[00:00] Hello and welcome to Bioinformatics for Biologists. I'm Stevie Bain, a researcher from the University of Edinburgh. In this video, we are going to run through how to do the BLAST search required for our workshop, Bioinformatics: Food Detective.

[00:16] It will be useful to have the accompanying handouts available as we progress through this short video. You can find all the web pages you need to access the BLAST search tool and the DNA sequences required in the handout.

[00:30] In this workshop, we use the NCBI database, the NCBI BLAST tool and DNA barcodes to identify species. We also interpret BLAST e-values and the reliability of our search results.

[00:43] For this activity, we DNA sequenced a sausage described as 100% pork to produce a number of DNA barcode sequences. DNA barcodes are regions of DNA that are common to all animals but vary between species. What we aim to do here is identify which species we find in this pork sausage based on these DNA sequences. We are also interested in analysing the reliability of our results.

[01:10] This is a subset of the DNA barcodes we use in this workshop. These can be found at 4273pi.org/schools.

[01:22] Follow the instructions on the handout to access the BLAST homepage. Here you will see that there are a number of different BLAST search tools. These compare nucleotide or protein sequences to sequences in the database and calculate the statistical significance. For this activity, we need the nucleotide BLAST tool. This will search for our nucleotide barcodes in the NCBI nucleotide database.

[01:51] When we click nucleotide BLAST, we are taken to a page that looks like this. Now we have to access our DNA barcode sequences and paste them into this large box at the top.

[02:06] As mentioned previously, we access these sequences at 4273pi.org/schools. You will find the sequences under the National 4/5 Biology workshop. When we click here you will see a page that looks like this. This has all the sequences we need for both tasks in this workshop. We simply highlight the first sequence and copy.

[02:35] Now we need to paste our sequence into the box at the top and scroll down to where it says program selection. Here we want to choose the option blastn as we are running a nucleotide blast. This ensures a more sensitive search of the database. We then click BLAST and our search will begin to run. This may take a few moments.

[03:05] Our results page looks like this. At the top, we have some information about the BLAST search we have just run. If we scroll down, we see a table titled sequences producing significant alignments. Here, all of our BLAST search results are listed. Under description, we find the names of the sequences in the database that best match sequence A. Importantly, this includes the species name. One thing to note is that BLAST uses the scientific or Latin names of species, not the common names.

[03:45] The table is ordered so that our best BLAST result is in the first row. This order is specified by 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. We will discuss this more in a moment. Although we don't directly ask for these other values in the worksheet: query coverage and percentage identity are also important to consider when doing a BLAST search.

[04:15] Query coverage is the percentage of our sequence –sequence A – that aligns to the sequence in the database. Percent Identity relates only to aligned regions and describes how similar the query sequence is to the sequence from the database i.e. how many bases in each sequence are identical?

[04:35] For this activity, we are most interested in the e-value as it allows us to determine how reliable our search results are. As we progress through the workshop, we compare the e-values we find in Task 1 to those we find in Task 2 in order to identify which set of results is most reliable. The e-value is how many times we would expect to see a match of this quality, between our sequence and the sequence in the database, by chance.

[05:05] If our e-value is high, we consider the match to be unreliable. If the e-value is low, we consider it reliable. If our e-value is 0, the lowest an e-value can be, that means the match is extremely reliable. There are no definite cut-offs for e-value, therefore in this activity we simply compare the e-values in one table to those in another.

[05:30] The BLAST output provides e-values in a way that you may not be familiar with, for example, 5.2e-15. This is simply a way of representing numbers with lots of digits without taking up too much space.

[05:52] You might like to pause this video and work through Task One of the work sheet now.

[05:58] Let's take a look at our first table of results: sequences A-H. Please note that the results you obtain may not be identical to ours, due to continual growth of the DNA database. Since these sausages are 100% pork, we may expect to only find pig DNA in our sample, however this is not the case as we find chicken, sheep, cattle and perhaps more surprisingly human DNA.

[06:25] If we take a look at the e-values, we can see that they are very low numbers. So, for example, in the top row 4e-49 is the same as 4 times 10 to the power of minus 49. These low numbers suggest that the results in this table are reliable.

[06:48] Indeed, there are good explanations for how DNA from animals other than pig may have found its way into our samples.

[06:58] The questions posed throughout this workshop should prompt discussion about food fraud and DNA contamination, but also the DNA extraction and sequencing processes. The other animal DNA most likely came from DNA cross-contamination from the butcher shop. The human DNA most likely came from the sausage making process or the lab.

[07:19] Even if benches and tools are thoroughly cleaned, traces of DNA can still linger behind and these will be picked up when the DNA is sequenced. So, although the presence of human DNA may be alarming at first, it is most likely due to human interaction with the sausage or the DNA sample. These are examples of DNA cross-contamination, not food fraud.

[07:46] You may now like to pause the video here to complete Task 2.

[07:50] In the second task, we find a more unexpected set of results. For example, frog, grapevine and bacteria. However, the e-values in this table are rather high. When we compare this table to the first table, we can reinforce the concept of using BLAST e-values as an indicator of result reliability. We said before that the lower the e-value the more reliable the match. We find that the results in Task 1 have lower e-values than those in Task 2 by many orders of magnitude. Hence, the results in Task 1 are more reliable. This makes sense when we look at the species we find in each table.

[08:31] We hope you found this video on Bioinformatics: Food Detective useful. For more information and more free resources, visit our website 4273pi.org and follow us on Twitter @4273pi.