**Bioinformatics:**

**Investigating Sequence Similarity – A Plant Biology Approach**

**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 this exercise, 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 Questions**

Bioinformatics techniques are used to address questions at the macro-level (e.g., How did birds lose their teeth? What is the historical distribution of a native species?) and the micro-level (e.g., Can organisms across all domains of life share DNA? What molecular changes increase the virulence of a virus?).

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? **(1)**

Consider the two objects in **Figure 1**. Are these objects similar? In what way(s) would you consider them to be similar? **(2)**

**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:Greek_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>

**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 plant samples that you examined last week:

*Bryophyta (Moss and Liverwort), Pteridophyta (Fern), Gymnosperm (Blue spruce), Angiosperm (Clover). What similarities and differences did you observe?*

**Similarity in Bioinformatics**

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, while the latter are generally scored as negative 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. **(4)**

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? **(5)**

Query: MGDVEKGKKIFIMKC

Subject 1: MGEVERGKKLFIMKC

If the query sequence is aligned to a different subject sequence (given below), what is the similarity score? **(6)**

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. **(7)**

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? **(8)**

**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.

Did the computationally calculated similarity scores match those that you manually calculated? **(9)**

EXERCISE 2

**Sequence Alignment to a Database of Sequences**

**Objectives**

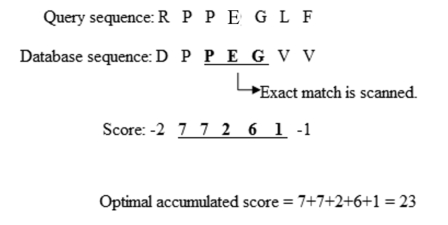
After completing this exercise, you should be able to:

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 above, 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 (class handout)**, 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. 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? **(1)**

What problems did you encounter that may have affected your calculated scores? **(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 histone protein sequences between different type of plants and between plants and fungi. In the first BLAST we will compare the sequence of an angiosperm (*Arabidopsis thaliana)* and moss (*Physcomitrella patens)* histone protein, H4. The second BLAST will be conducted to compare *A. thaliana’s* histone protein sequence with that of *Saccharomyces cerevisiae*, bread yeast.

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, meaning that mutations will have accumulated slowly, provided a useful measure for how closely or distantly related different organisms are to each other.

How similar do you expect the H4 proteins to be between two angiosperms versus an angiosperm and a moss? **(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 NP\_180441.1 (the accession number for *Arabidopsis thaliana -*  a small, well-studied flowering plant belonging to the mustard family) to search all databases for this protein. Click “Search.”
3. In the “Protein Sequence” box you should see “histone H4 [Arabidopsis thaliana], SELECT “FASTA” and observe the protein sequence. FASTA is a standard format, used for both genes and proteins, in which the first line begins with a > symbol, followed by any text.
4. 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.
5. On this page under “Database” select “RefSeq protein”. Then under “organism” start typing “*Physcomitrella patens*” (the scientific name for a type of moss). Let the full name fill in.
6. Click the “BLAST” button at the bottom of the page. The BLAST algorithm is currently looking for similarity by comparing the Arabidopsis Histone H4 sequence to all annotated *P. patens* protein sequences.
7. 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 in the box underneath the label “Sequences producing significant alignments.”
8. The first subject has a Max score of 199 and an identity (“Per. Ident”) of 100%, which means that the Arabidopsis and moss histone proteins have, basically, the same amino acid sequence. If we were to compare the *nucleotide* sequences for the gene encoding this protein between Arabidopsis and moss, do you think they would be identical? Explain.
9. Now repeat the BLAST by selecting “Edit and Resubmit.” Compare the Arabidopsis protein sequence to Saccharomyces cerevisiae (bread yeast, a species of Ascomycota fungi). How do you interpret the resulting Max Score and Identity score?