|X| - n$ characters of H are a suffix of X.

Next suppose that $|X| - n \in XH$, and consider the string $Z = z_1 \cdots z_{n+|H|} = x_1 \cdots x_{|X|} h_{|X|-n+1} \cdots h_{|H|} = x_1 \cdots x_n h_1 \cdots h_{|X|}$. Then any string that might be counted by the left side of (2.3) is a prefix of Z. However, since there is a unique first appearance of one of $A, \ldots, T$ in Z, we conclude that the left side of (2.3) equals 1, which proves our claim.

The proof of Theorem 2.1 can now be easily completed. We multiply (2.2) and (2.3) by z^{-n} and sum on n, the first for $n \geqslant |X|$ and the second for $n < |X|$. Since

$$\begin{aligned}\sum_{n \geqslant 0} \sum_{r \in GH} f_G(n+r)\, z^{-n} &= \sum_{r \in GH} z^r \sum_{n \geqslant 0} f_G(n+r)\, z^{-(n+r)} \\ &= zGH_z F_G(z),\end{aligned}$$

we obtain the remaining equations of Theorem 2.1.

As we mentioned in the introduction, we regard Theorem 2.1 (and Theorem 1.1, which is an immediate corollary obtainable by letting X be the empty string) as our main result. However, we will now outline how the method can be extended to count the number of strings with specified numbers of occurences of a given set of patterns. For simplicity we consider only the case of a single pattern A of length k, and we let $f_r(n)$ denote the number of strings of length n which contain exactly r appearances of A (overlapping copies of A being counted separately) and $g_r(n)$ denote the number of strings of length n which contain exactly $r + 1$ appearances of A, one of them at the very end. We let $F_r(z)$ and $G_r(z)$ be the corresponding generating functions. By (1.4), we know that

$$F_0(z) = \frac{zAA_z}{1 + (z - q)AA_z},$$

$$G_0(z) = \frac{1}{1 + (z - q)AA_z}.$$

Now suppose that $S = s_1 \cdots s_n$ is a string that contains exactly $r + 1$ A's. If the $(r+1)$ A occurs at position m, then $s_1 \cdots s_m$ is counted by $g_r(m)$, while $s_n s_{n-1} \cdots s_{m-k-1}$ is counted by $g_0(n-m+k)$. Hence

$$f_{r+1}(n) = \sum_m g_r(m)\, g_0(n-m+k),$$

and therefore

$$F_{r+1}(z) = G_r(z)\, G_0(z)\, z^k, \tag{2.4}$$

so that, for example,

$$F_1(z) = \frac{z^k}{(1+(z-q)AA_z)^2}.$$

Finally, if $S = s_1 \cdots s_n$ contains exactly rA's, then $s_1 \cdots s_n x$ is counted either by $f_r(n+1)$ or by $g_r(n+1)$. However, if $z_1 \cdots z_{n+1}$ is counted by $g_r(n+1)$, then $z_1 \cdots z_n$ may also be counted by $g_{r-1}(r)$. Hence

$$q f_r(n) = f_r(n+1) + g_r(n+1) - g_{r-1}(n+1),$$

and therefore

$$G_r(z) = G_{r-1}(z) + \frac{q-z}{z} F_r(z). \tag{2.5}$$

The expressions (2.4) and (2.5) can now be used to calculate all of the $F_r(z)$ and $G_r(z)$.

3. Unequal Probabilities

In this section we generalize our basic results to the case of a biased q-sided die. For simplicity we will only consider a generalization of Theorem 1.1. We first need to introduce some new notation. We consider a reduced set of q-ary patterns $A, \ldots, T$. We let $s(n)$ be the probability that none of $A, \ldots, T$ will occur in the first n rolls of the die, with $s(0) = 1$, and $s_H(n)$ be the probability that H comes up on the nth throw, but that none of $A, \ldots, T$ come up on any of the preceding throws. We define the generating functions

$$Q(z) = \sum_{n=0}^{\infty} s(n)\, z^{-n},$$

$$Q_H(z) = \sum_{n=0}^{\infty} s_H(n)\, z^{-n}.$$

Finally, in place of the old correlation polynomials GH_z, we introduce

$$c_{GH}(z) = \sum_{r \in GH} z^{r-1} Pr(h_{r+1} \cdots h_{|H|}),$$

where, just as in Section 2, $r \in GH$ means that the last r characters of G equal the first r characters of H, and, by our independence assumption, $Pr(h_{r+1} \cdots h_{|H|}) = Pr(h_{r+1}) \cdots Pr(h_{|H|})$. We will assume that each letter of our alphabet has a nonzero probability of appearing, as otherwise we might as well eliminate it from consideration.

If none of $A,..., T$ has occurred in the first n rolls, then either none will occur on the $(n+1)-st$, or else exactly one will. This observation leads to the recurrence

$$s(n) = s(n+1) + s_A(n+1) + s_B(n+1) + \cdots + s_T(n+1) \qquad \text{for } n \geqslant 0. \quad (3.1)$$

The probability that none of $A,..., T$ occurs in the first n throws and that the following $|H|$ throws will produce $H = h_1 \cdots h_{|H|}$ is just $s(n) Pr(H)$, and therefore, by an argument analogous to that of Section 2,

$$\begin{aligned} Pr(H)\, s(n) = {} & \sum_{r \in AH} s_A(n+r)\, Pr(h_{|H|-r+1} \cdots h_{|H|}) + \cdots \\ & + \sum_{r \in TH} S_T(n+r)\, Pr(h_{|H-n+1} \cdots h_{|H|}) \qquad \text{for} \quad n \geqslant 0. \, (3.2) \end{aligned}$$

If we multiply (3.1) (resp. (3.2)) by z^{-n} and sum on n, we obtain our main result:

THEOREM 3.3. *With notation as above, the generating functions $Q(z)$, $Q_A(z),..., Q_T(z)$ satisfy the following system of linear equations*:

$$\begin{gathered} (z-1)\, Q(z) + zQ_A(z) + \cdots + zQ_T(z) = z \\ Q(z) - zc_{AA}(z)\, Q_A(z) - \cdots - zc_{TA}(z)\, Q_T(z) = 0 \\ \cdots \\ Q(z) - zc_{AT}(z)\, Q_A(z) - \cdots - zc_{TT}(z)\, Q_T(z) = 0. \end{gathered}$$

4. Optimal Strategies in the Coin-Tossing Game

This section is devoted to a proof of Theorem 1.6. That is, we consider a two-person game played with an unbiased q-sided die. Given any sequence $A = a_1 \cdots a_k$, $k \geqslant 3$ (the case $k = 2$ can be easily disposed of), as the first player's choice, we will show that the optimal choice of length k for the second player is always of the form $B = ba_1 \cdots a_{k-1}$ for a suitable b. To be more precise, let $A' = a_1 \cdots a_{k-1}$, and let r be the basic period of A' (that is, the smallest nonzero shift that causes A' to overlap itself), so that $a_{r+1} \cdots a_{k-1} = a_1 \cdots a_{k-r-1}$. (If A' has the trivial autocorrelation $A'A' = 10 \cdots 0$, then we take $r = k - 1$.) Choose $\tilde{b}$ so that $\tilde{b} \neq a_r$, and set $\tilde{B} = \tilde{b} a_1 \cdots a_{k-1}$. This choice ensures that $\tilde{B}$ has only periods $> r$. What we will show is that the odds is favor of $\tilde{B}$ occurring before A are strictly greater than the odds is favor of any $B = b_1 \cdots b_k$ for which $b_2 \cdots b_k \neq A'$. Furthermore, we will show the odds in favor of $\tilde{B}$ are always greater than 1 and that they are $\geqslant q/(q-1) - o(1)$ as $k \to \infty$, no matter what the choice of A. (It can be shown that for suitable A, say $A = 10 \cdots 0$, the odds in favor of the second player can be held to $q/(q-1)$, no matter what sequence B he chooses. For $q = 2$, the smallest odds to which the second player can be held are 2 for $k = 3$, 9/5 for $k = 4$, 17/9 for $k = 5$, and 33/17 for $k = 6$. The optimal choices for the first player there are of the form $10 \cdots 011$ for $k \geqslant 4$.)

Since there are $q - 1$ choices for $\tilde{b}$, we see that $\tilde{B}$ is uniquely defined only for $q = 2$. It is clear in any event that for $q \geqslant 3$, there will in general be no single optimal choice for the second player; for example, if $A = 0 \cdots 0$, then $B = 10 \cdots 0$ or $B = 20 \cdots 0$ give equal advantage to the second player. (In general, different choices of $\tilde{b}$, $\tilde{b} \neq a_r$, will give different odds.) If $q = 2$, then $\tilde{B}$ is not always the optimal choice for the second player. For example, if $A = 0100$, then $r = 2$, $\tilde{b} = 0$, so $\tilde{B} = 0010$, and the odds in favor of $\tilde{B}$ are (by (1.5)) 3/2. However, the odds that $B = 1010$ will occur before A turn out to be, again by (1.5), $9/5 > 3/2$. In general, even for $q = 2$ there does not seem to be any simple rule for determining the best beater of A, except that it has to be of the form $ba_1 \cdots a_{k-1}$. Furthermore, for $q = 2$ the best beater always seems to be unique, but we have no proof of this in general.

Before proceeding with the proof we should discuss the underlying idea. The odds in favor of B appearing before A are given by formula (1.5). To maximize these odds, one should clearly chose B so as to make BB_q and AB_q small and BA_q large. It turns out that all of these criteria are met quite well by $B = \tilde{B}$, and we obtain the lower bound (4.3). On the other hand, if $B = b_1 \cdots b_k$ is any other choice with $b_2 \cdots b_k \neq A'$, then BA_q cannot be very large, with the result that $BB_q - BA_q$ is relatively large, and so even if $AB_q = 0$, the odds in favor of B are not very large (Inequality (4.4)). (There are some complications for $q = 2$ which require the consideration of special cases.)

The reason for the particular choice of $\tilde{b}$ in the definition of $\tilde{B}$ is that it makes the basic period t of $\tilde{B}$ (i.e., the smallest positive shift that causes $\tilde{B}$ to overlap itself) very large. We will assume that $t \leqslant k - 2$ (if $t \geqslant k - 1$, there is nothing to prove, as t will then automatically satisfy our definition of a large period). Since $\tilde{B} = \tilde{b}A'$, any period of $\tilde{B}$ is a period of A'. On the other hand, no period of $\tilde{B}$ that is $\leqslant k - 2$ can be a multiple of r, the basic period of A';

if $A^* = a_1 \cdots a_r$, then $A' = A^* \cdots A^* A^+$, where A^+ is a prefix of A^* (which may be the empty string), and since $\tilde{b} \neq a_r$, we have $\tilde{b} \neq a_{mr} = a_r$ for all $m \geqslant 1$, which is the desired conclusion. But now the fact that t is a period of A' which is not a multiple of the basic period r implies that $t + r \geqslant k$. This is a corollary of the general results in [8], but we will give an easier independent proof of this weak result. Suppose therefore that s is the smallest integer such that $s \leqslant k - 1 - r$, is a period of A', but s is not a multiple of r. We obtain a contradiction by proving that $s - r$ is a period of A'. Since s is a period, $a_{s+1} a_{s+2} \cdots a_{k-1} = a_1 \cdots a_{k-s-1}$, and so, since $k - s - 1 \geqslant r$, $a_1 \cdots a_r = a_{s+1} \cdots a_{s+r}$. But r is a period, so $a_{s+1-r} \cdots a_s = a_{s+1} \cdots a_{s+r} = a_1 \cdots a_r$, and therefore $a_1 \cdots a_{k-s-1+r} = a_1 \cdots a_r a_1 \cdots a_{k-s-1} = a_{s+1-r} \cdots a_s a_{s+1} \cdots a_{k-1}$, and thus $s - r$ is a period of A'. But $s - r$ is shorter than s and is not a multiple of r, which contradicts the minimality of s. Therefore there is no such s, and $t + r \geqslant k$ as claimed.

We have shown above that $t \leqslant k - 2$, then $t + r \geqslant k$. But $t \geqslant r + 1$, so $t \geqslant (k + 1)/2$. We will show next that we cannot have $t = (k + 1)/2$. For this equality to hold, we would need to have k odd, and $r = t - 1 = (k - 1)/2$, so that $A' = A^* A^*$, with $A^* = a_1 \cdots \mathrm{a}_r$. But since $t = r + 1$ is also a period, we must have $a_2 \cdots a_r = a_1 \cdots a_{r-1}$, which implies that $a_1 = a_2 = \cdots = a_r$, and so 1 is a period of A. Therefore we must have $r = 1$ (as r is the basic period of A) and $k = 3$. But we assumed that $t \leqslant k - 2$, which in this case equals 1, which contradicts the fact that $t > r$. Therefore we cannot have $t = (k + 1)/2$, and so $t \geqslant (k + 2)/2$.

We have now shown that if the basic period t of $\tilde{B}$ satisfies $t \leqslant k - 2$, then $t \geqslant (k + 2)/2$. We next wish to show that $t \geqslant (k + 2)/2$ in all cases (with the convention that if $\tilde{B}$ has no nontrivial period, then $t = k$). Since $k - 1 \geqslant (k + 2)/2$ for $k \geqslant 4$, we only have to consider $k = 3$, $t = 2$. But $1 \leqslant r < t = 2$ implies that $r = 1$, and so $A' = aa$, $\tilde{B} = baa$ for some $b \neq a$, and then $t \geqslant 3$, a contradiction.

Having shown that $t \geqslant [(k + 1)/2] + 1$ (where $[x]$ denotes the greatest integer less than or equal to x), we proceed to estimate the odds in favor of $\tilde{B}$. First of all, since $m \in A\tilde{B}$ implies $m - 1 \in \tilde{B}\tilde{B}$ or $m = 1$, we obtain

$$A\tilde{B}_q \leqslant 1 + q + q^2 + \cdots + q^{[k/2]-1} = \frac{q^{[k/2]} - 1}{q - 1}. \tag{4.1}$$

Next, since $m \in \tilde{B}\tilde{B}$, $m > 1$, implies that $m - 1 \in \tilde{B}A$, $\tilde{B}\tilde{B}_q - \tilde{B}A_q$ equals $q^{k-1} - q^{k-2}$ plus a sum of terms of the form $q^{m-1} - q^{m-2}$, where $m \in \tilde{B}\tilde{B}$, $1 < m < k$, plus possibly 1. Since the largest $m \in \tilde{B}\tilde{B}$ for which $m < k$ satisfies $m \leqslant [k/2] - 1$, we obtain

$$\tilde{B}\tilde{B}_q - \tilde{B}A_q \leqslant q^{k-1} - q^{k-2} - q^{[k/2]-2}, \tag{4.2}$$

unless $m \in \tilde{B}\tilde{B}$ for $m = 1, 2, \ldots, [k/2] - 1$. But if $m \in \tilde{B}\tilde{B}$ for $1 \leqslant m \leqslant [k/2] - 1$, then $\tilde{B} = C * D * C$, (the concatenation of C, D, and C again), where $C = \tilde{b}\tilde{b} \cdots \tilde{b}$ is of length $[k/2] - 1$, and $D = d_1 \cdots d_s$ has $s = 2$ if k id even and $s = 3$ if k is odd. Let $C' = \tilde{b} \cdots \tilde{b}$ be of length $[k/2] - 2$. We can assume that $k \geqslant 6$, since otherwise (4.2) will hold by our previous observations. Hence C' is not the empty string. Then $A' = C' * D * C$. Now $\tilde{b}$ was defined to be a character such that $\tilde{b} \neq a_r$, where r is the basic period of A'. Therefore C, which is a suffix of A', must also be prefix of A'; i.e., $d_1 = \tilde{b}$. But then $[k/2] \in \tilde{B}A$, and so

$$\tilde{B}\tilde{B}_q - \tilde{B}A_q \leqslant q^{k-1} - q^{k-2}$$

in this case. Hence (4.2) always holds.

Combining (4.1) and (4.2), we obtain the estimate

$$\frac{AA_q - A\tilde{B}_q}{\tilde{B}\tilde{B}_q - \tilde{B}A_q} \geqslant \frac{AA_q - (q^{[k/2]} - 1)/(q - 1)}{q^{k-1} - q^{k-2} + q^{[k/2]-2}}. \tag{4.3}$$

In particular, since $AA_q \geqslant q^{k-1}$, this estimate shows that the odds in favor of $\tilde{B}$ are at least $q/(q-1) - O(q^{-k/2})$ as $k \to \infty$.

Suppose now that $B = b_1 \cdots b_k$ is such that $b_2 \cdots b_k \neq A$. Then $BA_q \leqslant q^{k-3} + q^{k-4} + \cdots + 1$, and since $BB_q \geqslant q^{k-1}$, the odds in favor of B are

$$\frac{AA_q - AB_q}{BB_q - BA_q} \leqslant \frac{AA_q}{q^{k-1} - \dfrac{q^{k-2} - 1}{q - 1}}. \tag{4.4}$$

In order to show that the odds in favor of B are smaller than those in favor of $\tilde{B}$ it will therefore suffice to show that

$$\frac{AA_q}{q^{k-1} - (q^{k-2} - 1)/(q - 1)} < \frac{AA_q - (q^{[k/2]} - 1)/(q - 1)}{q^{k-1} - q^{k-2} + q^{[k/2]-2}} \tag{4.5}$$

for all A. (Unfortunately this will not be true in general, and we will need to consider special cases when $q = 2$.) Now (4.5) is equivalent to

$$\frac{q^k - q^{k-1} - q^{k-2} + 1}{q - 1}\,\frac{q^{[k/2]} - 1}{q - 1}$$
$$< AA_q \left(q^{k-2} - q^{[k/2]-2} - \frac{q^{k-2} - 1}{q - 1} \right).$$

If $q = 2$, then the right side above is in general negative. Hence let us assume that $q \geqslant 3$. Since $AA_q \geqslant q^{k-1}$, and $q^k - q^{k-1} - q^{k-2} + 1 < (q-1)\, q^{k-1}$, it will suffice to prove that

$$\frac{q^{[k/2]}-1}{q-1} \leqslant q^{k-2} - q^{[k/2]-2} - \frac{q^{k-2}-1}{q-1},$$

which is equivalent to

$$(1 + q^{-1} - q^{-2})\, q^{[k/2]} \leqslant (q-2)\, q^{k-2} + 2,$$

which clearly holds for $q \geqslant 3$ and $k \geqslant 3$. This then completes the proof of Theorem 1.6 when $q \geqslant 3$.

It remains for us to consider the case $q = 2$. First let us suppose that either $k-2 \notin BA$ or $k-3 \notin BA$ (if $k=3$, we require only that $k-3 \notin BA$). Then $BA_2 \leqslant 2^{k-3} + 2^{k-5} + 2^{k-6} + \cdots + 1 = 3 \cdot 2^{k-4} - 1$, and thus

$$\frac{AA_2 - AB_2}{BB_2 - BA_2} \leqslant \frac{AA_2}{2^{k-1} - 3 \cdot 2^{k-4} + 1} = \frac{AA_2}{5 \cdot 2^{k-4} + 1}.$$

Therefore in order to prove that B is not the best beater it will suffice (by (4.3)) to prove that

$$\frac{AA_2}{5 \cdot 2^{k-4} + 1} < \frac{AA_2 - 2^{[k/2]} + 1}{2^{k-2} + 2^{[k/2]-2}},$$

which is equivalent to

$$(2^{[k/2]} - 1)(5 \cdot 2^{k-4} + 1) < AA_2(2^{k-4} + 1 - 2^{[k/2]-2}). \tag{4.6}$$

Since $AA_2 \geqslant 2^{k-1}$, we find after a few rearrangements that it suffices to prove that

$$2^{2k-5} - 7 \cdot 2^{k+[k/2]-4} + 13 \cdot 2^{k-4} - 2^{[k/2]} + 1 > 0.$$

This inequality is easily seen to hold for $k = 3$ and all $k \geqslant 7$, and therefore B is not the best beater in these cases. The cases $4 \leqslant k \leqslant 6$ can be disposed of either by exhaustive search for the best beaters or by a rather tedious special case analysis which we omit.

To complete the proof we now only need to consider the case where $k-1 \notin BA$, $k-2 \in BA$, and $k-3 \in BA$ (and $k \geqslant 4$, as $k=3$ has already been dealth with). Since $k-2 \in BA$ implies that $b_1 \cdots b_{k-2} = a_3 \cdots a_k$, and $k-3 \in BA$ implies that $b_1 \cdots b_{k-3} = a_4 \cdots a_k$, we conclude that $b_1 = b_2 = \cdots = b_{k-2} = a_3 = a_4 = \cdots = a_k$. Let this common value be 0, say. Then $A = 0 \cdots 0 a_{k-1} a_k$, $B = b_1 b_2 0 \cdots 0$. We now need to consider cases. If $a_{k-1} = a_k = 0$, then $\tilde{B} = 10 \cdots 0$ gives odds of $2^k - 1$, whereas the odds in favor of B can be at most $AA_2/(2^{k-1}-(2^{k-3}+2^{k-4}+\cdots+1))=(2^k-1)/(2^{k-2}+1)$. The other cases are disposed of similarly, and we omit the details.

5. Worst Case Behavior of Pattern-Matching Algorithms

Before giving the proof of Theorem 1.7, we have to specify the kind of algorithm our result applies to. We assume that the q-ary string $S=s_1\cdots s_k$ which is to be searched for the occurrence of the pattern $A=a_1\cdots a_k$ is stored in a random access memory, so that the algorithm can examine any one of the characters in S. The algorithms we allow are of the very general "decision-tree" type. That is, such algorithms successively examine the characters of S, where the choice of the character to be read at stage r may depend on the outcomes of all the preceding $r-1$ examinations.

Suppose that there does exist an algorithm which, given any string S of length n, can determine whether A appears in S by examining $\leqslant$ n $-$ 1 characters of S. Suppose that for a particular string S, the algorithm examines only positions $i_1,\ldots,i_r$. Consider then the q^{n-r} strings $S'=s'_1\cdots s'_n$ which have $s_{i_j}=s_{i_j}$ for $1\leqslant j\leqslant r$. When searching one of these strings S' for occurrences of A, the algorithm would examine precisely the same positions as it does in searching S, and would reach the same conclusion about whether A does or does not occur. Therefore if the algorithm always examines $\leqslant n-1$ character for any string S, the number of strings S which do not contain A is a multiple of q. Our basic results, however, show that this cannot happen for too many consecutive values of n.

Let $f(n)$ be the number of q-ary strings that do not contain A, with $f(0)=1$. Then, by (1.3) we have

$$\sum_{n=0}^{\infty} f(n)\,z^{-n}=\frac{zAA_z}{1+(z-q)AA_z}. \tag{5.1}$$

Let

$$1+(z-q)AA_z=z^k-\sum_{j=0}^{k-1} h_j z^i. \tag{5.2}$$

Then (5.1) states that $f(n)$ satisfies the linear recurrence

$$f(n)=\sum_{j=0}^{k-1} h_j f(n-k+j) \qquad \text{for} \quad n\geqslant k.$$

The important fact here is that by (5.2), $h_0 \equiv 1 \pmod q$. Hence we can run the recurrence for $f(n)$ backwards modulo q;

$$f(n) \equiv f(n+k) - \sum_{j=1}^{k-1} h_j f(n+j) (\bmod q) \qquad \text{for} \quad n \geqslant 0. \qquad (5.3)$$

If for some n we had $f(n+1) \equiv f(n+2) \equiv \cdots \equiv f(n+k) \equiv 0 \pmod q$, then by (5.3) we would have $f(n) \equiv 0 \pmod q$, and by repeating the argument, also $f(0) \equiv 0 \pmod q$, which would contradict our definition $f(0) = 1$.

We conclude from the above discussion that among any k consecutive values of $f(n)$, at least one is not divisible by q. For such a value of n, any algorithm has to examine all n characters in some string S (which depends on the algorithm under consideration) of length n in order to determine whether A is present or not. If n is such that $f(n) \equiv 0 \pmod q$, then there is a value of m with $n-k+1 \leqslant m \leqslant n-1$ such that $f(m) \not\equiv 0 \pmod q$. Since the number of characters that have to be examined in the worst case among strings of length n is at least as large as that among strings of length m (given a string $s_1 \cdots s_m$ that requires r examinations, we can find characters $b_1 \cdots b_{n-m}$ such that A occur in $S' = s_1 \cdots s_m b_1 \cdots b_{n-m}$ only among the first m characters, and so at least r examinations would be required in S' as well), we conclude that at least $n-k+1$ characters have to be examined in some string. This concludes our proof.

The above argument shows that if for some n, it suffices to examine $n-r$ characters in any string of length n in order to determine whether A occurs in that string or not, then $f(n) \equiv 0 \pmod{q^r}$. Unfortunately the converse is not true; the fact that $f(n) \equiv 0 \pmod{q^r}$ does not imply the existence of an algorithm which requires $\leqslant n-r$ character examinations. For example, if $\alpha = 000$ and $q = 2$, then $f(5) = 24$, $2^3 | 24$, but Theorem 1.7 guarantees that we have to examine at least 3 characters in some strings.

The method we used above can sometimes be used to prove worst-case lower bounds for the problem of finding whether any one of a set of patterns is in a given string. Our main results show that the number $f(n)$ of strings of length n which contain none of a given set of patterns does satisfy a linear recurrence. If the last nonzero coefficient in that recurrence is relatively prime to q, then our method applies and at least $n-c$ characters have to be examined (where c is a constant depending on the set of patterns we are looking for). However, quite often that last coefficient is not relatively prime to q. In fact, it is possible for all of the coefficients (except for the leading one) of the characteristic polynomial of the recurrence to be divisible by q, so that the power of q dividing $f(n)$ goes to infinity with n. Such is the case if $q = 2$ and we exclude the two patterns $A = 01110$ and $B = 00110$, where $\phi(z) = z^{11} - 2z^{10} + 2z^7 - 2z^6$. In this case the method of this section fails to rule out the possibility of a sublinear algorithm.

3. A Class of Irreducible Polynomials

Several large classes of irreducible polynomials are known (see [18, Vol. 2, VIII. Abschn., Kap. 2, Sec. 3]). For example, if $a_1, \ldots, a_n$ are distinct integers, then

$$(x-a_1)(x-a_2)\cdots(x-a_n)-1$$

is irreducible, and so is

$$(x-a_1)(x-a_2)\cdots(x-a_n)+1 \tag{6.1}$$

unless $n=4$ and $a_1=a_2-1=a_3-2=a_4-3$ or $n=2$ and a_2+2. In this section we will prove Theorem 1.8 and some generalizations of it, dealing with the irreducibility of polynomials of the type

$$(x-a)\, f(x)+b,$$

where a and b are integers, and $f(x) \in Z[x]$. Our method relies on an analysis of the zeros of these polynomials and basic algebraic number theory. Our method is very similar to the one that was used in [1] to prove the irreducibility of a subclass of the polynomials of the form (6.1).

We now prove Theorem 1.8. Let $f(x)$ be a polynomial with coefficients 0 and 1, and let $q \geqslant 3$ be an integer. We wish to prove that

$$g(x)=(x-q)\, f(x)+1$$

is irreducible. Let d be the degree of $f(x)$, which we may assume is $\geqslant 1$. We first prove that $g(x)$ has exactly one zero ρ_0 in $|x-q| \leqslant 1$. Note that in the disk $|x-q| \leqslant 1$ we have $|x| \geqslant q-1$, so

$$\begin{aligned}
|f(x)| &= = |x|^d - |x|^{d-1} - \cdots - 1\\
&\geqslant |x|^{d-1}\,(|x|-1-(q-1)^{-1}-\cdots-(q-1)^{1-d})\\
&\geqslant (q-1)^{d-1}\,(q-2-(q-1)^{-1}-\cdots-(q-1)^{1-d})\\
&= \frac{(q-2)(q-1)^d+1}{q-2} \geqslant 1.
\end{aligned}$$

Moreover, equality can conceivably hold only if $q=3$ and $x=2$. In that case, however, $f(x) \geqslant 2^d \geqslant 2$. Hence we conclude that $|f(x)| > 1$ in $|x-q| \geqslant 1$. Therefore $|(x-q)\, f(x)| > 1$ on $|x-q|=1$, and so by Rouche's theorem $(x-q)\, f(x)$ and $g(x)=(x-q)\, f(x)+1$ have the same number of zeros in $|x-q|<1$. But $(x-q)\, f(x)$ has exactly the single zero $x=q$ in $|x-q|<1$, so we conclude that $g(x)$ has exactly one zero ρ_0 in $|x-q|<1$, and none on $|x-q|=1$.

Let ρ be any zero of $g(x)$. Since $g(x)$ is a monic polynomial with integer coefficients, ρ is an algebraic integer. Therefore so are $\rho - q$ and $f(\rho)$. But then $(\rho - q)\,f(\rho) = -1$ implies that $\rho - q$ is a unit. Now supose that ρ is not conjugate to ρ_0. Then $\rho - q$ is not conjugate to $\rho_0 - q$, and so all of the conjugates of $\rho - q$ are of absolute value > 1. But the fact that $\rho - q$ is a unit implies that the product of all its conjugates equals ± 1. This is a contradiction, and so ρ must be conjugate to ρ_0. But this means that all of the zeros of $g(x)$ are conjugate, and so $g(x)$ is irreducible. Q.E.D.

The above proof of the irreducibility of $g(x)$ breaks down when $q = 2$. The crucial fact we have used when $q \geqslant 3$ was that there was exactly one zero of $g(x)$ in $|x - q| \leqslant 1$. This fails for $q = 2$. In the first place, when $f(x)$ is a monomial, $f(x) = x^d$, $x = 1$ is a zero of $g(x)$. In this case, however, it can be shown that $g(x)/(x-1)$ has exactly one zero in $|x - 2| \leqslant 2$, and therefore $g(x)/(x-1)$ is irreducible. We can therefore restrict ourselves to $f(x)$ which are not monomials. In this case it can be shown that if $d = \deg(f(x)) \leqslant 10$, then $g(x)$ has exactly one zero in $|x - 2| \leqslant 1$, and is therefore irreducible. However,

$$g(x) = (x-2)(x^{19} + x^9 + x^8 + \cdots + x + 1) + 1$$

has three zeros in $|x - 2| \leqslant 1$. (It is irreducible, though.) The smallest d such that there exists a polynomial $f(x)$ with coefficients 0 and 1 and of degree d for which $g(x) = (x-2)\,f(x) + 1$ has more than one zero in $|x - 2| < 1$ is unknown. Also, no polynomials of that form are known that are reducible (except for those for which $f(x) = x^d$).

One might wonder whether there is anything special about the groups of polynomials of the form (6.1). A check of all the polynomials $g(x) = (x-2)\,f(x) + 1$ where $f(x)$ has coefficients 0 and 1, $d = \deg(f(x)) \leqslant 6$, and $f(x) \neq x^d$ showed that, as might be expected of a random sample of polynomials, almost all had the symmetric group S_{d+1} as their group, but there were several exceptions. Thus it seems that little can be said in general.

Our method for proving irreducibility can be easily generalized to cover other classes of polynomials. For example, our proof applies with no changes to $g(x) = (x-q)\,f(x) - 1$, provided $q \geqslant 3$ and $f(x)$ has coefficients 0 and 1. Furthermore we can relax the conditions on the coefficients of $f(x)$. If $f(x) \in Z[x]$ is monic, then our proof shows, mutatis mutandis, that $g(x) = (x-q)\,f(x) \pm 1$ is irreducible for q sufficiently large. One can go even further, and show that if $f(x) \in Z[x]$ is any polynomial and a any nonzero integer, then $g(x) = (x-q)\,f(x) + a$ has no nonconstant divisors in $Z[x]$.

7. Comparison of the Number of Strings Not Containing Given Patterns

In this section we will prove that if $AA_q > BB_q$, then for any length n, strings of length n that do not contain A are at least as likely as those that do not contain B. To be more precise, let $g_K(n)$ denote the number of q-ary strings of length n which do not contain K. Then we will show that if $AA_q > BB_q$, then $g_A(n) \geqslant g_B(n)$ for all $n \geqslant 1$. In fact we will show more. Theorem 1.1 implies (see (1.4)) that the generating function $G_K(z) = \sum_{n \geqslant 0} g_K(n) z^{-n}$ is of the form

$$G_K(z) = \frac{zKK_z}{1+(z-q)KK_z}.$$

We will consider rational functions of the form

$$H_f(z) = \frac{zf(z)}{1+(z-q)f(z)}, \tag{7.1}$$

where $f(z)$ is a polynomial with coefficients 0 and 1. We will show that if $H_f(z)$ is expanded as

$$H_f(z) = \sum_{n=0}^{\infty} h_f(n) z^{-n}, \tag{7.2}$$

then, for any two polynomials f and g with coefficients 0 and 1, $f(q) > g(q)$ implies that $h_f(n) \geqslant h_g(n)$ for all $n \geqslant 0$.

The general problem of deciding whether one linear recurrence sequence dominates another is very difficult. In our case we can solve the problem due to the very special form of our recuurences. The asymptotic analysis of the sequences $h_f(n)$ presents no special difficulties. It can be shown (cf. [9]) that $1+(z-q)f(z)$ has a single zero $\theta = \theta_f$ with $|\theta| > 1.7$, which satisfies (with $\deg f(z) = k-1$)

$$\theta = q - f(q)^{-1} - f'(q) f(q)^{-3} + O(k^2 q^{-3k}),$$

and that

$$h_f(n) = \frac{\theta^n}{1-(q-\theta)^2 f'(\theta)} + O((1.7)^n), \tag{7.3}$$

where the constants implied by the O-notation are independent of f and n. Since it can be shown [9] that θ is monotone increasing with $f(q)$, it is clear that if $f(q) > g(q)$ then $h_f(n) > h_g(n)$ for n large enough (in fact for $n \geqslant c \deg f(z)$ for some constant c). In any event, since θ can be easily computed to great accuracy, (7.3) gives a method of obtaining very accurate estimates of $h_f(n)$.

Although analysis does give good estimates of $h_f(n)$, it does not seem capable of proving that $f(q) > g(q)$ implies $h_f(n) \geqslant h_g(n)$ for all n, and we will therefore use more elementary methods. The first observation we make is that $f(q) > g(q)$ holds if and only if $f(2) > g(2)$, since f and g have coefficients 0 and 1. Therefore it will suffice to prove our result if $f(2) = g(2) + 1$ as the general result will then follow by transitivity.

Let deg $f(z) = k - 1$, and write

$$f(z) = \sum_{j=0}^{k-1} f_{k-1-j} z^j, \qquad f_m = 0 \text{ or } 1, \qquad f_0 = 1.$$

Let us note that the definition (7.2) implies that $h_f(m) = q^m$ for $0 \leqslant m \leqslant k - 1$, and $h_f(k) = q^k - 1$, and that for $n \geqslant k$, $h_f(n)$ satisfies the linear recurrence

$$h_f(n) = qh_f(n-1) + \sum_{j=1}^{k-1} f_j(qh_f(n-j-1) - h_f(n-j)) - h_f(n-k). \tag{7.4}$$

We will use the convention that $h_f(n) = 0$ for $n < 0$. In our proof we will use the following auxilliary result, whose proof we postpone until the end of this section.

LEMMA 7.5. *With notation as above, we have*

$$h_f(n) \leqslant qh_f(n-1) - h_f(n-k) \qquad \text{for} \quad n \geqslant 1, \tag{7.6}$$

$$h_f(n) \geqslant (q-1)\{h_f(n-1) + h_f(n-2) + \cdots + h_f(n-k+1)\} \qquad \text{for} \quad n \geqslant 0. \tag{7.7}$$

We now show how this lemma can be used to prove our result. Let

$$g(z) = \sum_{j=0}^{k-1} g_{k-1-j} z^j, \qquad g_m = 0 \text{ or } 1.$$

Since $g(2) = f(2) - 1$, deg $g(z) = k - 1$ or $k - 2$. We will assume here that deg $g(z) = k - 1$, since the other case can be dealt with similarly. Then $g_0 = 1$, and there is an integer $s \geqslant 1$ such that $g_s = 0$, $g_{s+1} = g_{s+2} = \cdots = g_{k-1} = 1$, $f_s = 1$, $f_{s+1} = \cdots = f_{k-1} = 0$, and $f_j = g_j$ for $j < s$. The recurrence satisfied by $h_g(n)$ is

$$h_g(n) = qh_g(n-1) + \sum_{j=1}^{k-1} g_j(qh_g(n-j-1) - h_g(n-j)) - h_g(n-k) \tag{7.8}$$

for $n \geqslant k$. As our first step we will show that for $n \geqslant k$,

$$h_f(n) \geqslant qh_f(n-1) + \sum_{j=1}^{k-1} g_j(qh_f(n-j-1) - h_f(n-j)) - h_f(n-k). \tag{7.9}$$

Indeed, since $h_f(n)$ satisfies (7.4), to prove (7.9) it suffices to prove that

$$\begin{aligned} &qh_f(n-s-1) - h_f(n-s) - h_f(n-k) \\ &\qquad \geqslant \sum_{j=s+1}^{k-1} (qh_f(n-j-1) - h_f(n-j)) - h_f(n-k) \\ &\qquad = -h_f(n-s-1) + (q-1) \sum_{j=s+2}^{k} h_f(n-j). \end{aligned}$$

By Lemma 7.5, however, the left side above is $\geqslant 0$, while the right side is $\leqslant 0$. Hence this inequality, as well as (7.9), is true.

Before concluding the proof, we need to make another remark. If we define $u(n) = u_g(n) = qh_g(n-1) - h_g(n)$ for $n \geqslant 1$, and $u(n) = 0$ for $n < 1$, then by Lemma 7.5 applied to g rather than f we see that $u(n) \geqslant 0$ for all n. Moreover, $u(n)$ satisfies the same recurrence (7.8) as $h_g(n)$ for all $n > k$, and $u(k) = 1$. Let a_n denote the difference of $h_f(n)$ and the quantity on the right side of (7.9). We have shown that $a_n \geqslant 0$. We will now show that

$$h_f(n) = h_g(n) + \sum_{m=k}^{\infty} a_m u(n-m+k) \qquad \text{for all} \quad n \geqslant 0. \tag{7.10}$$

Proof of Lemma 7.5. We first use induction to prove (7.7) and the following weakened form of (7.6):

$$h_f(n) \leqslant qh_f(n-1) \qquad \text{for} \quad n \geqslant 1. \tag{7.11}$$

For simplicity we will use $h(n)$ to denote $h_f(n)$. Both (7.7) and (7.11) are clearly true for $n \leqslant k$. Suppose they are true for all $n \leqslant m-1$. We use (7.4) with $n = m$. Since $h(r) \leqslant qh(r-1)$ for $1 \leqslant r \leqslant m-1$, Eq. (7.4) implies that

$$\begin{aligned} h(m) &\leqslant qh(m-1) + \sum_{j=1}^{k-1} \{qh(m-j-1) - h(m-j)\} - h(m-k) \\ &= qh(m-1) + (q-1) \sum_{j=2}^{k} h(m-j) - h(m-1) \leqslant qh(m-1) \end{aligned}$$

by applying the induction hypothesis (7.7) with $n = m-1$. This proves

(7.11) for $n=m$. Next, by setting $n=m$ in (7.4) and applying (7.11) to the terms in the sum, we see that

$$\begin{aligned} h(m) &\geqslant qh(m-1) - h(m-k) \\ &= (q-1)\,h(m-1) + h(m-1) - h(m-k), \end{aligned}$$

and by (7.7) applied with $n=m-1$ we deduce

$$\begin{aligned} h(m) \geqslant (q-1)\,h(m-1) + (q-1) \sum_{j=2}^{k-1} h(m-j) \\ + (q-2)\,h(m-k). \end{aligned}$$

which yields the desired result. Thus (7.7) and (7.11) are true for all $n \geqslant 1$. But then (7.6) follows immediately, since it clearly holds for $n \leqslant k$, and for $n \geqslant k+1$ is obtained from (7.4) by applying (7.11) to deduce that each term in the sum is nonpositive.

ACKNOWLEDGMENTS

The authors should like to acknowledge the use of the MACSYMA system, developed at MIT's Laboratory for Computer Science, and the programming help of Larry Masinter of Xerox PARC.

REFERENCES

1. N. C. Ankeny, R. Brauer, and S. Chowla, A note on the class-numbers of algebraic number fields, *Amer. J. Math.* **78** (1956), 51–61.
2. R. S. Boyer and J. S. Moore, A fast string searching algorithm, *Comm. ACM* **20** (1977), 762–772.
3. S. Collings, Improbable probabilities, to be published.
4. J. H. Conway, private communication.
5. W. Feller, "An Introduction to Probability Theorey and Its Applications," Vol. 1. 3rd ed., Wiley, New York, 1968.
6. M. Gardner, On the paradoxical situations that arise from nontransitive relations, *Sci. American* (October 1974), 120–125.
7. I. P. Goulden and D. M. Jackson, An inversion theorem for cluster decomposition of sequences with distinguished subsequences, to be published.
8. L. J. Guibas and A. M. Odlyzko, Periods in strings, *J. Comb. Theory* (*A*) **30** (1981), 19–42.
9. L. J. Guibas and A. M. Odlyzko, Maximal prefix-synchronized codes, *SIAM J. Appl. Math.* **35** (1978), 401–418.
10. L. J. Guibas and A. M. Odlyzko, A new proof of the linearity of the Boyer-Moore string searching algorithm, in "Proc. 18th Foundations of Computer Sci. Symp., IEEE, 1977," pp. 189–195; also *SIAM J. Comput.* **9** (1980), 672–682.
11. H. Harborth, Endliche 0–1-Folgen mit gleichen Teilblöcken, *J. Reine Angew. Math.* **271** (1974), 139–154.

12. K. H. Kim, M. S. Putcha, and F. W. Roush, Some combinatorial properties of free seimgroups, *J. London Math. Soc.* (2), **16** (1977), 397–402.
13. D. E. Knuth, J. H. Morris, Jr., and V. R. Pratt, Fast pattern matching in strings, *SIAM J. Comput.* **6** (1977), 323–350.
14. R. T. Leslie, Recurrent composite events, *J. Appl. Prob.* **4** (1967), 34–61.
15. S.-Y. R. Li, A martingale scheme for studying the occurrence of sequence patterns in repeated experiments, to be published.
16. P. T. Nielsen, On the expected duration of a search for a fixed pattern in random data, *IEEE Trans. Inform. Theory* **IT-19** (1973), 702–704.
17. W. Penney, Problem: penney-ante, *J. Recreational Math.* **2** (1969), 241.
18. G. Polya and G. Szegő, "Aufgaben and Lehrsätze aus der Analysis II, 4. Auflage," Springer-Verlag Berlin/Heidelberg/New York, 1971.
19. L. Ramshaw, private communication.
20. R. L. Rivest, On the worst-case behavior of string-searching algorithms, *SIAM J. Comp.* **6** (1977). 669–674.
21. S. W. Roberts, Properties of control chart zone tests, *Bell System Tech. J.* **37** (1958), 83–114.
22. S. W. Roberts, On the first occurrence of any of a selected set of outcome patterns in a sequence of repeated trials, unpublished manuscript (1963).
23. B. Saperstein, Note on a clustering problem. *J. Appl. Prob.* **12** (1975), 629–632.
24. A. D. Solov'ev, A combinatorial identity and its application to the problem concerning the first occurrence of a ratre event, *Theory Prob. Appl.* **11** (1966), 276–282.
25. R. L. Tenney and C. C. Foster, Nontransitive dominance, *Math. Mag.* **49** (1976), 115–120.
26. J. G. Wendel, private communication.

Printed in Belgium