Video Tutorial 17.1: Using a 16S rRNA Sequence to Identify a Bacterial Isolate

Students at Haverford College in the United States isolated bacteria from trees on their campus. Anderson, a University of Ibadan student who is studying bacteria that live on tropical plants in Nigeria, is interested in the results the American students obtained. He and Professor Iruka Okeke discuss one bacterial isolate that the American students found.

Anderson: What is this particular strain of bacteria, labeled F10, that is recorded as having yellow colonies?

Iruka: An undergraduate class at Haverford College in Pennsylvania, USA isolated it from a Japanese White Pine Tree. I don't know what it is called yet.

Anderson: How do we determine the species name for this bacterial isolate?

Iruka: As you know, portions of the RNA subunits that are critical for ribosomal structure and function are highly conserved, while other sections are not under strong selection and vary among species. Therefore, the genes encoding variable portions of ribosomal RNA can be used to identify species.

In bacteria, the gene encoding the 16S ribosomal RNA subunit is often used for this purpose. It can be thought of as a kind of bar code to identify a species.

The students performed PCR using primers that anneal to the 16S ribosomal RNA gene of *E. coli* at base positions 27 and 1492. They obtained a 1.5 kb product from this reaction and sent this DNA out to be sequenced from its 5' and 3' ends using the Sanger sequencing technique. Let's look at the chromatograms that resulted from the sequencing.

As you can see, the majority of the run has high and distinct peaks on the chromatogram, indicating that this portion of the sequence is reliable. We can read the DNA sequence from these sections.

Anderson: I can see peaks marked A, C, G and T - each has its own color. Sometimes there are peaks that are ambiguous on the chromatogram, and those would be designated N, which could represent any nucleotide.

Iruka: Those are peaks that are ambiguous on the chromatogram. The letter N could represent any nucleotide while a W is either an A or a T. It is important to assess the quality of the raw data before you begin to analyze it or draw conclusions. As you can see, the beginning of the sequence is unclear, since the peaks of the chromatogram are lower, and sometimes two peaks overlap. Therefore we will only be able to rely on the accuracy of the sequence in the sections where there are tall and distinct peaks.

Several different runs were obtained from the amplified DNA; these sequences overlap each other and we have used computer programs to assemble them into one continuous sequence. We have more confidence in the identity of each nucleotide if we have sequenced each segment more than once.

Anderson: When we have looked at the chromatograms, decided on the reliable sequences and assembled them, how can we use this information?

Iruka: We can compare the assembled sequence to the sequences of other 16S ribosomal RNA genes in the Genbank database. We can use a program known as BLAST, which stands for Basic Local Alignment Search Tool. You can access BLAST from a number of computer servers in different countries. We will use the server of the NCBI - National Center of Biotechnology Information at the US Library of Medicine. As it says on their web page, "The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance." The NCBI website contains more information and tutorials about BLAST, which you should explore.

There are different BLAST algorithms that allow you to query nucleotide or protein databases. Which should we be using here?

Anderson: Well, the sequence we have for the 16S rRNA gene is DNA, a nucleic acid. The gene is never translated into protein so I don't see a reason to query a protein database.

Iruka: Absolutely! We'll be running a nucleotide BLAST, referred to as a BLAST-N search, against a nucleotide database. Let us begin.

The entry page is set up to easily allow us to perform a standard nucleotide BLAST. I am now copying our sequence in plain text or FASTA format and pasting it into the query box. I am going to select the database to query from this list.

Anderson: That would be a database of all known 16S ribosomal RNA sequences from bacteria and archaea.

Iruka: Yes. The default parameters for the program and the algorithm are correct for us, but if we wanted to change them, we could. We should record all the settings we are using in your notebook.

Now we are ready to press the BLAST button and the program will search through all the existing sequences on the NCBI server to find matches to our query sequence.

Here are our results. At the top of the page is the title of our query and the version of BLAST we used: we should record that too. The graphic shown here allows us to quickly see that we have a lot of high-scoring matches and many of them cover most or all of the query sequence.

Anderson: What else does this figure tell us?

Iruka: Not much. We need to look at the summary table and alignments before we can come to any conclusions. The results or "hits" are listed on the left in order of decreasing homology with our query, and the data are on the right. They give us a sense of how good each match is. The matches are actually listed in order of decreasing score.

Anderson: Which information should I pay the most attention to?

Iruka: The E-value or the expect value. This tells us the probability that we obtained this match by chance alone. If the e-value is 0.001, there is a one in a thousand chance that the match we are seeing is due to chance rather than due to homology.

Anderson: That's a low number.

Iruka: Not really. If there are 200 million sequences in the database we are using, an E-value of 0.001 would mean that we could get 200 thousand hits due to chance.

Anderson: Wow.

Iruka: We usually look for E-values that are multidigit negative exponentials – those indicate significant results.

Anderson: Like one times ten raised to the power of -100. [Anderson actually said "raised to the power of 100", but that is not correct.]

Iruka: Exactly, or even smaller. Here, E-values at the top of our list are zero so we know that our matches are due to homology.

The query coverage tells us how much of the sequence is actually analyzed in the match. It is 90% so the values you are seeing here do not include the regions at the very ends of the sequence.

The identities column tells us what proportion of nucleotides in our query, that is, the sequence we put in, and the subject, the match we pulled out, are identical.

Anderson: It's 99%. Does this mean that we can conclude that strain F10 is *Curtobacterium pusillum*?

Iruka: It very well might be but we need to review the evidence first. If we click on our match of interest, we can see the pairwise alignment that generated the summary data in the table. The 99% identity figure was computed from 1465 matching bases out of a total of 1478, which is high. The alignment looks good and I don't see any Ns or any non-nucleotide letters in the sequence that could skew the identities value. I'd say that we probably do have *Curtobacterium pusillum*, although our strain is not 100% identical to the one in the database. If you see about 98% identity or greater, there is a high probability that you are looking at a member of the same species, but if the identity is below 97%, the two sequences would most probably be from different species.

Anderson: What if our match had only been 95%?

Iruka: Over all or most of the 16S gene, and with good quality sequence?

Anderson: Yes.

Iruka: Then I would hypothesize that we might have found a new species, certainly one that is not represented in the database, and that would be exciting.

Anderson: But here, we seem to have *Curtobacterium pusillum*. So Professor, what is this species?

Iruka: Anderson, I am afraid that I haven't the faintest idea. If we click on the Sequence ID, we find out more about this organism.

We can see that the genus *Curtobacterium* belongs to the family *Microbacteriaceae*, which also includes the genus *Clavibacter* – which I've heard of.

Anderson: Yes but you still haven't told me anything about *Curtobacterium pusillum*.

Iruka: Okay. We had better click on the link to the nucleotide database record for *Curtobacterium pusillum* here. But there isn't much other information about this species beyond its taxonomy.

Let's go back and look at the next best match on our results table. It is *Curtobacterium flaccumfaciens*, also with 99% identity over 1461 nucleotides. In addition to information on the sequence and the strain's taxonomy, there is also a link to a Medline-indexed abstract on Pubmed.

Anderson: Look at the title - it is a paper on grass-associated bacteria!

Iruka: Well, since you are interested in plant-associated bacteria, perhaps you could lead a journal club on this paper and we can learn a bit more about the genus *Curtobacterium*. There are several species mentioned in the paper.

Anderson: Great idea - but I have one more question on our results. Strain F10 was 99% identical to *C. pusillum* and *C. flaccumfaciens.* I am pretty certain that these will be very similar to each other…

Iruka: And you could compare their 16S sequences by performing a pairwise analysis using the Align tool in BLAST.

Anderson: But if our strain is 99% identical to two different species, to what species does it actually belong?

Iruka: There are quite a few possibilities here. *Pusillum* and *flaccumfaciens* might actually be a single species that has been named twice by different researchers. Or they might be two distinct species that have very similar 16S genes, although this is uncommon. If so, F10 could belong to either species or even a third species. Additional experimentation comparing the sequence of other genes among *Curtobacterium pusillum* and *Curtobacterium flaccumfaciens* strains, and of course including F10, might shed more light on this. In the interim, we can tentatively name our species by its best match and see what future inquiry unfolds. These comparisons don't always identify the species with certainty, but they do allow you to locate an organism on a particular branch in the tree of life.