Strings with maximally many distinct subsequences and substrings

Abraham Flaxman
Department of Mathematical Sciences
Carnegie Mellon University
Pittsburgh PA, USA
abie@cmu.edu

Aram W. Harrow
Department of Physics
Massachusetts Institute of Technology
Cambridge MA, USA
aram@mit.edu

Gregory B. Sorkin
Department of Mathematical Sciences
IBM Research
Yorktown Heights NY, USA
sorkin@watson.ibm.com

Submitted: Nov 18, 2003; Accepted: Dec 9, 2003; Published: Jan 5, 2004
MR Subject Classifications: 68R15, 05D40, 05A15, 05A16

Abstract

A natural problem in extremal combinatorics is to maximize the number of distinct subsequences for any length-n string over a finite alphabet Σ; this value grows exponentially, but slower than 2^n. We use the probabilistic method to determine the maximizing string, which is a cyclically repeating string. The number of distinct subsequences is exactly enumerated by a generating function, from which we also derive asymptotic estimates. For the alphabet $\Sigma = \{1, 2\}$, $(1, 2, 1, 2, \dots)$ has the maximum number of distinct subsequences, namely $\mathrm{Fib}(n+3) - 1 \sim \left((1+\sqrt{5})/2\right)^{n+3}/\sqrt{5}$.

We also consider the same problem with sub*strings* in lieu of sub*sequences*. Here, we show that an appropriately truncated de Bruijn word attains the maximum. For both problems, we compare the performance of random strings with that of the optimal ones.

1 Introduction

In this article we consider a natural problem in the extremal combinatorics of strings, namely to find a string whose number of subsequences is as large as possible, and to determine the number. Strings and texts are themselves one of the basic combinatorial structures, and the sorting, searching, and compression of strings is even more important with strings comprising one of the most important facets of the World-Wide Web (and the only facet currently indexable). We would thus have expected such an elementary question already to have been considered, but we have been unable to find the problem or its solution in print.

While the problem is not especially difficult, its solution is quite pretty. The string maximizing the number of distinct subsequences is utterly regular (and unique, up to the trivial symmetry among the characters of the language), yet the probabilistic method provides an elegant way of establishing this fact, while giving no information about the number itself. Once the maximizing string is known, however, the number of subsequences is described by a simple recursion relation; for binary strings, this is essentially the Fibonacci recursion $\mathrm{Fib}(n) = \mathrm{Fib}(n-1) + \mathrm{Fib}(n-2)$ [FoP02], and the number of distinct subsequences is $\mathrm{Fib}(n+3) - 1$, which is asymptotically equal to $\phi^{n+3}/\sqrt{5}$ where $\phi = (1+\sqrt{5})/2$ is the so-called golden ratio (attributed by [Hor61] to Daniel Bernoulli, 1732, or by [Mil60], via [Ait27], to Bernoulli, by 1728). For strings over larger alphabets, the recursion is analogous to the tribonacci numbers, tetranacci numbers, and similar generalizations of the Fibonacci numbers; again the growth is asymptotically exponential; and we give tight bounds on the base, which is the largest root of an explicit polynomial.

The probabilistic argument also shows that, for any alphabet size, "everything can be maximized at once": there is a single (and essentially unique) infinite string whose n-long prefixes are the maximizing strings, and each n-prefix not only maximizes the number of subsequence, but simultaneously maximizes the number of m-long subsequences for every $m \leq n$.

We also consider producing a string maximizing the number of distinct sub*strings*, or the number of distinct m-long substrings. Here we exhibit such a string for each n using a modified de Bruijn word [dB46]. For $d \geq 3$ there is an infinite string where each n-long prefix is a substring-maximizing string, but for $d = 2$ no such infinite string exists.

2 Strings with maximally many distinct subsequences

Let Σ be a finite alphabet of size d; without loss of generality we take $\Sigma = [d]$. Let $A = (a_1, a_2, \ldots, a_n) \in \Sigma^n$ be an n-long string over Σ. A string B is a *subsequence* of A, $B \preccurlyeq A$, if there is a set of indices $i_1 < i_2 < \cdots < i_m$ such that

$$B = (a_{i_1}, a_{i_2}, \ldots, a_{i_m}).$$

The empty string B, with $|B| = 0$, is a subsequence of any string. We define the set of all subsequences of A as $\mathrm{subseq}(A) = \{B : B \preccurlyeq A\}$.

Aho [Aho03] poses the natural question, "What string A of length n has a largest set of distinct subsequences?" We will generalize this slightly and also ask for an n-long string having the maximum number of m-long subsequences, for any $m \leq n$. Accordingly, with $\Sigma = [d]$, we define the maximum number of distinct subsequences any length-n string may have by

$$f_d(n) := \max_{A \in \Sigma^n} |\mathrm{subseq}(A)|,$$

and the maximum number of distinct *m-long* subsequences any length-n string may have by

$$f_d(m,n) := \max_{A \in \Sigma^n} |\text{subseq}(A) \cap \Sigma^m|.$$

Note that $f_d(m,n) \le f_d(n) \le 2^n$, since the multiset of all subsequences (not necessarily distinct) is of size 2^n.

We first dispense with a triviality: the minimization rather than maximization of the number of distinct subsequences of fixed or arbitrary length.

Remark 1 *Let $\Sigma = [d]$, and let $A \in \Sigma^n$. Then for any $0 \le m \le n$,*

- *the number of distinct m-long subsequences of A satisfies $|\text{subseq}(A) \cap \Sigma^m| \ge 1$;*
- *for any m with $0 < m < n$, the lower bound is achieved uniquely (up to symmetry over the alphabet) by the string $A = (1, 1, \ldots, 1)$;*
- *this string (uniquely) minimizes the number of distinct subsequences, giving*

  $$|\text{subseq}(A)| = n + 1;$$

- *and thus (uniquely up to symmetry) the single infinite string $(1, 1, \ldots)$, truncated to length n, simultaneously minimizes all the quantities considered.*

All the statements in the above Remark are self-evident; what is surprising is that they are largely paralleled for maximization, as per the following theorem.

Theorem 2 *Let $\Sigma = [d]$, and let $A \in \Sigma^n$. Then for any $0 \le m \le n$,*

- *the maximum number of distinct m-long subsequences $|\text{subseq}(A) \cap \Sigma^m|$ is achieved (and for $m \ge 2$ achieved uniquely, up to symmetry over the alphabet) by the string $A_n^\star = (1, 2, \ldots, d,\ 1, 2, \ldots, d,\ \ldots, a_n)$, where $a_n = n \mod d$;*
- *this string (uniquely) maximizes the number of distinct subsequences $|\text{subseq}(A)|$;*
- *and thus (uniquely up to symmetry) the single infinite string*

  $$(1, 2, \ldots, d,\ 1, 2, \ldots, d,\ \ldots),$$

  truncated to length n, simultaneously maximizes all the quantities considered.

Before commencing the proof, we recall that the obvious "greedy alignment" algorithm suffices to determine if $B = (b_1, \ldots, b_m)$ is a subsequence of $A = (a_1, \ldots, a_n)$; see for example [CR94]. That is, we find the first appearance of character b_1 in A, then find the first appearance *after that* of the second character b_2 in A, and so forth; $B \preccurlyeq A$ if and only if we can match all the characters of B before "running off the end" of A. Formally, for $0 \le j \le m$, define $I_j(A,B)$ by $I_0(A,B) = 0$ and

$$I_j(A,B) = \min\{i\colon\ I_{j-1} + 1 \le i \le n,\ a_i = b_j\}, \tag{1}$$

with the min defined to be $n+1$ if no such value j exists. Then $B \preccurlyeq A$ if and only if $I_m(A,B) \le n$. When the arguments are clear, we will write I_j in lieu of $I_j(A,B)$.

Proof of Theorem 2. We will use a probabilistic argument to show that, for any m,

$$A_n^\star = (1, 2, \dots, d,\ 1, 2, \dots, d,\ \dots, a_n),$$

with $a_n = n \mod d$, maximizes $|\mathrm{subseq}(A) \cap \Sigma^m|$.

Fix any string $A = (a_1, a_2, \dots, a_n) \in \Sigma^n$, and let $B = (b_1, b_2, \dots, b_m) \in \Sigma^m$, B be a random string, where the b_j are chosen independently, uniformly at random. Note that the probability B is a subsequence of A is given by

$$\mathbb{P}[B \in \mathrm{subseq}(A)] = \frac{|\mathrm{subseq}(A) \cap \Sigma^m|}{d^m}. \tag{2}$$

For convenience, extend A to any infinite sequence $\bar{A}$ in which every character appears infinitely often. Through Eq. (1), each (random) B defines a corresponding random sequence $I_0, I_1, \dots, I_m$, where $I_j = I_j(\bar{A}, B)$, and $B \preccurlyeq A$ if and only if $I_m \le n$.

Define the "waiting time" to see b_j by

$$W_j = I_j - I_{j-1},$$

so $B \preccurlyeq A$ if and only if $\sum_{j=1}^m W_j \le n$. That is, Eq. (2) is equivalent to

$$|\mathrm{subseq}(A) \cap \Sigma^m| = d^m\ \mathbb{P}\Big[\sum_{j=1}^m W_j \le n\Big]. \tag{3}$$

The key to our result is showing that the waiting times W_j are dominated by i.i.d. random variables which are uniformly distributed on $[d]$, and have exactly this distribution when $A = A_n^\star$. To this end, let Y_j denote the number of *distinct* values of a_i, $I_{j-1} + 1 \le i \le I_j$, observed during the jth waiting period:

$$Y_j = |\{\bar{a}_i\colon\ I_{j-1} + 1 \le i \le I_j\}|.$$

Necessarily, $Y_j \le I_j - I_{j-1} = W_j$, and thus the right-hand side of Eq. (3) is

$$\le d^m\ \mathbb{P}\Big[\sum_{j=1}^m Y_j \le n\Big]. \tag{4}$$

For a random string B, the sequence $Y_1, \dots, Y_m$ has the same distribution as a sequence $Z_1, \dots, Z_m$ of i.i.d. unif$[d]$ random variables. To see this, observe that once character b_{j-1} has been matched, the number of distinct characters seen until b_j is matched is 1 if b_j matches $\bar{a}_{I_{j-1}+1}$, 2 if b_j matches the first distinct character after that, 3 if it is the second such distinct character, etc. Each of these "next distinct characters" is equally likely to be b_j, and every character is guaranteed to come up eventually in $\bar{A}$. Thus, expression (4) is

$$= d^m\ \mathbb{P}\Big[\sum_{j=1}^m Z_j \le n\Big], \tag{5}$$

where

$$Z_j \sim \text{unif}[d]$$

are a set of i.i.d. random variables. Thus Eq. (5), which is independent of A or $\bar{A}$, provides an upper bound on (3).

For the sequence $A = A_n^\star$, $Y_j \equiv W_j$: no character is seen twice during any waiting period. Thus $A = A_n^\star$ gives equality in inequality (4); and expression (3) achieves the upper bound given by (5), *proving a main part of the theorem*. That is, for any m, $A_n^\star$ maximizes $|\text{subseq}(A) \cap \Sigma^m|$, and it immediately follows that $A_n^\star$ also maximizes the number of distinct subsequences of every length.

We wish also to show that, up to symmetry between the characters of Σ, $A_n^\star$ is the unique string maximizing the number of subsequences. We will do so by assuming that the string A is not cyclic, and proving that inequality (4) is strict. Since over the set of strings B the event that $\sum_{j=1}^m W_j \leq n$ is a subset of the event that $\sum_{j=1}^m Y_j \leq n$, it suffices to demonstrate any string B for which the second event holds but the first does not. Since A is not cyclic, it has some d-long substring S_2 in which some character σ_2 fails to appear; working now in the extension $\bar{A}$, extend S_2 to S_2' which includes the first appearance of σ_2, and write $\bar{A}$ as the concatenation S_1, S_2', S_3 where of course S_3 is an infinite string.

Let $\bar{B} = S_1, \sigma_2, S_3$. By construction, all the values of Y_i are 1 except the $S_1 + 1$st, which by definition of Y can be at most d, so

$$\sum_{i=1}^{|S_1|+1} Y_i \leq |S_1| + d \leq (n-d) + d = n,$$

and thus there exists some value $m \geq |S_1| + 1$ for which $\sum_{i=1}^m Y_i = n$. For this value of m, let B be the m-long prefix of $\bar{B}$. Then $W_{|S_1|+1} > Y_{|S_1|+1}$, and for every i, $W_i \geq Y_i$, so $\sum_{i=1}^m W_i > \sum_{i=1}^m Y_i = n$. This B demonstrates that inequality (4) is strict for the non-cyclic string A, so expression (3) cannot achieve the bound given by expression (5). $\square$

A simple corollary holds for maximizing over a pair of strings.

Corollary 3 *Let $\Sigma = [d]$. For any $m \leq n$, $\max_{A \in \Sigma^n, B \in \Sigma^m} |\text{subseq}(A) \cap \text{subseq}(B)| = f_d(m)$.*

Proof. Trivially, $|\text{subseq}(A) \cap \text{subseq}(B)| \leq |\text{subseq}(B)| \leq f_d(m)$. If B is the cyclic sequence $A_m^\star$ then the second inequality is tight; and if A is any extension of B (for example if $A = A_n^\star$) then $\text{subseq}(A) \supseteq \text{subseq}(B)$, the first inequality is also tight, and the bound is attained. $\square$

It remains to compute the value of $f_d(n)$, which we now know to be given by the string $A_n^\star$.

Remark 4 *The maximum number of distinct subsequences $f_d(n)$ of any n-long string satisfies the recurrence*

$$f_d(n) = 1 + f_d(n-1) + f_d(n-2) + \cdots + f_d(n-d), \tag{6}$$

with initial conditions $f_d(n) = 2^n$ *for* $n = 0, \ldots, d-1$.

Proof. We exploit the regular structure of $A_n^\star$. For any first character b_1 of B, and corresponding value of W_1, there are exactly $f_d(n - W_1)$ ways to choose the remainder of B so that $B \preccurlyeq A_n^\star$. (If $n < 0$, we define $f_d(n) = 0$.) Allowing also the case that B is the empty string, $|B| = 0$, which has no first character, Eq. (6) follows.

The initial conditions follow from observing that if $n \leq d-1$ (in fact, if $n \leq d$), then all 2^n subsequences, given by independently accepting or rejecting each character, are distinct. □

It follows that for $d = 2, 3, 4, \ldots$, $f_d(n) + 1/(d-1)$ obeys the recurrence relations for the Fibonacci numbers, tribonacci numbers, tetranacci numbers, etc. (see for example the citations in [SP95]), although the boundary conditions are different for $d > 2$ (and are offset for $d = 2$).

A generating-function characterization of the numbers $f_d(n)$ and $f_d(m, n)$ is given by the following theorem.

Theorem 5 *Generating functions for* $f_d(m, n)$ *and* $f_d(n)$ *are given by*

$$F_d(x, y) := \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} f_d(m, n) x^n y^m = \frac{1}{1 - x - y - yx(1 - x^d)}, \text{ and} \tag{7}$$

$$F_d(x) := \sum_{n=0}^{\infty} f_d(n) x^n = \frac{1}{1 - 2x + x^{d+1}}. \tag{8}$$

Proof sketch. The waiting-time characterization of subsequences (after (1)) means that $F_d^m(x) := \sum_n f_d(m, n) x^n$ is obtained by summing x^n over all $W_1, \ldots, W_m$ and all n such that $1 \leq W_j \leq d$ and $n \geq \sum_j W_j$. Summing $F_d^m(x) y^m$ gives $F_d(x, y)$, and setting $y = 1$ yields $F_d(x)$. The details are standard "generatingfunctionology". □

The generating functions enumerate the subsequences exactly, but the asymptotic growth rate may be useful and is given by the following theorem.

Theorem 6 *For any* d*, there exists a constant* $2 - 2^{-d+1} < \phi_d < 2$ *such that*

$$\lim_{n \to \infty} \left(f_d(n) + 1/(d-1) - C_1^{(d)} \phi_d^n \right) = 0, \tag{9}$$

with $(1 + 1/\phi_d)^{-d} \leq C_1^{(d)} \leq (1 - 1/\phi_d)^{-d}$.

Proof sketch. Generalizing work of Miles [Mil60] and Miller [Mil71], Wolfram [Wol98, Corollary 3.5] gives a solution to the generalized Fibonacci recurrence relation (our (6) without the "1+"). This shows that $f_d(n) + 1/(d-1) = \sum_{i=1}^{d} C_i r_i^n$, where r_i are the roots of the characteristic equation $W(x) = x^d - \sum_{i=0}^{d-1} x^i = 0$, they are all distinct, the root r_1 of largest modulus is the dth generalized golden ratio ϕ_d and satisfies $2 - 2^{-d+1} < r_1 = \phi_d < 2$ [Wol98, Lemma 3.6], and the other roots have modulus $|r_i| < 1$. This proves (9).

Consider (8). Since $1 - 2x + x^{d+1} = x^{d+1}(1/x - 1)W(1/x)$, its $d+1$ roots are $1/r_i$, and $r_0 = 1$. Since they are distinct, partial-fraction expansion gives $F_d(x) = \prod_{i=0}^{d} \frac{1}{1-r_i x} = \sum_{i=0}^{d} \frac{c_i}{1-r_i x}$. This gives $f(n) = [x^n]F_d(x) = \sum_{i=0}^{d} c_i r_i^n$, so comparison with the previous paragraph shows $c_i = C_i$. Next, $(1 - r_1 x)F_d(x) = \prod_{i\neq 1} \frac{1}{1-r_i x} = C_1 + \sum_{i\neq 1} \frac{C_i}{1-r_i x}$, and evaluating at $x = 1/r_1$ yields $\prod_{i\neq 1} \frac{1}{1-r_i/r_1} = C_1$. From (9), C_1 must be a positive real, so $C_1 = |C_1| = \prod_{i\neq 1} \frac{1}{|1-r_i/\phi_d|}$; $1 - 1/\phi_d \leq |1 - r_i/\phi_d| \leq 1 + 1/\phi_d$ completes the proof. $\square$

For example, for $d = 2$, $\phi_2 = (1+\sqrt{5})/2$, the golden ratio. At the other extreme, as $d \to \infty$, ϕ_d approaches 2 exponentially quickly, since $2 - 2^{-d+1} < \phi_d < 2$. This corresponds to the case in which almost any subsequence, indicated by the presence or absence of each character, is distinct. Note that $(2/3 - \epsilon_d)^d \leq C_1^{(d)} \leq (2 + \epsilon_d)^d$, for some $\epsilon_d \to 0$.

3 Strings with maximally many distinct substrings

We close with a solution to a simpler problem, choosing an n-long string A with a maximum number of sub*strings* rather than sub*sequences*.

To avoid introducing further notation, within this section we will redefine the same notation we used before. A string B is a *substring* of A, $B \preccurlyeq A$, if there is an offset i such that

$$B = (a_{i+1}, a_{i+2}, \ldots, a_{i+m}).$$

The empty string B, with $|B| = 0$, is a substring of any string. We define the set of all substrings of A as $\mathrm{substr}(A) = \{B : B \preccurlyeq A\}$, and we redefine $f_d(n)$ and $f_d(m,n)$ to be the maximum number of substrings (respectively m-long substrings) an n-long string over $\Sigma = [d]$ may have:

$$f_d(n) := \max_{A\in\Sigma^n} |\mathrm{substr}(A)|,$$
$$f_d(m,n) := \max_{A\in\Sigma^n} |\mathrm{substr}(A) \cap \Sigma^m|.$$

Once again, the problem of minimization rather than maximization is trivial, and the following remark needs no proof.

Remark 7 *Let $\Sigma = [d]$, and let $A \in \Sigma^n$. Then for any $0 \leq m \leq n$: the number of distinct m-long substrings of A satisfies $|\mathrm{substr}(A) \cap \Sigma^m| \geq 1$; for any m with $0 \leq m \leq n$, the lower bound is achieved uniquely (up to symmetry over the alphabet) by the string $A = (1, 1, \ldots, 1)$; this string (uniquely) minimizes the total number of distinct substrings, giving $|\mathrm{substr}(A)| = n + 1$; and thus (uniquely up to symmetry) the single infinite string $(1, 1, \ldots)$, truncated to length n, simultaneously minimizes all the quantities considered.*

We turn our attention back to the maximization problem.

Theorem 8 *Let $\Sigma = [d]$, and let $A \in \Sigma^n$. Then for any $0 \leq m \leq n$,*

- *the number of distinct m-long substrings of A satisfies* $|\mathrm{substr}(A) \cap \Sigma^m| \leq \min\{d^m, n-m+1\}$*;*

- *for all m with $0 \leq m \leq n$, these upper bounds are simultaneously achieved by a modified de Bruijn word;*

- *thus this string maximizes the number of distinct substrings, giving* $|\mathrm{substr}(A)| = \frac{d^{k+1}-1}{d-1} + \binom{n-k+1}{2}$ *where* $k = \lfloor \log_d n \rfloor$.

- *For $d \geq 3$ there is an infinite string whose prefixes simultaneously maximize all the quantities considered. However, for $d = 2$ no such infinite string exists.*

There are two contrasts with the previous cases. First, our modified de Bruijn word is not unique: de Bruijn words [dB46] correspond to Eulerian tours of a certain graph and many different tours will work in our construction. Second, when $d = 2$ there is not a single infinite string whose n-long prefixes are the maximizing solutions: different values of n require modifying different de Bruijn words. But when $d \geq 3$ there is such a infinite string.

Proof. Only the second and fourth points require proof, and we take them together. Recall that a de Bruijn *graph* G_k has a vertex for each $(k-1)$-long string over $[d]$, and for each k-long string, has a directed edge from the string's $(k-1)$-prefix vertex to its $(k-1)$-suffix vertex. G_k is Eulerian, and fixing any Euler tour T, the cyclic string defined by the first letter of each edge, in order of visitation, is a cyclic de Bruijn word: it contains every k-long string. Cutting this cyclic word anywhere and concatenating its $(k-1)$-prefix gives a $(d^k + k - 1)$-long string A_k which is evidently "best possible" for $n = d^k + k - 1$: all k-long and shorter strings are present as substrings, and all $(k+1)$-long and longer substrings are distinct.

To extend this to a similar string A_{k+1}, interpret the d^k k-long substrings of A_k (which were the *edge* labels of the Eulerian tour T of G_k) as *vertex* labels in G_{k+1}, defining a Hamilton path H. G_{k+1} is $(d-1)$-connected, so for $d > 2$, deleting the edges in H from G_{k+1} leaves it connected, implying that H may be extended to an Euler tour of G_{k+1}: call it T'. Now T' defines a d^{k+1}-long cyclic de Bruijn word which can be cut anywhere and its k-prefix concatenated to give a best-possible string A_{k+1} for $n = d^{k+1} + (k+1) - 1$. Cutting the cyclic word at the original starting point (before the $(k+1)$-prefix of A_k) yields such a string A_{k+1} whose $(d^k + k - 1)$-prefix is A_k. Thus the n-prefix of A_{k+1} is best possible for all n in the range $d^k + k - 1 \leq n \leq d^{k+1} + (k+1) - 1$. Repeating the process results in an infinite string $A^\star$ each of whose prefixes is best possible for its length.

For $d = 2$, however, deleting a path H can isolate the vertices $(1, \ldots, 1)$ and $(2, \ldots, 2)$; indeed it is shown in [O'B01] that (for $k > 1$) no de Bruijn word A_k can be extended to length $2^{k+1} + (k+1) - 1$. In this case, choose A_k to end in $(1, \ldots, 1)$, so that $(1, \ldots, 1)$ is the last vertex visited by the Hamilton path H. Then H can be extended to a circuit which traverses every edge except the self-loop at $(2, \ldots, 2)$. The string associated with this circuit, having length $2^{k+1} + (k+1) - 2$, is again best possible. That is, for any k we can find a string whose n-prefix is optimal for any n in the range $2^k + k - 1 \leq n \leq 2^{k+1} + (k+1) - 2$

(which ranges partition the natural numbers), but no string can bridge two such ranges (and in particular no infinite string works for all n). □

4 Comparison with random strings

In extremal problems of any sort, an appropriate random structure is always a good candidate for consideration. For both problems considered here, random strings are *not* extremal, but it is interesting to see how close they come.

For the subsequence problem, reasoning as in the proof of Theorem 2, where the "waiting times" in a cyclic string A are uniformly distributed in $[d]$ and have mean $(d+1)/2$, the waiting times in a random string A have geometric distribution with parameter d and thus mean d. Perhaps surprisingly, this does not mean that a random string must be twice as long as a cyclic one to have the same number of substrings. For a random string A of length n, the probability that a random string B of length m is a subsequence is precisely $\sum_{n' \leq n} \binom{n'-1}{m-1}(1/d)^m(1-1/d)^{n'-m}$, as may be seen either from first principles or by noting that the sum of geometrically-distributed random variables is beta-distributed. The number of m-long strings B is d^m, so the expected number of m-long subsequences is $\sum_{n' \leq n} \binom{n'-1}{m-1}(1-1/d)^{n'-m}$. Summing over all m, this is dominated by $n' = n$ and by $m = cn$ for some fixed c. Substituting cn for m, taking logarithms, dividing by n, and differentiating with respect to c yields $c = d/(2d-1)$, and that the logarithm of the total number of subsequences is about $n \ln(2-1/d)$. For $d = 2$ this is $n \ln(3/2)$ as opposed to $n \ln(\phi)$ for a cyclic string A, a significant difference. For large d, though, $n \ln(2-1/d)$ versus a cyclic string's value of between $n \ln(2-2^{-d+1})$ and $n \ln(2)$ is not so dramatic. To summarize: both a cyclic string and a random one have exponentially many subsequences; the base of the exponent is larger for the cyclic string than for the random one, but for large d both bases tend towards 2; and the factor by which a random string needs to be longer than a cyclic one to have the same number of subsequences is more than 1 but asymptotically at most $\ln(2-2^{-d+1}) \,/\, \ln(2-1/d)$, which tends to 1 as $d \to \infty$.

For the substring problem, a random string's performance is even better: the expected number of distinct substrings of an n-long string is asymptotically maximal. In fact, for each $m \geq 2 \log_d n$, the probability that two m-long substrings (defined by starting and ending indices in A) are equal is exponentially small in their length, and so the expected number of m-long substrings is asymptotically maximal. Also, for any $c < 1$, a simple calculation shows that each string of length $m \leq c \log_d n$ will occur as a substring of n with high probability (probability $\exp(-n^{1-c})$). In summary, an n-long random string A gives an expected number of m-long substrings that is asymptotically optimal *except* for m between about $\log_d n$ and $2 \log_d n$, thus giving asymptotically the right number of substrings in all (summed over $m = 0, \ldots, n$).

Finally, since the maximal number of subsequences is given by Fibonacci numbers and related series, we remark that there is a notion of a *Fibonacci string*. These strings, with $A_0 = (2)$, $A_1 = (1)$, and $A_i = (A_{i-1}, A_{i-2})$ (so $A_2 = (12)$, $A_3 = (121)$, $A_4 = (12112)$, *etc.*) are the extremal examples for the Periodicity Lemma on strings (see [FW65] and for example [CR94]), and are natural candidates for other extremal properties. However, they are not extremal for the number of distinct subsequences, nor for the number of distinct substrings.

Acknowledgments

We thank Al Aho for suggesting the subsequence question, Don Coppersmith for helpful conversations, and an anonymous referee for a helpful and remarkably expeditious review.

References

[Aho03] Alfred Aho, personal communication, 2003.

[Ait27] A. C. Aitken, *On Bernoulli's numerical solution of algebraic equations*, Proc. Roy. Soc. Edinburgh Sect. A **46** (1927), 289.

[CR94] Maxime Crochemore and Wojciech Rytter, *Text algorithms*, The Clarendon Press Oxford University Press, New York, 1994. With a preface by Zvi Galil. MR 96g:68038

[dB46] N. G. de Bruijn, *A combinatorial problem*, Koninklijke Nederlandse Akademie v. Wetenschappen **49** (1946), 758–764.

[FoP02] Leonardo Fibonacci of Pisa, *Liber abaci*, 1202.

[FW65] N. J. Fine and H. S. Wilf, *Uniqueness theorems for periodic functions*, Proc. Amer. Math. Soc. **16** (1965), 109–114. MR 30 #5124

[Hor61] A. F. Horadam, *A generalized Fibonacci sequence*, Amer. Math. Monthly **68** (1961), 455–459. MR 23 #A847

[Mil60] E. P. Miles, Jr., *Generalized Fibonacci numbers and associated matrices*, Amer. Math. Monthly **67** (1960), 745–752. MR 23 #A846

[Mil71] M. D. Miller, *On generalized Fibonacci numbers*, Amer. Math. Monthly **78** (1971), 1108–1109.

[O'B01] Matthew J. O'Brien, *De Bruijn graphs and the Ehrenfeucht-Mycielski sequence*, Master's thesis, Mathematical Sciences Department, Carnegie Mellon University, 2001.

[SP95] N. J. A. Sloane and Simon Plouffe, *The encyclopedia of integer sequences*, Academic Press Inc., San Diego, CA, 1995. With a separately available computer disk. MR 96a:11001

[Wol98] D. A. Wolfram, *Solving generalized Fibonacci recurrences*, Fibonacci Quarterly **36** (1998), no. 2, 129–145. MR 99c:11015