

The probability of a random query matching a sequence in a database

We let z denote the size of the alphabet, and $p = 1/z$ is the probability of randomly picking a character from the alphabet. We assume that a random query sequence S of length m is made up by randomly picking each character in S . Similarly, a random database D of length n is made up by randomly picking each character in D . We think of D as one long sequence of length n . Note that n is vastly larger than m .

We want to know what the probability is that S exactly matches at least one substring in D .

The probability that S matches to D starting at a particular point i is p^m , so the probability that it does not match there is $1 - p^m$.

Now ignoring the issue of dependence between starting points, the probability that S exactly matches no substring in D , is

$$(1 - p^m)^{n-m}$$

which is about

$$(1 - p^m)^n = (1 - np^m/n)^n \approx e^{-np^m} = 1/e^{np^m}.$$

So the probability that S matches at least once somewhere in D is

$$1 - 1/e^{np^m}.$$

Now when does this probability get so large that an exact match really offers no evidence that there is biological relationship between S and the sequence in D it matches to? Certainly if the probability of a random match is at least 0.63, then a match could easily be explained as a random event.

Now $1 - 1/e^{np^m} = 0.63$ when $1/e^{np^m} = 0.37 = 1/e$, (which you all remember I'm sure). So $1 - 1/e^{np^m} = 0.63$ when $np^m = 1$, or $n = 1/p^m = z^m$. Now 0.63 is a relatively large probability, and this gives rise to the rule-of-thumb, that exact matches could certainly be the product of random happenstance when the size of the database is as large or larger than the alphabet size raised to the length of the query. To even think of using an exact match as evidence of a biological connection, you need to be on the correct side of this rule.

Another way to write the rule is if m is smaller than $\log_z n$, then the probability of a match by random is too large to ascribe any significance to an exact match.

Now, depending on your application, and your tolerance for false positive matches versus false negative results, a cutoff of 0.63 may seem much too large or much too small. The result above simply gives you a signpost - a rule of thumb relating the size of the database to the size of the query and the size of the alphabet, to the probability of 0.63 that a match will be found in random strings.

Good exam questions can be created from this exposition by asking how specific changes in the model or assumptions, change the the qualitative outcome of the analysis.