[00:00] Welcome to Bioinformatics for Biologists. I'm Stevie Bain, a researcher from the University of Edinburgh.

[00:09] In this video, we are going to run through some of the key activities for Tasks A and B in our workshop, Bioinformatics: The Power of Computers in Biology. It will be useful to have the accompanying handouts available as we progress through this short video. These can be found at 4273pi.org/teacher-resources. To complete these tasks, we will use the NCBI database – specifically the BLAST search tool.

[00:35] This workshop provides an opportunity to gain practical experience in bioinformatics and highlights the link between DNA sequencing and computation. Here, we use a bioinformatics tool – the BLAST search tool - to explore mutations, evolution and nutrition.

[00:52] Let's begin with Task A – the identification of 'mystery' sequence R using the NCBI BLAST tool. First, we must access the BLAST search tool on the NCBI website – you can find the address for this website in the accompanying handout. Once here we can see that there are a number of different BLAST search tools. The BLAST program compares nucleotide or protein sequences to sequence databases. The program we need for this task is BLASTx. This will search for matches to our nucleotide sequence in the NCBI protein database.

[01:26] We click on BLASTx and this takes us to our BLAST search form. It is here that we need to paste in Sequence R. But first, let's retrieve sequence R.

[01:40] Sequence R can be retrieved on the 4273pi website – 4273pi.org – under the heading schools. Look for Bioinformatics: The Power of Computers in Biology and click on the link. We highlight and copy the entire sequence including the defline at the top. We then go back to the BLAST search page and paste sequence R into the large box at the top where it says: "Enter Query Sequence". We leave all other settings as default and click BLAST.

[02:28] This BLAST search should only take a few moments. The important thing is not to refresh the page. You may want to pause the video here and run your own BLAST search for Task A.

[02:40] When the BLAST search is complete, we see a results page that looks like this. In the top section of the page we have some information about the query sequence – in this case, sequence R – and the search that we've just conducted. For the purposes of this exercise, we are most interested in the table lower down the page titled 'Sequences producing significant alignments'. This table contains the proteins in the database that best match sequence R.

[03:10] This table is ordered so that our best result is in the top row. The order is determined by the E-value a statistic that describes the number of hits one can expect to see by chance when searching a database of a particular size. Therefore, the lower this number the more reliable the match. Zero is the most reliable a match can be. Under the description column we find the name of the protein and in square brackets the species that protein comes from.

[03:40] For this activity, we want to find out more about our best BLAST result. So, we can click on the name in the description column and it will give us more information about this match. At the top, we can see the name of the protein, L- gulonolactone oxidase, and the species it comes from, *Mus musculus*. If we want to find out more, we can click on sequence ID.

[04:09] On this page, we get some detailed information about this protein including the accession number which is how the NCBI database is catalogued. This page also has information about the source of the protein – including the common name of the species. If we scroll down to 'source' we see *Mus musculus* also known as the house mouse. There is also a list of scientific literature related to this protein. By looking at the titles here we can begin to understand the biological role of this protein. However, we also recommend doing a web search of the protein name for a more concise description of its function.

[04:50] Now let's go back to the previous page and take a look at the alignment. An alignment is basically a way of arranging the query sequence – in this case sequence R – with a sequence in the database – called the 'subject' - to identify regions of similarity that may be a consequence of structural, functional and evolutionary relationships. Everywhere we see 'Sbjct' that's the sequence from the database. Everywhere we see 'Query' that's Sequence R. In between the query and the subject, we have the consensus sequence.

[05:24] We will take a look at sequence alignments in much more detail in Task B. We now know that Sequence R codes for the mouse GULO gene that codes for the protein L-gulonolactone oxidase - an enzyme involved in vitamin C production. We now want to find out if humans also have a functional copy this gene or a non-functional pseudogene.

[05:45] To do this we go back to the BLAST homepage and look for the BLAST genomes heading. Here we see a text box with a few species' names underneath. We click on human and this takes us to a BLAST search tool similar to the one we used previously, however, this time, instead of BLAST-searching the whole database, this tool will specifically search for sequence R in the human genome database.

[06:15] Again, we paste sequence R into the box at the top where it says enter query sequence and this time, we optimise for somewhat similar sequences in the program selection. This is because we are now doing a nucleotide BLAST – BLAST searching for a nucleotide sequence in a nucleotide database and we need to run a more sensitive search. We then click BLAST and wait for our results. You may want to pause the video here and run this search.

[06:45] Once again, we patiently wait a few minutes for the results page to appear. When it does, we can see that the top of the page tells us some more information about the search we have just conducted. If we scroll down to the table of 'Sequences producing significant alignments", we will see the best matches for sequence R in the human genome. As with the previous search, the best match is in the first row.

[07:15] In the Description column, we can see that our best matching sequence is on *Homo sapiens* chromosome 8. If we look at the e-value, we can see that this is a very low number, 6e-45. This is the same as $6x10^{-45}$. This is a reliable match.

[07:35] Let's click on the name in the description column to view the alignment of sequence R with the matching sequence in the *Homo sapiens* database. As mentioned previously, an alignment is basically a way of arranging the query sequence with a sequence in the database to identify regions of similarity. These alignments allow us to spot mismatches and gaps between the two sequences that correspond to mutations.

[08:00] The accompanying handouts explain in detail how BLAST annotates mutations in alignments. If we look along this first alignment here, we can see that where the base in the query matches the base in the subject, there is a little vertical line between the two. Where they do not match, there is no line – this is an example of a substitution mutation.

[08:30] If we look at this point here, we something a little different. In the subject sequence, instead of a base, we see a little dash or a hyphen. This represents an insertion or a deletion mutation. Here it is an example of a one-base frameshift mutation and therefore we can assume that humans have a 'pseudogene' - a segment of DNA that resembles a functional gene but has mutations that render it not functional.

[08:58] But how do we know if this mutation represents an insertion in the mouse sequence or a deletion in the human sequences? Well, usually from this output alone, we would not be able to say. However, in this case, we do have prior knowledge about sequence R. We know that this codes for a functional gene in the house mouse. Therefore, this must be a deletion in the human sequence.

[09:20] If we take a look at this part of the alignment, we have an insertion/deletion with a length of three bases. This is not an example of a frameshift mutation as three bases make up a codon. Removing three bases does not shift the reading frame of the sequence in the way that removing 1 or 2 bases does.

[09:41] Now that we have completed Task A and Task B, this opens up the opportunity for discussion about the biological role of vitamin C, how humans can get vitamin C and what happens if they do not get enough.