Escuela Profesional de
Ciencia de la Computación

Bioinformatics

Sequence alignment

MSc. Vicente Machaca Arceda

Universidad Nacional de San Agustín de Arequipa

June 13, 2020

Table of Contents

Table of Contents

Introduction

Objectives

- Understand the importance of sequence alignment in Bioinformatics.

vmachacaa@unsa.edu.pe

Introduction

Objectives

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

vmachacaa@unsa.edu.pe

Introduction

Motivation

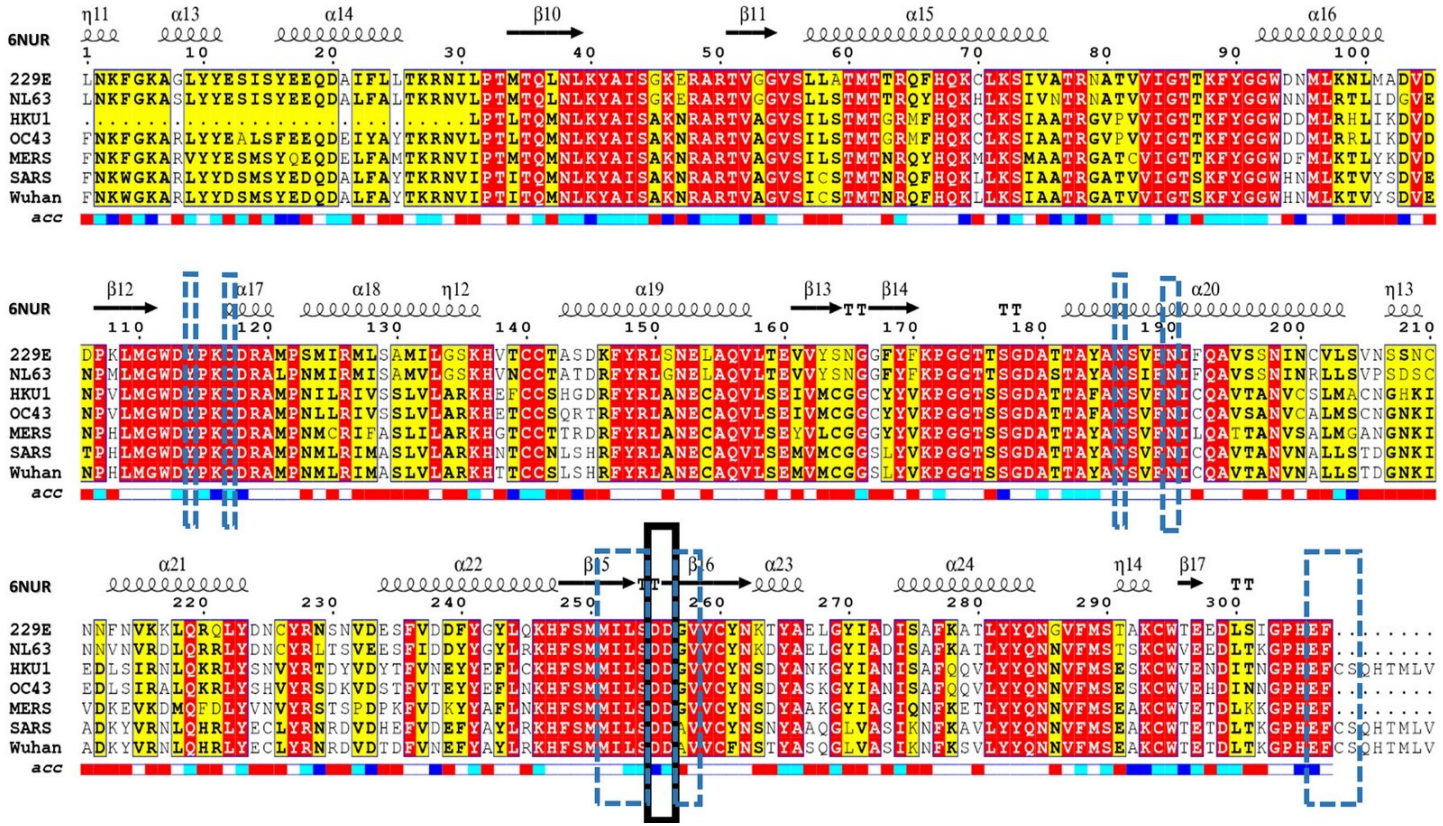

Figure: The SARS HCoV RdRp is the closest strain to the COVID-19, this in- formation is important for drug designers [1].

vmachacaa@unsa.edu.pe

Introduction

Motivation

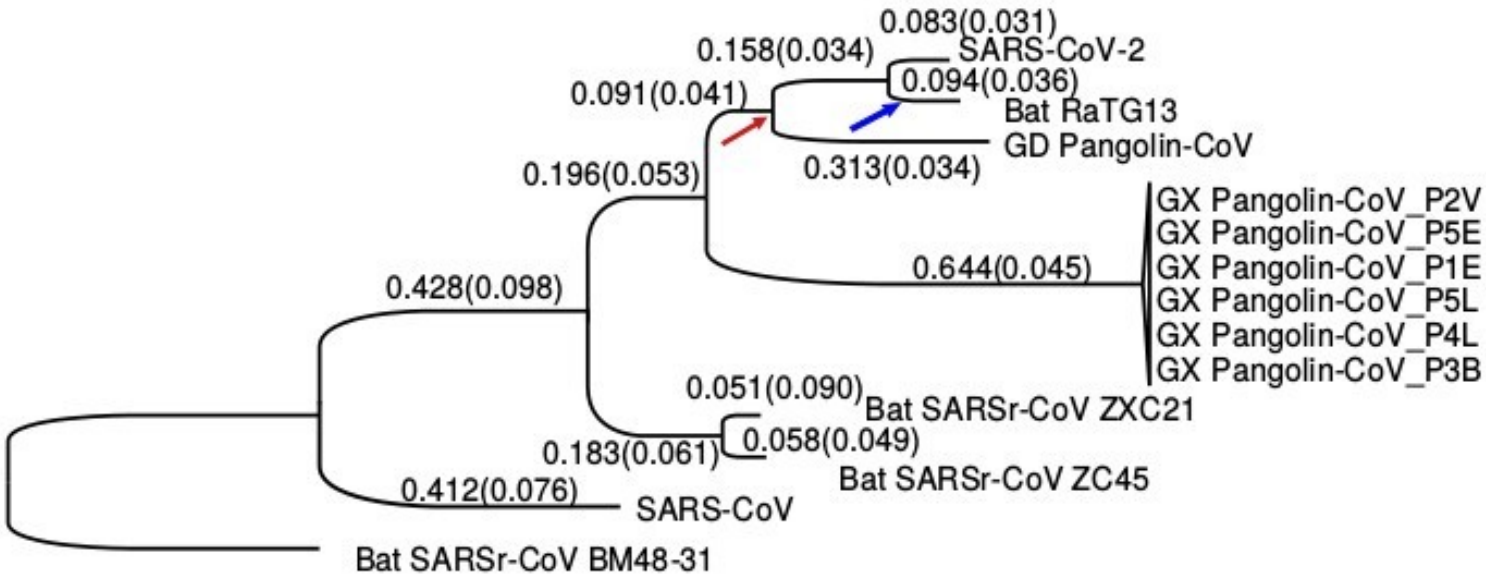

Figure: The phylogenetic tree of SARS-CoV-2 (COVID-19) and the related Coronaviruses [2].

vmachacaa@unsa.edu.pe

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

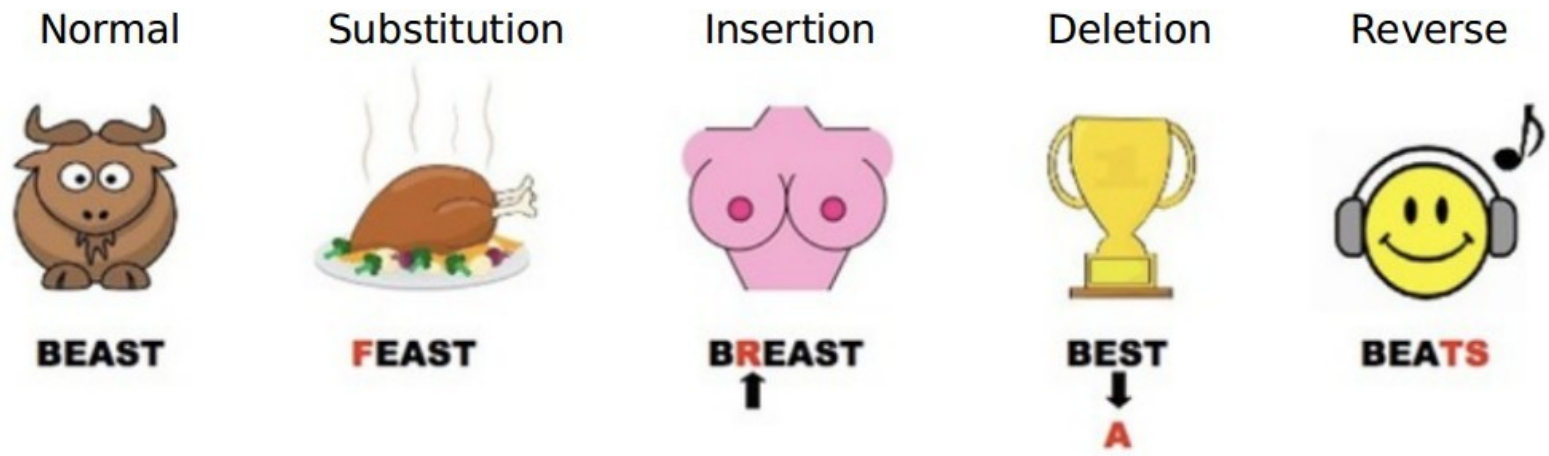

Figure: Overview of the Different Types of Point Mutations.

vmachacaa@unsa.edu.pe

Previous concepts

Genomic variations

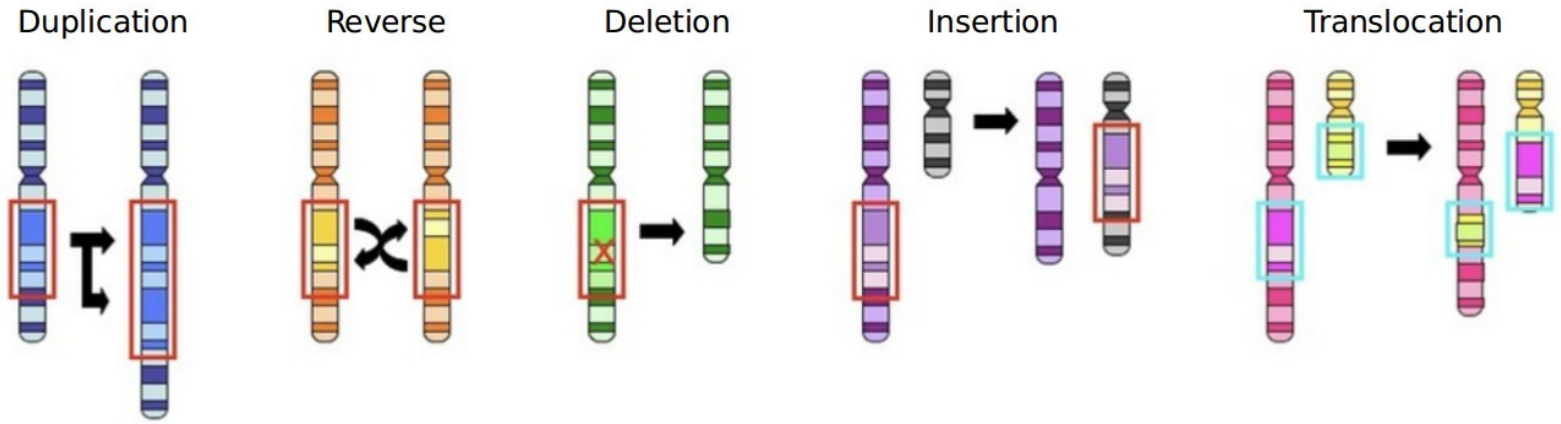

Figure: Overview of the Different Types of Block Mutations.

vmachacaa@unsa.edu.pe

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

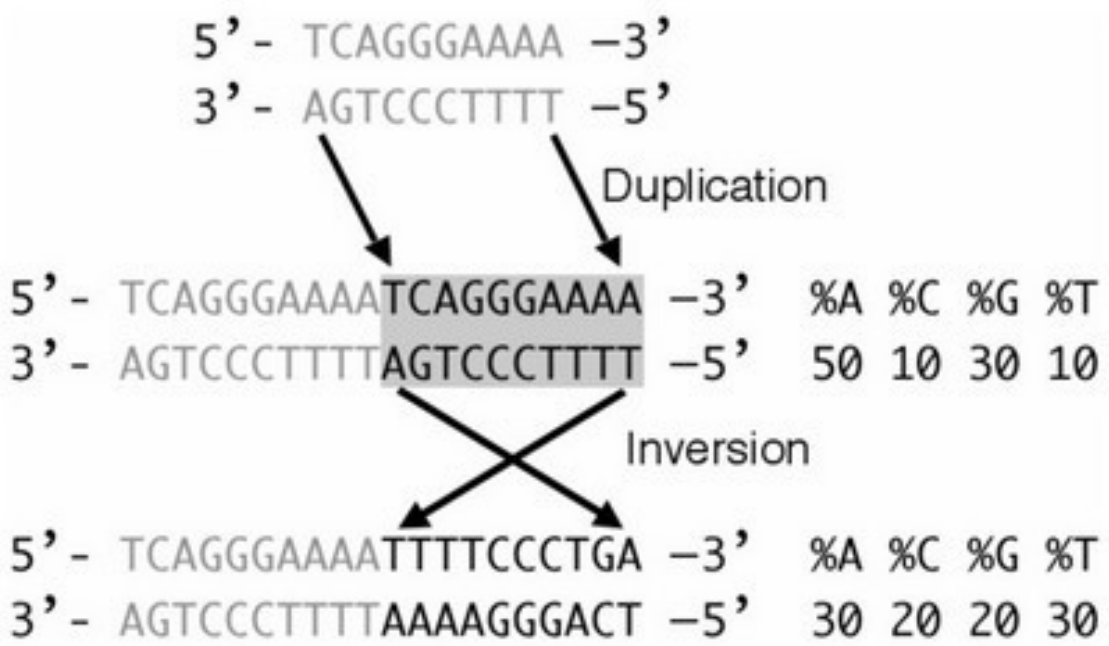

Figure: Inversion example.

vmachacaa@unsa.edu.pe

Previous concepts

Genomic variations

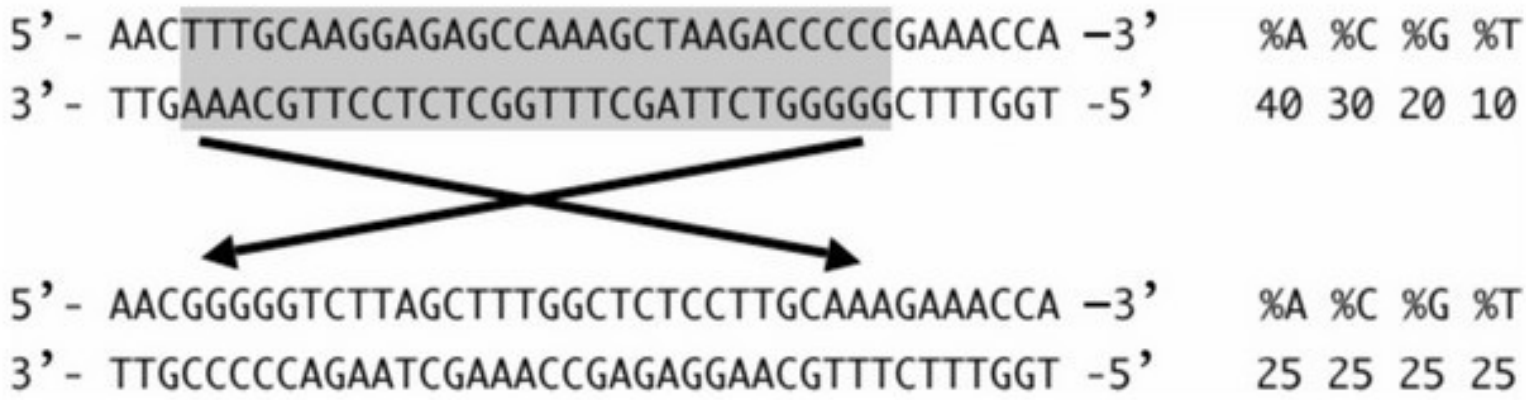

Figure: Inversion example.

vmachacaa@unsa.edu.pe

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

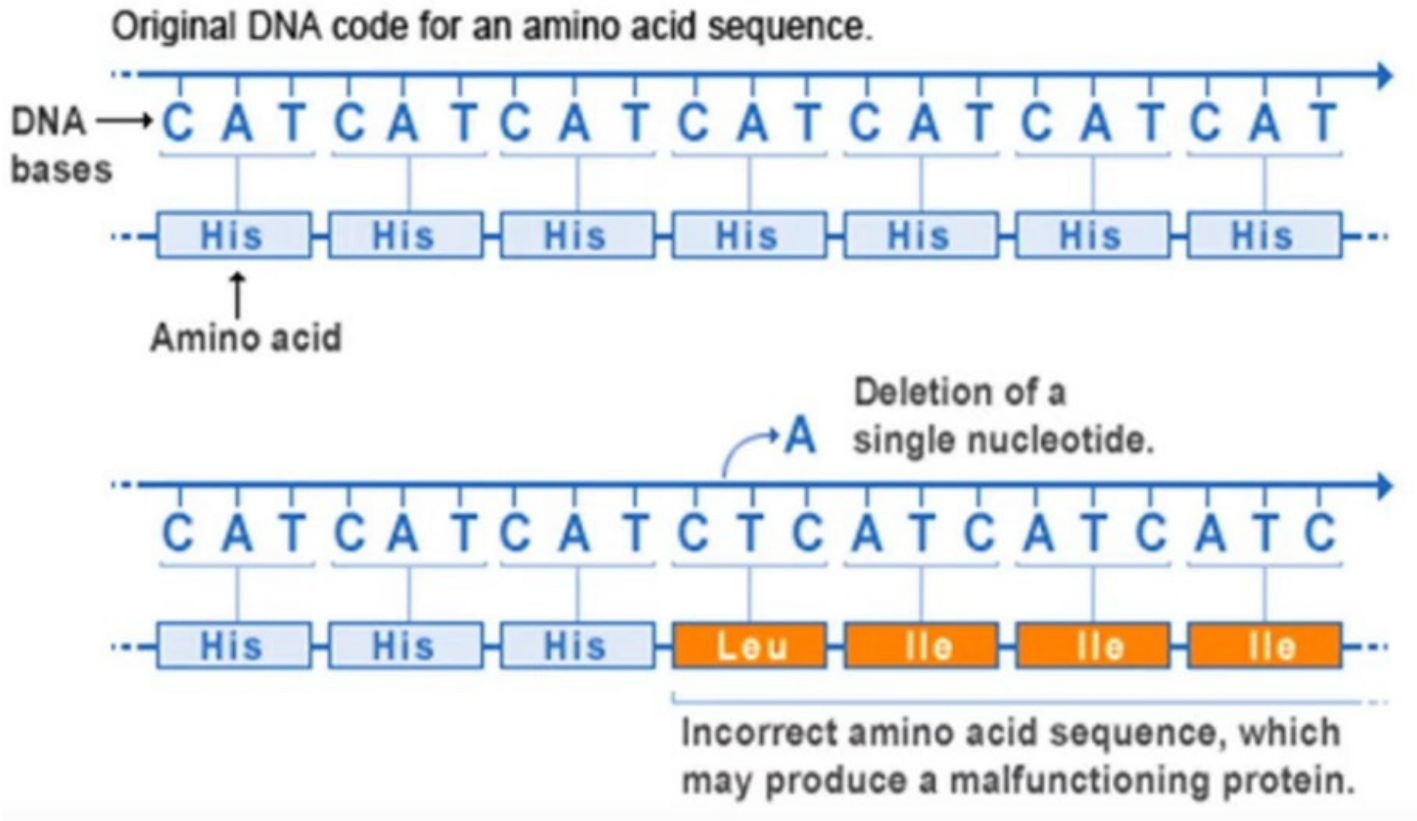

Figure: A frameshift mutation cause an incorrect amino acid sequence, which may produce a malfunctioning protein.

vmachacaa@unsa.edu.pe

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

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).

vmachacaa@unsa.edu.pe

Sequence alignment

Examples

```
No alignment

 CGATGCTAGCGTATCGTAGTCTATCGTAC
              |     ||
 ACGATGCTAGCGTTTCGTATCATCGTA

Alignned

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

Figure: No alignment versus alignment.

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

vmachacaa@unsa.edu.pe

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
```

Figure: Alignment and gaps.

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

vmachacaa@unsa.edu.pe

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 missmatches.
- Number of gaps.
- Length of the gaps.

vmachacaa@unsa.edu.pe

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

vmachacaa@unsa.edu.pe

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

Figure: Global Alignment and Local Alignment.

vmachacaa@unsa.edu.pe

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.

vmachacaa@unsa.edu.pe

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

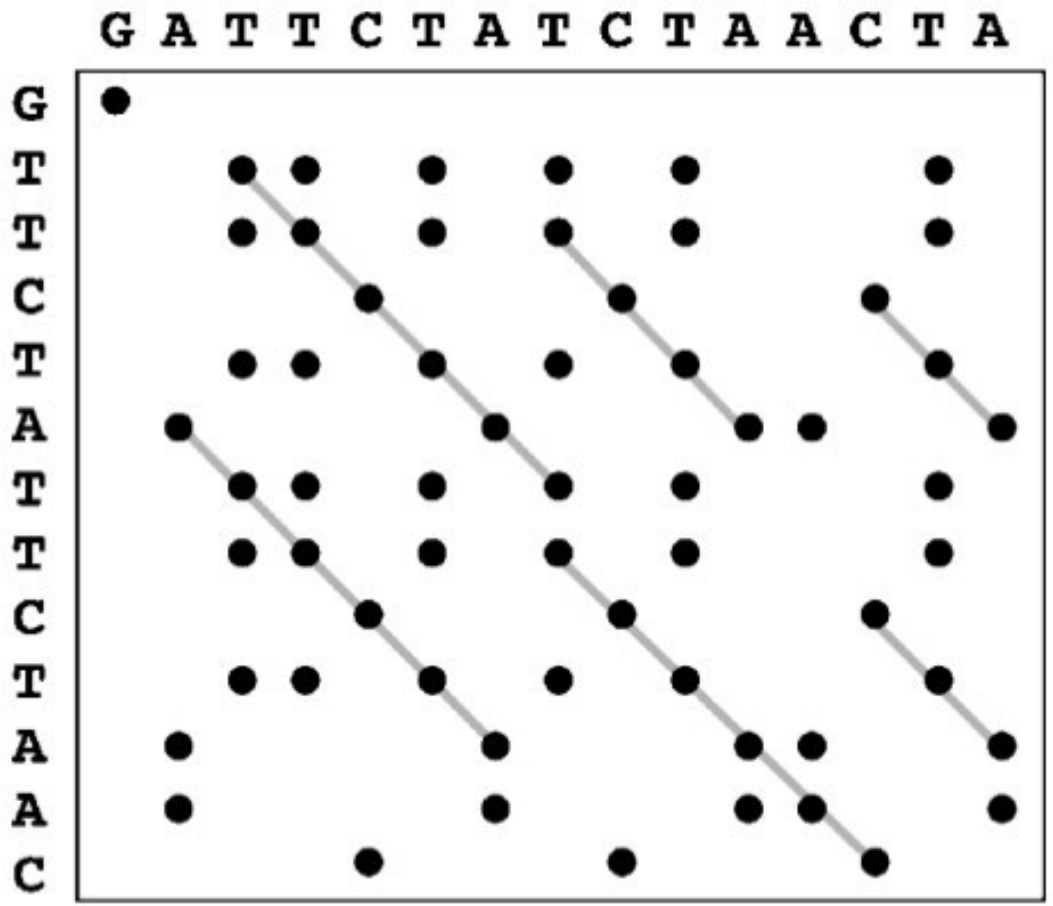

Figure: Dot matrix example

vmachacaa@unsa.edu.pe

Sequence alignment

Dot matrix

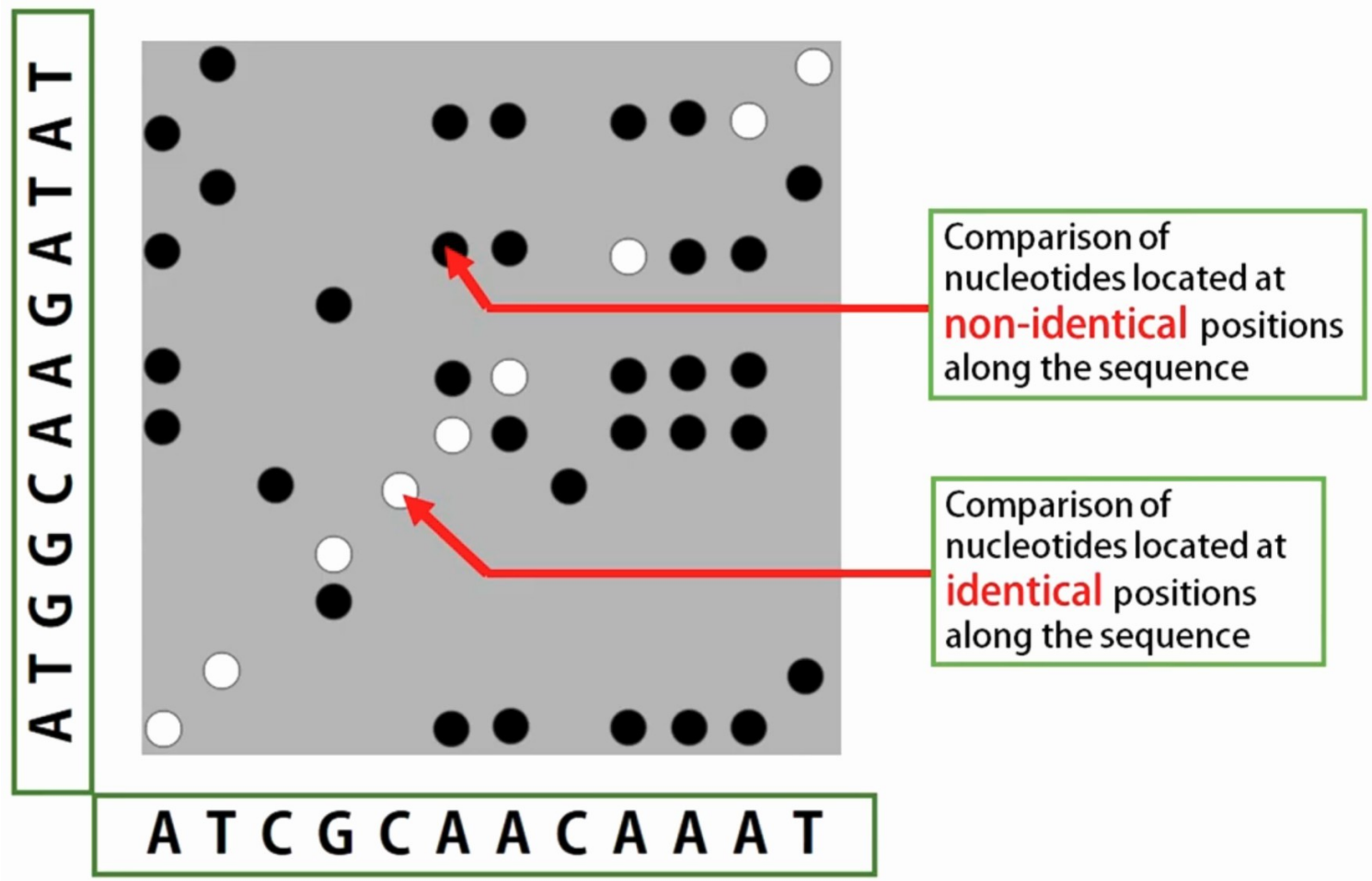

Figure: Dot matrix example

vmachacaa@unsa.edu.pe

Sequence alignment

Dot matrix

What we could conclude from the Dot plot?

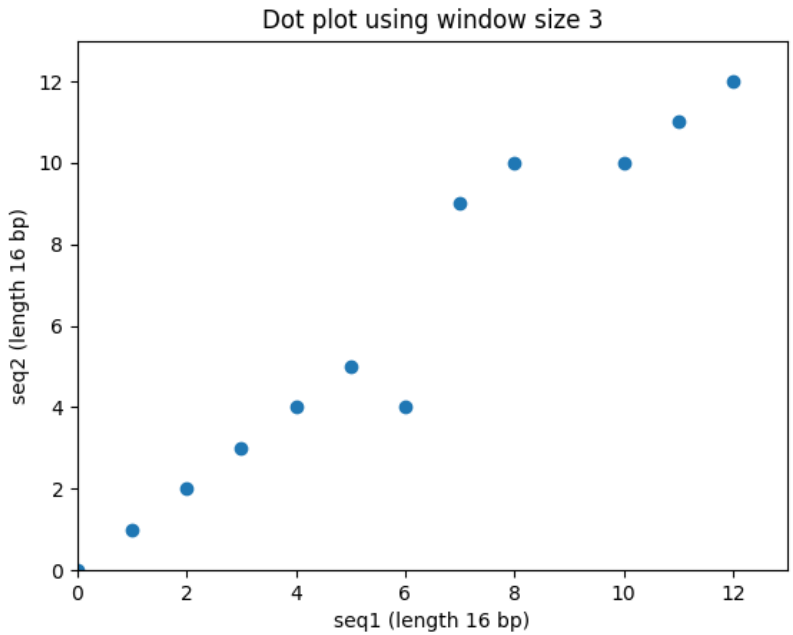

Figure: Dot matrix example. *seq*1 = *ACCTGAGAGTGTGGCT* and *seq*2 = *ACCTGAGACAGTGGCT*

vmachacaa@unsa.edu.pe

Sequence alignment

Dot matrix

What we could conclude from the Dot plot?

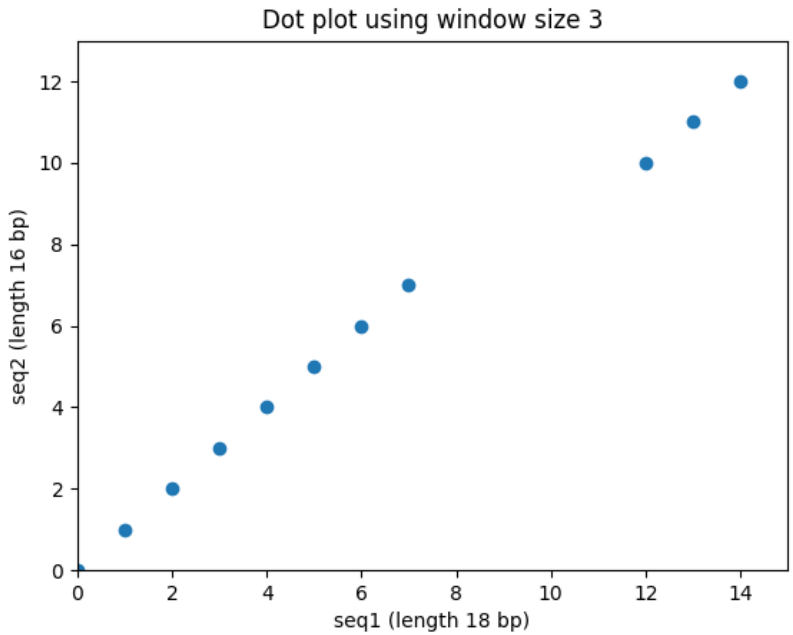

Figure: Dot matrix example. *seq*1 = *ACCTGAGACATTGTGGCT* and *seq*2 = *ACCTGAGACAGTGGCT*

vmachacaa@unsa.edu.pe

Sequence alignment

Dot matrix

What we could conclude from the Dot plot?

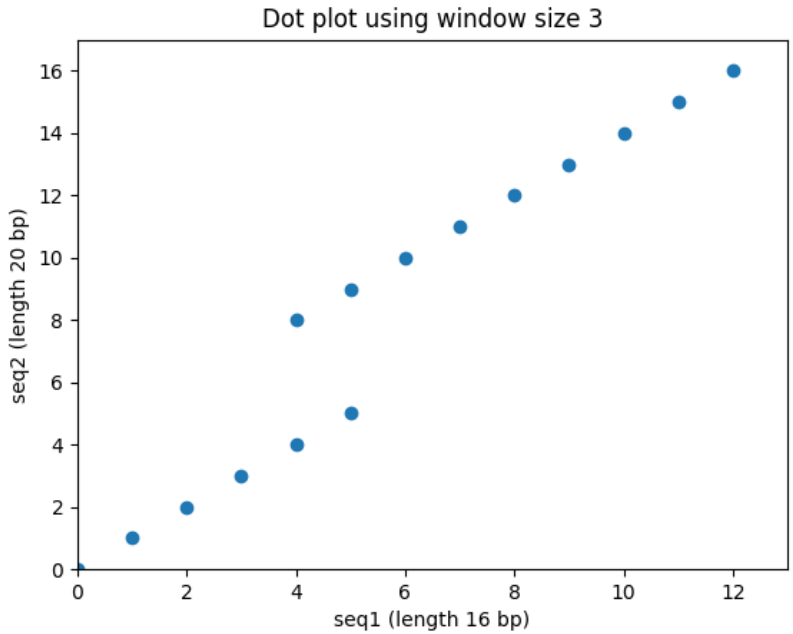

Figure: Dot matrix example. *seq*1 = *ACCTGATACAGTGGCT* and *seq*2 = *ACCTGATAGATACAGTGGCT*

vmachacaa@unsa.edu.pe

Sequence alignment

Dot matrix

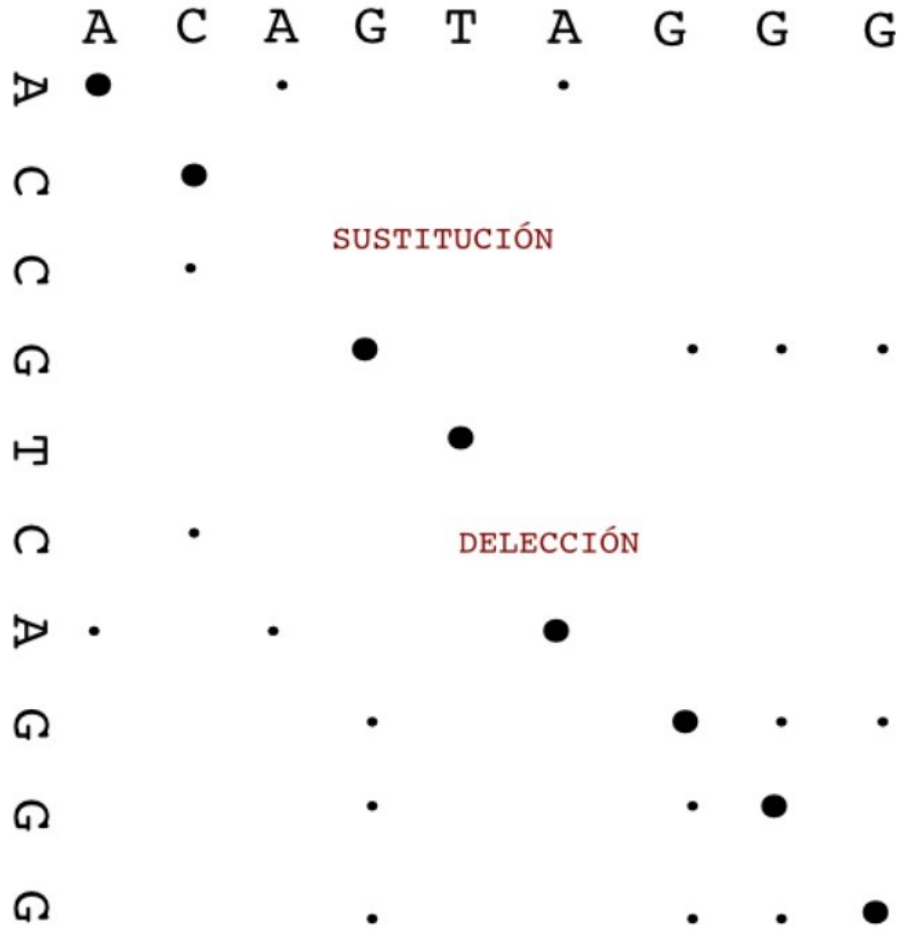

Figure: Dot matrix example

vmachacaa@unsa.edu.pe

Sequence alignment

Dot matrix

Figure: Deletion/insertion example in Dot matrix.

vmachacaa@unsa.edu.pe

Sequence alignment

Dot matrix

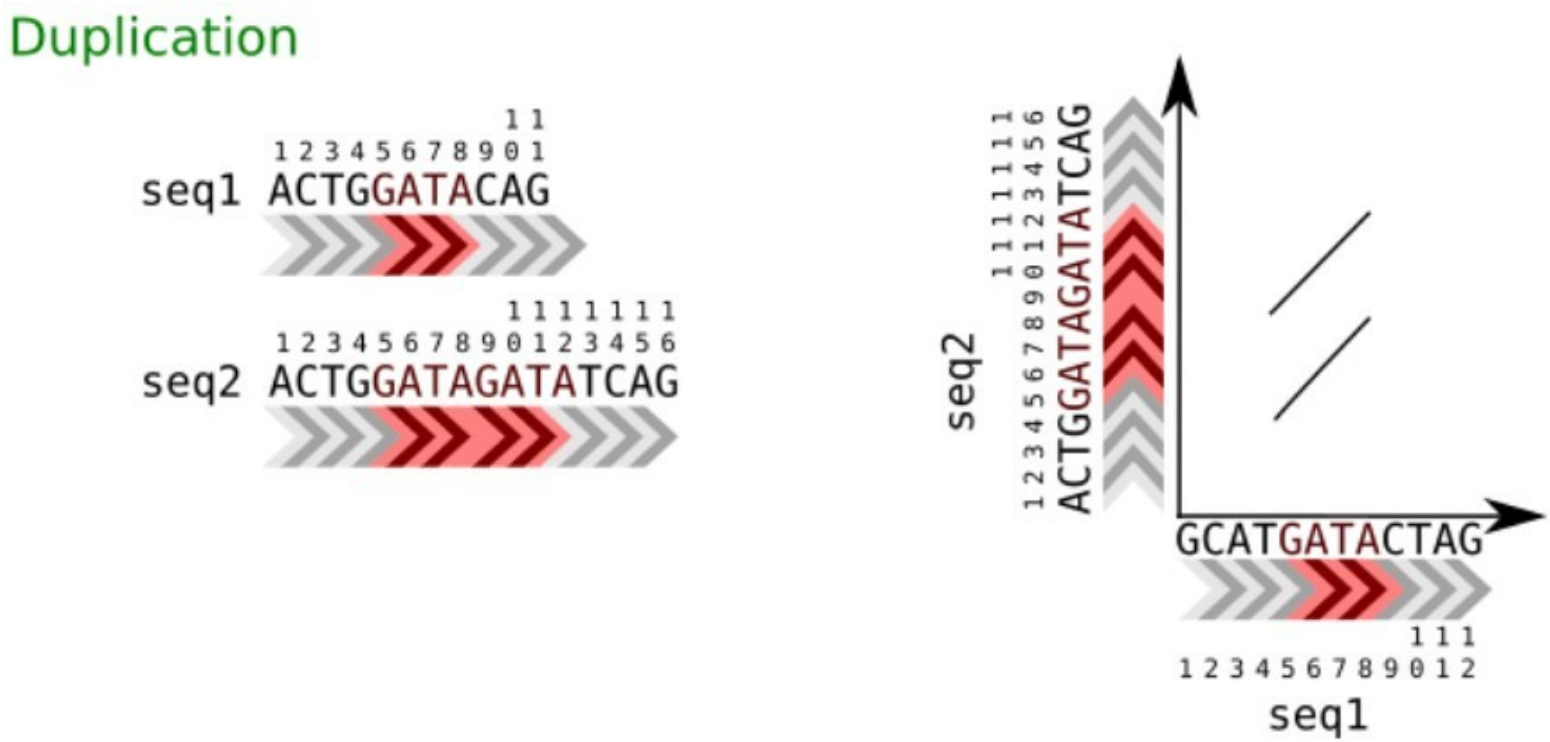

Figure: Duplication example in Dot matrix.

vmachacaa@unsa.edu.pe

Sequence alignment

Noise in Dot matrix

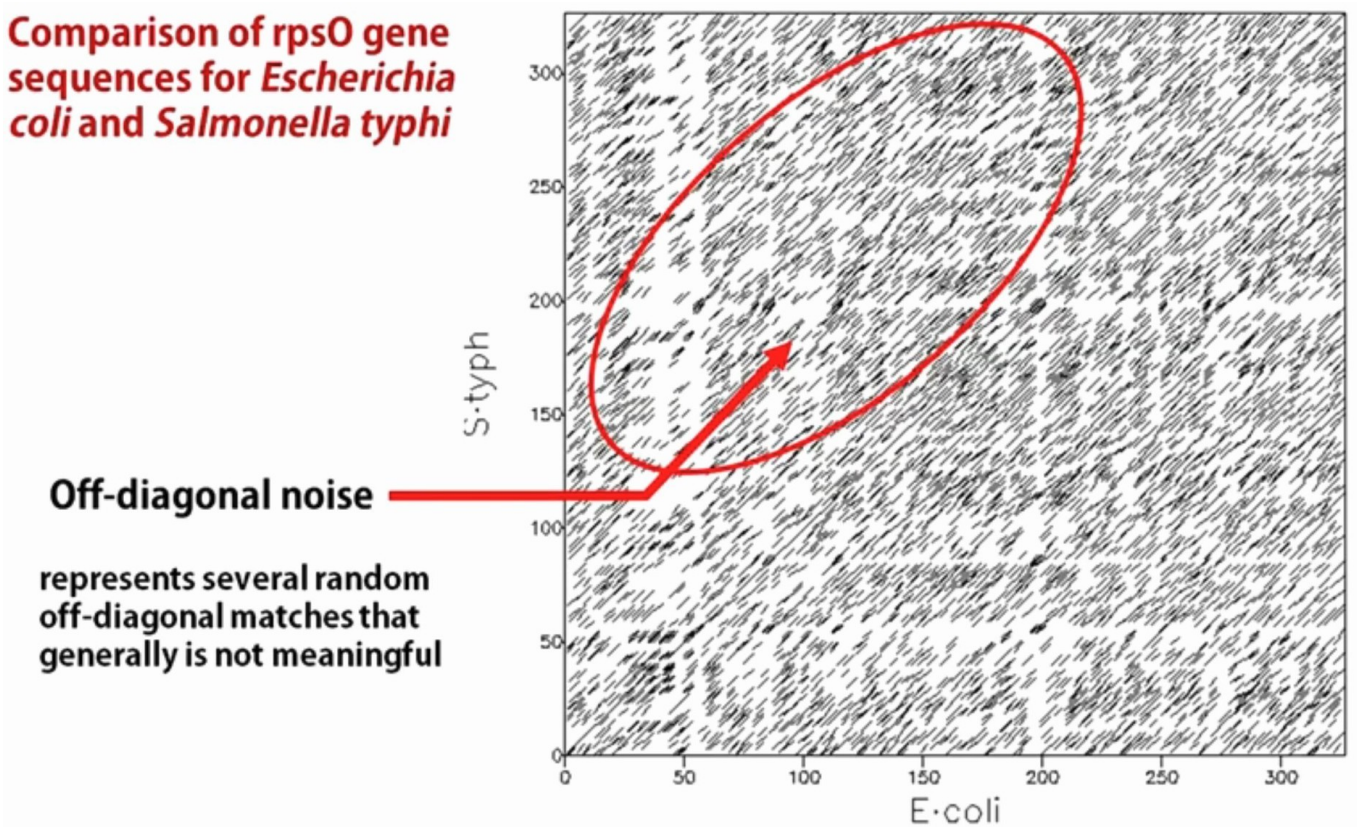

Figure: Several off-diagonal matches that generally is not meaningfull.

vmachacaa@unsa.edu.pe

Sequence alignment

Dot matrix with window size

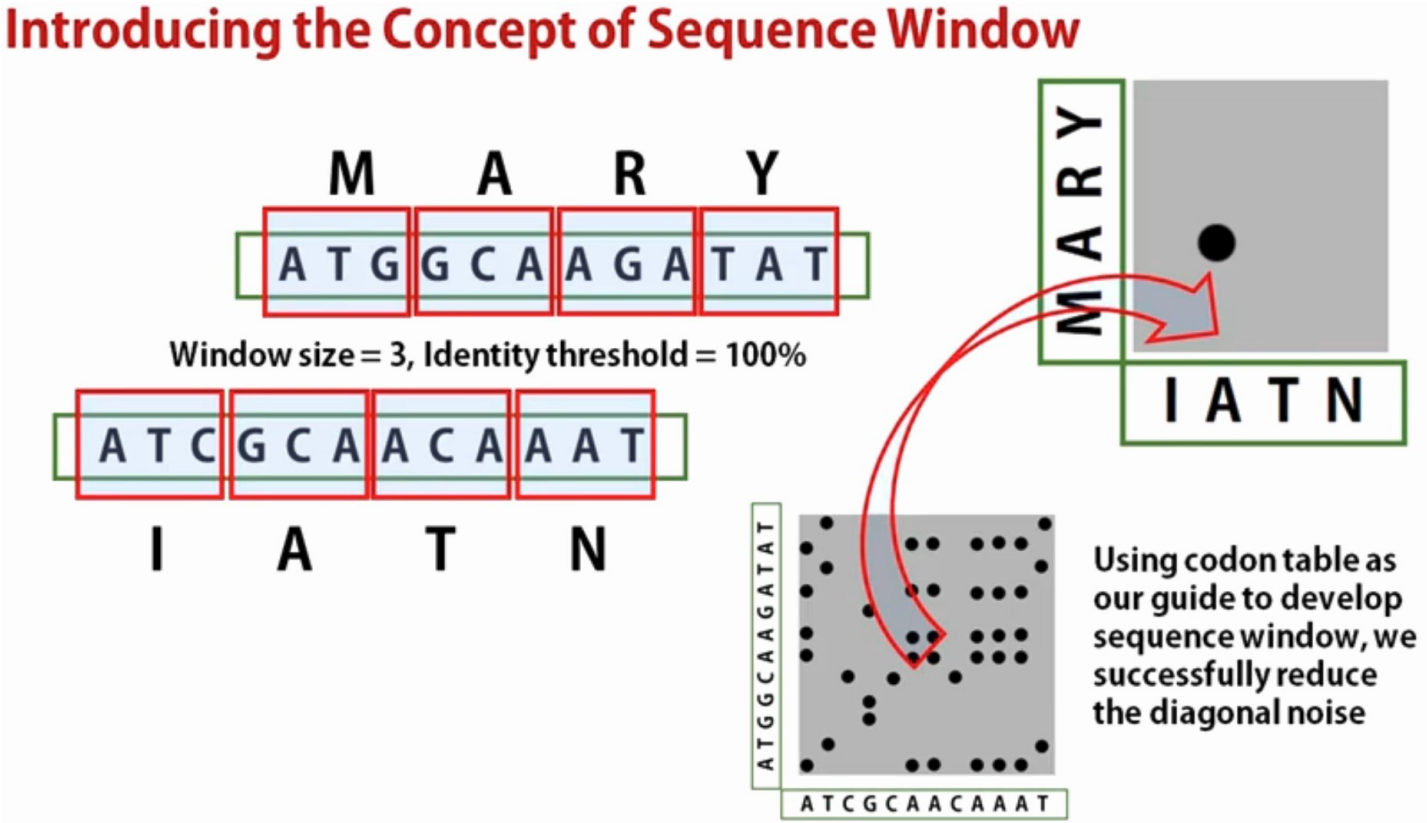

Figure: Dot plot with windows successfully reduce the diagonal noise.

vmachacaa@unsa.edu.pe

Sequence alignment

Dot matrix with window size

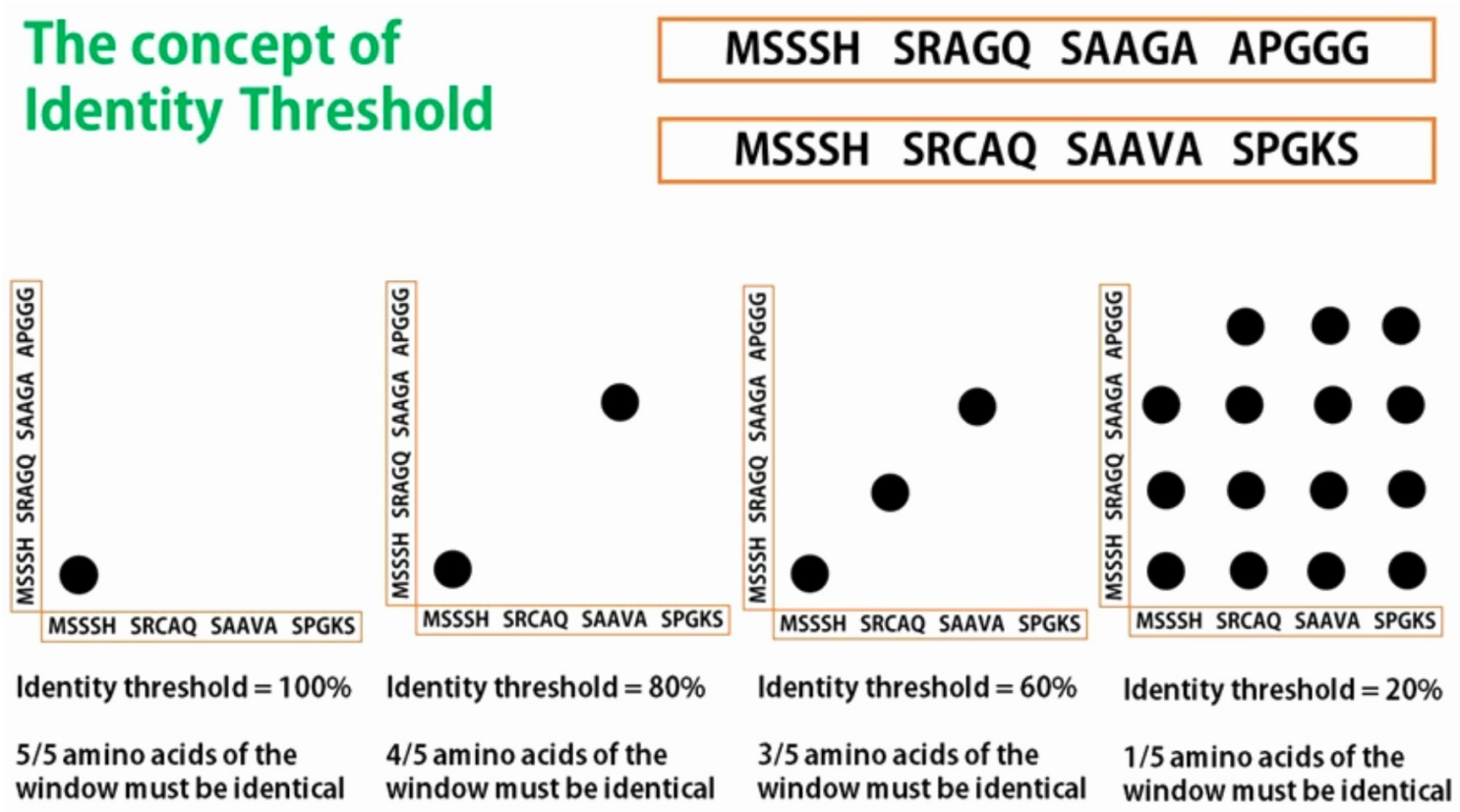

Figure: According to identity threshold, different dot plots are obtained.

vmachacaa@unsa.edu.pe

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.

vmachacaa@unsa.edu.pe

Sequence alignment

Practice

Visit this website in order to download the genomes: UniProt

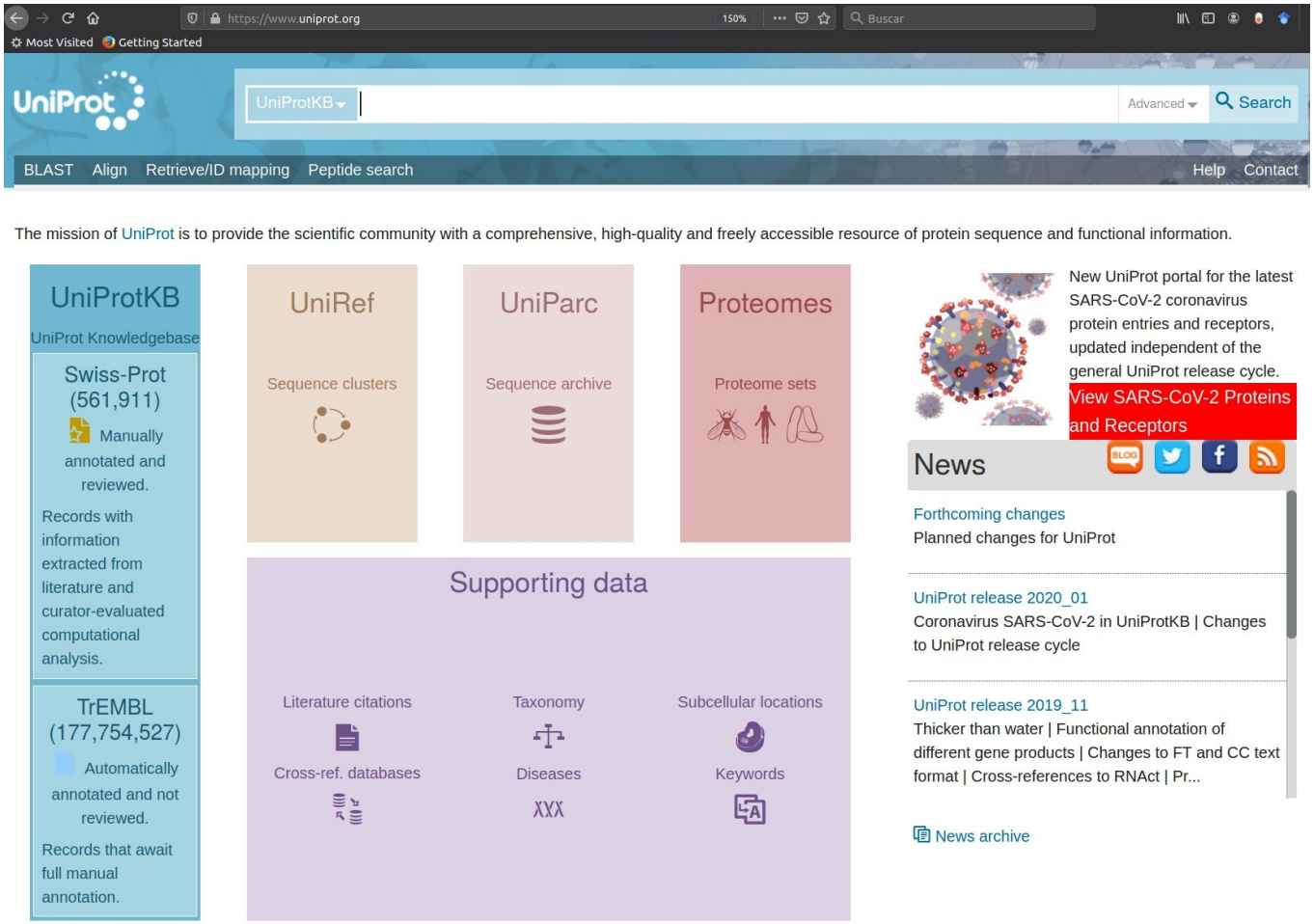

Figure: UniProt: Database of proteins and genomes.

Sequence alignment

Practice

Search the Filamin-A protein of a human species.

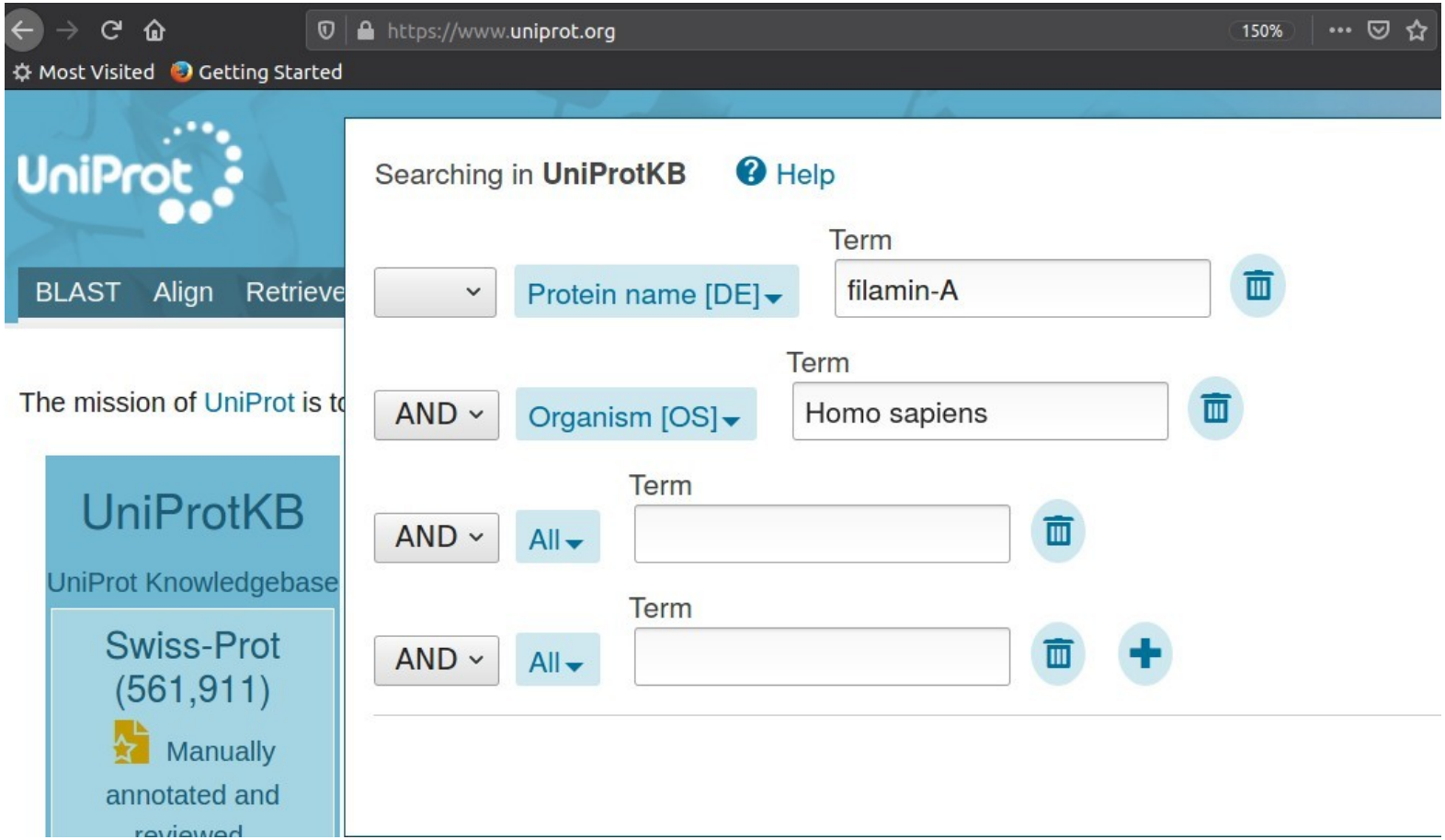

Figure: Searching the Filamin-A protein of a human species

vmachacaa@unsa.edu.pe

Sequence alignment

Practice

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

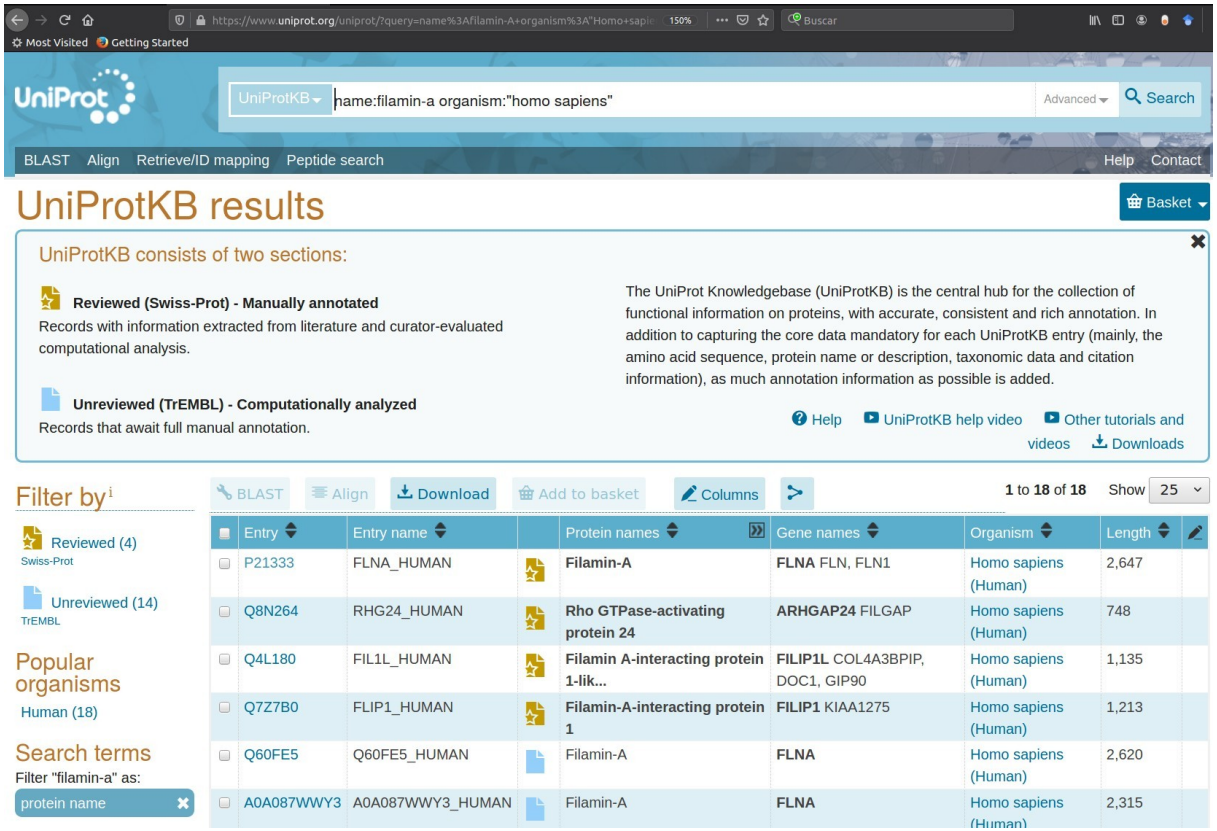

Figure: List of results

vmachacaa@unsa.edu.pe

Sequence alignment

Practice

Look for the download button (Download isoform 1).

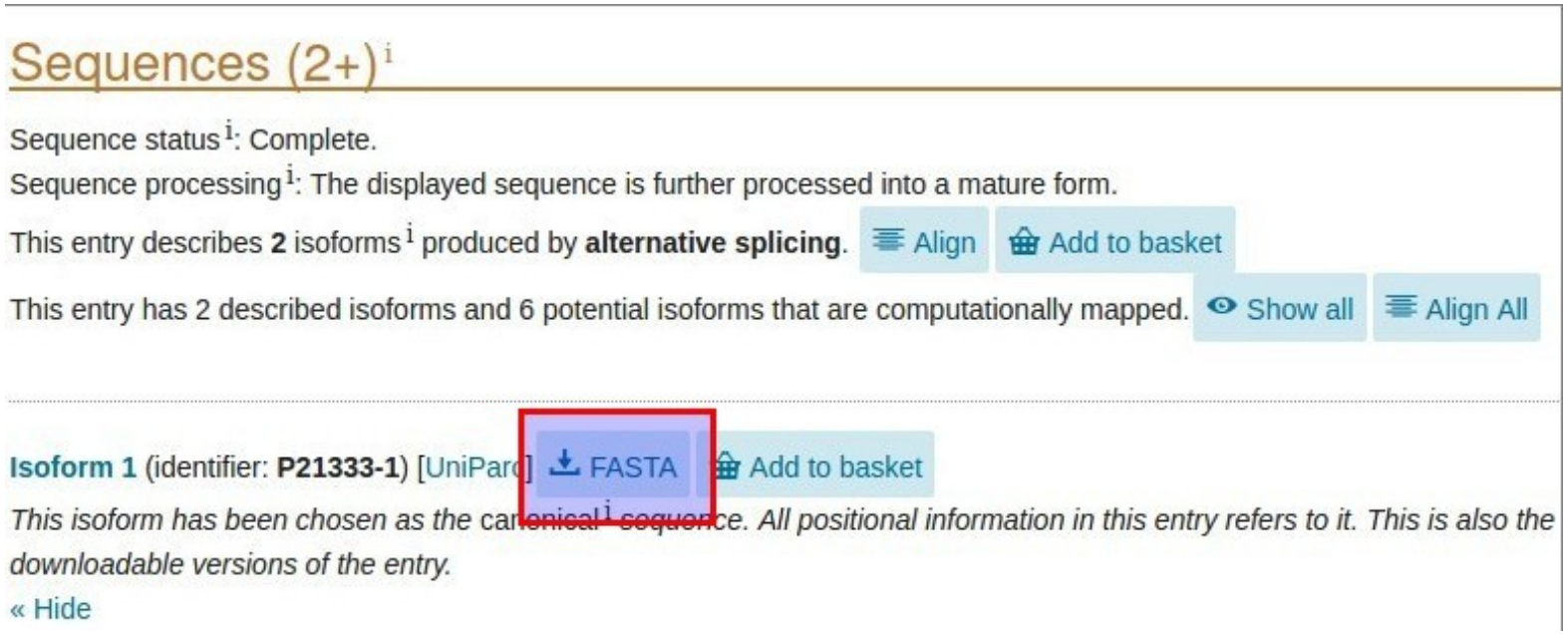

Figure: Download the sequence

vmachacaa@unsa.edu.pe

Sequence alignment

Practice

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

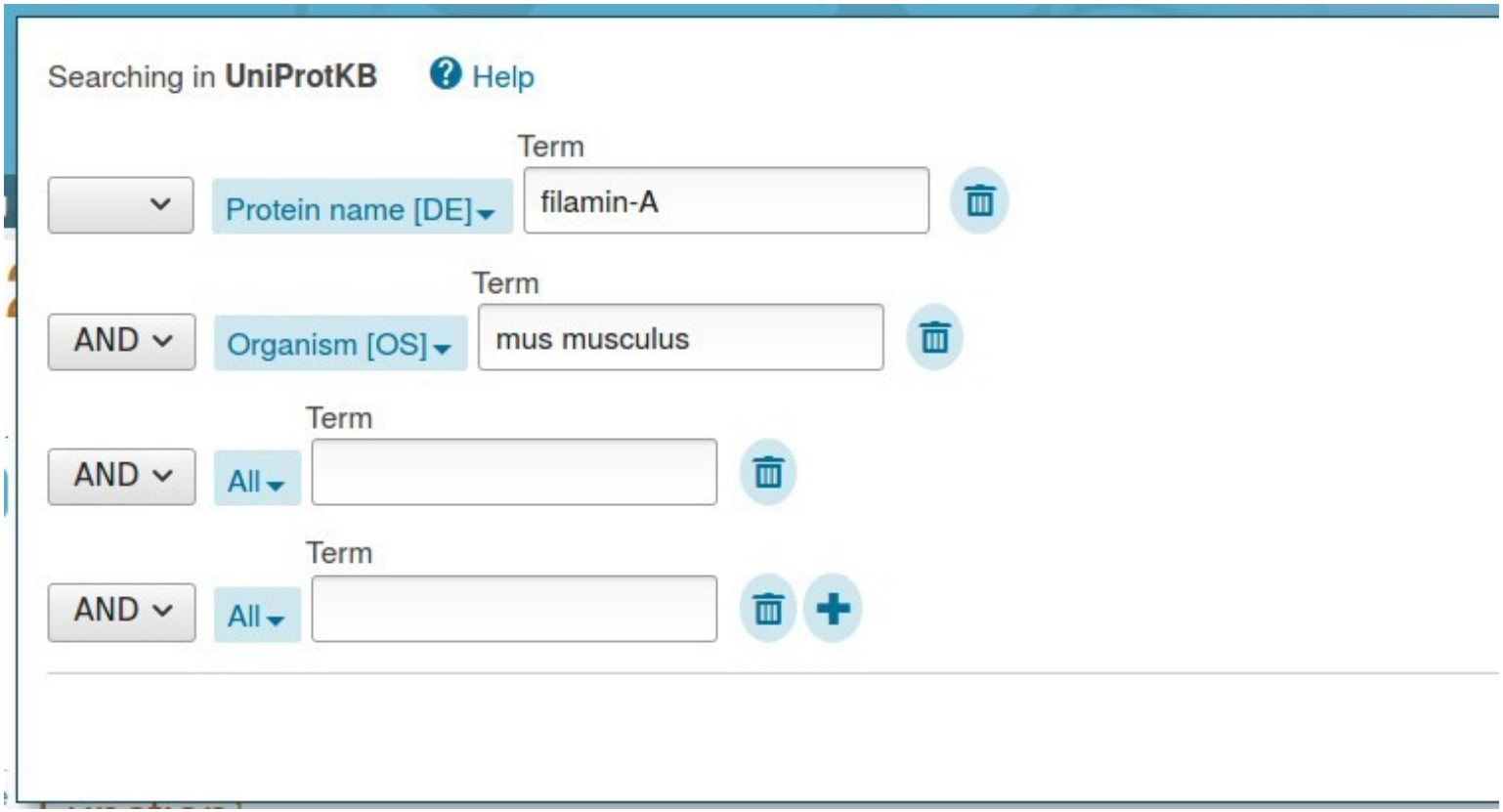

Figure: Searching the Filamin-A protein of a mouse species

vmachacaa@unsa.edu.pe

Sequence alignment

Practice

There are several online tool to process Dot matrix:

- DotMatcher.
- EMBOSSS.

vmachacaa@unsa.edu.pe

Sequence alignment

Practice

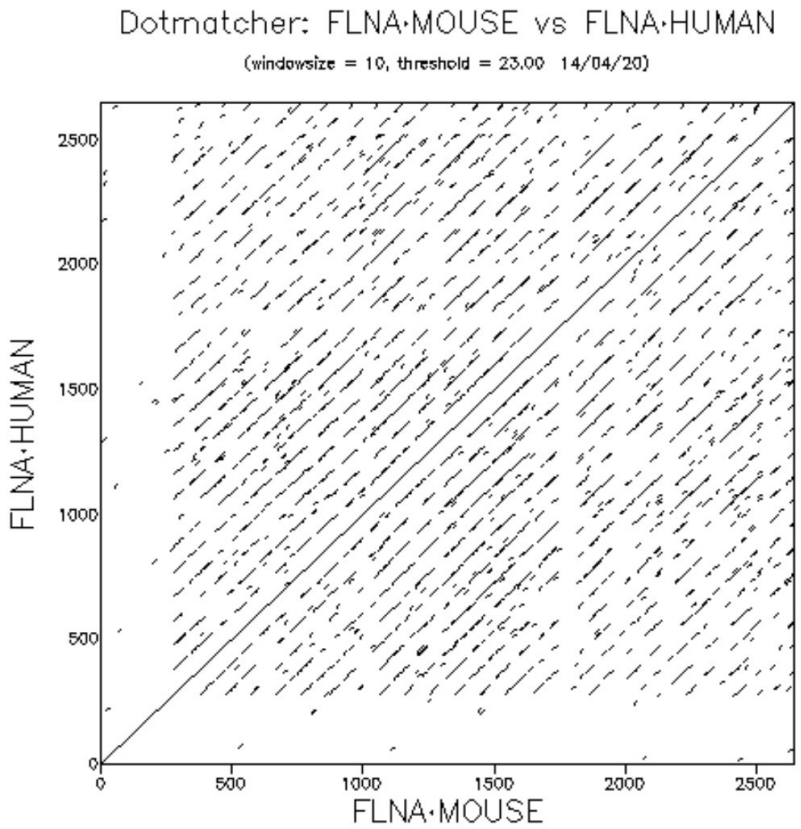

Figure: Dot matrix of Filamin-A protein in human and mouse species.

vmachacaa@unsa.edu.pe

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.

vmachacaa@unsa.edu.pe

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.

vmachacaa@unsa.edu.pe