

Genomics
Education
Partnership
thegep.org


Sequence Similarity Introduction

Katie M. Sandlin, M.S.
Last Update: 10/26/2023

1

Bioinformatics


- Involves the use of **computers** to store, retrieve, analyze, and compare the composition of biological molecules, specifically DNA and protein sequences



GEP Genomics Education Partnership

2

Bioinformatics



- Bioinformatic tools allow scientists (including YOU!) to access the abundant genomic and protein sequences that are available in databases via the **internet**.

GEP Genomics Education Partnership

Photo: Huge monitor meme generator

3

Overview

- 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.

GEP Genomics Education Partnership

4

Objectives

After completing this exercise, you should be able to:

- Define similarity in a non-biological and biological sense.
- Quantify the similarity between two sequences.
- Explain how a substitution matrix is used to quantify similarity.
- Calculate amino acid similarity scores using the BLOSUM 62 substitution matrix.
- Explain how BLAST detects similarity between two sequences.
- Explain how to use BLAST and interpret the alignments.

GEP Genomics Education Partnership

5

Q1. Investigation of Similarity

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

GEP Genomics Education Partnership

6

Q2. Investigation of Similarity

- Are these objects similar? If so, in what way(s) would you consider them to be similar?



GEP Genomics Education Partnership

Martinus Kibbi, CC BY-SA 3.0, via Wikimedia Commons.
pasrah-art, CC BY-SA 2.5, via Wikimedia Commons.

7

Similarity

- Defined as a resemblance or likeness; related in appearance or nature; or having a corresponding aspect or feature.

GEP Genomics Education Partnership

8

Similarity

- 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).

GEP Genomics Education Partnership

9

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

Passage 1

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 2

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.

GEP Genomics Education Partnership

Dr. Seuss, 1960

10

Similarity in Bioinformatics

- "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**.

GEP Genomics Education Partnership

11

Similarity in Bioinformatics

- It is important to note that sequence similarity does not always ensure sequence homology, but that sequence similarity is an expected consequence of homology.

GEP Genomics Education Partnership

12

Homology

- Similarities to the mouse gene (*Pax6*) are highlighted in green.

Mouse *Pax6* gene:
GTATCCAACGGTTGTGTGAGTAAATCTCGGGCAGGTATTACGAGACGGCTCCATCAGA

Fly *eyeless* gene: Genetic similarity to mouse: 76.66%
Protein similarity to mouse: 100%
GTATCAAATGGATGTGTGAGCAAAATCTCGGGAGGTATTATGAAACAAGGATACGA

Shark eye control gene: Genetic similarity to mouse: 85%
Protein similarity to mouse: 100%
GTCTCCAACGGTTGTGTGAGTAAATCTCGGGCAGATACTATGAAACAGGATCCATCAGA

Squid eye control gene: Genetic similarity to mouse: 78.33%
Protein similarity to mouse: 100%
GTCTCCAACGGCTGC-GTTAGCAAGATTCTCGGACGGTACTAGAGACGGCTCCATAAGA

Flatworm eye control gene: Genetic similarity to mouse: 71.66%
Protein similarity to mouse: 100%
GTGTCTAATGGTTGTGTAGTAAATCTTGGCGATATTATGAAACAGGTCTATTAA

<https://evolution.berkeley.edu/why-the-eye/homologous-genes/>

GEP Genomics Education Partnership

13

Identifying Similarity

- Imagine that you have **identified a new** gene or protein. What questions might you be asking?
 - What is the function of this protein?
 - 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.

GEP Genomics Education Partnership

14

Identifying Similarity

- 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.

GEP Genomics Education Partnership

15

Scoring Similarity

- 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 and the relative frequency of such substitutions in naturally occurring proteins.

GEP Genomics Education Partnership

16

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

GEP Genomics Education Partnership

17

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.

GEP Genomics Education Partnership

18

Scoring Similarity

- How positive or negative a substitution score is depends on the relative similarity in residue chemical properties.
- BLOSUM 62 is a commonly used substitution matrix.

19

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

Query: MGDVEKGKKIFIMKC
Subject: MGEVERGKKLIFIMKC

20

- What is the total similarity score for these two aligned sequences?
 - Query: MGDVEKGKKIFIMKC
 - Subject: MGEVERGKKLIFIMKC

21

- Finish Questions 6-9
- Read the text on pages 6-8

22

Query Word

Query Sequence: R P P E G L F
Database Sequence: D P P E G V V

What's our query word (i.e., scan for an exact match that is 3 amino acids long)?

23

Query Word

Query Sequence: R P P E G L F
Database Sequence: D P P E G V V
Score: 7 5 6

24

Q11.

- What was the query for this search?
- What species database did you search for a hit to the query?

31

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.

32

Exercise 2

- Complete the Computational Procedure for the Nucleotide BLAST.
- Answer Q13-15.

33