**Worksheet 2 - Probing Yeti artifacts**

**Learning Objectives**

*In this section, students will use the provided DNA sequence to*

* explain the general process and purpose of performing BLAST analysis of sequences
* describe and be able to create a FASTA format
* perform a BLASTn search and interpret the results
* identify the source of the sample DNA
* determine if any of the sequences might be from an unknown species

Using the instructions below, complete the questions and fill out the tables in this document. Save your document.

**Expedition**

Having learned more about how DNA sequences can be used as DNA barcodes, you finish packing for your research excursion, making certain to pack the warm, heavy clothes that your mother insists upon. Upon arriving with Dr. Shipton and the rest of your team at the Tribhuvan International Airport in Katmandu, Nepal, your team takes the airport shuttle with all of your research gear to a local hotel to acclimate and recover from your travel experiences. At the hotel, Dr. Shipton informs you that you are awaiting the final permit from The Right Honourable Bidhya Devi Bhandari, who is the current president of Nepal.

Dr. Shipton anticipates the signed document will be delivered within the next several days. She also states that the permit only allows for sampling of small amounts of materials that are made available by collectors and any samples that might be obtained from the exploration of Makalu-Baron National Park (Figure 1). Any collected materials, after being tested, belong to the government of Nepal and will be turned over to the Sagarmantha National Park Museum for storage and possible display.



Figure 1. Welcome to Makalu-Baron National Park.

Three days after arriving in Katmandu, Dr. Shipton schedules a lunch meeting for you and the remainder of the research team. Dr. Shipton begins by explaining that the signed permit has been received and that the group will travel from Katmandu in the morning to set up the research laboratory in Chaurkharka, just outside of the Makalu-Baron National Park. Dr. Shipton also introduces Jamling Norgay, an experienced sherpa. Mr. Norgay is the son of Tenzing Norgay, one of the sherpas that accompanied Sir Edmund Hillary on the first summit of Mount Everest in 1953. Mr. Norgay places a box on the table and proceeds to explain that the contents of the box are on loan from several collectors and will be used for sampling when the research laboratory has been established in Chaurkharka.

**Setting up the laboratory**

After a long and bumpy ride to Chaurkharka, you arrive at the research field station that will serve as your laboratory for the next several weeks. While Dr. Shipton prepares some members of the research group for the first field expedition, you are instructed to set up the laboratory and to begin to collect DNA. New sequencing technologies allow you to set up a USB-style sequencer on your laptop, so you really do not need too much setup time. You proceed to start collecting small samples from the artifacts Mr. Norgay has provided, taking precautions to not damage any of the artifacts and to keep your samples as clean as possible. Several days after the research expedition group has left, you begin to collect your first sequence results. You have run a short amplification using part of the 12S rDNA sequence. The first six sequences that you collect are shown in Figure 1. You set about trying to identify the sequences that you obtained from the artifacts.

**BLAST analysis**

The sequences can be individually entered into the query box or as a group of sequences. One common method of presenting multiple sequences is in a file format called FASTA. In this format, the title of the sequence file is entered with a > before the name (see Figure 2). On a return line, the sequence is entered. Multiple sequences can be entered on their own lines in this format. Many bioinformatics programs (in addition to BLAST) can utilize the FASTA format. The sequence format provided in Figure 2 is in the FASTA format.

>Yeti1

CTTAGCCTTAAACATAAATAATTTATTAAACAAAATTATTCGCCAGAGAACTACTAGCAACAGCTTAAAACTCAAAGGACTTGGCGGTGCTTTAAACCCTCCTA

>Yeti2

CTTAGCCTTAAACATAAATAATTTATTAAACAAAATTATTCGCCAGAGAACTACTAGCAACAGCTTAAAACTCAAAGGACTTGGCGGTGCTTTAAACCCTCCTA

>Yeti3

CTTAGCCCTAAACATAGATAATTTTACAACAAAATAATTCGCCAGAGGACTACTAGCAATAGCTTAAAACTCAAAGGACTTGGCGGTGCTTTATATCCCTCTA

>Yeti4

CTTAGCCCTAAACATAAATAATTGTAAAAACAAAATTATTCGCCAGAGTACTACCGGCAACAGCCCAAAACTCAAAGGACTTGGCGGTGCTTTATATCCATCTA

>Yeti5

CTTAGCCCTAAACACAGATAATTACATAAACAAAATTATTCGCCAGAGTACTACTAGCAACAGCTTAAAACTCAAAGGACTTGGCGGTGCTTTATATCCTTCTA

>Yeti6

CTTAGCCCTAAACTCTAATAGTTACATTAACAAAACCATTCGCCAGAGTACTACAAGCAACAGCTTAAAACTCAAAGGACTTGGCAGTGCTTTATATCCCTCTA

Figure 2. Sequences from presumptive Yeti artifacts in FASTA format. *For a description of the FASTA format, please refer to the Box below*.

Dr. Shipton was nice enough to provide you with more instructions on how to identify the source of the artifacts using the sequence information that you collect. Hopefully, you will identify some sequence that cannot be identified.

In this part, you will be looking at the sample sequences to identify the source of the DNA (if possible) while doing a BLAST comparison. While samples may be damaged and sequencing does have errors, assume that the sequences that you have to analyze have been reproduced in the lab and are 100% accurate within the sample.

1. Using the FASTA sequence provided (Figure 2), you will perform a BLAST search to determine the genus (and possibly species) of the origin of the sample.
   1. Open browser window to https://blast.ncbi.nlm.nih.gov.

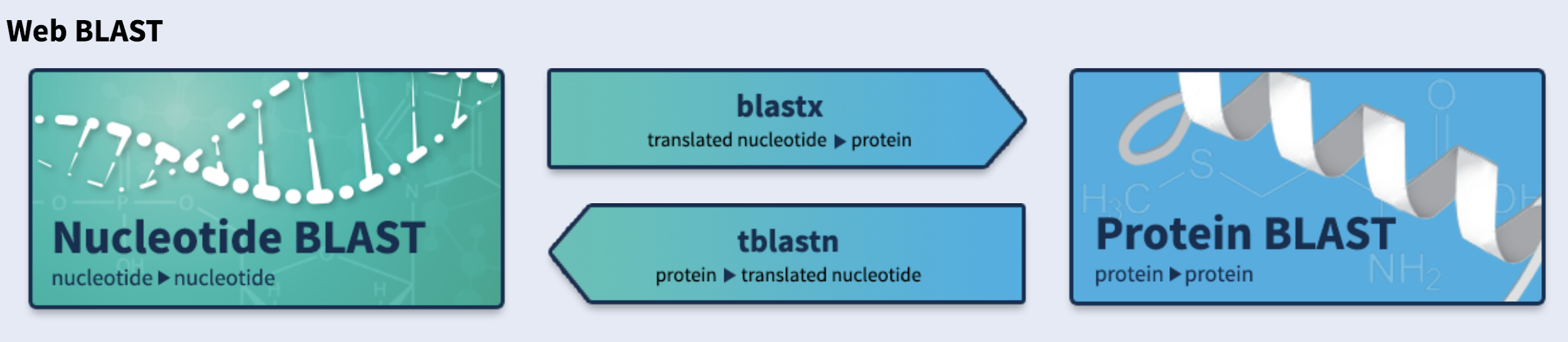


Figure 3. BLAST window.

* 1. In the window, click on Nucleotide BLAST on the left hand side of the window. *The other ‘buttons’ are alternative versions of the BLAST program that either use amino acid sequences for the search or search the database in different manners. You just need the basic BLAST program that compares nucleotides to the nucleotide database. There are additional parameters that can be used for the search, but for this assignment all you need to do is use the default options.*
  2. Copy and paste your FASTA file (from Figure 3) into the white box.

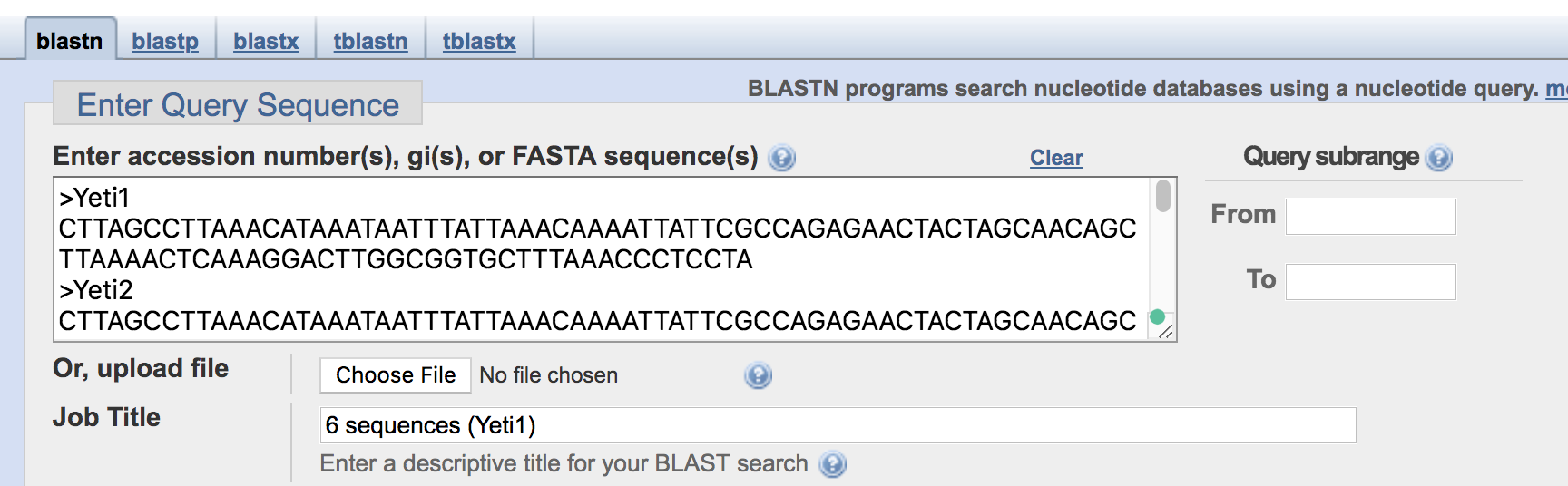


Figure 4. BLASTn window with Yeti sequences.

* 1. Scroll down and click on the BLAST button (blue).

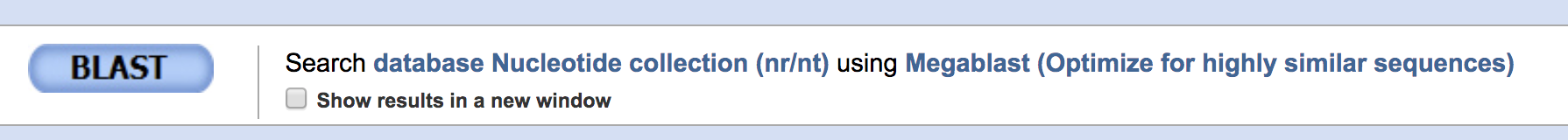


Figure 5. BLAST search button

* 1. The computer algorithm will run for a short period of time and provide a new screen with the information needed. Near the top of the screen, there is a ‘Results for’ window with the identity of each query sequence entered. You will need to select different query sequences to complete the assignment. Each time you change the query in the search, it may take a short time to upload the results of the selected sample.
  2. The next section is titled “sequences producing significant alignments”. This has some of the information. The computer program will (by default) provide up to 100 matches (best matches and newest matches appear at top of list).

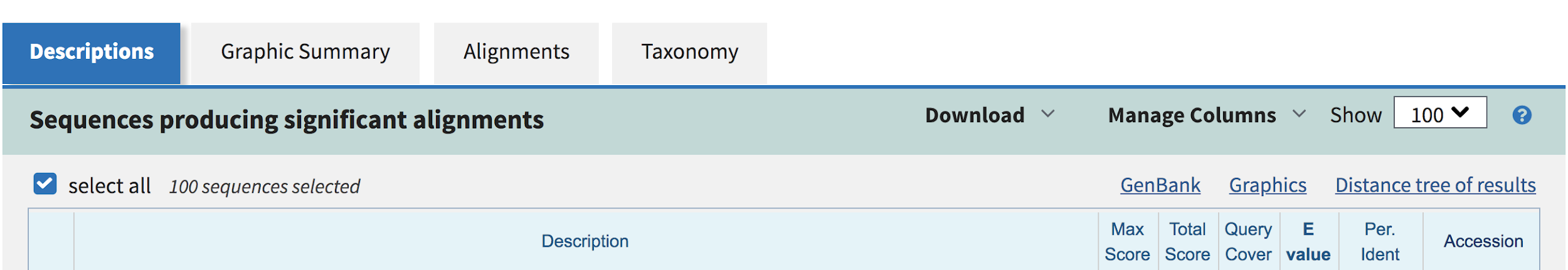


Figure 6. Descriptions tab in the resulting BLASTn search. The general description of the GenBank file will be provided in the section below this header on the web page.

**Box: GenBank and BLAST information**

FASTA – this sequence format (applicable to both nucleic acid and amino acid sequences) includes an identifier line that begins with a caret (>) followed by a description. Spaces can be used, but nothing after the space will be carried over to subsequent formatting. After the description and a line return, the sequence is added (without numbers or spaces) onto subsequent lines. Multiple sequences can be placed in one file (like in Figure 2) by starting a new entry on a new line with a description (starting with >).

GenBank - a database of nucleic acid and amino acid sequences. This database contains millions of sequences and is updated frequently.

BLAST – stands for basic local alignment search tool. This is a program that runs a computer algorithm to search a sequence of interest (query) against the database.

Query – nucleic acid or amino acid sequence used in the BLAST algorithmic search (e.g., your sequence file)

Subject – database file identified in the search.

Description – this is a brief description of the databank file that has been identified. This is a hyperlink to the alignment.

Alignment – this is the result of the alignment between the query sequence and the databank subject sequence.

E (expect) value – this is a calculated score of the quality of the alignment. The value is based on the length of the query sequence, the size of the database, and the degree of the alignment.

Accession – this is a unique identifier for the GenBank file. The hyperlink will allow you to go to the specific file and access the information (which is more than just the sequence) in the file.

Identities – in the alignment, this is the number of identical nucleic acids (or amino acids if you run a different BLAST program) between the query and subject sequences. The number is represented as both a number out of the entire alignment (e.g., 98/100) and a percent (e.g., 98%). This also provides you with the overall length of the alignment which you can also get from the alignment file below.

Gaps –this represents the number of gaps in the sequence alignment that allow for optimal alignment results.

* 1. Using the results of the BLAST search, fill out the table below. To get the Identities information, you can either click on the Description (link) or scroll down to the first sample. (After you get the information for Yeti1, go to the top of the page, to the “Results for” and select the next query sequence. Repeat to complete the table.). Use the first identified entry (ignore synthetic or unidentified entries) .

**Question 1.** Using the BLAST search and NCBI information, prepare a table of the results for Dr. Shipton.

|  |  |  |  |
| --- | --- | --- | --- |
| Sample | Accession | Description (record just genus and species here) | Identities (record as #/#) |
| Yeti1 |  |  |  |
| Yeti2 |  |  |  |
| Yeti3 |  |  |  |
| Yeti4 |  |  |  |
| Yeti5 |  |  |  |
| Yeti6 |  |  |  |

|  |  |
| --- | --- |
| **Question 2:** Are any of the samples of unknown description (origin)? |  |
| **Question 3:** Do any of the samples match the same genus/species description? |  |
| **Question 4:** What is the most likely conclusion that you can make from the artifacts provided by Mr. Norgay? |  |

**Question 5:** What are the common names of the origin of the samples? *Hint: Use Google.com if you don’t know the common name.*

|  |  |  |
| --- | --- | --- |
| **Sample** | **Genus/species (copy and paste from table above)** | **Common name** |
| Yeti1 |  |  |
| Yeti2 |  |  |
| Yeti3 |  |  |
| Yeti4 |  |  |
| Yeti5 |  |  |
| Yeti6 |  |  |

In this section, you will be using BLAST to perform a pairwise comparison of the sequences to determine if any of the sequences are identical. A pairwise comparison looks at two sequences aligned against each other. In identifying the samples in the previous section, you performed pairwise analysis, but you were looking to identify the sample, not compare samples. Pairwise analysis is useful when you are looking for sequence differences between samples. In the next section, you will perform a multiple sequence alignment which will allow you to look for commonalities and differences amongst the samples. This information can provide indications of where the constant sequence regions are (that may be useful in amplification of the region) and where the variable sequences are (that may help identify differences in the species sequences.

1. Now we can see how different these sequences are from each other. There are two different manners in which this can be done. One way is to compare each sequence to the others. We can do this using the BLAST program.
   1. At the top of the BLAST page that you are on, select ‘Edit search’ near the top left side. This will bring you back to the start window.

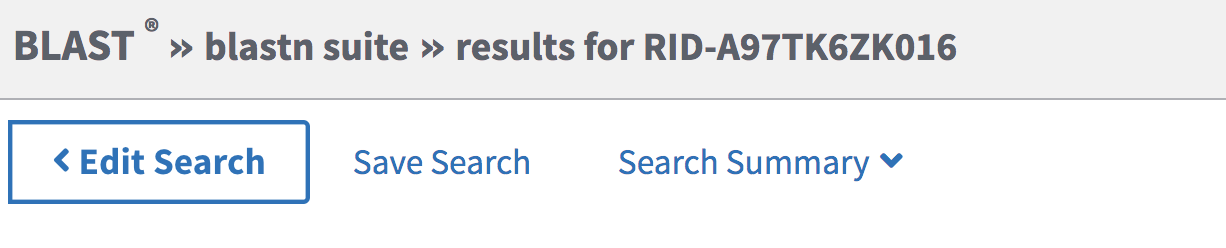


Figure 7. Near the top of the page will be the Edit Search option as shown.

* 1. Below the window where you entered your FASTA file is a small box next to “Align two or more sequences”. Click on that button (a checkmark should appear). Paste your FASTA file into the second window (make sure that it is still in the first window).

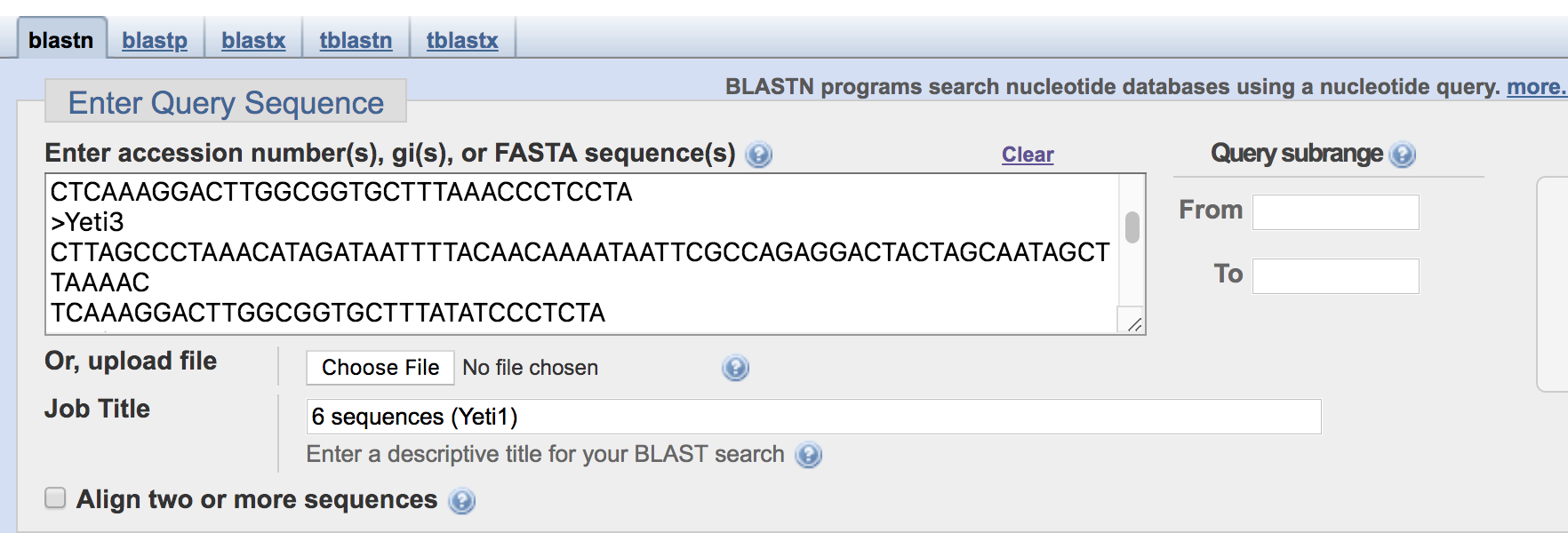


Figure 8. Blastn search window showing the location of the “Align two or more sequences”.

* 1. Run BLAST again. This time, you are comparing your query files against themselves.
  2. When the program has run, make note of which sequence is in the “Results for” box at the top of the page.

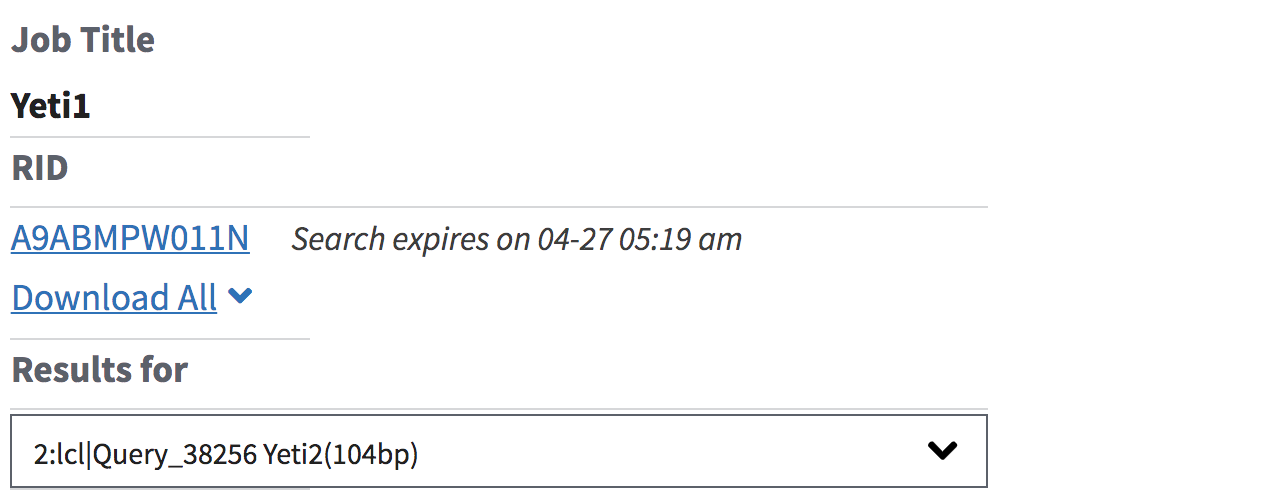


Figure 9. Resulting BLASTn search window comparing Yeti1 sequence with Query Yeti2 sequence.

* + 1. i. Scroll down to the alignments and find the other sequences.
    2. Record your ‘identities’ for each OTHER sample in the gray boxes of the table below.
    3. The table below has been partially filled out since Yeti1 will be a 100% match to Yeti1 (and so on).
    4. Some of the matches may not cover the entire sequence entered, but you can ignore this for now.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Yeti1 | Yeti2 | Yeti3 | Yeti4 | Yeti5 | Yeti6 |
| Yeti1 | 100% |  |  |  |  |  |
| Yeti2 |  | 100% |  |  |  |  |
| Yeti3 |  |  | 100% |  |  |  |
| Yeti4 |  |  |  | 100% |  |  |
| Yeti5 |  |  |  |  | 100% |  |
| Yeti6 |  |  |  |  |  | 100% |

**Question 6:** In some cases, there may not be a match provided. Why do you think this might be?

**Question 7:** Based on the results of the table above and your knowledge of the animal origin of the samples, explain why the percent similarity of the results makes sense (or not). Is there anything unexpected in the results?