Sequence Similarity Introduction

Katie M. Sandlin Adapted from: Kleinschmit*, A. J., Brink, B., Roof, S., Goller, C. C., Robertson, S. (2021). Sequence Similarity: An inquiry based and "under the hood" approach for incorporating molecular sequence alignment in introductory undergraduate biology courses. CourseSource, (Version 1.0). QUBES Educational Resources. doi:10.24918/cs.2019.5

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 internet. 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 (ancestry), 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 and the alignment of amino acid and nucleotide sequences.

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 the BLOSUM 62 substitution matrix.
5. Explain how BLAST detects similarity between two sequences.
6. Explain how to use BLAST and interpret the alignments.

Q1. *What do we mean when we describe two objects as being similar?*

Q2. *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 and* Andre Karwath, CC BY-SA 2.5, 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 obvious similarities among 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). Consider the following passages (Seuss, 1960):

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.

Q3. *How could the similarity between these two passages be quantified? What must be done prior to determining the similarity of these passages?*

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 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 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 chemically dissimilar substitutions 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 and 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; very small
Arginine	Arg	R	polar (positively charged), large
Asparagine	Asn	N	polar (uncharged); small
Aspartate	Asp	D	polar (negatively charged); small
Cysteine	Cys	C	polar (uncharged); small; sulfur containing
Glutamate	Glu	E	polar (negatively charged), medium sized
Glutamine	Gln	Q	polar (uncharged), medium sized
Glycine	Gly	G	non-polar; very small
Histidine	His	H	polar (positively charged); aromatic, medium sized
Isoleucine	Ile	I	non-polar large
Leucine	Leu	L	non-polar large
Lysine	Lys	K	polar (positively charged), large
Methionine	Met	M	non-polar; sulfur containing, large
Phenylalanine	Phe	F	non-polar; aromatic, very large
Proline	Pro	P	non-polar; small
Serine	Ser	S	polar (uncharged); very small
Threonine	Thr	T	polar (uncharged); small
Tryptophan	Trp	W	non-polar; aromatic, very large
Tyrosine	Tyr	Y	polar (uncharged); aromatic, very large
Valine	Val	V	non-polar; medium sized

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**

Q4. *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.*

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 top row 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 left column 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 of each sequence is glycine (G) which scores 6. The third amino acid of the query sequence is

aspartate (D) but in the subject sequence it's glutamate (E) which scores 2. Therefore, our total score thus far of 13.

Q5. *What is the total similarity score for these two aligned sequences?*

Query: `MGDVEKGKKIFIMKC`

Subject 1: `MGEVERGKKLFIMKC`

Q6. *If the query sequence is aligned to a different subject sequence (given below), what is the similarity score?*

Query: `MGDVEKGKKIFIMKC`

Subject 2: `MCDVWKGKSIFIMKC`

Q7. *Explain why the similarity scores calculated above are different. Consider and refer to information provided in Table 1 as part of your explanation.*

Q8. *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?*

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 output subject sequences, by 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 a given subject 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 subject sequence in the database. 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(es) 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**).

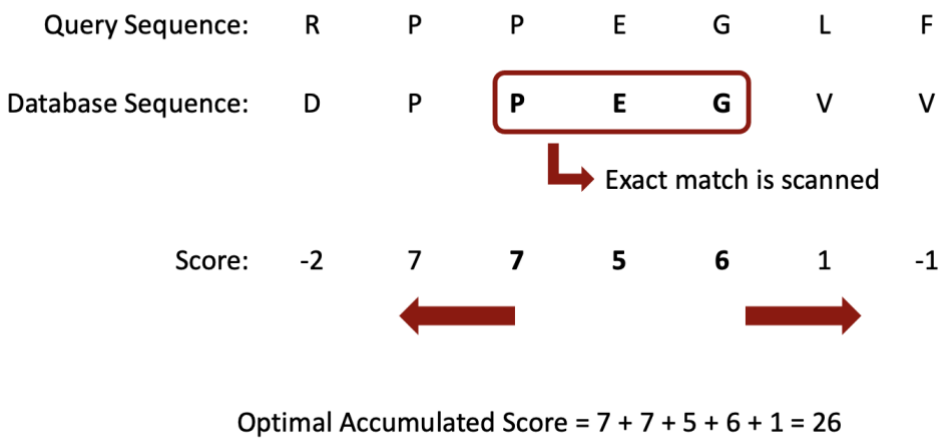

Figure 2. An exact query word match alignment and extension followed by

alignment truncation before negative substitution values using a simplified BLAST algorithm.

To illustrate BLAST, we will use the following amino acid sequence as our query, with the 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 **BLAST Alignment Template**, 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.

Q9. *Which of the proteins from the database is most similar to our query? Which is the least similar?*

Protein BLAST (*blastp*)

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 RNA polymerases. Due to their critical role, histone proteins are highly conserved in their amino acid sequence, structure, and function. Humans (*Homo sapiens*) and chimpanzees (*Pan troglodytes*) last shared a common ancestor between 5 and 8 million years ago.

Q10. *How similar do you expect their H4 proteins to be?*

Computational Procedure:

1. Go to the NCBI home page at ncbi.nlm.nih.gov. This website 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" to view 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. 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 for "Database," select "refseq_protein." Then under "organism" type "*Pan troglodytes*" (the scientific name for the chimpanzee) and select "taxid:9598."
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 Descriptions section will show you the hits or similarity search results found. The first aligned database record has a score of 203 and an identity of 100%, which means that the human and chimpanzee H4 histone protein has the same amino acid sequence.
10. The Graphic Summary will show you a graphical comparison of the query and subject sequences.
11. Click on "Alignments" to see the amino acid alignments.

Q11. *What was the query for this search? What species database did you search for a hit to the query?*

Q12. *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.*

Nucleotide BLAST (*blastn*)

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" to view hits from your search within a nucleotide sequence database.
4. Near the top of the page, click on the first "Homo sapiens histone cluster 4, H4, mRNA." A new screen will appear that gives you a lot of information about this 354 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 and select "taxid:9598." 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, click on the "Graphic Summary." At the bottom of the box, you will see five bright red lines. The fact that the lines are bright red indicates that these are very close matches. Hover over the lines, and it will tell you, in technical terms, which chimpanzee gene your query (human histone H4) matches.
10. Click on the "Descriptions" tab and then click on the Accession of the top hit. On the following page, view the information under "source," below the subheading "FEATURES" to determine the chromosome on which your hit lies.

Q13. *On what chromosome is the gene for histone H4 located in chimpanzees?*

10. On the BLAST results page, click on “Alignments.” The “Query” is human histone H4. The “Subject” is chimpanzee histone H4. Notice that wherever the genes are identical, there is a vertical line between the identical base pairs.

Q14. *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.*

Q15. *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?*

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403-410. doi: 10.1016/S0022-2836(05)80360-2

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences, 89(22), 10915-10919.

Seuss, D. (1960). One fish, two fish, red fish, blue fish. New York: Beginner Books, a division of Random House