

Bioinformatics

Sequence alignment

MSc. Vicente Machaca Arceda

Universidad Nacional de San Agustín de Arequipa

June 13, 2020

Table of Contents

- 1 Introduction
 - Objectives
 - Motivation
 - Previous concepts

- 2 Sequence alignment
 - Definition
 - Dot matrix
 - Practice

Table of Contents

- 1 Introduction
 - Objectives
 - Motivation
 - Previous concepts
- 2 Sequence alignment
 - Definition
 - Dot matrix
 - Practice

Introduction

Objectives

- Understand the importance of sequence alignment in Bioinformatics.

Introduction

Objectives

- Understand the importance of sequence alignment in Bioinformatics.
- Implement the most relevant sequence alignments algorithms.

Introduction

Motivation

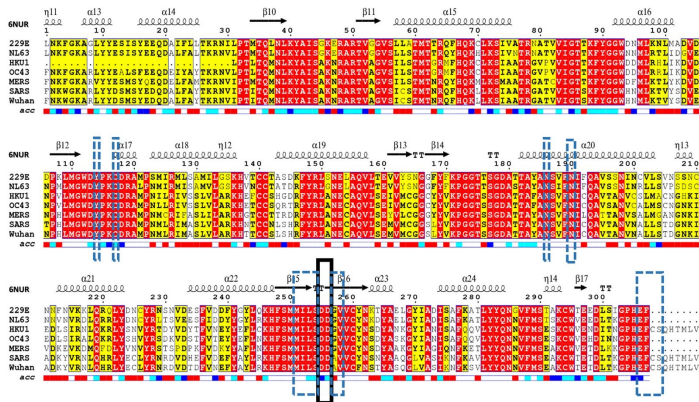


Figure: The SARS HCoV RdRp is the closest strain to the COVID-19, this information is important for drug designers [1].

Introduction

Motivation

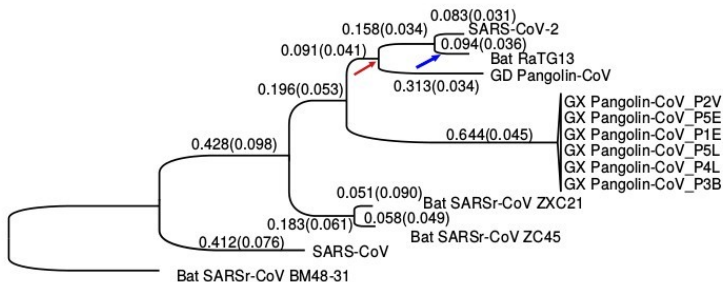


Figure: The phylogenetic tree of SARS-CoV-2 (COVID-19) and the related Coronaviruses [2].

Previous concepts

Genomic variations

Mutations

Many sources of mutation exist that can alter the genome of a cell during its life span, or during replication. Various mutations can affect anything from single base pairs (point mutations), to large genomic regions containing multiple genes.

Previous concepts

Mutations types

- **Somatic mutations**, occurs in a single cell and cannot be inherited (cancer).
- **Germline mutations**, occurs in germ cells (sperm and ovum), mutations in these cells can be passed on to offspring.

Previous concepts

Mutations types

- **Point mutations**, are changes to one base in the DNA.
- **Block mutations**, are changes to segments of a chromosome, resulting in large scale changes in the DNA.

Previous concepts

Genomic variations

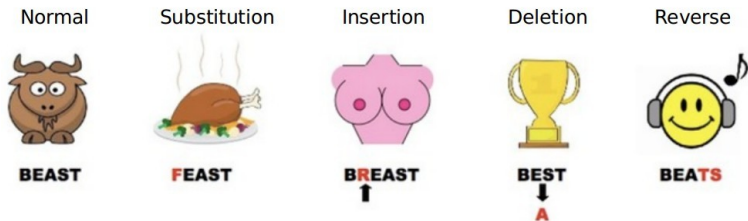


Figure: Overview of the Different Types of Point Mutations.

Previous concepts

Genomic variations

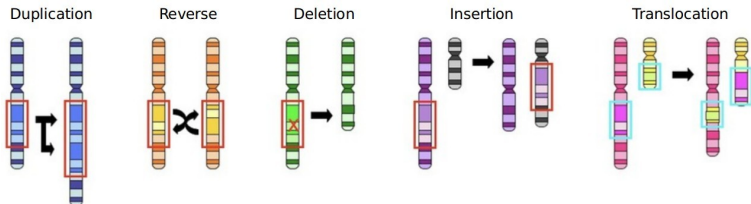


Figure: Overview of the Different Types of Block Mutations.

Previous concepts

Genomic variations

Inversion

A sequence change where, compared to a reference sequence, more than one nucleotide replacing the original sequence are the **reverse complement** of the original sequence.

Previous concepts

Genomic variations

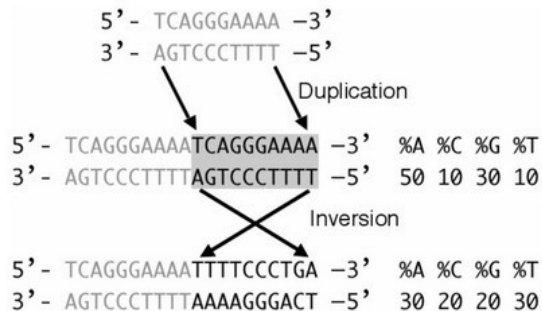


Figure: Inversion example.

Previous concepts

Genomic variations

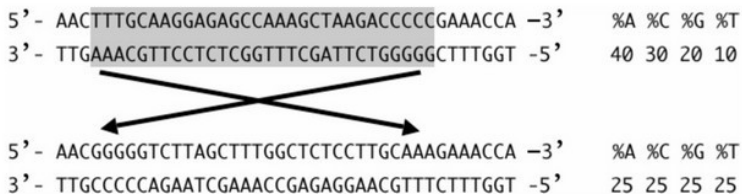


Figure: Inversion example.

Previous concepts

Genomic variations

Frameshift mutation

Also called a framing error or a reading frame shift. It is a genetic mutation (insertions or deletions) of nucleotides that is not divisible by three.

Previous concepts

Genomic variations

Frameshift mutation

Also called a framing error or a reading frame shift. It is a genetic mutation (insertions or deletions) of nucleotides that is not divisible by three.

Due to the triplet nature of gene expression by codons, the insertion or deletion can change the reading frame (the grouping of the codons), resulting in a completely different translation from the original.

Previous concepts

Genomic variations

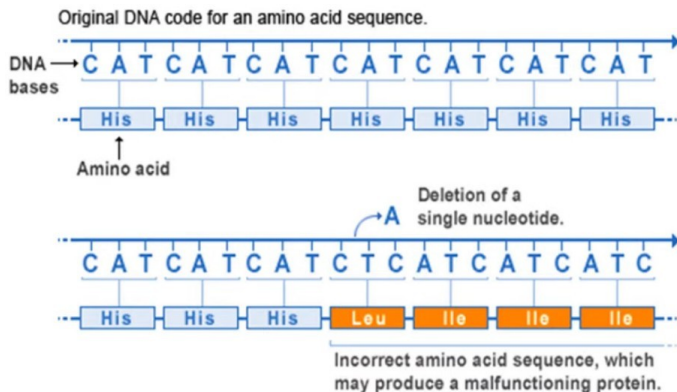


Figure: A frameshift mutation cause an incorrect amino acid sequence, which may produce a malfunctioning protein.

Previous concepts

Genomic variations

Frameshift mutations are apparent in severe genetic diseases such as Tay–Sachs disease (destruction of nerve cells). Also, they increase susceptibility to certain cancers [3].

Table of Contents

- 1 Introduction
 - Objectives
 - Motivation
 - Previous concepts
- 2 Sequence alignment
 - Definition
 - Dot matrix
 - Practice

Sequence alignment

Definition

Pairwise sequence alignment

This is the process by which sequences are compared by searching for common character patterns and establishing residue–residue correspondence among related sequences [4].

It is an important first step toward structural and functional analysis of newly determined sequences [4].

Sequence alignment

Sequence homology versus sequence similarity versus sequence identity

According to Xiong [4]:

- When two sequences are descended from a common evolutionary origin, they have homologous relationship or share **homology**.
- Sequence similarity is the **percentage** of aligned residues that are similar.
- Sequence similarity and sequence identity are synonymous for nucleotide sequences but different in a protein sequence. **Sequence identity** refers to the **percentage** of matches of the same amino acid residues; **sequence similarity** refers to the **percentage** of aligned residues that have similar physicochemical characteristics (size, charge, and hydrophobicity).

Sequence alignment

Examples

No alignment

```
CGATGCTAGCGTATCGTAGTCTATCGTAC
      |      ||
ACGATGCTAGCGTTTCGTATCATCGTA
```

Aligned

```
-CGATGCTAGCGTATCGTAGTCTATCGTAC
||||| ||||| ||||| ||||| |||||
ACGATGCTAGCGTTTCGTA-TC-ATCGTA-
```

In the alignment process there could be substitutions, changes of residues and gaps. Gaps could cause by insertions or deletions.

Figure: No alignment versus alignment.

Sequence alignment

Examples

No gaps (10 matches)

```
a:  ATATTGCTACGTATATCAT
      |||||
b:  ATATATGCTACGTATCAT
```

With one gap (14 matches)

```
a:  ATAT-TGCTACGTATATCAT
      |||  |||||
b:  ATATATGCTACGTATCAT
```

With two gaps (16 matches)

```
a:  ATAT-TGCTACGTATATCAT
      |||  |||||  |||||
b:  ATATATGCTACG--TATCAT
```

Algorithms should take into account the possibility of introducing gaps. **Several alignments can be constructed** between two sequences.

Figure: Alignment and gaps.

Sequence alignment

Evaluating the alignments

To compare alignments we can score them. The main features taken into account are usually:

- Number of matching residues.
- Number of mismatches.
- Number of gaps.
- Length of the gaps.

Sequence alignment

Evaluating the alignments

We can devise different scoring schemes. For instances:

- scoring schema 1: match +1, mismatch: 0, gap creation: -1
gap extension: -1
- scoring schema 1: match +1, mismatch: -1, gap creation:
-1 gap extension: 0

Sequence alignment

Global Alignment and Local Alignment

In **Global alignment**, two sequences to be aligned are assumed to be generally similar over their entire length. Alignment is carried out from beginning to end [4].

Local alignment, does not assume that the two sequences have similarity over the entire length. It only finds local regions with the highest level of similarity [4].

Sequence alignment

Global Alignment and Local Alignment

Global Alignment:

```
--AGATCCGGATGGT--GTGACATGCGAT--AAG--AGGCGTT
      ||| | | | ||||| ||||| ||| | ||
GTCCATCTG--TCTTGGGTGAC-TGCGATAACAAGTTA--CCTT
```

Local Alignment:

```
--AGATCCGGATGGT--GTGACATGCGATA--AG--AGGCGTT
                        ||||| |||||
GTCCATCTG--TCTTGGGTGAC-TGCGATACAAGTTA--CCTT
```

Figure: Global Alignment and Local Alignment.

Sequence alignment

Alignment Algorithms

Global and local, are fundamentally similar and only differ in the optimization strategy used in aligning similar residues, the algorithms can be based on one of the three methods:

- The dot matrix method.
- The dynamic programming method.
- The word method.

Sequence alignment

Dot matrix

Dot matrix

The most basic sequence alignment method is the **Dot matrix** method, proposed by Gibbs and McIntyre (1970) [5], also known as the **Dot plot** method. It is a graphical way of comparing two sequences in a two dimensional matrix [4].

Sequence alignment

Dot matrix

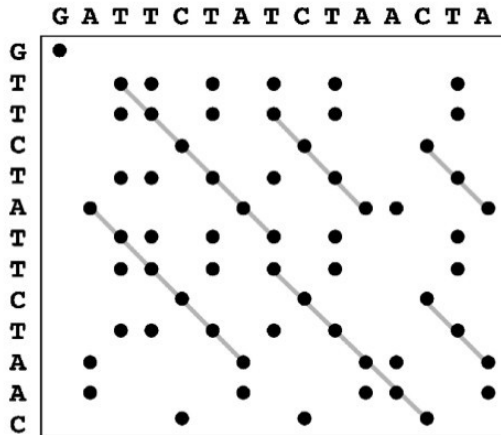


Figure: Dot matrix example

Sequence alignment

Dot matrix

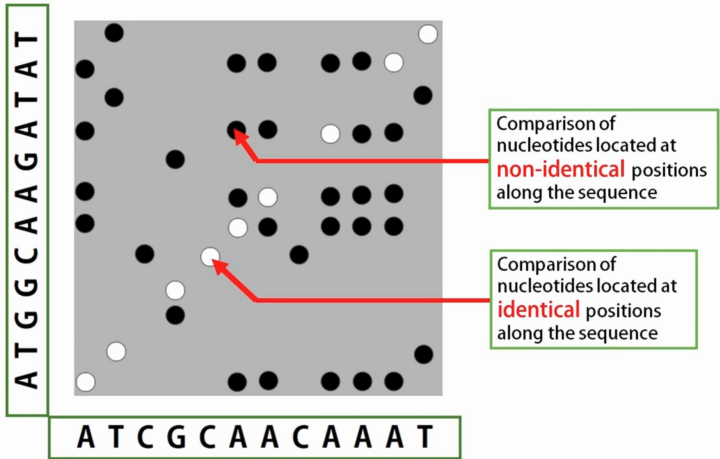


Figure: Dot matrix example

Sequence alignment

Dot matrix

What we could conclude from the Dot plot?

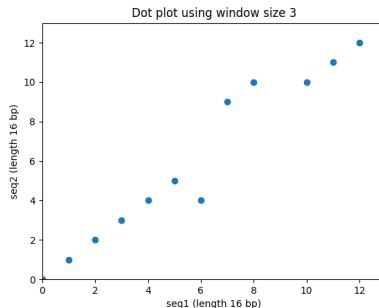


Figure: Dot matrix example. *seq1* = ACCTGAGAGTGTGGCT and *seq2* = ACCTGAGACAGTGGCT

Sequence alignment

Dot matrix

What we could conclude from the Dot plot?

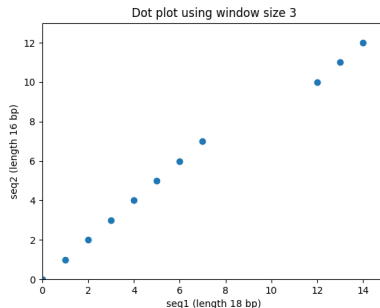


Figure: Dot matrix example. *seq1* = ACCTGAGACATTGTGGCT and *seq2* = ACCTGAGACAGTGGCT

Sequence alignment

Dot matrix

What we could conclude from the Dot plot?

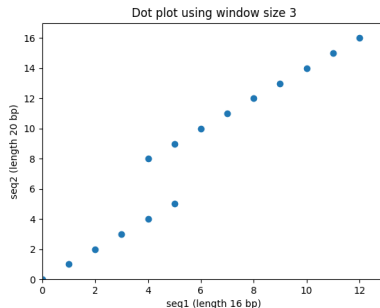


Figure: Dot matrix example. *seq1* = ACCTGATACAGTGGCT and *seq2* = ACCTGATAGATACAGTGGCT

Sequence alignment

Dot matrix

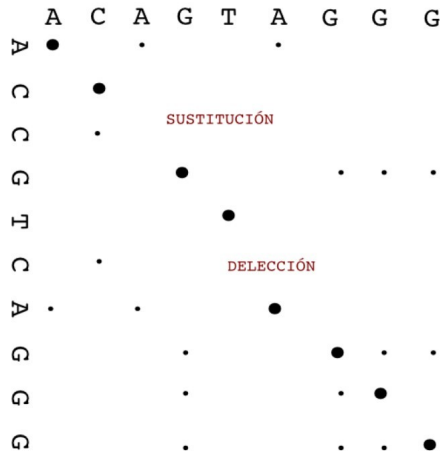


Figure: Dot matrix example

Sequence alignment

Dot matrix

Deletion / insertion

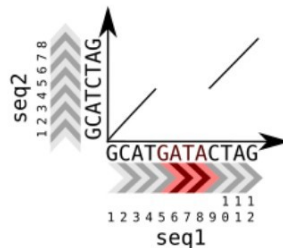


Figure: Deletion/insertion example in Dot matrix.

Sequence alignment

Dot matrix

Duplication

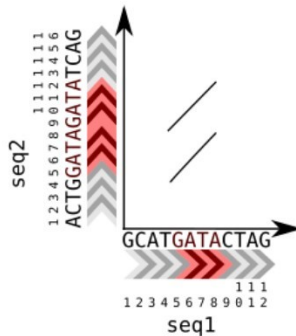
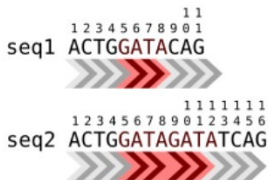


Figure: Duplication example in Dot matrix.

Sequence alignment

Noise in Dot matrix

Comparison of *rps0* gene sequences for *Escherichia coli* and *Salmonella typhi*

Off-diagonal noise

represents several random off-diagonal matches that generally is not meaningful

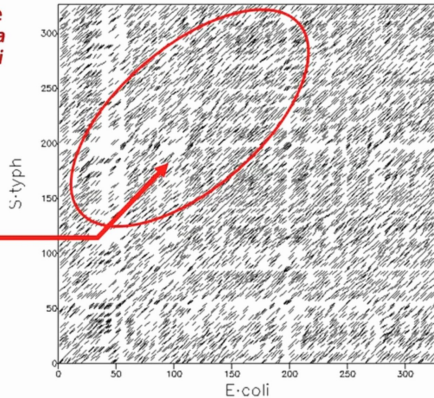


Figure: Several off-diagonal matches that generally is not meaningful.

Sequence alignment

Dot matrix with window size

Introducing the Concept of Sequence Window

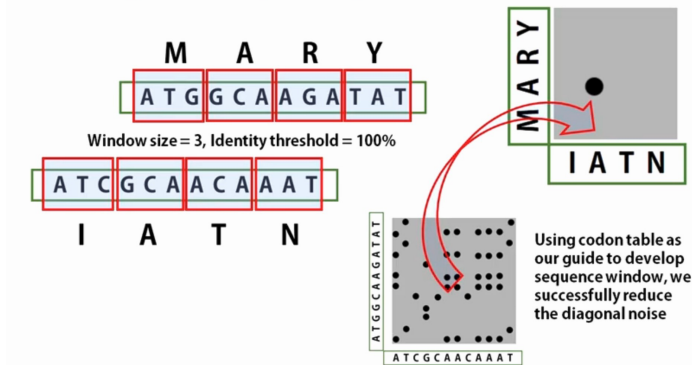


Figure: Dot plot with windows successfully reduce the diagonal noise.

Sequence alignment

Dot matrix with window size

The concept of Identity Threshold

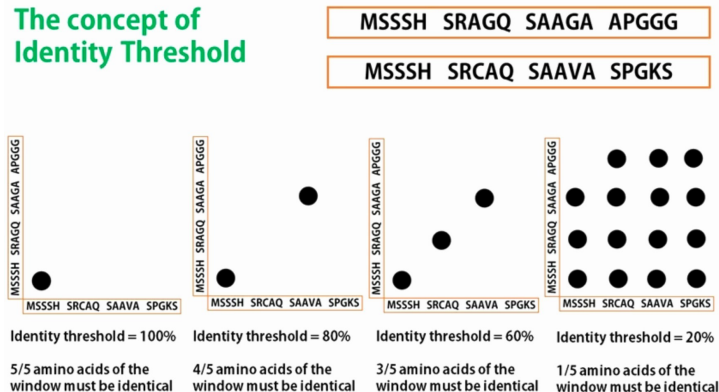


Figure: According to identity threshold, different dot plots are obtained.

Sequence alignment

Practice

Now, we are going to see how Dot matrix is compute using online tools. Follow the next steps:

- Download sample genomes.
- Visit the online tool.
- Interpret the results.

Sequence alignment

Practice

Visit this website in order to download the genomes: UniProt

The screenshot shows the UniProt website homepage. At the top, there's a navigation bar with links for BLAST, Align, Retrieval/ID mapping, and Peptide search. A search bar is prominently displayed with a dropdown menu set to 'UniProtKB'. Below the navigation bar, a mission statement reads: 'The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.'

The main content area is divided into several sections:

- UniProtKB** (UniProt Knowledgebase):
 - Swiss-Prot (561,911)**: Manually annotated and reviewed. Records with information extracted from literature and curator-evaluated computational analysis.
 - TrEMBL (177,754,527)**: Automatically annotated and not reviewed. Records that await full manual annotation.
- UniRef**: Sequence clusters.
- UniParc**: Sequence archive.
- Proteomes**: Proteome sets.
- Supporting data**:
 - Literature citations
 - Cross-ref. databases
 - Taxonomy
 - Diseases
 - Subcellular locations
 - Keywords
- News**:
 - Upcoming changes**: Planned changes for UniProt.
 - UniProt release 2020_01**: Coronavirus SARS-CoV-2 in UniProtKB | Changes to UniProt release cycle.
 - UniProt release 2019_11**: Thicker than water | Functional annotation of different gene products | Changes to FT and CC text format | Cross-references to RNaC | Pr...
 - News archive**

A red banner on the right side of the page reads: 'New UniProt portal for the latest SARS-CoV-2 coronavirus protein entries and receptors, updated independent of the general UniProt release cycle. View SARS-CoV-2 Proteins and Receptors'.

Figure: UniProt: Database of proteins and genomes.

Sequence alignment

Practice

Search the Filamin-A protein of a human species.

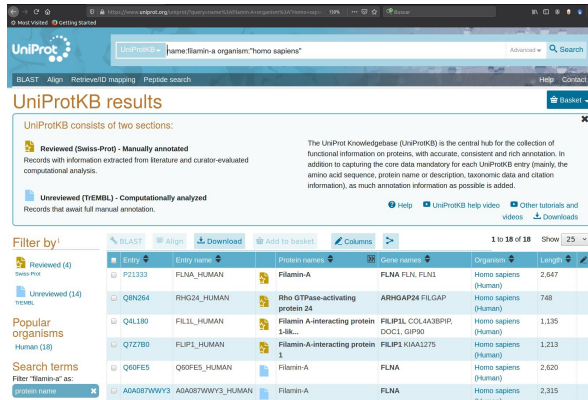
The screenshot shows the UniProt search interface in a web browser. The address bar displays <https://www.uniprot.org>. The page header includes the UniProt logo and navigation links: BLAST, Align, and Retrieve. A sidebar on the left contains the text "The mission of UniProt is to" and a box for "UniProtKB" (UniProt Knowledgebase) with "Swiss-Prot (561,911)" and a note "Manually annotated and reviewed". The main search area is titled "Searching in UniProtKB" with a "Help" link. It features three search criteria sections, each with a dropdown menu, a text input field, and a trash icon. The first section has a dropdown set to "Protein name [DE]" and the input field contains "filamin-A". The second section has a dropdown set to "AND", a dropdown set to "Organism [OS]", and the input field contains "Homo sapiens". The third section has a dropdown set to "AND", a dropdown set to "All", and an empty input field. A fourth section is partially visible below, also with a dropdown set to "AND", a dropdown set to "All", and an empty input field. A plus icon is located to the right of the fourth section.

Figure: Searching the Filamin-A protein of a human species

Sequence alignment

Practice

Select the sequence P21333 (its length is ~2.6).



UniProtKB results

UniProtKB consists of two sections:

- Reviewed (Swiss-Prot) - Manually annotated**
Records with information extracted from literature and curator-evaluated computational analysis.
- Unreviewed (TrEMBL) - Computationally analyzed**
Records that await full manual annotation.

The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added.

Filter by:

- Reviewed (4)
- Unreviewed (14)
- Popular organisms
- Human (18)
- Search terms
- Filter "filamin-a" as:
- protein name

1 to 18 of 18 Show 25

Entry	Entry name	Protein names	Gene names	Organism	Length
P21333	FLNA_HUMAN	Filamin-A	FLNA FLN, FLN1	Homo sapiens (human)	2,647
Q8N264	RHG24_HUMAN	Rho GTPase-activating protein 24	ARHGAP24 FILGAP	Homo sapiens (human)	748
Q4L180	FLN1_HUMAN	Filamin A-interacting protein 1-8k...	FILIP1L COL4A3BP1P, DOC1, GIP90	Homo sapiens (human)	1,135
Q7Z7B0	FILP1_HUMAN	Filamin-A-interacting protein 1	FILIP1 KIAA1275	Homo sapiens (human)	1,213
Q60FE5	Q60FE5_HUMAN	Filamin-A	FLNA	Homo sapiens (human)	2,620
ADA087WWY3	ADA087WWY3_HUMAN	Filamin-A	FLNA	Homo sapiens (human)	2,315

Figure: List of results

Sequence alignment

Practice

Look for the download button (Download isoform 1).

Sequences (2+)ⁱ

Sequence statusⁱ: Complete.

Sequence processingⁱ: The displayed sequence is further processed into a mature form.

This entry describes 2 isoformsⁱ produced by **alternative splicing**. [Align](#) [Add to basket](#)

This entry has 2 described isoforms and 6 potential isoforms that are computationally mapped. [Show all](#) [Align All](#)

Isoform 1 (identifier: **P21333-1**) [UniParc] [FASTA](#) [Add to basket](#)

This isoform has been chosen as the canonicalⁱ sequence. All positional information in this entry refers to it. This is also the downloadable versions of the entry.

[« Hide](#)

Figure: Download the sequence

Sequence alignment

Practice

Do the same operations for mouse species (Download the Q8BTM8 sequence, its length is ~2.6).

The screenshot shows the UniProtKB search interface. At the top, it says "Searching in UniProtKB" with a help icon. Below this, there are four search criteria, each with a dropdown menu for the search type, a text input for the term, and a trash icon to clear the entry.

- Criteria 1: Search type "Protein name [DE]", term "filamin-A".
- Criteria 2: Search type "AND", search type "Organism [OS]", term "mus musculus".
- Criteria 3: Search type "AND", search type "All", term (empty).
- Criteria 4: Search type "AND", search type "All", term (empty).

Each criteria has a trash icon to its right. The fourth criteria also has a plus icon to its right, indicating that more criteria can be added.

Figure: Searching the Filamin-A protein of a mouse species

Sequence alignment

Practice

There are several online tool to process Dot matrix:

- DotMatcher.
- EMBOSSS.

Sequence alignment

Practice

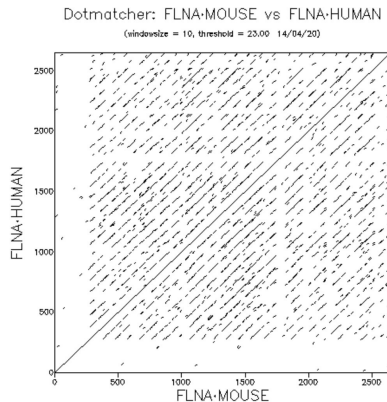






Figure: Dot matrix of Filamin-A protein in human and mouse species.

References I

-  A. A. Elfiky, “Anti-hcv, nucleotide inhibitors, repurposing against covid-19,” *Life sciences*, p. 117477, 2020.
-  X. Tang, C. Wu, X. Li, Y. Song, X. Yao, X. Wu, Y. Duan, H. Zhang, Y. Wang, Z. Qian *et al.*, “On the origin and continuing evolution of sars-cov-2,” *National Science Review*, 2020.
-  P. A. Zimmerman, A. Buckler-White, G. Alkhatib, T. Spalding, J. Kubofcik, C. Combadiere, D. Weissman, O. Cohen, A. Rubbert, G. Lam *et al.*, “Inherited resistance to hiv-1 conferred by an inactivating mutation in cc chemokine receptor 5: studies in populations with contrasting clinical phenotypes, defined racial background, and quantified risk,” *Molecular medicine*, vol. 3, no. 1, pp. 23–36, 1997.
-  J. Xiong, *Essential bioinformatics*. Cambridge University Press, 2006.

References II



A. J. Gibbs and G. A. McIntyre, "The diagram, a method for comparing sequences: Its use with amino acid and nucleotide sequences," *European journal of biochemistry*, vol. 16, no. 1, pp. 1–11, 1970.