**Bioinformatics:**

**Investigating Sequence Similarity**

**Overview**

**Bioinformatics** involves the use of computational science to store, retrieve, analyze, and compare the composition of biological molecules, specifically DNA and protein sequences. It is a field of science that encompasses biology, chemistry, computer science, and mathematics. The tools of this field permit scientists and students to access the abundant genomic and protein sequences that are available in databases via the world-wide web. The ability to utilize these resources is of great importance for understanding genomic sequences (*e.g.,* assigning probable functions to genes), identifying previously unknown microorganisms, investigating phylogenetic relationships, and tracking disease outbreaks.

Many of the tools used in bioinformatics (*e.g.,* BLAST) are based on the ability to search for either nucleotide or amino acid sequences that share some degree of similarity. In the following three exercises you will be introduced to the idea of similarity, the alignment of amino acid and nucleotide sequences, and the use of basic bioinformatics tools to construct a molecular phylogram for homologous sequences.

**Inquiry-Based Investigative Question**

We are doing these exercises to showcase the use of bioinformatics tools for furthering our understanding of virology. After completing the three guided exercises that we will do as a class, you will be asked to investigate the following question:

***Is the 2018-2019 Ebola outbreak in North Kivu Province of Democratic Republic of Congo caused by one of the known Ebola strains or a new strain?***

Before tackling this question, you will complete three exercises meant to make you familiar with bioinformatics tools and approaches.

EXERCISE 1

**Similarity and Sequence Alignment**

**Objectives**

After completing this exercise, you should be able to:

1. Define similarity in a non-biological and biological sense.
2. Quantify the similarity between two sequences.
3. Explain how a substitution matrix is used to quantify similarity.
4. Calculate amino acid similarity scores using various matrices.

**Investigation of Similarity**

What do we mean when we describe two objects as being similar?

Consider the two objects in **Figure 1**. Are these objects similar? In what way(s) would you consider them to be similar? **(worksheet question 1)**

 

**Figure 1.** Compare the objects and determine their similarity.
Credit: *Matthias Kabel, CC BY-SA 3.0, via Wikimedia Commons.* [*https://commons.wikimedia.org/wiki/File:Greek\_vase\_Dionysos\_attica\_520\_bC.jpg*](https://commons.wikimedia.org/wiki/File%3AGreek_vase_Dionysos_attica_520_bC.jpg) Andre Karwath, CC BY-SA 2.5, via Wikimedia Commons. [https://commons.wikimedia.org/wiki/File:Chinese\_vase.jpg](https://commons.wikimedia.org/wiki/File%3AChinese_vase.jpg)

**Similarity** is defined as a resemblance or likeness; related in appearance or nature; or having a corresponding aspect or feature. In addition to obvious similarities amongst objects with the same function, written works can also display similarity. When two passages are highly similar, it is considered plagiarism. This implies a common origin to the passages (i.e., the second passage was copied from the first). Likewise, seeing an "excessive" (i.e., more than one would expect based on chance) amount of similarity between two organisms implies a common ancestry. This implication also holds true for biological sequences, which can be more easily quantified than anatomical or behavioral traits. Consider the following passages with apologies to Dr. Seuss (Seuss, 2001):

Passage one:

 One fish, two fish,

 red fish, blue fish.

 Black fish, blue fish,

old fish, new fish.

This one has a little star.

This one has a little car.

Say! what a lot of fish there are.

Passage two:

 One sheep, two sheep,

 black sheep, blue sheep.

 Red sheep, blue sheep,

old sheep, new sheep.

This one has a little bell.

That one drank from a well.

Wow! what loads of sheep there are.

How could the similarity between the two passages above be quantified? What must be done prior to determining the similarity of these passages? **(worksheet question 2)**

**Similarity in Bioinformatics**

Likewise, seeing an "excessive" (i.e., more than one would expect based on chance) amount of physical similarity between two organisms implies a common ancestry. This implication also holds true for biological sequences. Shared ancestry between two organisms or sequences is known as **homology**. It is important to note that sequence similarity does not always insure sequence homology, but that sequence similarity is an expected consequence of homology.

Imagine that you have identified a new gene or protein. One of the first questions you might ask is “What is the function of this protein?” or “What type of protein is encoded by this gene?” A first step in answering these questions would likely include a search of nucleotide and/or protein databases for a known gene or protein that is similar to your recently identified sequence. A search of these databases is based on finding a sequence that can be aligned with your sequence of interest and then the similarity of the sequences can be calculated using a suitable scoring matrix.

Several scoring matrices for amino acid sequence comparisons (*e.g.,* BLOSUM, PAM) have been developed by scientists. These matrices take into account the substitution of chemically and/or physically similar amino acids as well as the relative frequency of such substitutions in naturally occurring proteins. The 20 commonly occurring amino acids are represented by a single letter or three letter abbreviations within the table and each possible substitution is given a numerical score that is associated with how similar or different the substituted amino acid properties are (**Table 1**). Amino acid substitutions between residues that are similar in size and/or polarity are generally scored as positive values in a substitution matrix. Amino acid substitutions between residues that are very different in size and/or polarity are generally scored as negative values in a substitution matrix. How positive or negative a substitution score is depends on the relative similarity in residue chemical properties. A commonly used matrix called BLOSUM-62 (Henikoff, 1992) is shown in **Table 2**.

**Table 1.** Standard amino acid abbreviations.Both the three- and one-letter abbreviations are given along with the chemical properties of the amino acids.

|  |  |  |  |
| --- | --- | --- | --- |
| **Amino Acid** | **Three-letter Abbreviation** | **Single-letter Abbreviation** | **Chemical Properties** |
| Alanine | Ala | A | Non-polar; tiny |
| Arginine | Arg | R | Polar (positively charged) |
| Asparagine | Asn | N | Polar (uncharged); small |
| Aspartate | Asp | D | Polar (negatively charged); small |
| Cysteine | Cys | C | Polar (uncharged); tiny; Sulphur containing |
| Glutamate | Glu | E | Polar (negatively charged) |
| Glutamine | Gln | Q | Polar (uncharged) |
| Glycine | Gly | G | Non-polar; tiny |
| Histidine | His | H | Polar (positively charged); aromatic |
| Isoleucine | Ile | I | Non-polar |
| Leucine | Leu | L | Non-polar |
| Lysine | Lys | K | Polar (positively charged) |
| Methionine | Met | M | Non-polar; Sulphur containing |
| Phenylalanine | Phe | F | Non-polar; aromatic |
| Proline | Pro | P | Non-polar; small |
| Serine | Ser | S | Polar (uncharged); tiny/small |
| Threonine | Thr | T | Polar (uncharged); small |
| Tryptophan | Trp | W | Non-polar; aromatic |
| Tyrosine | Tyr | Y | Polar (uncharged); aromatic |
| Valine | Val | V | Non-polar; small |

**Table 2.** BLOSUM-62 substitution matrix. The twenty amino acids are given in both the left column and in the uppermost row of the table. The single letter amino acid abbreviations are used.

|  |
| --- |
| **BLOSUM-62 Substitution Matrix** |
|  | C | S | T | P | A | G | N | D | E | Q | H | R | K | M | I | L | V | F | Y | W |
| C | **9** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| S | -1 | **4** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| T | -1 | 1 | **5** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| P | -3 | -1 | -1 | **7** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| A | 0 | 1 | 0 | -1 | **4** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| G | -3 | 0 | -2 | -2 | 0 | **6** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| N | -3 | 1 | 0 | -2 | -2 | 0 | **6** |  |  |  |  |  |  |  |  |  |  |  |  |  |
| D | -3 | 0 | -1 | -1 | -2 | -1 | 1 | **6** |  |  |  |  |  |  |  |  |  |  |  |  |
| E | -4 | 0 | -1 | -1 | -1 | -2 | 0 | 2 | **5** |  |  |  |  |  |  |  |  |  |  |  |
| Q | -3 | 0 | -1 | -1 | -1 | -2 | 0 | 0 | 2 | **5** |  |  |  |  |  |  |  |  |  |  |
| H | -3 | -1 | -2 | -2 | -2 | -2 | 1 | -1 | 0 | 0 | **8** |  |  |  |  |  |  |  |  |  |
| R | -3 | -1 | -1 | -2 | -1 | -2 | 0 | -2 | 0 | 1 | 0 | **5** |  |  |  |  |  |  |  |  |
| K | -3 | 0 | -1 | -1 | -1 | -2 | 0 | -1 | 1 | 1 | -1 | 2 | **5** |  |  |  |  |  |  |  |
| M | -1 | -1 | -1 | -2 | -1 | -3 | -2 | -3 | -2 | 0 | -2 | -1 | -1 | **5** |  |  |  |  |  |  |
| I | -1 | -2 | -1 | -3 | -1 | -4 | -3 | -3 | -3 | -3 | -3 | -3 | -3 | 1 | **4** |  |  |  |  |  |
| L | -1 | -2 | -1 | -3 | -1 | -4 | -3 | -4 | -3 | -2 | -3 | -2 | -2 | 2 | 2 | **4** |  |  |  |  |
| V | -1 | -2 | 0 | -2 | 0 | -3 | -3 | -3 | -2 | -2 | -3 | -3 | -2 | 1 | 3 | 1 | **4** |  |  |  |
| F | -2 | -2 | -2 | -4 | -2 | -3 | -3 | -3 | -3 | -3 | -1 | -3 | -3 | 0 | 0 | 0 | -1 | **6** |  |  |
| Y | -2 | -2 | -2 | -3 | -2 | -3 | -2 | -3 | -2 | -1 | 2 | -2 | -2 | -1 | -1 | -1 | -1 | 3 | **7** |  |
| W | -2 | -3 | -2 | -4 | -3 | -2 | -4 | -4 | -3 | -2 | -2 | -3 | -3 | -1 | -3 | -2 | -3 | 1 | 2 | **11** |

Considering amino acid residue chemical properties, explain why an alanine substituted with a Serine is assigned a score of 1, while an Alanine substituted with a Tryptophan is assigned a score of -3 in the BLOSUM-62 substitution matrix. **(worksheet question 3)**

To illustrate the use of this matrix, let’s say you have isolated a protein with the following amino acid sequence (this will be our **query** sequence): MGDVEKGKKIFIMKC. We want to compare the similarity of this sequence to the following sequence that was found in a database of protein sequences: MGEVERGKKLFIMKC. The arrangement of two sequences to identify regions of similarity is termed **sequence alignment**.

To do this, find the first amino acid of the query **in the left column** of the BLOSUM-62 matrix, then look for the first amino acid of the **subject** sequence (the sequence being compared to the query) in the top row of the matrix and locate the box at the intersection of these two amino acids in the table. The number that corresponds to this pairing is noted and added to the value for each of the subsequent pairings. For example, the first amino acid of each sequence is methionine (M) which scores 5 and the second amino acid in each sequence is glycine (G) which scores 6, giving a total score thus far of 11.

What is the total similarity score for these two aligned sequences? **(worksheet question 4)**

Query: MGDVEKGKKIFIMKC

Subject 1: MGEVERGKKLFIMKC

If the query sequence is aligned to a different subject sequence (given below), what is the similarity score? **(worksheet question 5)**

Query: MGDVEKGKKIFIMKC

Subject 2: MCDVWKGKSIFIMKC

Explain why the similarity scores calculated above are different. Consider and refer to information provided in Table 1 as part of your explanation. **(worksheet question 6)**

When the query sequence is compared to itself, a similarity score of 80 is obtained. Considering this, why are the two scores you calculated above different despite having the same number of identical amino acids? Which of the two subject sequences most likely diverged evolutionarily longer ago from the query sequence? **(worksheet question 7)**

**Computational Procedure:**

Similarity scores can also be determined using tools available on the Internet. One collection of sequence analysis tools can be found on the Sequence Manipulation Suite website (<http://www.bioinformatics.org/sms2/>). Go to this website and find the link to the sequence alignment tool called “Pairwise Align Protein” and click on it. On this page, you will need to enter both the query and subject sequences (these have been provided to you electronically). The default matrix is BLOSUM-62. Once both sequences have been entered click on the “submit” button and wait for your results.

Report the similarity scores from the computational scoring. Did the computationally calculated similarity scores match those that you manually calculated? **(worksheet question 8)**