Computational search for gene sequences
Blast Mystery #1: Do sharks have blood?

You may believe that your calling is pipets and microscopes. You may consider that the best use of a computer is to hold open a door. Nonetheless, if you have any contact at all with bioinformation (and if you're a biologist, you have or you will), then you will most likely make contact with Blast.

Blast is a program that makes it possible to answer one of the most common questions biologists pose: "Here I am looking at my favorite [gene, protein, sequence fragment]… what similar has been seen before, and how similar is it?" You may want to identify an unknown gene or to place a known gene within an evolutionary context. You might want to learn what parts of a protein are conserved and what parts are variable.

Wan-Ling turned to Blast because she wanted to know what genes are expressed in the symbiotic glands of *Gunnera manicata*, as judged by what genes are found in a large set of sequenced mRNA transcripts. We'll play with those actual sequences in the research simulation, but for now, let's examine how Blast can be used for such purposes. We'll do so by trying to answer a different question

Do sharks have blood?

Dark denizens of the deep – sharks! Cold bloodless killers! Well, they may be cold blooded, but are they actually *bloodless*? Maybe they're getting a bad rap. We'll find out, using Blast to search in shark sequences for sequences similar to a gene we know to encode hemoglobin.

1. Get a sequence for human hemoglobin (Might as well start with a sequence we know)
    a. Go to the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/)
    b. In the ***Search*** box, click the down arrow and then select ***Nucleotide***
    c. In the ***for*** box, enter `hemoglobin` and click ***Go***
    Thousands of sequences! We've go to find a way to cut down the possibilities.
    d. Modify the ***for*** box: `hemoglobin AND human [organism]`. This confines the search to hemoglobin nucleotide sequences from humans.
    Now hundreds! Many of them are termed "variants". I wish they would go away. Also, I want only complete sequences…
    e. Modify the ***for*** box:
    `hemoglobin NOT variant AND complete AND human [organism]`
    Now dozens! Maybe this is good enough. Scanning a few items down you might find what we're looking for: Homo sapeins hemoglobin, alpha 1 [i.e. alpha hemoglobin], mRNA. Click the entry

2. Examine the GenBank entry for human alpha hemoglobin
    a. It was not easy finding the gene we wanted just from an English description! So we don't have to do this again, **write down** the Accession Number (BC…), the unique identifier for this sequence.

b. Notice (from the definition) that the sequence is of mRNA. That's good. It won't have any introns to complicate matters. Notice also that the record provides links to many articles regarding alpha hemoglobin. This is quite a resource if you're looking for information about the gene.

c. Scroll down and you'll find a field labeled "CDS" (which I believe illogically stands for *CoD*ing *S*equence). The numbers on that line give you the beginning and end coordinates for the gene. **Write them down** and scroll down to the nucleotide sequence itself at the end. Locate the first three nucleotides and the last three nucleotides of the gene, according to the coordinates. Are the first three a start codon? Are the last three a stop codon?

3. Copy the nucleotide sequence of human alpha hemoglobin

a. You found the sequence, but it would be simpler if you could get rid of the spaces and coordinate numbers. You can! Scroll back to the top of the screen and click FastA. FastA format is pure sequence, except for the first line (identified by ">") which gives information about the sequence.

b. Copy the sequence (along with the informational first line) and paste it somewhere, e.g. into Notepad.

4. Find DNA sequences similar to human alpha hemoglobin

a. Go back to the NCBI home page (click the NCBI icon at the upper left part of the page).

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

c. Click nucleotide blast, since we have a nucleotide sequence and we want to find another similar nucleotide sequence, specifically one in sharks.

d. Paste the human alpha hemoglobin sequence into the big white box (you could also just type in the accession number if you want). In the **Database** section you'll probably find "Human Genomic + transcript" checked. For some reason NCBI believes most people want to search only human sequences. We don't. Check "Others", so we have a chance of finding shark 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, indicating LOTS of very similar sequences. Scroll down to the descriptions of these sequences. Unfortunately they're all from homo sapiens. That figures. Blast orders the results by most-to-least similar, and of course human alpha hemoglobin is most similar to other human alpha hemoglobins. We need a different strategy.

5. Find DNA sequences similar to human alpha hemoglobin

a. Go back to the nucleotide BLAST home page for another try. You may have to paste the sequence back in the box again. Make sure the **Database** and **Program Selection** are as before.

b. This time type *sharks* into the **Organism** box. You'll find many suggesed categories come up. Click "sharks and rays".

c. Run the BLAST again. This time very few hits! And what are they? "Procollagen"? "Immunoglobin"? What do these things have to do with hemoglobin? Look also at the extent of the hits in the **Alignments** section. Tiny! No more than 19 nucleotides (out of 494 possible). This is garbage! Maybe sharks really have no blood?

6. Examine the DNA sequences and translate it

Maybe we made a strategic error, searching for hemoglobin DNA rather than hemoglobin protein sequences. We can recover.

a. Go into BioBIKE (any flavor) and DEFINE a variable that contains the nucleotide sequence of human alpha hemoglobin. You can do this by putting into the *value* box the SEQUENCE-OF function, and typing into the *entity* box the accession number of human alpha hemoglobin (in quotes). (Aren't you glad you wrote down the accession number back in 2a?). Finally, click the Options icon and select FROM-GENBANK, to tell BioBIKE that the thing in quotes is a GenBank accession number. Execute the DEFINition, and you should have the sequence in computer-readable form.

b. Translate the DNA in all six possible reading frames by bringing down READING-FRAMES-OF and putting into the *entity* box the variable you just defined. Executing the function should give you two lines of nucleotide sequence (labeled "sequence" and "complement") and six lines of amino acid sequence (labeled "translation-frame-1" through "6"). What do they mean?

c. Locate the beginning and end of the amino acid sequence of human alpha hemoglobin using the coordinates you so carefully wrote down in 2c. What symbol is at the beginning of the sequence? Why? What symbol is at the end? Why? Which of the six reading frames is the one used by the alpha hemoglobin gene?

7. Find protein sequences similar to human alpha hemoglobin

Suppose you forgot to write down the coordinates of the gene. Or better, suppose you had no way of knowing them. You *suspected* that the DNA sequence had a gene within it, but you didn't know where. You could take all six translations and BLAST each one against the set of all known proteins. One of them would work, and five of them would fail. That's a pretty tedious approach. There's a simpler way.

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 DNA sequence of human alpha hemoglobin, select "sharks and rays", as before. Click BLAST.

b. Now many hits! What are they? How long are the hits? Are they as long as the gene?

c. Do sharks have hemoglobin?

The mystery

- **Why did finding a hemoglobin gene through protein similarity work while finding it through DNA similarity did not?**
- **What moral can you draw from this story about how to find similar genes in distantly related organisms?**