Computational search for gene sequences
Blast Mystery #2: Anthrax attack or not?

Now that you are familiar with how you can use Blast to find similar sequences, let's try it out on another problem:

Be a hero and thwart a microbial attack

You are working for the Center for Disease Control and a blood sample crosses your desk from one of many paitents who have rapidly and mysteriously expired after having exhibited symptoms of anthrax poisoning. Your first thought is to check to see whether there is evidence of *Bacillus anthracis*, the causative agent of anthrax. You can test this idea by attempting to amplify an internal portion of the *lef* (lethal factor) gene encoding the anthrax toxin, found on a the plasmid pXO1 of *B. anthracis*. Your PCR amplification is successful, and you sequence the PCR product. The sequence is available at the DG29 link. Is the sequence similar to portions of genes that encode toxin?

1. Find DNA sequences similar to DG29

   You suspect that the amplification fragment will be an exact match to the *B. anthracis* toxin gene, so you go straight to nucleotide Blast.

   a. Go to the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/)

   b. Go to the BLAST home page (click BLAST on the bar near the top of the page)

   c. Click nucleotide blast.

   d. Paste the DG29 sequence (obtainable from the module web page) into the large white box. In the **Database** section check "Others", so we have a chance of finding *Bacillus* sequences. In the **Program Selection** section, click "Somewhat similar sequences". Finally, press **BLAST** at the bottom of the page.

   e. In a few seconds you should see a page filled with red lines,… No! Just lots of little matches. Scroll down to see what they are. *Bacillus anthracis* genes? Why not? What went wrong?

2. Find protein sequences similar to DG29

   In Blast Mystery #1, you were able to find hemoglobin in sharks by using a version of Blast that translates a given DNA sequence in all six reading frames. Maybe you should try this strategy again, on the theory that the PCR-amplified fragment is only distantly related to known anthrax genes. Or maybe it's not related at all.

   a. Return to the NCBI Blast page. This time click BlastX. This flavor of Blast automatically translates your DNA sequence in all six possible reading frames and compares each one against a protein database. Paste in the DG29 DNA sequence if necessary, and click BLAST.

   b. Now many hits! What are they? How long are the hits? Are they as long as the PCR fragment? How good are the hits?

   c. Most peculiar! BlastX didn't give distantly related protein sequences, it gave ***identical*** protein sequences! Why then didn't BlastN (nucleotide Blast) turn up anything? Shouldn't the gene encoding the identical protein sequence be very similar to DG29?

3. Obtain the sequence of a gene encoding a protein sequences similar to DG29

Having found a protein in GenBank, you'd think it would be easy to find the corresponding gene, i.e. a DNA sequence encoding the protein. It's not. I'll save you lots of pain and suffering through short cuts.

a. Scroll down the list of protein matches, looking specifically for a hit that: (a) is a good hit, (b) mentions *Bacillus anthracis* (and not some other *Bacillus*), (c) mentions "lethal factor", and (d) mentions pXO1. The first one I find is gb|AAA22569.1 (i.e., GenBank sequence with the accession ID AAA22569, version 1). Click on its link.

b. Scroll down. You'll see that the record gives you the amino acid sequence of a lethal factor protein. We want its gene. You can get it by examining the ***database source*** (DBSOURCE) of the protein sequence. That field (near the top of the record) indicates that the protein sequence was derived from M30210.1. ***Write down*** that accession number and follow the link to confirm that it leads to the gene. It does.

4. Compare DG29 with the sequence of the gene encoding the lethal factor

According to nucleotide Blast, they should not be at all similar. According to BlastX, they should be nearly identical. Who is right?

a. Go into BioBIKE (any flavor) and DEFINE a variable containing the sequence of DG29. To do this, bring down DEFINE and name the variable anything you want. Then click the *value* box and bring down SEQUENCE-OF. Click Multiline Input from the green Action Icon of the *entity* box, giving you more room to play with. Type in white box two double quotes, and paste between the quotes the sequence of DG29 (only the sequence, not the FastA header). Click Enter. Execute the DEFINition.

b. Now DEFINE a variable containing the sequence of the gene encoding lethal factor. To do this, bring down DEFINE and name the variable anything you want. Then click the *value* box and bring down SEQUENCE-OF. Click the *entity* box and type the accession number of the gene ("M30210.1"), between double quotes. Then choose the FROM-GENBANK option from the option menu of the function to specify that that "M30210.1" is an accession number and not a sequence itself. Execute the DEFINition.

c. Align the two sequences, bringing down ALIGNMENT-OF. This function will align a list of sequences, as many as you like. Note that the argument is labeled *sequence-list.* So you need to provide a list. Click *sequence-list* and bring down the LIST function. Your sequence-list will contain two sequences: DG29 and the lethal factor gene. So you need two holes. Use the Option menu of LIST to add another *item.* In the first item put in the first variable you defined, and in the second put in the second variable. Execute the alignment.

d. The lethal factor gene is much bigger than the small PCR fragment you isolated, so you'll have to scroll well down the alignment to find the region where the two coincide. But when you get there… are the two sequences similar? Could you by eye have picked them out as similar?

The mystery

- **Why couldn't nucleotide Blast find the similarity between these two sequences?**