Surprising Strings

Dennis E. Shasha describes an interesting problem in his *Puzzling Adventures* column entitled "You Don't Say." It appears on page 122 of the December 2003 issue of *Scientific American*.

He calls a string over an alphabet *surprising* if there are no two symbols x and y of the alphabet such that there are two pairs of occurrences of these symbols where x precedes y by the same distance. For instance, the string AAXYBB is not surprising since there are two occurrences of the symbols A and B where A precedes B by 4 symbols. Likewise BCBABCC is not surprising since there are two occurrences of the symbols B and B where the first B precedes the second B by 2 symbols (so overlaps do count).

Shasha gives the reader three challenges. The first is to construct the longest surprising string you can that is composed from 5 distinct symbols. The other two are to do the same for 10 and 26 symbols.

Some results. The following table describes the longest surprising strings for an alphabet of size n up through $n = 9$. For each n the length s_n of the longest surprising string is given, along with the quantity of surprising strings of that length and an example surprising string of that length. When counting surprising strings, two strings are considered the same if one can be found by permuting the letters of the alphabet. For instance, the string 0010 and the string 1101 are considered to be the same. The example string that is listed is the first one in lexicographic order of that length.

n	s_n	quantity	example
1	2	1	00
2	4	3	0010
3	7	4	0012102
4	10	2	0112032310
5	12	212	001232410431
6	15	770	001231452503410
7	18	1630	001231456264035102
8	21	1396	010234563742761154032
9	24	312	012334546785281076142053

These results came from an exhaustive computer search. The algorithm that was used was a straightforward search in which surprising strings were extended one symbol at a time.

Some bounds on the length of the longest surprising string. Various arguments can be given on the length s_n of the longest surprising string for an alphabet of size n.

Here's an argument that s_n is less than or equal to $n^2 + 1$.

> Consider the symbols that follow a symbol x in a surprising string. Any such symbol may occur, but each at most one time. Since there are n of them, that means that x itself can occur at most n times in the string, unless it's the last symbol in the string, in which case it could occur $n + 1$ times. Altogether, there are n different symbols in the alphabet, and each can occur at most n times, except the one that occurs last, so that gives at most $n^2 + 1$ symbols in the string.

There is however, a much better bound on s_n, namely $3n$. And here's an argument for that bound. First, turn the definition of surprising string around a bit. It was stated as follows: a string is not surprising if there are two symbols x and y of the alphabet such that there are two

pairs of occurrences of these symbols where x precedes y by the same distance. You could say instead that a string is not surprising if there are two symbols x and y each occur twice in the string and the occurrences of each are separated by the same distance, that is, the two xs occur the same distance apart as the two ys. In other words, each distance can be the distance between only one pair of occurrences of the same symbol. Now if a string has length l, then the only possible distances between symbols of the string are 1, 2,..., l, − 1 so there are exactly l − 1 distances possible, and each one can be the distance between at most one pair of occurrences of the same symbol.

Now, suppose that a symbol occurs 3 times in the string. Then there are 3 distances between these 3 occurrences, using up three of the possible l − 1 distances. If every symbol occurred exactly 3 times, then $3n$ distances would be used up out of a possible $3n$− 1 distances, which is impossible. Therefore, not every symbol can occur exactly 3 times.

But if a symbol occurs more than 3 times, even more distances are used up. If a symbol occurs 4 times, then there are 4 choose 2, which equals 6, distances between these 4 occurrences. So 3 more distances are used up, but the length of the string only increases by 1. And a symbol that occurs 5 times uses up 5 choose 2, which equals 10, distances. It looks like triple occurrence should be the best you can do to find long surprising strings.

In fact, this intuitive feeling can be proved.

> Let a be the number of letters that occur once, b the number that occur twice, c the number that occur three times, and so forth. Since each letter occurs some number of times,
>
> $$n = a + b + c + d + \ldots$$
>
> and the length of the string l is
>
> $$l = a + 2b + 3c + 4d + \ldots .$$
>
> To show that $l < 3n$, we need to show that
>
> $$a + 2b + 3c + 4d + \ldots < 3a + 3b + 3c + 3d + \ldots,$$
>
> that is, we need to show that
>
> $$d + 2e + 3f + \ldots < 2a + b.$$
>
> Now, if a string has length l, then there are only l − 1 different distances between the letters of the string. Each letter that occurs once uses none of these distances; each that occurs twice uses 1; each that occurs three times uses 3; each that occurs 4 times uses 6 (which is 4 choose 2); each that occurss 5 times uses 10 (which is 5 choose 2); etc. Therefore,
>
> $$b + 3c + 6d + 10e + 15f + \ldots < l = a + 2b + 3c + 4d + \ldots .$$
>
> That means
>
> $$2d + 5e + 9f + \ldots < a + b.$$
>
> Now the coefficients 2, 5, 9, ... of the left hand side grow much faster than 2, 4, 6, ... (since the k-th coefficient is (k choose 2) − k, and that's less than 2(k − 3)), therefore,
>
> $$2d + 4e + 6f + \ldots < a + b.$$
>
> That implies

$$d + 2e + 3f + \ldots < (a + b)/2 \leq 2a + b$$

which is what was to be shown.

Thus, we know the length of the longest surprising string for an alphabet of size n is less than $3n$.

In his column, Shasha notes that "this length doesn't increase very fast, and I believe that even for 26 symbols, the longest surprising sequence is less than 100 letters long. Indeed, the bound of $3n$ shows it has to be less than 78 symbols long.

Back to Clark's Math Problem Solving Team page

This page is located at http://alepho.clarku.edu/~djoyce/mpst/surprising/
David E. Joyce, Nov. 14, 2003
Department of Mathematics and Computer Science
Clark University
Worcester, MA 01610