

Expected length of the longest common substring of two strings

1) Suppose we are given two random strings S_1 and S_2 , of length n each, where the characters are drawn uniformly from an alphabet of size z . That is, at each position in either string, any character in the alphabet appears at that position with probability $p = 1/z$.

Then given a specific position i in S_1 , and a specific position j in S_2 , and a number m , we note that the probability that the two strings have the same m characters (completely match for m positions) starting from position i in S_1 and position j in S_2 is exactly p^m . They might match further to the right for more than those m positions, but here we are only concerned with matching those m positions.

2) Next we define the variable $X_{i,j}(m)$ to have the value 1 if the two strings completely match for m positions starting from positions i and j respectively, and to have the value 0 otherwise. Then

$$C_m = \sum_{i=1}^{i=n} \sum_{j=1}^{j=n} X_{i,j}(m)$$

is the number of pairs of starting positions from which the two strings completely match for m positions. Let us call that number C_m for that pair of strings.

Now we can see that $E(C_m)$, the expected number of pairs of starting positions from which two random strings match for m positions is

$$E(C_m) = E\left(\sum_{i=1}^{i=n} \sum_{j=1}^{j=n} X_{i,j}(m)\right) = \sum_{i=1}^{i=n} \sum_{j=1}^{j=n} E(X_{i,j}(m)) = (n - m + 1)^2 \times p^m$$

which is about $n^2 p^m$, when n is large compared to m . So $E(C_m)$ is roughly $n^2 p^m$. Note that for a fixed n , the number $E(C_m)$ falls (and rather rapidly) with increasing m . That is, the bigger the m , the smaller the expected number of such pairs. This should make intuitive sense.

3) Now let r be largest integer such that $E(C_r) \geq 1$. That is, for $m > r$, $E(C_m) < 1$. Certainly, there is such an integer number r , although it is not clear at this point why we are interested in it. Of course, $E(C_r) = n^2 p^r$, so $n^2 p^r \geq 1$.

To solve for r in terms of n and p , we take the log base p of both sides, so

$$\log_p 1 = 0 \geq \log_p n^2 + \log_p p^r = 2 \log_p n + r.$$

Remember (even though I forgot in class) that when the base of a log is less than one, as p is, then the log is a negative number, so the log is a *decreasing* function. That is why $\log_p 1 \geq \log_p n^2 p^r$, even though $1 \leq n^2 p^r$.

Hence

$$r \leq -2 \log_p n = 2 \log_{1/p} n = 2 \log_z n.$$

This is the conclusion I showed in class, but what has made it work (now that we corrected for the base of the log being less than one) is that I defined here r to be the largest integer such that $E(Cr) \geq 1$, rather than the smallest integer such that $E(Cr) \leq 1$.

4) Now when $m > r$, the expected number of pairs of matches of length m is less than one, so "on average" we will not see even one match of length $m > r$. So "on average" the length of the longest common substring is not going to be larger than r . This argument is somewhat heuristic and unsatisfying, but it suggests that the expected length of the longest common substring in two random strings is no more than r , hence no more than $2 \log_z n$.

In summary, the expected length of the longest common substring, as a function of the alphabet size z , and the string length n , is at most $2 \log_z n$. We didn't derive an equality, but since log is a very slowly growing function, the inequality we derived is still of interest, because it says that the expected length is less than or equal to a slowly growing function, and hence is small compared to n . That is a very different conclusion than we saw for the expected longest common subsequence.

Now in fact, a fully rigorous argument can establish that the expected length of the longest common substring is precisely $2 \log_z n$. But that is beyond the scope of this class.

So, given two random strings of length n each, it is not surprising to find a common substring of length $2 \log_z n$. For example, if $z = 4$, and $n = 1,000,000$ then it is not surprising that the longest common substring of two random strings of length 1,000,000 each is 20. However, the probability that the longest common substring is of length 40 or more is miniscule (how do you use what we have done here to conclude that?)

There are two major points that come out of this discussion. First, the expected length of the longest common substring grows rather slowly as a function of n , in great contrast to the expected length of the longest common subsequence, which grows linearly with the length of n . So, conclusions that were not justified when using longest common subsequence (say that it has length $n/3$), might be justified when using longest common substring (say if you find a common substring of length $n/3$). Second, you need to know how the expected length of the longest common substring grows as a function of n if you use the length of the longest common substring as an indication of the biological relatedness of two sequences. You need to have a longest common substring that is larger than $2 \log_2 n$ to even think you should conclude (from that evidence alone) that the two strings share a relationship that comes from some cause other than random happenstance.