(10 Mar 2015)

Introduction to Bioinformatics
Problem Set 5: Blast

1. Why did the search for FMRP in *Drosophila* work well when human FMRP protein was used as the query but failed so abysmally when the corresponding human gene was used as the query? Now's the time to find out.

1a. Hearken back to the tour *Search for FMRP in Drosophila* and consider item 15, the comparison of the human and *Drosophila* proteins. The comparison begins with a stretch of very high amino acid similarity, from the initial M (methionine) up to YK (tyrosine-lysine). How many amino acids are in that stretch?

1b. How many nucleotides are required at the beginning of the gene to encode those amino acids (M through YK)?

Let's get those nucleotides, from the human gene and from the *Drosophila* gene. If the amino acid sequences are extremely similar, then so should be the nucleotide sequences, no? The protein sequence found by Blast comes with a link to the corresponding gene, with the GenBank ID of NM_169324.2. You determined during the tour that the GenBank sequence does not begin with the FMR gene but includes upstream sequence. You'll recall that the gene begins at nucleotide 423.

1c. In BioBIKE (any instance will do), DEFINE a variable (maybe `fly-seq`) as the beginning of the fly gene, using SEQUENCE-OF and specifying the FROM-GENBANK option. The argument should be the GenBank ID (in quotes). Also specify values for FROM and TO, so that you bring in only those nucleotides from the beginning of the gene to the end of the portion of the gene encoding Y and K. To check if you were successful, get the TRANSLATION-OF `fly-seq` and compare it to the amino sequence you considered in **1a**.

1d. Retrace your steps in the tour and find the GenBank ID for the human FMR gene and what nucleotide coordinate the gene starts with. DEFINE a variable (maybe `human-seq`) as in 1c, but using the appropriate GenBank ID.

1e. Align these two DNA sequences, using the ALIGNMENT-OF function. The argument should be a LIST consisting of two items: `fly-seq` and `human-seq`.

1f. Does the alignment look as good as the amino acid alignment? Why not? Do something to test your theory.

2. The gene D29p32 from mycobacteriophage D29 presents an anomaly. Using PhAnToMe/BioBIKE (Ph/BB), find the description of this gene, perhaps using DESCRIPTION-OF or by going to the gene within the Sequence Viewer. You'll see that the gene is called an integrase, i.e. a protein responsible for the integration of the phage genome into the host genome to initiate lysogeny. But then find the DESCRIPTION-OF D29. There's way too much information, but fortunately, the identifiers are in alphabetical order. Look for the lifestyle of the phage… it's described as lytic![*] What is a lytic phage doing with an

[*] For those who did not do Phage Lab,... Bacterial viruses (phages) come in two varieties: (1) those that are lytic (always lysing – killing – their hosts) and (2) lysogenic (capable of lysis but also of integrating into the genome of the host bacterium). A lytic phage would have no use for an integrase protein.

integrase? Maybe the description of the gene is in error? Or maybe the description of the phage?

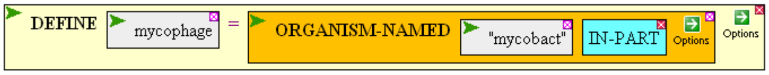
2a. Is there similarity between D29p32 and perhaps known integrases? Check that by blasting the sequence against all known protein. You could do this at the NCBI site as you have in the past, but why not enjoy the one-stop-shopping afforded by BioBIKE? In Ph/BB, use SEQUENCE-SIMILAR-TO to access Blast, giving the protein of D29p32 (remember the p- convention) as the query and *GENBANK* (obtainable from the DATA menu) as the target. This tells Ph/BB to ask NCBI to perform the Blast search using all proteins contained in the GenBank database. Having absorbed the lesson of the previous problem, you'll undoubtedly use the sequence of the protein, not the gene. Specify PROTEIN-VS-PROTEIN as the type of sequence comparison, and execute the function.

In a few seconds you should get the results. Note the identity of the best hit (note also that NCBI's name of the gene differs from that in BioBIKE… I hope you're getting used to this!). The other hits are mostly annotated as integrases. One is of particular interest. You'll soon read an article regarding the genes of mycobacteriophage Che12, so keep an eye on its genes. Note the E-value for the match between D29p32 and the gene from Che12. Write it down.

2b. You would think that most of the search performed by NCBI in **2a** was wasted effort. After all, most of the GenBank database consists of eukaryotic sequences. Let's try it again with a more reasonable target, just phage sequences. Repeat the execution of SEQUENCE-SIMILAR-TO, but this time using as the target `all-phage` (a set that includes all phages known to Ph/BB).[†] Also use the BYPASS-LOOKUP option, to force BioBIKE to actually run Blast rather than merely look up a precomputed value.

The format of the results may change, but the message is about the same, no? Why does the first Blast but not the second show hits to FRAT1 and Eagle? Again, note the best hit and also jot down the E-value for the match to the protein from Che12.

2c. Even the confined search of **2b** was largely wasted effort. The closest matches to a mycobacteriophage protein would figure to be other mycobacteriophage proteins. So re-execute SEQUENCE-SIMILAR-TO, this time with a target of only a subset of all phages, the mycobacteriophages. To do this, DEFINE a variable consisting of all organisms with "mycobact" as part of their name, like so:



DEFINE mycophage = ORGANISM-NAMED "mycobact" IN-PART Options Options

(actually, IN-PART is the default behavior, so the option isn't necessary). You can confined the list to phages by using the PHAGE-ONLY option.

How do the results compare with those of **2b**? Now what is the E-value for the match against the Che12 protein?

[†] Ph/BB attempts to speed up the execution of SEQUENCE-SIMILAR-TO by pre-running every possible Blast of proteins it knows about against all other proteins it knows about. Then when called upon to do the blast, it simply looks up the results in a massive table. Fast, yes, but not always reliable. If you get an error message complaining that a file doesn't exist (probably an element of the lookup table), then bypass this feature, using the BYPASS-LOOKUP option. Then Ph/BB will perform a conventional Blast.

2d. If we're so interested in Che12, then why not do a directed search just of that genome? Do a last execution of SEQUENCE-SIMILAR-TO, using Che12 as the target. What is the E-value for the match against the Che12 protein?

2e. Consider all the E-values you've noted for matches of p-D29p32 against the Che12 protein. Why aren't they all the same? Develop a hypothesis that ***quantitatively*** accounts for the differences in E-values you observed and test that hypothesis.

3. Execute SEQUENCE-SIMILAR-TO D29p32 IN Che12 using all five flavors of Blast (DNA-VS-DNA, etc), one at a time. Also use the BYPASS-LOOKUP option (see footnote on previous page), as doing so will ensure a more fair comparison.

3a. One of the five searches gives an E-value much worse than the others. Which one and why?

3b. One of the five searches gives far more hits than the others. Why?

3c. The search you just identified in **3b** has 6 hits that are far better than the rest. Why? Why is the best hit in that search so much better than hits 2 through 6?