

1 Introduction to Statistical Issues in Database Search

When we use a target sequence to search a database of sequences, either by local or global alignment, we want to know if the high scoring alignments are biologically meaningful. Can we learn something useful about our target sequence from the high scoring alignments and what is known about the sequences involved in those alignments? This is of course an empirical question about the validity of the whole approach of trying to expose important biology from comparing sequences by aligning them under some objective function. The answer has been Yes (at least often enough), or otherwise we would not be so invested in database search and sequence alignment. However, it is not always true that the database sequence that gives the best alignment to the target sequence is actually related to the target sequence in any meaningful way. Since we often have no good intuition about whether an alignment is meaningful or not (that is, whether it reveals significant biology), we often use probability and statistics to help identify alignments that might be biologically important, or more often, to exclude those that are not.

Suppose you align a target sequence S_1 to every sequence in a database D and the best similarity value found is S , and that is obtained by aligning S_1 with S_2 . How high does S have to be before you should get excited? Is the function, structure, or other properties of S_2 likely to be carry over to S_1 ? Only additional work (perhaps lab work) will resolve this. But we would like some cheap way to tell whether additional work is likely to be profitable.

The rough philosophy that is followed is to think about aligning S_1 to each sequence in a database the size of D consisting of randomly generated strings, where the probability of any particular character is its frequency in D . Let D' denote this randomly generated database and let S' be the largest similarity value of S_1 to the strings in D' . If S' is "much smaller" than S , then one feels more confident that the alignment of S_1 to S_2 has some true biological meaning, and its worth investing additional effort to confirm this. Conversely, if S' is larger than S , then one feels much less confident (at least without some additional evidence or intuition) that S_2 has any valuable relationship to S_1 . (There really is a philosophical point here that runs all through science and statistics, but I am having a hard time making it explicit: Phenomena that could easily occur as the result of

random acts, should not be taken as evidence that a non-random process is being observed, even though that may be the truth. This is actually a large debate, but we will leave it here – lacking any other reasons to believe that S_1 and S_2 are related in a meaningful way, S must be "close to" or larger than S' to be considered as good evidence of a relationship.)

So much for philosophy. How do we make this precise? One approach is take every sequence in D and randomly permute its characters. The resulting database is taken to be D' . Then align S_1 to every string in D' . Because we compute the similarity of S_1 to every string in D' , we have the complete distribution of the similarity values, so we can see where S falls in that distribution. Let f be fraction of similarity values that are above S . We might pick some small cutoff C , say .001, and use the test that if $f > C$, then S is not large enough to be considered significant. That is, if one in one-thousand of the strings in D' are more similar (have a higher similarity value) to S_1 than does S_2 , then a similarity value of S seems like weak evidence for a true relationship between S_1 and S_2 .

An alternative is to randomly permute S_1 and then align that string to D . What are the advantages and disadvantages of doing that?

Now what actual cutoff C should be used? That depends entirely on ones tolerance for false-positive results versus false-negative results. What are you most concerned about, wasting time and money checking out a sequence that is not actually related to S_1 or missing an important clue to solving some important biological problem? If you have cheap ways to check out the reported sequences then you can set C high, but if not, you set C low. In the end you decide if you want to use database search as a *filter* or an *oracle*.

Now what I've outlined above doesn't change if we use a score other than similarity (local or global), as long as we can get an empirical distribution of those scores when comparing S_1 to every sequence in D' (or a large enough sample of D'). And this kind of "Jumbling" or "Shuffling" test is actually used in some search applications. But this approach is not feasible for Blast or Fasta, because so many sequence are filtered out before their score is computed. In that case, analytical results are needed. For concreteness, let S now denote the best Blast score obtained by Blasting S_1 against D . We want a *formula* for the probability that a score of S or better would be obtained from Blasting S_1 against D' , i.e. a database of random sequences, where the size of that database is the same as D , and where its character composition is the same as D . Next week, we will discuss such a formula.