**Sequence Similarity –Bioinformatics**

EXERCISE 1

**Investigating Sequence Similarity**

**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.

**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 includes 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 important for understanding the functions of 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 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**

You will learn to use the actual websites, tools, and databases that scientific researchers utilize to study genes and proteins from various organisms, including the SARS-CoV 2 virus that has been instrumental in the development of COVID vaccines.

**Investigation of Similarity**

Thought question: 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?



**Figure 1.** Compare the objects and determine their similarity.   
Credit: *Matthias Kabel, CC BY-SA 3.0, via Wikimedia Commons*

**Similarity** is defined as a resemblance or likeness; related in appearance or nature; or having a corresponding aspect or feature. In addition to apparent 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 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,

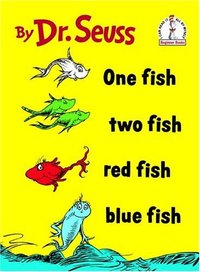
red fish, blue fish.

Black fish, blue fish,

old fish, new fish.

This one has a little star.

This one has a little car.



Credit: Seussville.com

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.

**Question 1:**

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

**Similarity in Bioinformatics**

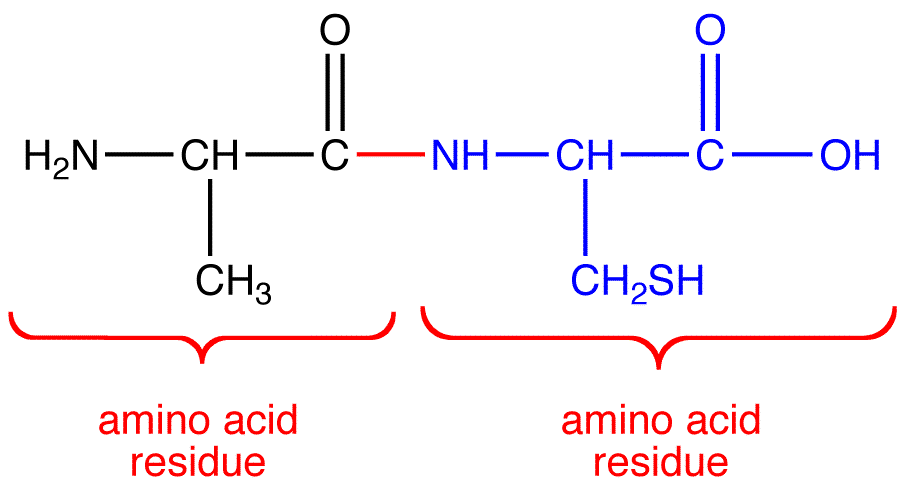
Like the passages above, seeing an "excessive" (i.e., more than one would expect based on chance) amount of physical similarity between two organisms implies a common ancestry. This implication also holds for biological sequences. Shared ancestry between two organisms or sequences is known as **homology**. It is important to note that sequence similarity does not always ensure 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. Then the similarity of the sequences can be calculated using a suitable scoring matrix.

The 20 commonly occurring amino acids are often represented by a single letter or three letter abbreviations, as shown in **Table 1**. The chemical properties of the amino acids affect how they interact with each other and determine the structure of the protein formed by a chain of amino acids.

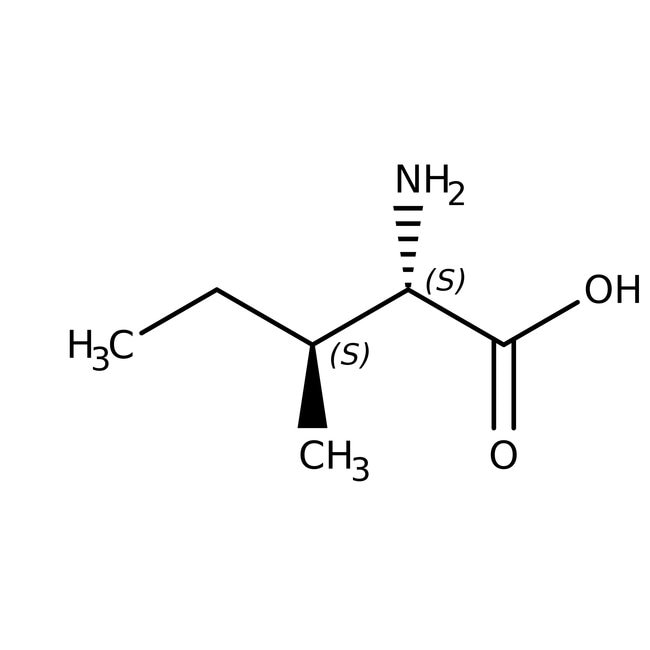
**Table 1.** Standard amino acid abbreviations.Both the three- and one-letter abbreviations are given along with the chemical properties of the amino acids.

|  |  |  |  |
| --- | --- | --- | --- |
| **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 ring; large |
| 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 ring; large |
| Tyrosine | Tyr | Y | Polar (uncharged); aromatic ring; large |
| Valine | Val | V | Non-polar; small |

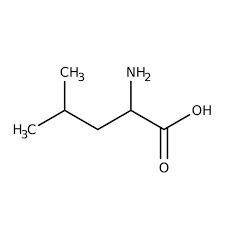
When amino acids are joined by a peptide bond, they lose a water molecule. The amino acid that is left is called a **residue**.

In a substitution matrix, each possible substitution is given a numerical score that is associated with how similar or different the substituted amino acid properties are (**Table 2**). **Amino acid residues that are substituted with new residues that are similar in size and/or polarity are generally scored as positive values in a substitution matrix. Substitutions between residues that are very different in size and/or polarity are typically scored as negative values in a substitution matrix.**

How positive or negative a substitution score is depends on the similarity between the chemical properties of the amino acids. For example, substituting a leucine for an isoleucine is a conservative replacement and will have little effect on the protein because they are both non-polar.

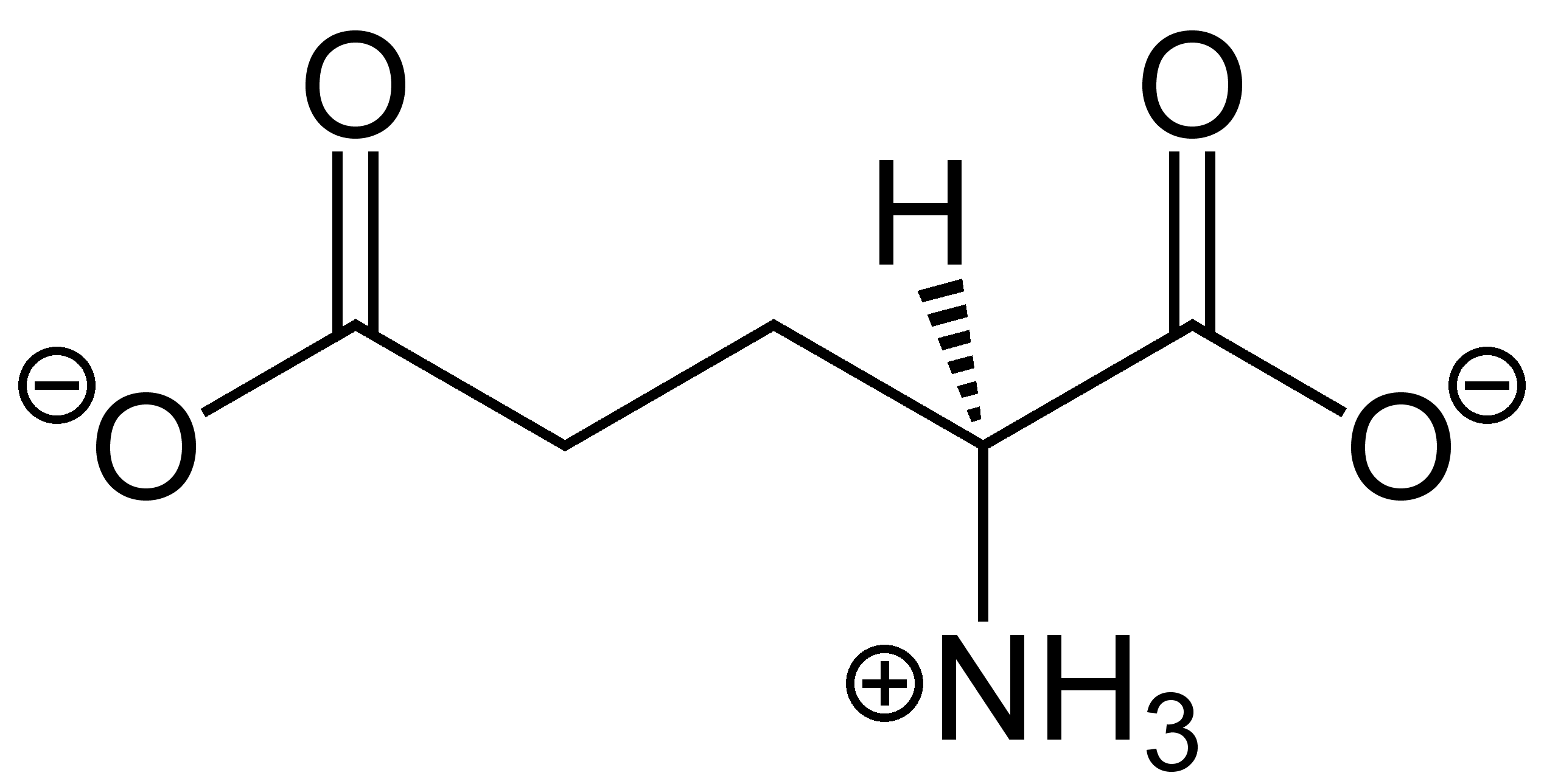


Isoleucine



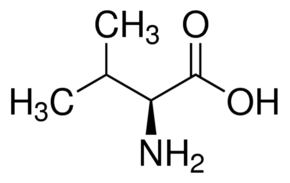
Leucine

However, the substitution of a glutamate (negatively charged) with a valine (uncharged) is a radical replacement that can have significant effects on the structure of the protein, as seen in sickle cell anemia.



Glutamate

Valine



Several scoring matrices for amino acid sequence comparisons (*e.g.,* BLOSUM, PAM) have been developed by scientists. These matrices consider the substitution of chemically and/or physically similar amino acids as well as the relative frequency of such substitutions in naturally occurring proteins.

A commonly used matrix called BLOSUM-62 (Henikoff, 1992) is shown in **Table 2** on the next page. You will use this matrix to determine the similarity between two sequences.

**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** |

**Question 2:**

Considering amino acid residue chemical properties, explain why an alanine substituted with a serine is assigned a score of 1, while an alanine replaced with a tryptophan is assigned a score of -3 in the BLOSUM-62 substitution matrix. **(Answer this question on your worksheet)**

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):

Query: MGDVEKGKKIFIMKC

We want to compare the similarity of this sequence to the following sequence that was found in a database of protein sequences (this will be our **subject** sequence):

Subject: MGEVERGKKLFIMKC

The arrangement of two sequences to identify regions of similarity is termed **sequence alignment** and is shown below.

Query: MGDVEKGKKIFIMKC

Subject: MGEVERGKKLFIMKC

To calculate a similarity score between the two sequences, find the intersection point of the amino acid from the query sequence and the amino acid from the subject sequence in the BLOSUM-62 substitution matrix that yields a numeric value. 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. Each amino acid pair is evaluated using the matrix and all the scores are added together to calculate the total similarity score for two aligned sequences.

**Question 3:**

What is the total similarity score for these two aligned sequences? **(Answer this question on your worksheet)**

Query: MGDVEKGKKIFIMKC

**Subject 1**: MGEVERGKKLFIMKC

**Question 4:**

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

Query: MGDVEKGKKIFIMKC

**Subject 2**: MCDVWKGKSIFIMKC

**Question 5:**

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

**Question 6:**

Some species have a common ancestor but have evolved to have different characteristics over time. This process is known as divergent evolution. Which of the two subject sequences most likely diverged evolutionarily longer ago from the query sequence?

What evidence did you use to determine your answer?

**(Answer these questions on your worksheet)**

**Computational Procedure:**

1. 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/>).
2. Go to this website and find the link on the left side of the page to the sequence alignment tool called “Pairwise Align Protein” and click on it.
3. On this page, click the “Clear” button below the boxes to remove the current text.
4. Enter the Query sequence shown above into the first box and the Subject 1 sequence into the second box (see Question 4).
   1. Copy and paste the sequences from above.
   2. Keep the default alignment matrix - BLOSUM-62.
   3. Once both sequences have been entered, click on the “submit” button and wait for your results. Record these results in Question 8 on your worksheet.
5. Repeat for Subject 2 sequence (see Question 5).

**Question 7:**

Record the similarity (alignment) scores from the computational scoring for Subject 1 sequence and Subject 2 sequence on your **worksheet**.

Did the computationally calculated similarity scores match those that you manually calculated? If not, work through the steps again and identify what might have caused the difference.

**(Answer this question on your worksheet)**