**BIOL 102 –Bioinformatics**

EXERCISE 2

**Sequence Alignment to a Database of Sequences**

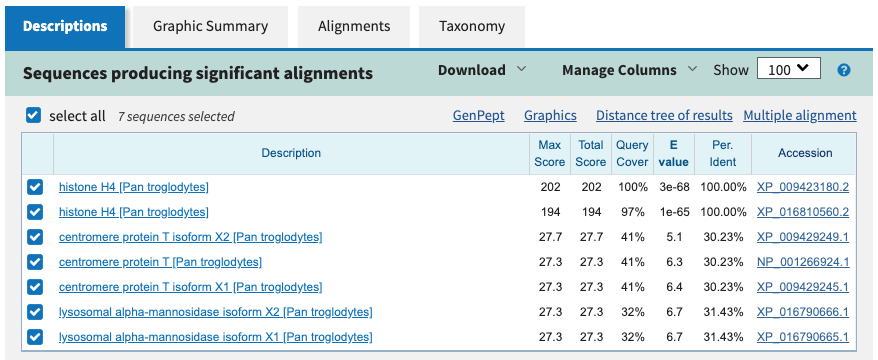
**Objectives**

After completing this exercise, you should be able to:

1. Explain how similarity is used to perform a BLAST search.
2. Explain the BLAST search algorithm.
3. Evaluate the results of a BLAST search.

In the examples used in Exercise 1, the query and subject sequences were of the same length and had very few substitutions, making a direct comparison of the sequences easily accomplished. An alignment approach that attempts to align all residues between two sequences is termed a **global alignment**. In reality, when a newly identified amino acid sequence is used to query a database of amino acid sequences, there will likely be considerable differences in the sequence length and/or in the number of amino acid substitutions, unless the protein is highly conserved. **Local alignments** can find sequence similarity between divergent sequences of different lengths, often using a subset of the query sequence. Thus, to find known sequences that are similar to the query sequence, the query sequence must be aligned with all possible sequences and similarity scores calculated.

Aligning a sequence against a database also allows the user to infer homology between the query and the search output subject sequences when considering statistical metrics associated with alignments. This alignment of amino acid or nucleotide sequences is based on pattern matching and is often carried out using a local alignment tool called **B**asic **L**ocal **A**lignment **S**earch **T**ool (BLAST). The **Expect value (E value)**, shown in Figure 1, is a statistic that **represents the number of times that one can expect the alignment in question to arise randomly between the Query and Subjects within the database**. Very low or zero E-values indicate the match is significant.



**Figure 1**: A BLAST search will provide an E value that indicates how often the alignment would be expected to occur by chance.

It is important to note that **query length and database size will influence E values**. Shorter query sequences and larger databases will make it more likely that the Query will randomly align with a database sequence. Alignments with relatively small E values are more likely to be significant and biologically interesting. It is important to note that similarity between sequences does not imply homology, but **similarity is an expected consequence of homology**. Thus, if a query sequence returns a long list of small E value hits that correspond to a described gene in several different species, it is likely that you have identified a homologous sequence. Example of homologous sequences are shown in **Figure 2**.

**Figure 2**: Homologous protein sequences in different species. Credit: Conserved Sequence; Wikipedia



In the local alignment process, a sequence is “cut” into short segments (**query words**) that can be used to locate a match(s) within the database. BLAST takes this approach, rather than aligning the full-length query sequence, to reduce the amount of computational time needed to return database hits and avoid searching for possible alignments that are unlikely to be biologically relevant. The increase in speed using this strategy comes with the tradeoff of being less accurate than other alignment algorithms, but the results are still quite robust. Once a match to this query word is found, further matching between the query and target sequences is determined (Altschul et al., 1990). BLAST relies on a user-defined scoring threshold when choosing query words and extending alignments. In this activity, we will model a simplified version of BLAST using a single query word followed by the construction of alignments started by an exact query word match that extends in both directions until a negative substitution value is aligned on both sides (**Figure 3**).

Query word: P E G

Query sequence: R P P E G L F

Database sequence: D P P E G V V

Scoring: -2 7 7 5 6 1 -1

Accumulated score = 7+7+5+6+1 = 26

**Figure 3**. An exact query word match alignment and extension followed by alignment truncation prior to negative substitution values using a simplified BLAST algorithm.

To illustrate BLAST, we will use the following amino acid sequence as our Query, with the highlighted (boxed) triplet (MPY) as the **query word**:

STWGERGLMPYRGLACEGHI

Let’s assume that a search of the database revealed four protein sequences with possible similarity. Using the instructor provided **BLAST Alignment Template (on the worksheet)**, align the query word with each of the protein sequences, and extend the match in both directions. *Note: You must find the entire query word*. Then calculate a similarity score using the BLOSUM-62 matrix (**on the worksheet**).

To calculate the similarity score, add the similarity matrix values for the query word and each continuously aligned amino acid in both directions until a negative value is encountered. Continue adding the matrix values until you encounter a negative score on each side, which will terminate the local alignment on both sides. *(Hint: examine the protein sequences carefully. There might be more than one Query word.)*

**Question 1:**

Which of the proteins from the database is most similar to our Query? Which is the least similar? **(Answer this question on your worksheet)** *Remember –* ***add number values in each direction until you find a negative numbe****r – then stop.* ***Don’t include the negative number in the Total Score****.*

**Question 2:**

What problems did you encounter that may have affected your calculated scores? **(Answer this question on your worksheet)**

**Protein BLAST**

Now that you have been introduced to the process of aligning sequences and scoring their similarity let’s use BLAST to locate and compare two protein sequences. In this example, we will be comparing the sequence of the human and chimpanzee histone protein, H4.

Histones are small, basic proteins that are used by all eukaryotes to package their DNA within the nucleus. The histone proteins also play a role in regulating gene expression through their modification and the subsequent effect on the accessibility of the DNA to polymerases. Due to their critical role, histone proteins are highly conserved in their amino acid sequence, structure, and function. Humans and chimpanzees last shared a common ancestor between 5 and 8 million years ago.

**Question 3:**

How similar do you expect their H4 proteins to be? Briefly explain the reason for your answer. **(Answer this question on your worksheet)**

**Computational Procedure:**

1. Go to the NCBI home page at <https://www.ncbi.nlm.nih.gov/>
   1. or Google search NCBI – National Center for Biotechnology Information
   2. This web site provides access to DNA and protein databases as well as BLAST.
2. At the top of the page, enter HIST4H4 (the abbreviation for Histone H4) to search “**All Databases**” for this protein. Click “Search.”
3. Under the “**Proteins**” heading, click on “**Protein**” – hits from your search within a protein sequence database.
4. One of the first few results will be for the **HIST4H4 protein of *Homo sapiens (Accession: AAI11094.1 GI: 109730275)***.
   1. Click on this link to take you to a page that will provide details about the 103-amino acid protein, including its sequence.
5. At the top of this page, click on the “FASTA” link.
   1. This link will display the amino acid sequence in a simple format known as FASTA.
   2. This standard format, used for both genes and proteins, begins with a > symbol, followed by any text.
   3. The text is sometimes longer than one line. This difference is OK, so long as there is no “return” in the text. After this one line of text, there is a return, followed by the nucleotide sequence of the gene or the amino acid sequence of the protein.
6. In the upper right-hand corner of the FASTA page, click on “Run BLAST” under the “Analyze this sequence” heading. This link will take you to the BLAST home page.
7. On this page, under “Database” select “Reference proteins (refseq protein).”
8. Then under “Organism,” type “*Pan troglodytes*” (the scientific name for the chimpanzee).
   1. A list of related names will pop up. Select “Pan troglodytes (taxid:9598) to get the correct results.
9. Click the “BLAST” button at the bottom of the page.
   1. The BLAST algorithm looks for similarity by comparing the human Histone H4 sequence to all annotated chimpanzee protein sequences.
10. When the search process is completed, scroll to the middle of the page to the table.
    1. The Descriptions tab provides the scores for each alignment.
    2. Alignments with 100% Query Cover, a low E value, and 100.00% Identity will have the same sequence.
11. The first alignment has a score of 202 and an identity of 100%, which means that the human and chimpanzee H4 histone proteins have the same amino acid sequence.
12. Click on the Graphic Summary tab. This page will show you a graphical comparison of the query and subject sequences.
    1. Hover over a line and click to see the title of the alignment, scores, and a link to the alignment.
13. Click on the Alignments tab to view the amino acid alignments of the Query with each subject sequence.

**Question 4:**

If we were to compare the *nucleotide* sequences for the gene encoding this protein between humans and chimps, do you think they would be identical? Explain. **(Answer this question on your worksheet)**

**Nucleotide BLAST**

A comparison of nucleotide sequences can also be made using BLAST. In this exercise, we will compare the human and chimpanzee H4 gene sequences.

**Computational Procedure:**

1. Return to the NCBI home page <https://www.ncbi.nlm.nih.gov/>
   1. or Google search NCBI – National Center for Biotechnology Information
2. At the top of the page, enter HIST4H4 (the abbreviation for Histone H4) to search all databases. Click “Search.”
3. Under the “Genomes” heading, click on “Nucleotide” – hits from your search within a nucleotide sequence database.
4. Near the top of the page, click on “HIST4H4 (Homo sapiens histone cluster 4, H4, mRNA) (Accession: BC111093.1 GI: 109730274).”
   1. A new screen will appear that gives you a lot of information about this 354 bp transcript sequence.
   2. We are interested in obtaining the nucleotide sequence in a format that can be compared to that of the chimpanzee.
5. At the top of this page, click on the “FASTA” link. This link will display the nucleotide sequence in the simple FASTA format.
6. In the upper right-hand corner of the page, click on “Run BLAST” under the “Analyze this sequence” heading. This will take you to the BLAST home page.
7. On this page, under “Database” confirm that “Nucleotide collection (nr/nt)” is selected.
8. Under “Organism,” type “*Pan troglodytes*” into the text box. In the pop up menu, choose: Pan troglodytes (taxid:9598)
   1. Defining the specific organism will allow you to compare the human histone H4 gene only with the chimpanzee genome and not with the entire NCBI database.
   2. This action saves time because histones have been sequenced for many organisms. Other BLAST searches could allow you to compare your sequence with the entire NCBI database or with different subsets of it.
9. Scroll to the bottom of the page and click on “BLAST.”
10. When the search process is completed, scroll to the middle of the page to the table.
    1. The Descriptions tab provides the scores for each alignment.
11. Click on the tab for “Graphic Summary.”
12. On the resulting page, scroll down and look at the Graphic Summary. In the box, you will see bright red lines indicating the top matches. The bright red color of the lines indicates that these are very close matches. Mouse over the lines to see the name of the chimpanzee gene that your Query (human histone H4) matches.
13. Return to the “Descriptions” tab and click on the **Accession** links on the far right side of the table for one of the matches.
    1. Scroll down to the information under “FEATURES” and look for the subheading “source.”
    2. Notice that the match is on Chromosome 6 of the chimpanzee.
14. Return to the table containing tabs for Descriptions, Graphic Summary, etc. Click on the “Alignments” tab.
    1. The “Query” is human histone H4. The “Subject” is chimpanzee histone H4.
    2. Notice that whenever the nucleotides of the genes are identical, there is a vertical line between the identical base pairs.
15. Take a **screenshot** of the alignment between your Query (human histone) and the Subject (chimpanzee) titled *“Pan troglodytes histone H4 (LOC115945179), mRNA”*. This alignment should be the first on in the list.
16. Paste the alignment screenshot into the worksheet under the heading **HIST4H4 Nucleotide Alignment screenshot**

**Question 5:**

How many differences did you find between the Query and Subject? What percent identity is there between these sequences? Is this at all surprising to you? Explain. **(Answer this question on your worksheet).**

1. Return to the NCBI homepage (<https://www.ncbi.nlm.nih.gov/>) and search for the human histone record by typing “HIST4H4 homo sapiens” into the search box. Click “Search.”
2. Find the record that appears at the top of the search results page: Gene: H4-16 – H4 histone 16
   1. H4-16 – H4 histone 16 is another name for the HIST4H4 protein in humans
3. Click on the link to open this record. Scan through the page.
   1. Identify the chromosome on which the human HIST4H4 protein is located.

**Question 6:**

On which chromosome is the gene for human histone H4 located? **(Answer this question on your worksheet.)**

**Question 7:**

Imagine that a colleague asks you to align a conserved metabolic protein-coding gene sequence from a dog to its human homolog. Which sequence type - DNA or protein - would you expect to exhibit the highest percentage of identities? Why? **(Answer this question on your worksheet.)**