7.02 BLAST TUTORIAL ANSWER KEY
Spring 2005

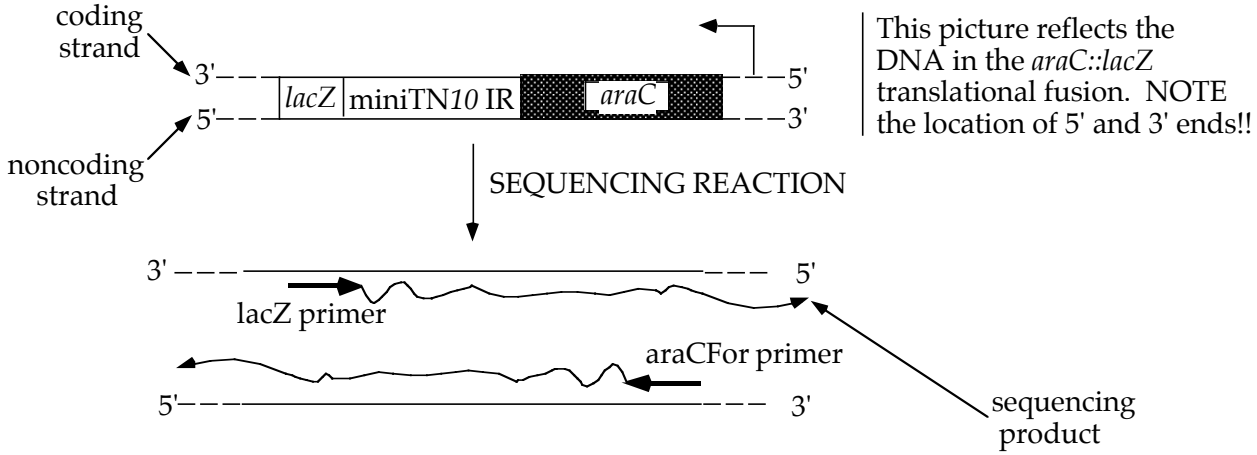
Disclaimer: This answer key was correct based on a BLAST tutorial completed by Kate Bacon Schneider on November 11th, 2004. Because new sequences are added to the BLAST nucleotide/protein databases daily, the numbers associated with a particular query may be slightly different than what is seen here. For example, in Question 2, the *araC* sequence may be between 94-177 instead of 94-184 when you complete the BLAST tutorial.

That said, it is more important to understand HOW to read a BLAST nucleotide or protein output than the numbers themselves.

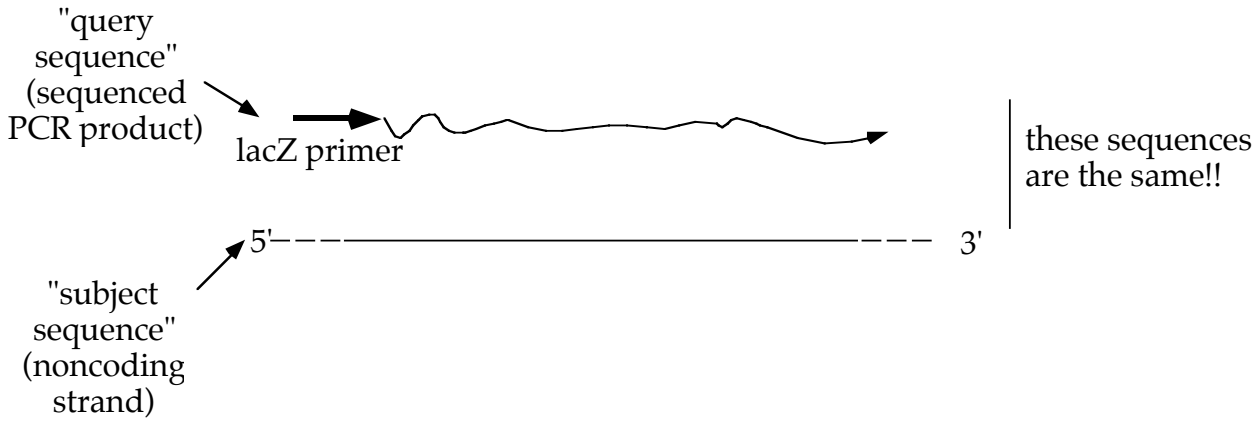
1. The sequence of segment 4 is as follows:
CGAGTATCCCGGCAGCAGGGGATCATTTTG

2. The *araC* sequence is between nucleotides ~94-184 on the diagram, the miniTn*10* inverted repeat is between 22-93, and *lacZ* is between 1-21. Note that the small amount of *lacZ* sequence present on the PCR product was not enough to produce any significant alignments by BLAST, but we can guess where it is based on what an *araC-lacZ* translational fusion looks like.

3. The range of the subject sequence is "237…147." The numbering of the "subject sequence" starts high and goes lower because it is actually the REVERSE COMPLEMENT of the "query sequence" (your sequenced PCR product). To make this more clear, let's look at where the "query sequence" came from:

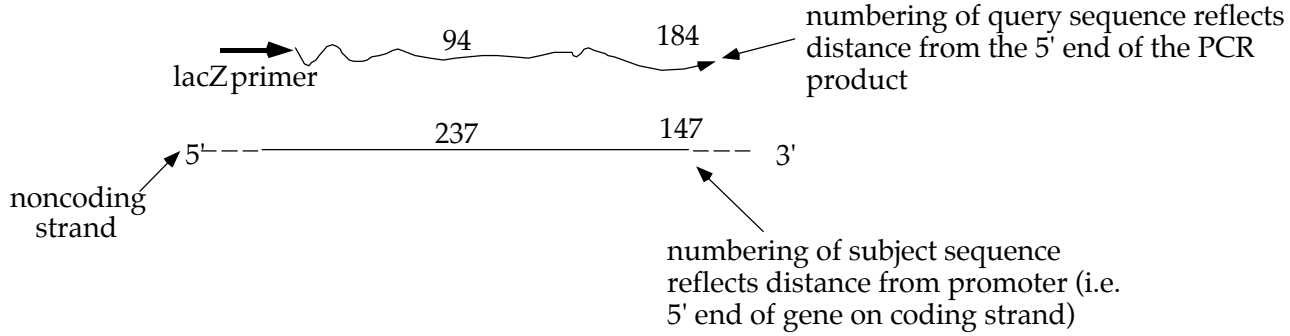


Note that the sequencing product and the NONCODING strand are going to have the same sequence—because both are COMPLEMENTS of the coding strand.



(3. continued)

Now, where do these numbers of the "subject" and "query" sequence come from? The numbering of the "query sequence" is based on the location of the PCR primer (the 5′ end of the PCR product); therefore, nucleotides closer to the primer have lower numbers than those farther away. The subject sequence is numbered from the 5′ end of the gene (i.e., **the end closest to the promoter on the <u>coding strand</u>**). Therefore, this is how the two sequences are numbered (and why you get the BLAST results you observe!!):



4. The accession number of the sequence is: V00259.

5. The sequence is 1172 base pairs (based on what is written next to "source" in the GenBank entry). If you look under ORIGIN on the GenBank entry, you see a sequence of nucleotides numbered from 1 to 1172.

6. The *araC* gene is 879 nucleotides (0.88 kb) in length. In the lab manual, *araC* is described as being 0.9 kb. These lengths correspond.

7. The coding region of the *arabinose C* gene is between nucleotide 165 (nt 165 is the "A" of the "ATG" start codon) and nucleotide 1043 (nt 1043 is the final "A" of the TAA stop codon). Based on the size in nucleotides (879) above, you'd predict that this would make a protein of 293 amino acids. However, remember that the STOP codon doesn't encode an amino acid. So really the protein has 292 amino acids. You can also find this information (# of amino acids) by clicking on the "protein ID" link (CAA23508.1) in the CDS section of the GenBank entry.

Note also that the sequence given in a GenBank entry is that of the **<u>coding strand</u>** of the particular gene (written 5′ to 3′). Thus, if we want to amplify *araC* by PCR, we'd design a forward primer (binds to noncoding strand) with the SAME sequence as that given in the GenBank entry, and a reverse primer (binds to coding strand) whose sequence is the REVERSE COMPLEMENT of that given in the GenBank entry.

8. Your query sequence includes nucleotides 147-237, of which 165-237, or 72 nucleotides, are in the translational fusion. This means that 24 *araC* amino acids are in the translational fusion.

9. The "first four sequences with significant homology" that we are interested are the four "AraC homologs" from different bacterial species (*E. coli, Citrobacter freundii, Salmonella typhimurium, and Erwinia chrysanthemi*). These four sequences have the highest "scores" of all that come up on the protein BLAST search. (By the way, you can find out the species info by clicking on the sequence name link and looking next to "organism" on the resulting window.)

PocR from *Salmonella typhimurium* is a known, non-AraC protein that has significant homology to AraC. You may guess that AraC binds DNA because many of its homologs are known DNA binding proteins and are transcriptional activators. Proteins with similar sequences/structures often have similar functions. Of course, the DNA binding activity of any new protein must be tested experimentally (as all the "known DNA binding proteins" were once tested experimentally). This is because not all proteins that "look like" DNA binding proteins actually bind DNA *in vitro* or *in vivo*!!