**Bioinformatics:**

**Investigating Sequence Similarity**

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 length, 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. The Expect value (E value) is a statistic that represents the number of times that one can expect the alignment in question to randomly arise between the query and subjects within the database. 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.

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). In this 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, opposed to 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 imparted by 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 through the use of 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 2**).



**Figure 2**. An exact query word match alignment and extension followed by alignment truncation prior to negative substitution values using a simplified BLAST algorithm. Credit: *Adapted from DISP, Public Domain, via Wikimedia Commons.* [*https://upload.wikimedia.org/wikipedia/commons/8/87/Extension\_process.jpg*](https://upload.wikimedia.org/wikipedia/commons/8/87/Extension_process.jpg)

To illustrate BLAST, we will use the following amino acid sequence as our query, with the highlighted (boxed) triplet 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. 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, which will terminate the local alignment on both sides.

Which of the proteins from the database is most similar to our query? Which is the least similar? **(Worksheet question 1)**

What problems did you encounter that may have affected your calculated scores? **(Worksheet question 2)**

**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 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.

How similar do you expect their H4 proteins to be? Briefly explain the reason for your answer. **(Worksheet question 3)**

**Computational Procedure:**

1. Go to the NCBI home page at ncbi.nlm.nih.gov. 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*. 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. This will display the amino acid sequence in a simple format. FASTA is a standard format, used for both genes and proteins, in which the first line begins with a > symbol, followed by any text. The text is sometimes longer than one line. This 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 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” select “RefSeq protein”. Then under “organism” type “*Pan troglodytes*” (the scientific name for the chimpanzee).
8. Click the “BLAST” button at the bottom of the page. The BLAST algorithm is currently looking for similarity by comparing the human Histone H4 sequence to all annotated chimpanzee protein sequences.
9. When the search process is completed, a Graphic Summary box will appear towards the top of the page. This will show you a graphical comparison of the query and subject sequences and as you scroll down you will see the amino acid alignments.
10. The first subject has a score of 200 and an identity of 100% which means that the human and chimpanzee H4 histone protein has the same amino acid sequence.

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. **(Worksheet question 4)**

**Nucleotide BLAST**

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

**Computational Procedure:**

1. Return to the NCBI home page (ncbi.nlm.nih.gov).
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).” A new screen will appear that gives you a lot of information about this 412 bp transcript sequence. At this point we are only 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 will display the nucleotide sequence in a simple 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. Under “Organism” near the middle of the page, type “*Pan troglodytes*” into the text box. This will allow you to compare the human histone H4 gene with the chimpanzee genome, and not with the entire NCBI database. This saves time, because histones have been sequenced for innumerable organisms. Other BLAST searches could allow you to compare your sequence with the entire NCBI database or with different subsets of it.
8. Scroll to the bottom of the page and click on “BLAST.”
9. On the resulting page, scroll down and look at the Graphic Summary. At the bottom of the box, you will see seven bright red lines of two classes indicating that there are two basic matches (one of the matches has many transcript isoforms transcribed from the same genetic locus). The fact that the lines are bright red indicates that these are very close matches. Mouse over the lines, and it will tell you, in technical terms, which chimpanzee gene your query (human histone H4) matches. Click on the accession # link, followed by navigating to the information under “source,” and a select the subheading under “Features.” Notice that one of the two classes of matches is on the number 12 chromosome of the chimpanzee, while the other is on chromosome 6.

On what chromosome is the gene for human histone H4 located?

1. Scroll down to “Alignments.” The “Query” is human histone H4. The “Subject” is chimpanzee histone H4. Notice that whenever the genes are identical, there is a vertical line between the identical base pairs.
2. Now take a **screen shot** of one of each of the two variants of subject sequence aligned to your query. Use Microsoft Snipping Tool found by clicking on the Microsoft icon on the desktop. This can be found by searching through the Microsoft start menu
3. Paste each of the two alignment screen shots into the worksheet.

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. **(Worksheet question 5)**

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? **(Worksheet question 6)**