JOURNAL OF COMBINATORIAL THEORY, Series A 30, 19–42 (1981)

Periods in Strings

LEO J. GUIBAS

Xerox Palo Alto Research Center, Palo Alto, California 94305 and Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213

AND

ANDREW M. ODLYZKO

Bell Laboratories, Murray Hill, New Jersey 07974

Communicated by the Managing Editors

Received October 2, 1978

In this paper we explore the notion of *periods* of a string. A period can be thought of as a shift that causes the string to match over itself. We obtain two sets of necessary and sufficient conditions for a set of integers to be the set of periods of some string (what we call the *correlation* of the string). We show that the number of distinct correlations of length n is independent of the alphabet size and is of order $n^{\log n}$. By using generating function methods we enumerate the strings having a given correlation, and investigate certain related questions.

1. INTRODUCTION

Let X be a string of length n over some finite alphabet Σ of size q. We will index the elements of X from left to right by 0 through $n-1$, and write $X[i]$ for the ith element. A non-negative integer p, $p < n$, will be called a *period* of X if we have $X[i] = X[i+p]$, for i in $[0, n-p)$. Pictorially, p is a period if a second copy of X, shifted right by p positions and placed over the original copy, matches in the overlapping part. The set of all such p will be called the set of periods of X.

Our aim in this papers is to obtain necessary and sufficient conditions for a set of integers to be the set of periods of some string. In Section 2 we introduce an alternative notation that is somewhat easier to work with. Sections 3 and 4 contain the statements of necessary and sufficient conditions, along with discussion of related results. Section 5 contains the main result of the paper, namely, a proof that the conditions in Sections 3 and 4 actually characterize the sets which are periods of strings. In Section 6

0097-3165/81/010019-24$02.00/0

we discuss the enumeration of the sets of periods of strings of a given length, and in Section 7 we count the number of strings having a given set of periods.

This problem arose in connection with our work on string searching algorithms [1, 4, 5, 8]. Such algorithms work by attempting to match a pattern string at various places in a text string. The more sophisticated of these algorithms extract information from an unsuccessful match and use it to rule out other matches which have no chance of succeeding. These decisions invariably require knowledge of how the pattern matches over itself, that is, knowledge of the periods of the pattern, or of its prefixes or suffixes. The structure of the periods of the pattern or parts thereof can crucially affect the performance of such string searching algorithms.

Properties of the periods of strings have been investigated previously by Schützenberger and others in [2, 9, 11]. The already referenced work [5] also contributes to the problem. These papers prove and make use of the so-called GCD rule for periods, which will follow from our theory. Results related to those we exhibit in Section 7 were previously obtained by Harborth [7], who studied the related problem of the enumeration of strings of a certain length with a given minimal period. An extensive set of applications of the notion of correlation is described in [5].

2. Basic Definitions and Notations

If X and Y are two strings we will define the *correlation* of X over Y to be binary vector of the same length as X, composed as follows. The ith bit (from the left) of the correlation is determined by placing Y under X so that Y's leftmost character is under the ith character of X (from the left). Then, if all pairs of characters that are directly over each other match, the ith bit of the correlation is 1, else it is 0. For example, if $\Sigma = \{H, T\}$, $X =$ "HHTTHH" and $Y =$ "THHTHH," then the correlation is "000100," as depicted below:

```
X       H H T T H H
Y       T H H T H H.....0
          T H H T H H .....0
            T H H T H H .....0
              T H H T H H .....1
                T H H T H H .....0
                  T H H T H H .....0
```

We denote the correlation of X over Y by XY, a notation first introduced in [3]. Note that in general $XY \neq YX$. Let v denote the correlation of X over itself, i.e., $v = XX$. Let $n = |X|$ be the length of X, and number the bits of v from 0 to $n-1$. Then note that $v_p = 1$ if and only if p is a period of X. Thus the (auto-)correlation v of X provides a convenient encoding of the set of periods of X.

The smallest non-zero period of a string X will be called its *basic period*. By convention, the basic period of a string X of correlation $XX = 10 \ldots 0$ (i.e., no non-zero periods) is $|X|$. Two periods p and q of X will be called *independent* if neither is a multiple of the other. (Note trivially that multiples of a period are themselves periods.) If X has period p, then we will often use the fact that X can be written as $X = UU \cdots UU'$, where $|U| = p$, and U' is a prefix of U. Given any binary vector v, indexed [0, n), we will denote by $\pi(v)$ the smallest positive i such that $v_i = 1$. If no such i exists, we set $\pi(v) = n$. Thus $\pi(XX)$ denotes the basic period of X.

Note that if $|X| = |Y|$, but $X \neq Y$, then the correlation XY has the form $0 \cdots 0z$, where z is the autocorrelation of the first match Z of Y into X, i.e., longest Z such that $X = UZ$ and $Y = ZV$. Conversely, any bit vector of the form $0 \cdots 0z$, where z is the autocorrelation of some string Z, can obviously arise as the correlation of two easily constructed strings X and Y. For this reason we will confine ourselves from now on to the properties and characterization of autocorrelations. For brevity we will use the term correlation synonymously with autocorrelation.

3. The Propagation Rules

In this section we define two abstract properties of binary vectors that reflect necessary conditions satisfied by the sets of periods of strings. This latter fact will not be proved until Section 5, but for convenience of language we will continue to use the terminology of the previous section. We show that these two properties imply a previously known and very useful result on periods, the so-called GCD rule.

The forward propagation rule essentially asserts the transitivity of matching (or equality): if X has periods p, q, with $p < q$, then it also has $q + (q - p)$ as a period. As we will see, correlations satisfy this rule, as well as the backwards propagation rule described below.

Definition 3.1 (Forward Propagation Rule). A bit vector $v = (v_0, v_1, \ldots, v_{n-1})$ of length n satisfies the forward propagation rule if, whenever we have $v_p = v_q = 1$, with $p < q$, we also have $v_t = 1$, for all t of the form $p + i(q - p)$, $i = 0, 1, 2, \ldots$, and t in the range $[p, n)$.

The backward propagation rule asserts that if we follow the arithmetic progression defined by periods p and q to the *left*, and find that $p-(q-p)$ is *not* a period, then in proceeding to the left we must encounter at least as many 0's as we encountered 1's (really, full periods) in going to the right from q on (unless we fall off the beginning).

DEFINITION 3.2 (Backward Propagation Rule). A bit vector $v=(v_0, v_1,..., v_{n-1})$ of length n satisfies the backward propagation rule if the following condition holds. For every p and q such that $p<q\leqslant 2p$ with $v_p=v_q=1$, but $v_{2p-q}=0$, let $s=\lfloor (n-p)/(q-p)\rfloor$. Then for all t in the range $[0, 2p-q]$ of the form $p-i(q-p)$, $i=1, 2,..., s$, we have $v_t=0$.

The propagation rules indicate local conditions, in the sense that if v satisfies them, the so does any substring (i.e., set of consecutive elements) of v. In the sequel we will need only a special case of this observation.

LEMMA 3.1. *If v satisfies the forward and backward propagation rules, then so does any prefix or suffix of v.*

We now prove the GCD rule. For additional discussion of this remarkable result on the periods of strings see [2, 8].

THEOREM 3.1 (The GCD Rule). *Let $v=(v_0, v_1,..., v_{n-1})$ be a non-empty bit vector of length n satisfying the forward and backward propagation rules, and having $v_0=1$. Consider a pair of indices p and q in $[1, n)$ and let $t=\mathrm{GCD}(p, q)$. If $v_p=v_q=1$ and $p+q\leqslant n+t$, then we also have $v_t=1$.*

Proof. Without loss of generality we can assume $q\geqslant p$. Note that neither of the propagation rules will ever imply anything except about those bit positions of v whose index is a multiple of $t=\mathrm{GCD}(p, q)$. Thus we might as well assume $t=1$ (or equivalently confine our attention to the vector $v'=(v_0, v_t, v_{2t},..., v_{lt})$, where $l=\lfloor n/t\rfloor-1$). By the above lemma, every prefix w of v also satisfies the conditions of the theorem. As a consequence we can confine our attention to the special case $n=p+q-1$, for any larger n will a fortiori imply the same conclusion.

Thus we have reduced our problem to the case $t=1$, $n=p+q-1$. We must show that for such a v, $v_1=1$, or (equivalently, by forward propagation), that all elements of v are 1. We proceed by induction on q (the larger of p, q). Starting the induction is trivial. Suppose now that we know the result of the theorem for all pairs (p', q') with $p'<q'$, and $q'<q$. Write $q=mp+r$, with $0\leqslant r<p$. By Lemma 3.1, the vector $u=(u_0, u_1,..., u_{p+r-2})=(v_{mp}, v_{mp+1},..., v_{n-1})$ satisfies the forward and backward propagation rules. Vector u starts with a 1 ($v_{mp}=1$), and, if $r>1$, has periods p and r (i.e., $v_{(m+1)p}=v_q=1$). Further $\mathrm{GCD}(r, p)=$

$\mathrm{GCD}(p, q) = 1$, u has length $p + r - 1$, and $r < p < q$. The inductive hypothesis applies to u and allows us to conclude that u (and therefore v, to the right of and including index mp) is composed entirely of 1's. If $r = 1$, we trivially obtain the same conclusion.

But now to the right of (and including) index mp in v there are at least p 1's (since v has length $(m + 1)p + r - 1$). The backward propagation rule implies that no z in the interval $[(m - 1)p, mp)$ such that $v_z = 0$ exists, since $v_{(m-1)p} = 1$. For to the right of the rightmost such z we have at least p consecutive 1's (apply the rule to $z + 1$ and $z + 2$), but to the left we cannot have p consecutive 0's. In the same fashion, we can argue next that v is all 1's in the interval $[(m - 2)p, (m - 1)p)$, and so on, all the way to $[0, p)$. Thus we have obtained the desired conclusion that v is composed entirely of 1's. ∎

4. The Recursive Definition

We now introduce a recursive predicate on binary vectors which, as the next section shows, also turns out to be equivalent to the condition that the binary vector is a correlation.

Definition 4.1 (The Recursive Predicate Ξ). Let v be a bit vector of length n. Define $p = \pi(v)$. The vector v satisfies predicate Ξ iff v is empty (equivalently, $n = 0$), or v can be written as $v = (v_0, v_1, \ldots, v_{n-1})$ and satisfies the following constraints:

(1) $v_0 = 1$, and

(2) one of the following two mutually disjoint conditions holds:

Case (a). If $n/p \geqslant 2$, then let $r = n - p(\lfloor n/p \rfloor - 1)$. In this case we must have $v_i = 0$ for i in $[1, n - r)$, except at multiples of p (where $v_i = 1$). Further, if we let $w = (w_0, w_1, \ldots, w_{r-1}) = (v_{n-r}, \ldots, v_{n-1})$, then

(i) $w_p = 1$ or $r = p$,

(ii) if $\pi(w) < p$ then $\pi(w) > (r - p) + \mathrm{GCD}(p, \pi(w))$, and

(iii) w satisfies predicate Ξ.

Case (b). If $n/p < 2$, then let $r = n - p$. In this case we must have $v_i = 0$ for i in $[1, p)$ and, if we let $w = (w_0, w_1, \ldots, w_{r-1}) = (v_p, \ldots, v_{n-1})$, then w satisfies predicate Ξ.

Note that in case (a) we are choosing r so that $p \leqslant r < 2p$ and $n - r$ is a multiple of p. Because of the GCD rule, condition (ii) is equivalent to the requirement that $\pi(w)$ not be a proper divisor of p.

The procedure *Ksi* defined below in an ALGOL-like notation implements the above predicate. The predicate Ξ is true of vector v iff *Ksi*$[0, n]$: 1 (e.g., the first component of the returned record) is true. (We must remember to set the boundary condition $v[n] = 1$.) The second component of the returned record is the length of the basic period of v. Note that the number of bit position examinations done by *Ksi* is bounded by $2n$ (and thus *Ksi* runs in linear time). This follows since each bit position is examined at most once before the recursive call is made. *Ksi* is then invoked recursively on a virgin substring consisting of bits which have not yet been examined. There is an additional bit examined in the final test, but the total number of such examinations is certainly bounded by the number of recursive calls, which is bounded by n.

```
PROCEDURE Ksi[INTEGER l, r] RETURNS RECORD[BOOLEAN; INTEGER];
  BEGIN INTEGER i, p, q, s, newp; BOOLEAN flag;
  IF r = l THEN RETURN[TRUE, 0];
  IF v[l] = 0 THEN RETURN[FALSE, UNDEFINED];
  p ← 1;
  WHILE l + p < r AND v[l + p] = 0 DO p ← p + 1;
  COMMENT now p is the basic period of v[l..r];
  s ← (r − l) DIV p;
  IF s ⩾ 2 THEN
    BEGIN COMMENT case (a);
    q ← p * (s − 1);
    flag ← TRUE;
    FOR i IN [1..q) DO
      BEGIN
      IF (i MOD p) = 0 THE flag ← flag AND (v[l + i] = 1)
        ELSE flag ← flag AND (v[l + i] = 0);
      END;
    IF NOT flag THEN RETURN[FALSE, UNDEFINED]
    [flag, newp] ← Ksi[l + q, r];
    RETURN[flag AND (v[l + q + p] = 1) AND
        ((p = newp) OR (newp > r − l − q − p + GCD(newp, p))), p];
    COMMENT here we made the additional bit examination;
    END COMMENT case (a);
  ELSE BEGIN COMMENT casr (b);
    q ← p;
    [flag, newp] ← Ksi[l + q, r];
```

```
    RETURN[flag, p];
    END COMMENT case (b);
END.
```

Thus the procedure gives us a linear time algorithm for testing if a given bit vector can arise as the correlation of some string.

5. Proof of Equivalence

This section contains our main result, the proof of equivalence of the three characterizations of the sets of periods. Note that condition (1) below refers to *binary* strings. This implies the non-obvious fact that an alphabet of size two gives rise to all sets of periods that can arise with strings over an alphabet of arbitrary size.

THEOREM 5.1 (Characterization of Periods). *Let* $v = (v_0, v_1, \ldots, v_{n-1})$ *be a non-empty bit vector. Then the following four statements are equivalent*:

(1) v *is a correlation of a binary string,*

(1′) v *is a correlation of some string,*

(2) $v_0 = 1$ *and* v *satisfies the forward and backward propagation rules, and*

(3) v *satisfies predicate* Σ.

Proof. We will prove equivalence by showing that $(1) \Rightarrow (1') \Rightarrow (2) \Rightarrow (3) \Rightarrow (1)$.

$(1) \Rightarrow (1')$: Obvious.

$(1') \Rightarrow (2)$: Assume that v is the correlation of string Z. It immediately follows that $v_0 = 1$. Let now p and q be such that $v_p = v_q = 1$, with $p < q$. Then p is a period of Z and we can write $Z = XY$, where $|X| = p$ and Y is a prefix of Z. Since q is also a period of Z, and $q > p$, we can conclude by the transitivity of matching that $q - p$ is a period of Y. Thus Y can be written as $U \cdots UU'$, where $|U| = q - p$. Therefore all t of the form $p + i(q - p)$, $i = 0, 1, 2, \ldots$ that are in the range $[p, n)$ are also periods of Z. This proves that v satisfies the forward propagation rule.

We argue in a similar fashion for the backward propagation rule. Let p and q be such that $p < q \leqslant 2p$, and $v_p = v_q = 1$. Since p is a period we can write $Z = XWY$, where $|X| = p - (q - p)$, $|W| = q - p$, and Y is a prefix of Z. From period q we can conclude that $Y = U \cdots UU'$, $|U| = q - p$. Now the condition $v_{2p-q} = 0$ means that the periodic structure of Y either cannot be continued by $q - p$ steps to the left when Y is viewed as a suffix of Z, or it

cannot be continued by $q - p$ steps to the right when Y is viewed as a prefix of Z. The arguments for the two cases are identical, so we confine our attention to the former, which is equivalent to the condition $W \neq U$. Note that $s = \lfloor (n-p)/(q-p) \rfloor$ counts the number of U's in Y. If for some i in $(1, s]$ we have $v_{p-i(q-p)} = 1$, then we can also write $Z = TYR$, with $|T| = p - i(q-p)$. But in that case we have $TU \cdots UU'R = XWU \cdots UU'$ and one of the U's on the left-hand side must coincide exactly with the W on the right-hand side, implying $U = W$, a contradiction.

(2) $\Rightarrow$ (3): As we remarked in Lemma 3.1, if v satisfies the forward and backward propagation rules, then so does every suffix of v. To prove that v satisfies predicate Ξ we use induction on the length n of v. Consider $p = \pi(v)$. If $n/p \geqslant 2$, then we are in case (a) of the recursive predicate. From the GCD rule it follows that any t such that $v_t = 1$ must either be a multiple of p, or else satisfy $t > p(\lfloor n/p \rfloor - 1)$. It only remains to check the conditions on w (in the notation of Definition 4.1). Note that w starts with a 1 and is a suffix of v. It satisfies condition (i) because we can apply the forward propagation rule to the 1's at positions 0 and p. By Theorem 3.1 w satisfies the GCD rule, and so condition (ii) follows. Finally from our indictive hypothesis we conclude that condition (iii) is also valid. A similar analysis can be done when $n/p < 2$ and we are in case (b) of the predicate. To start off the induction, note that the null vector and the one element vector (1) both satisfy the propagation rules, as well as predicate Ξ.

(3) $\Rightarrow$ (1) (this is the hard one): We must now show that a bit vector v satisfying predicate Ξ does arise as the correlation of a binary string. Again we proceed by induction. We will in fact prove something stronger: Let v be any bit vector satisfying the recursive predicate, and let W be any binary string whose correlation is the w referred to in the predicate (W exists by induction). Then we will find a string V with suffix W and correlation v.

If we are in case (a) of the predicate, matters are simple. The string W has period p and can therefore be written as $W = PR$, where $|P| = p$, $|W| = r$. The new string V can now be obtained by just preceding W with $\lfloor n/p \rfloor - 1$ copies of the period P. It is certain that V has all the periods indicated by v, but we must check that no additional ones have been introduced. This is tantamount to showing that if t is not a multiple of p, and $t < n - r$ (notation as in Definition 4.1), then t is not a period of V. Assume the contrary, and consider the smallest such t which is a period. The GCD rule on V implies that t must properly divide p. But such a t is also a period of W, $t \leqslant p/2$, and therefore $\pi(w)$ divides t. Thus $\pi(w)$ divides p, a contradiction to condition (ii).

We next consider the difficult case (b) of the recursive defintion. Inductively, let W be a binary string with correlation w. Our task is to determine a binary string X so that $V = WXW$, and has correlation v. Note that $|W| =$

$r (r < n/2)$. Let $x = |X| = n - 2r > 0$ be the length of the sought string and $t = \pi(w)$ be the basic period of W.

We must distinguish two cases. If $x \geqslant t$ then it suffices to set $X =$ "$aa \cdots a$," where a is the complement of the leftmost bit of $\pi(w)$. A shift of V in the range $[t, r)$ will cause t consecutive characters of W to be placed under X, thus guaranteeing a mismatch, since among them there must be the leftmost character of the basic period of W. Finally a shift in the range $[1, t)$ or $[r, p)$ will also obviously mismatch. Thus the correlation of V is $100 \cdots 0w$, as desired.

Suppose now that $x < t$. Write $r = mt + y$, with $0 \leqslant y < t$. First we deal with the case $m > 1$. Then we can conclude from the GCD rule (or the recursive property) that any period q of W which is not a multiple of t must satisfy $q > (m-1)t + y$. Note first that a shift of V in the range $[r, n-r)$ cannot result in a match, no matter what X is, since it would violate the constraint that t is the basic period of w.

Let z be the smallest shift in $[1, r)$ which can be a period of $Z = WXW$. From the GCD rule applied to z and $r + x$ as periods of Z, it follows immediately that z must divide $r + x$. Thus $r + x$ is a multiple of a period z of Z, which is a fortiori a period of W. Since $r + x$ is in the range $[r, r + t)$, it must be the $(m+1)$-st multiple of the basic period t of W. Can it also be the multiple of some other period of W, say q, which is not a multiple of t? Recall that such a q satisfies $(m-1)t + y < q \leqslant mt + y = r$. If $r + x = kq$, then

$$k(m-1)t + ky < (m+1)t + y, \qquad \text{for} \quad k \geqslant 2,$$

which, since $m > 1$, can only hold if $m = 2$ and $k = 2$. Furthermore it is clear that such a q is unique. We conclude that (1) either $r + x = p$ is a multiple only of period t and certain of its multiples, or (2) $r + x$ is a multiple of t and of another independent period q, in which case we have that $m = 2$ and $r + x = 2q$.

We are now ready to determine X. In case (1) we can simply let $X =$ "$** \cdots *a$," where a is the complement of the rightmost characters of the period t of W, and $*$ is arbitrary. Then any shift u in $[1, r)$ which can match must be a multiple of t, and thus will fail to match over X. In case (2) we must have $3t = 2q$, $1 \leqslant t < q = 3t/2$, and thus $t > 1$. Now $x = t - y$, and if $y = 0$, then $x = t > 1$. Else, if $y \geqslant 1$, from the lower bound $q > t + y$, we get $3t/2 > t + y$, and thus $t > 2y$. Therefore in this case also $x = t - y > 2y - y = y \geqslant 1$. We conclude that $|X| \geqslant 2$. Now we can set the rightmost bit of X to the complement of the rightmost bit of t, and the bit next to the rightmost to the complement of the corresponding bit of q (recall $q > 1$, so the "next to rightmost" bit of period q exists). The rest of X can be filled in arbitrarily. As above, we can easily check that this X will cause any candidate shifts in $[1, r)$ to mismatch.

Finally we must return to the case $m = 1$. Then we have $r = t + y$ and any other period q of W must satisfy $q > t$. A multiple $ku = r + x$ of some period u of W can be in the interval $[t + y, 2t + y)$ only if

$$kt < 2t + y,$$

and so we must have $k = 2$. It follows that at most one such period u exists. If such a u exists then, as above, we set the last character of X to the complement of the rightmost bit of u (the other characters can be arbitrary). If no u exists, the contents of X are immeterial.

This completes the proof. ∎

COROLLARY 5.1 (alphabet size independence). *Any alphabet of size at least two will give rise to the same set of correlations.*

It would be an instructive exercise for the reader at this point to modify the procedure *Ksi* given in the previous section so that when vector v satisfies predicate Ξ, this procedure will return a string with v as correlation. The modified procedure can also be made to run in time linear in n, the length of v.

6. COUNTING THE CORRELATIONS

In this section we use the recursive predicate Ξ to obtain bounds on the number of distinct correlations of length n.

THEOREM 6.1 (the number of correlations). *The number $\kappa(n)$ of distinct correlations of length n satisfies*

$$\left(\frac{1}{2 \ln 2} + o(1)\right) \ln^2 n \leqslant \ln \kappa(n) \leqslant \left(\frac{1}{2 \ln(3/2)} + o(1)\right) \ln^2 n,$$

as $n \to \infty$.

Proof. If we just consider the correlations given to us by case (b) of the recursive predicate Ξ, then we get

$$\kappa(n) \geqslant \sum_{0 \leqslant r < n/2} \kappa(r), \qquad \kappa(0) = 1.$$

If we let $g(n)$ be defined by the recurrence

$$g(n) = \sum_{0 \leqslant r < n/2} g(r), \qquad g(0) = 1,$$

then we obviously have

$$\kappa(n) \geqslant g(n), \qquad n = 0, 1, 2, \ldots.$$

If we now consider $\tilde{g}(x)$, the continuous analog of g, defined by

$$\tilde{g}(x) = \int_0^{(x/2)-2} \tilde{g}(r)\, dr, \qquad \tilde{g}(t) = 1, \qquad \text{for} \quad t \in [0, 6), \tag{1}$$

then from the monotonicity of $\tilde{g}$ it follows that

$$g(n) \geqslant \tilde{g}(n).$$

It is easy to check that asymptotically the solution to the integral equation (1) has the form

$$g(x) = \exp\left(\frac{1}{2 \ln 2}(1 + o(1)) \log^2 x\right),$$

from which the lower bound asserted by the theorem follows. For the upper bound we proceed in an analogous fashion. We claim that predicate Ξ implies that

$$\kappa(n) \leqslant 2 \sum_{0 \leqslant r < 2n/3} \kappa(r).$$

Certainly $\sum_{0 \leqslant r < 2n/3} \kappa(r)$ is an upper bound on the number of correlations of length n having a basic period $p > n/3$. But if $p \leqslant n/3$ then we are in case (a) of the predicate Ξ. The suffix w of length $n - p(\lfloor n/p \rfloor - 1) \leqslant 2n/3$ completely defines the correlations, and since it also has to be a correlation, the above inequality follows. The rest of the argument for the upper bound is exactly analogous to that for the lower bound, and is therefore omitted. ∎

We conjecture that $\ln \kappa(n)$ is in fact asymptotic to a constant times $\ln^2 n$, but we have been unable to prove this. A table of some value of $\kappa(n)$ follows.

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$\kappa(n)$	1	2	3	4	6	8	10	13	17	21	27	30	37	47	57	62	75	87	102	116

n	25	30	35	40	45	50	55
$\kappa(n)$	220	392	664	1005	1552	2240	3226

Remark. In a variant of the Knuth–Morris–Pratt string searching algorithm [8], an array is constructed which records the basic period of all prefixes of the pattern string. If that string has length m, Galil and Seiferas [6] have shown how (by omitting information about "long" periods) to compress this array of nominal size equal to the pattern length, so size $O(\ln m)$ locations, each location holding $O(\ln m)$ bits, or a total of $O(\ln^2 m)$ bits. We claim that *if two patterns have the same basic period table, then they have the same correlation.* We prove this inductively. Suppose we know the assertion to be true for patterns of length up to m. Let us be given the basic period table for a pattern of length $m+1$, and let p be the period of the pattern. We only have to provide a unique answer to the question: Is q, $q > p$, a period of the pattern? Let kp be the largest multiple of p less than or equal to q. Then note that q is a period of the pattern if and only if $q - kp$ is a period of the prefix of the pattern of length $m + 1 - kp$. But this is a smaller problem which we can inductively solve.

It follows from the above that there are at least $\kappa(m)$ basic period tables of length m. Thus to encode one of them we need at least $\Omega(\ln^2 m)$ bits (by Theorem 6.1). This establishes that the Galil–Seiferas compression cannot be uniformly exceeded by more than a constant factor.

7. Populations

In this section we count the number of strings of length n over an alphabet of size q which have a given (auto)-correlation. We will do this by obtaining a recurrence of n for $L_n(C)$, the population of strings with correlation $K = 100 \cdots 00C$, where K is of length n and consists of a 1 followed by all 0's until the final suffix C, which is itself a correlation and is assumed fixed. As usual, c will denote the length of C. For simplicity of notation we will often write L_n instead of $L_n(C)$, C being implicitly understood. Before we can state our result we need one additional definition: Let ψ denote a sequence (depending on C) defined for all integers by

$$\begin{aligned} \psi_k &= 0, && \text{for } k > c; \\ &= C[c-k], && \text{for } 1 \leqslant k \leqslant c; \\ &= q^{-k}, && \text{for } k \leqslant 0. \end{aligned}$$

Thus in the range $1 \leqslant k \leqslant c$, we see that ψ_k is equal to 1 or 0 depending on whether $c - k$ is a period in C, or not.

Our first theorem states the recurrence on L_n.

THEOREM 7.1 (basic population recurrence). *The number L_n of q-ary strings of length n which have correlation $10\cdots 0C$ satisfies the recurrence*

$$L_n + \sum_{\substack{k,l \\ n+l=2k}} L_k \psi_l = 2\psi_{2c-n} L_c, \tag{1}$$

where we set $L_n = 0$ for $n < c$.

Proof. We distinguish two cases.

Case 1. $n > 2c$. The strings of correlation $10\cdots 0C$ are a subset of the set $\mathscr{S}$ of strings which have length n and a correlation with C as a suffix. We claim that both sides of (1) equal $2|\mathscr{S}|$.

Any string in $\mathscr{S}$ can be obtained as XYX, where X is any of the L_c strings of correlation C, and Y is arbitrary. Note that since $n > 2c$ we have $\psi_{2c-n} = q^{n-2c}$ = number of possible choices of Y. Thus the RHS of (1) equals $2|\mathscr{S}|$.

We can now view the LHS as classifying the strings in $\mathscr{S}$ according to their *longest period* which is less than or equal to $n-c$. Of course L_n counts the strings in $\mathscr{S}$ which have no period shorter than $n-c$. Now consider the summation. The term $L_k\psi_l$ will count those strings in $\mathscr{S}$ whose longest period less than $n-c$ is $n-k$, for $k > c$. For $k = c$ we just get $L_c\psi_{2c-n}$, which is equal to $|\mathscr{S}|$. Note also that $k \leqslant \lfloor (n+c)/2 \rfloor$, as $\psi_l = 0$ for $l > c$. This corresponds to the fact that the longest period p, $p < n-c$, of a string Z in $\mathscr{S}$ cannot be less than $n - \lfloor (n+c)/2 \rfloor$. This is so, because $2p$ would then also be a period of Z, $2p > p$, and $2p \leqslant 2n - 2\lfloor (n+c)/2 \rfloor - 2 < n-c$, a contradiction.

Consider now a string Z in $\mathscr{S}$ whose longest period less than $n-c$ is equal to $n-k$, where $c < k < \lfloor (n+c)/2 \rfloor$. Write Z as $Z = XY$, where $|Y| = k$. Note first that Y has correlation of the form $10\cdots 0C$, since any period of Y shorter than $k-c$ would imply a period of Z longer than $n-k$ but less than $n-c$. Thus the number of possible Y's is L_k. Given Y, we can obtain Z as follows. If $k \leqslant \lfloor n/2 \rfloor$, then $Z = YTY$, where T is of length $n-2k$ and arbitrary. But if $k > \lfloor n/2 \rfloor$ then at most one Z can possibly exist. Such a Z will exist exactly if $n-k$ is a period of Y. Note that $2k - n < c$, and so this is equivalent to asserting that $2k-n$ is in C, or that $\psi_{2k-n} = 1$. Thus in both cases ψ_{2k-n} counts the number of ways to obtain Z given Y. This completes the argument.

Case 2. $n \leqslant 2c$. Note first that if $n < c$, then both sides of (1) are 0. The RHS is 0 because $2c - n > c$ and so $\psi_{2c-n} = 0$. The LHS is also zero, as $L_n = 0$, and for each term $L_k\psi_l$ og the sum we must have $k < c$ or $l > c$, so $L_k = 0$ or $\psi_l = 0$. We similarly easily check that (1) holds when $n = c$, so from now on we assume that $c < n \leqslant 2c$ (we regard c itself as a period). We claim that for n in $(c, 2c]$, $L_n = L_c$ or 0 according as to whether $x = n - c$ is

a primitive period in C or not. (Recall that a period is primitive if it is not the multiple of a smaller period.) A binary vector of the form $10 \cdots 0C$ and length $n = c + x \leqslant 2c$ is not a correlation unless x is a primitive period in C. The forward propagation rule implies that x must be a period in C. However, x must also be be primitive, as otherwise $10 \cdots 0C$ would violate the backward propagation rule. Thus if x is not a primitive period in C, then $L_n = L_{c+x} = 0$. If x is a primitive period, then $10 \cdots 0C$ is a valid correlation and $L_n = L_{c+x} = L_c$, as the strings of correlation $10 \cdots 0C$ can be uniquely obtained from their last c characters. Thus the above assertion is proved.

Now back to proving (1) for $n = c + x$, $x \in (0, c]$. When do we have non-zero terms in the sum? We must have $k = c + y$, $y \geqslant 0$, y a primitive period in C (we allow 0 as a period). Furthermore, $l = c - t$, $0 \leqslant t \leqslant c$, t a period in C, and $n + l = 2k$, or $c + x + c - t = 2(c + y)$, or $x - t = 2y$ as well. Now if x is not a period in C, then $\psi_{2c-n} = \psi_{c-x} = 0$ and the RHS of (1) is 0. But so is the LHS, since $L_{c+x} = 0$ as we saw above, and $x = t + 2y$ is impossible, as a sum of periods is also a period. (This follows from the transitivity of matching.) Next, if x is period in C, but not a primitive one, then the RHS of (1) equals $2L_c$. On the left-hand side we have $L_n = L_{c+x} = 0$. However, the sum contains exactly two non-zero terms. One is obtained by taking $y = 0$, $x = t$, giving the term $L_c \psi_{c-x} = L_c$. Any other term must have $y > 0$, but since $2y = x - t \leqslant c$, it follows that y must be the basic period in C. By the GCD Rule we must then have $t = ry$ for some integer $r > 1$, and $x = (r + 2)y$. Furthermore, once x is given, r, and therefore t, are completely determined. So there is exactly one other non-zero term, namely, $L_{c+y}\psi_{c-ry} = L_c$ since y is a primitive period. Finally we consider the situation when x is a primitive period in C. Again the RHS of (1) equals $2L_c$, but now $L_n = L_{c+x}$ on the LHS is non-zero and equal to L_c. Thus it remains to show that the sum contains exactly one non-zero term, which must equal L_c. As we saw above, if $y > 0$ in $x - t = 2y$, then x is not primitive. Thus we must take $y = 0$, $x = t$ and this gives the unique non-zero term $L_c \psi_{c-x} = L_c$. The argument is complete. ∎

Remark. The above proof essentially contains the argument of Theorem 5.1.

To continue with our analysis, we will need to introduce the generating functions of L_n and ψ_k. Let

$$L(z) = \sum_{n=0}^{\infty} L_n z^{-n},$$

and analogously

$$\Psi(z) = \sum_{n=0}^{\infty} \psi_n z^{-n}.$$

In this notation the result of Theorem 7.1 can simply be stated as the functional equation

$$L(z) + \Psi(z)\, L(z^2) = 2L_c\, \Psi(z)\, z^{-2c}. \tag{2}$$

As is already clear from Theorem 7.1, L_c divides L_n for all n, and so it will also be convenient to introduce the normalized generating function

$$\tilde{L}(z) = \frac{L(z)}{L_c}.$$

Thus we can rewrite (2) as

$$\tilde{L}(z) + \Psi(z)\, \tilde{L}(z^2) = 2\Psi(z)\, z^{-2c}.$$

We are now ready to discuss the asymptotics of L_n as $n \to \infty$ with C fixed.

THEOREM 7.2 (asymptotics on the populations). *The number L_n of q-ary strings of length n and correlation $10 \cdots 0C$ has the asymptotic value*

$$\frac{L_n}{L_c} = \left(\frac{2}{q^{2c}} - \tilde{L}(q^2)\right) q^n + O(q+\varepsilon)^{n/2},$$

where $\tilde{L}(z)$ satisfies the functional equation

$$\tilde{L}(z) + \Psi(z)\, \tilde{L}(z^2) = 2\Psi(z)\, z^{-2c}.$$

The above expansion is valid as $n \to \infty$, for any positive ε, and the implied O constant is independent of the underlying correlation C (but depends on ε).

Proof. We claim that we can write $\tilde{L}(z)$ as

$$\tilde{L}(z) = \frac{q\beta}{z-q} + H(z),$$

where $\beta = 2q^{-2c} - \tilde{L}(q^2)$ is constant, and $H(z)$ is analytic on the disk $|z| > q^{1/2}$. To see this, set momentarily

$$M(z) = \tilde{L}(z) - z^{-c},$$

and note that $M(z)$ satisfies the equation

$$M(z) = z^{-2c}\Psi(z) - z^{-c} - \Psi(z)\, M(z^2). \tag{3}$$

If we set $\Theta(z) = z^{-2c}\Psi(z) - z^{-c}$, then it follows from iterating equation (3) that

$$M(z) = \sum_{k=0}^{m} (-1)^k \Theta(z^{2^k}) \prod_{r=0}^{k-1} \Psi(z^{2^r}) + (-1)^{m+1} M(z^{2^{m+1}}) \prod_{r=0}^{m} \Psi(z^{2^r}). \quad (4)$$

(As usual, we take an empty product to equal 1.) To eliminate the last term on the right-hand side above we will let $m \to \infty$. First of all, note that $\Psi(z)$ and $\Theta(z)$ are both rational functions, analytic except possibly at $z = 0$ and $z = q$, and that for $|z| > 2q$, say,

$$\Psi(z) = z^c(1 + O(|z|^{-1}))$$

and

$$\Psi(z) = z^{s-2c}(1 + O(|z|^{-1})),$$

where $s < c$ is the largest power of z in the polynomial part of $z^{2c}\Theta(z) = \Psi(z) - z^c$. On the other hand, if $|z| > 2q$, then

$$M(z) = O(|z|^{-c-1}).$$

Therefore for any complex z such that $|z| > 1$ and $z^{2^k} \neq q$ for any k we have

$$M(z^{2^{m+1}}) \prod_{r=0}^{m} \Psi(z^{2^r}) = O\left(\frac{(|z|^{1+2+4+\cdots+2^m})^c}{|z|^{(c+1)2^{m+1}+1}}\right) = O(|z|^{-2^{m+1}}),$$

as $m \to \infty$. Hence if we let $m \to \infty$ in (4), we obtain

$$M(z) = \sum_{k=0}^{\infty} (-1)^k \Theta(z^{2^k}) \prod_{r=0}^{k-1} \Psi(z^{2^r}).$$

We claim that this infinite summation defines an analytic function in the disk $|z| > q$. Note that the Θ or Ψ terms only have singularities at points of the form $z = q^{1/2^k}$, or $z = 0$, and so it suffices to show that the series converges absolutely. The ratio of the $(k+1)$st to the kth term of the series is

$$\rho_{k+1} = -\frac{\Theta(z^{2^{k+1}})\,\Psi(z^{2^k})}{\Theta(z^{2^k})}.$$

If we set $x = |z|^{2^k}$ and take k sufficiently large, then we see that the above ratio is asymptotic in absolute value to x^{-c+s}, where s is as defined earlier. Since $s < c$, it follows that eventually

$$|\rho_{k+1}| < (1 + \delta)|z|^{-2^k},$$

for some positive δ, which proves the convergence of our summation. Note also that

$$|\Theta(z)| \leqslant \left(\frac{|z|^c - 1}{|z| - 1} + \frac{|z|}{|z - q|} \right) |z|^{-2c},$$

which, for fixed z, $|z| > q$, is bounded above by an absolute constant independent of C (or c). We show next that $M(z)$ has a simple pole at $z = q$. We have

$$\lim_{z \to q+} \frac{(z - q) M(z)}{q} = q^{-2c} - \Theta(q^2) + \Psi(q^2)\,\Theta(q^4) - \cdots$$

$$= 2q^{-2c} - \tilde{L}(q^2),$$

which, as we have already shown, exists. Let us set

$$\beta = 2q^{-2c} - \tilde{L}(q^2).$$

The above expansion provides us with a very efficient way of numerically computing β, given the correlation C. In fact the argument regarding ρ_{k+1} can be strengthened to show that the terms of the alternating series ($-$ alternating since $\Theta(x)$ and $\Psi(x)$ are positive real for positive real x)

$$\sum_{k=1}^{\infty} (-1)^k\, \Theta(q^{2^k}) \prod_{r=1}^{k-1} \Psi(q^{2^r})$$

always decrease in absolute value, and thus the partial sum of this series can be used to bound β from above and below. This also applies to the initial term, so in particular $\beta > 0$. To see this it suffices to show that

$$\Psi(x) \leqslant \frac{\Theta(x)}{\Theta(x^2)}, \qquad \text{for} \quad x \geqslant q^2 \geqslant 4.$$

Note that $\Psi(x) \leqslant x^{c+1}$, so it is enough to show

$$(x^{-4c}\,\Psi(x^2) - x^{-2c})\, x^{c+1} \leqslant x^{-2c}\Psi(x) - x^{-c}, \qquad \text{or}$$

$$x\Psi(x^2) - x^{2c+1} \leqslant x^c \Psi(x) - x^{2c}. \tag{5}$$

Now if $c = 0$, then this inequality becomes

$$\frac{x^3}{x^2 - q} - x \leqslant \frac{x}{x - q} - 1, \qquad \text{or}$$

$$\frac{xq}{x^2 - q} \leqslant \frac{q}{x - q},$$

which follows, since $x > 1$.

For $c > 0$, we have

$$\frac{x^3}{x^2-q} \leqslant \frac{x^2}{x-q} \leqslant \frac{x^{c+1}}{x-q}. \tag{6}$$

Also

$$\begin{gathered} x \sum_{\substack{k \in C \\ k \neq c}} x^{2k} \leqslant \sum_{\substack{k \in C \\ k \neq c}} x^{k+c}, \qquad \text{or} \\ x(x^2 C_{c^2}) - x^{2c+1} \leqslant x^c C_x - x^{2c}. \end{gathered} \tag{7}$$

Here $k \in C$ means that k is the position of a 1 in C, counting from the right, with the rightmost bit being in position 1. Furthermore, C_z indicates the correlation C viewed as a polynomial in z. (That is, $zC_z = \sum_{k \in C} z^k$.) Inequality (5) now follows by adding (6) and (7) and this completes the proof of the above remark.

We can therefore write

$$\tilde{L}(z) = \frac{q\beta}{z-q} + H(z), \tag{8}$$

where $H(z)$ is just $M(z)$ with the simple pole removed and is easily seen to be analytic for $|z| > q^{1/2}$, by arguments exactly analogous to those used for $M(z)$. In fact, as for $M(z)$, the function $H(z)$ is bounded on the circle $|z| = (q+\varepsilon)^{1/2}$ by a constant which is independent of C. If we multiply (8) by z^{n+1} and integrate around the circle $|z| = (q+\varepsilon)^{1/2}$, we obtain by Cauchy's theorem

$$\tilde{L}_n = \beta q^n + O((q+\varepsilon)^{n/2}), \qquad \text{as} \quad n \to \infty.$$

So we have shown

$$\frac{L_n}{L_c} = \left(\frac{2}{q^{2c}} - \tilde{L}(q^2)\right) q^n + O((q+\varepsilon)^{n/2}),$$

as $n \to \infty$, with the implied O-constant independent of c. ∎

Remark. It should be clear that we can continue the above asymptotic expansion as far as we please. The next order expansion would be

$$\frac{L_n}{L_c} = \beta q^n + k_1 q^{n/2} + k_2(-1)^n q^{n/2} + O((q+\varepsilon)^{n/4}), \qquad \text{etc.}$$

In fact, the same technique can be used recursively to compute the asymptotics of L_c in terms of a still smaller correlation C' ($C = 10 \cdots 0C'$),

and so, given a symbolic manipulation system, we could carry out this further to obtain good estimates for L_n itself, and/or a closed form formula that allows explicit numerical evaluation.

We repeat here two definitions introduced in the above proof which will be useful to us later. These are

$$\Theta(z) = z^{-2c}\Psi(z) - z^{-c},$$

and

$$\beta = 2q^{-2c} - \tilde{L}(q^2) = q^{-2c} - \sum_{k=1}^{\infty} (-1)^{k-1}\,\Theta(q^{2^k}) \prod_{r=1}^{k-1} \Psi(q^{2^r}).$$

As already remarked, the above formula gives us a very efficient way of computing β. We list some interesting values below:

	C	β	$L_c\beta$
$q=2$	Λ	0.26771654	0.26771654
	1	0.150203882	0.300407764
	10	0.055000309	0.110000618
	11	0.04445766	0.08891532
$q=3$	Λ	0.55697974	0.55697974
	1	0.094234491	0.282703474
	10	0.0121190452	0.072714272
	11	0.0109175415	0.0327526242
$q=24$	Λ	0.95659723	0.95659723
	1	0.001732971	0.041591309

Here Λ denotes the empty string. Thus approximately 27% of all binary strings have the trivial correlation [10]. Slightly more, 30% have the correlation $10\cdots01$, which can be shown to be the most popular correlation in base 2. Our results extend the bounds derived by Harborth in [7]. Note that for $q=3$ or larger, the trivial correlation is the most popular. For large q, the above results show that this fraction is essentially $(q-2)/(q-1)$. For $q=24$ more than 95% ($22/23 = 0.956...$) of all strings have that correlation!

We now prove a result that allows us to compare β's corresponding to C's of the same length.

THEOREM 7.3 (asymptotic comparison). *Let A, B denote correlations of length c. If $A_q \geqslant B_q$, then $\beta_A \leqslant \beta_B$. Here T_q denotes the correlation T, viewed as a number to the base q.*

Proof. We will in fact prove the above assertion for all binary strings of length c beginning with a "1." By transitivity over all binary strings of length c, it suffices to prove the assertion for two strings A and B that "are different by 1," i.e., of the form

$$A = 1xxxx10\cdots 0,$$

$$B = 1xxxx01\cdots 1,$$

where the "$xxxx$" part of A and B is the same. Let l be the position of the rightmost 1 in A (counting from the right). We have, using an obvious notation,

$$\beta_A = q^{-2c} - \sum_{k=1}^{\infty} (-1)^{k-1}\Theta_A(q^{2^k}) \prod_{r=1}^{k-1} \Psi_A(q^{2^r})$$

$$= q^{-2c} - \sum_{k=1}^{\infty} (-1)^{k-1} a_k,$$

and similarly

$$\beta_B = q^{-2c} - \sum_{k=1}^{\infty} (-1)^{k-1} b_k.$$

To prove our result it suffices therefore to show that

$$\sum_{k=1}^{\infty} (-1)^{k-1} a_k \geqslant \sum_{k=1}^{\infty} (-1)^{k-1} b_k,$$

which will follow if we can show that

$$a_{2k-1} - a_{2k} \geqslant b_{2k-1} - b_{2k}.$$

(From the proof of the previous theorem we know that both these differences are non-negative.) Let t represent either the a or the b sequence. The difference $t_{2k-1} - t_{2k}$ is equal to

$$\prod_{r=1}^{2k-2} \Psi(q^{2^r})(\Theta_T(x) - \Theta_T(x^2)\,\Psi_T(x)), \tag{9}$$

where $x = q^{2^{2k-1}} \geqslant q^2$. We must show that the above expression for $T = A$ is greater than or equal to the same expression with $T = B$.

Since $A_q \geqslant B_q$ implies $A_r \geqslant B_r$ for any r, $r \geqslant q$ (why?), and $\Psi_T(z) = zT_z + z/(z-q)$ is monotonic in T, it follows that the product terms of (9)

compare the right way. For the remaining term we find, using the definition $\Theta_T(z) = z^{-2c}\Psi(z)_T - z^{-c}$, that

$$\Omega_T(x) = \Theta_T(x) - \Theta_T(x^2)\ \Psi_T(x) = 2x^{-2c}\Psi_T(x) - x^{-4c}\Psi_T(^2) - x^{-c},$$

and thus, if we can show that

$$\Omega_A(x) - \Omega_B(x) = 2x^{-2c}(\Psi_A(x) - \Psi_B(x)) - x^{-4c}(\Psi_A(x^2) - \Psi_B(x^2)) \quad (10)$$

is non-negative, we are done. But for our A and B we have

$$\Psi_A(z) - \Psi_B(z) = z(z^{l-1} - z^{l-2} - z^{l-3} - \cdots - 1)$$
$$= z\,\frac{z^l - 2z^{l-1} + 1}{z - 1},$$

where l is as defined above and $l < c$. The argument can now be completed by substituting this into (10) and using elementary algebra. ∎

Remarks. 1. The number of strings of correlation C always divides the number of strings of correlation $10\cdots 0C$.

2. The result of Theorem 7.3 is only true asymptotically. For example, if $q = 2$, then

$$L(10^{11}100011)/L(100011) = 62,$$

but

$$L(10^{11}100100)/L(100100) = 63.$$

(By 0^{11} we mean the string of 11 zeros).

3. Consider the correlation $10\cdots 0C$ of length n, with $n > 2c$. If $n < 3c$, and $3c/2 - n/2 > c - \pi(C)$, then

$$L(C) = L_c q^{n-2c}.$$

Proof. The remark follows from the observation that no other correlations with suffix C of length n exist. To see this, let b be the basic period of a different correlation, and write $kb < (n-c) \leqslant (k+1)b$. By forward propagation, $(n-c) - kb \geqslant \pi(C) > (n-c)/2$, and so $2kb < n - c \leqslant (k+1)b$, i.e., k must be zero. ∎

4. A "good guess" on the outcome of the comparison $L(C_1) \geqslant L(C_2)$ for correlations C_1, C_2 of length n is the outcome of the comparison

$$\sum_{r \in C_1} r \leqslant \sum_{r \in C_2} r.$$

TABLE 1

Correlations and their Population over a Binary Alphabet

Correlation	Population	Correlation	Population
1000000000000000000	.281076	1000000000100001111	8
1000000000000000001	.315322	1000000000100010001	.40
1000000000000000010	.115226	1000000000100010011	8
1000000000000000011	93146	1000000000100100100	.16
1000000000000000100	63568	1000000000100100101	8
1000000000000000101	29874	1000000000101010101	8
1000000000000000111	24234	1000000000111111111	6
1000000000000001000	24318	1000000001000000000	284
1000000000000001001	23940	1000000001000000001	318
1000000000000001010	.7612	1000000001000000010	110
1000000000000001111	.6094	1000000001000000011	.88
1000000000000010000	12276	1000000001000000100	.64
1000000000000010001	10190	1000000001000000101	.30
1000000000000010010	.4024	1000000001000000111	.24
1000000000000010011	.2004	1000000001000001000	.24
1000000000000010101	.1918	1000000001000001001	.18
1000000000000011111	.1528	1000000001000001010	6
1000000000000100000	.5080	1000000001000001111	6
1000000000000100001	.5588	1000000001000100010	6
1000000000000100010	.1518	1000000001000100011	6
1000000000000100011	.1512	1000000001001001001	6
1000000000000100100	.1000	1000000010000000010	232
1000000000000100101	500	1000000010000000011	182
1000000000000101010	476	1000000010000000111	.46
1000000000000111111	382	1000000010000001010	.16
1000000000000001000000	.2560	1000000010000001111	.12
1000000000000001000001	.2432	1000000010000010010	4
1000000000000001000010	.1024	1000000010000010011	2
1000000000000001000011	768	1000000010000011111	2
1000000000000001000100	512	1000000010000100011	6
1000000000000001000101	128	1000000010001000111	2
1000000000000001000111	126	1000000100000001000	.90
1000000000000001001001	378	1000000100000001001	.90
1000000000000001010101	120	1000000100000001010	.30
1000000000000001111111	.96	1000000100000001111	.24
1000000000010000000	.1184	1000000100000011111	6
1000000000010000001	.1312	1000000100000100000	.40
1000000000010000010	390	1000000100000100001	.44
1000000000010000011	330	1000000100000100010	.12
1000000000010000100	256	1000000100000100011	.12
1000000000010000101	128	1000000100000100100	8
1000000000010000111	.90	1000000100000100101	4
1000000000010001000	.90	1000000100000101010	4

Table continued

TABLE 1 (*continued*)

Correlation	Population	Correlation	Population
10000000000010001001	90	10000001000000111111	 2
10000000000010010010	60	10000010000010000010	26
10000000000010010011	30	10000010000010000011	22
10000000000010101010	30	10000010000010000111	 6
10000000000011111111	24	10000100001000010000	12
10000000000100000000	 592	10000100001000010001	10
10000000000100000001	 616	10000100001000010010	 4
10000000000100000010	 240	10000100001000010011	 2
10000000000100000011	 184	10000100001000010101	 2
10000000000100000100	 112	10001000100010001000	 6
10000000000100000101	48	10001000100010001001	 6
10000000000100000111	48	10010010010010010010	 4
10000000000100001000	48	10010010010010010011	 2
10000000000100001001	48	10101010101010101010	 2
10000000000100001010	16	11111111111111111111	 2

5. Table 1 gives the populations, over a binary alphabet, of all legal correlations of length 20. There are 116 different legal correlations, each followed by the number of *binary strings* giving rise to it. The reader should be able to check many of our results using this table.

ACKNOWLEDGMENTS

The authors wish to thank Lyle Ramshaw for many valuable comments on the manuscript. Some of the computations in Section 7 were checked using the MACSYMA system at MIT.

REFERENCES

1. S. B. BOYER AND J. S. MOORE, A fast string searching algorithm, *Comm. Assoc. Comput. Mach.* **20** (1977), 762–771.
2. N. J. FINE AND H. S. WILF, Uniqueness theorems for periodic functions, *Proc. Amer. Math. Soc.* **16** (1965), 109–114.
3. M. GARDNER, On the paradoxical situations that arise from nontransitive relations, *Sci. Amer.* (October 1974), 120–125.
4. L. J. GUIBAS AND A. M. ODLYZKO, A new proof of the linearity of the Boyer–Moore string searching algorithm, *in* "Proceedings, 18th Annual Symposium on Foundations of Computer Science Proceedings," IEEE Computer Society, New York, 1977. Revised version *SIAM J. Comput.*, in press.
5. L. J. GUIBAS AND A. M. ODLYZKO, String overlaps, pattern matching, and nontransitive games, *J. Combinatorial Theory Ser. A*, in press.
6. Z. GALIL AND J. SEIFERAS, Saving space in fast string matching, *in* "Proceedings, 18th Annual Symposium on Foundations of Computer Science Proceedings," IEEE Computer Society, New York, 1977.

7. H. HARBORTH, Endliche 0-1-Folgen mit gleichen Teilblöcken, *Crelle's J.* **271** (1974), 139–154.
8. D. E. KNUTH, J. H. MORRIS, AND V. R. PRATT, Fast pattern matching in strings, *SIAM J. Comput.* **6** (1977), 323–350.
9. R. C. LYNDON AND M. P. SCHÜTZENBERGER, The equation $a^M = b^N c^P$ in a free group, *Michigan Math. J.* **9** (1962), 289-298.
10. T. P. NIELSEN, On the expected duration of a search for a fixed pattern in random data, *IEEE Trans. Inform Theory* (1973), 702–704.
11. M. P. Schützenberger, A property of finitely generated submonoids of free monoids, "Actes de Colloque sur le semigroups à Szeged, 1976," North-Holland, Amsterdam, in press.