Identification of genomes: Clustal Omega and BLAST: One introduction

Carlos Navarro Venegas *

*Faculty of Veterinary and Livestock Sciences (FAVET). University of Chile, Chile.

Abstract
Knowing the sequence of a pathogen isolated in our laboratory is currently an accessible task, and there are no excuses to start it.

The remarkable occurrence of Kary Mullis and the existence of computer programs has helped a lot in this task and in this introductory article we give an account of that and invite those who still do not dare to this adventure.

Keywords: Genomes; Nucleotide sequence; Clustal Omega; BLAST

1. Introduction
The genome is generally DNA except for one type of virus: the Ribovirus whose genome is RNA (existing single-stranded and double-stranded).

Knowing the foregoing and the existence of the Genbank® database (https://www.ncbi.nlm.nih.gov/genbank/ ), today there are no excuses for the detection of any pathogen or for identifying which genotype (for example) would belong [1, 2, 3, 4].

The identification of the genome begins with its detection, using the incredible occurrence of Kary Mullis (5), obtaining double-stranded DNA fragments by means of the appropriate PCR (RT-PCR in the case of most RNA viruses).

2. Material and Methods
Once these fragments have been obtained, it is necessary to know their nucleotide sequence and due to the inevitable errors inherent in the laboratory, it is preferred to send each sample to be sequenced in triplicate (minimum).

3. Results
Thus, for each sample three (at least) sequences are received. So, it is necessary to know the “average” sequence known as the consensus sequence and for that the CLUSTAL OMEGA software (https://www.ebi.ac.uk/Tools/msa/clustalo/ ) can be used. When entering the sequences into the program, an alignment is obtained, which will indicate by means of a "*" where there is 100% nucleotide identity for each nucleotide position, for example (Table 1).
Table 1 Results of alignment from Clustal Omega use

<table>
<thead>
<tr>
<th>Sample</th>
<th>Sequence</th>
<th>Alignment Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3</td>
<td>TTAATATGCTGGAGTGAAAATATTAAAGTAGTTGTGTAATAATATGGATTAACCT</td>
<td>60</td>
</tr>
<tr>
<td>S1</td>
<td>TTAATATGCTGGAGTGAAAATATTAAAGTAGTTGTGTAATAATATGGATAAAACT</td>
<td>60</td>
</tr>
<tr>
<td>S2</td>
<td>TTAATATGCTGGAGTGAAAATATTAAAGTAGTTGTGTAATAATATGGATAAAACT</td>
<td>60</td>
</tr>
</tbody>
</table>

Here, S1, S2 and S3 correspond to the sequences of the same sample and the number of asterisks show that there is "a great resemblance" between the sequences (obviously, asterisks should appear in all positions: it is the same sample). We must build the consensus sequence by assigning a nucleotide in the place where the "*" does not appear, considering the one that is repeated the most in that position, in this case it corresponds to:

Consensus sequence: TTAATATGCTGGAGTGAAAATATTAAAGTAGTTGTGTAATAATATGGATAAAACT

This program (CLUSTAL OMEGA) can align several sequences at once and each sequence can be of variable length.

The constructed consensus sequence represents the work done in the laboratory, but... What does it correspond to? Does it correspond to the genome we want to detect? ... Let’s see and let nothing be left up in the air (minute 5.5)!!!

The next stage contemplates the use of another free online software up to now: BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome) a very powerful tool that indicates the percentage of nucleotide identity (PIN) of the entered sequence with respect to ALL the available database. That is, it tells us with which genome it has the greatest “coincidence”. In this output, the program shows the first 100 alignments and each PIN.

For example, (Figure 1) inputting the newly constructed consensus sequence to BLAST yields the following output:

![Figure 1 Results from BLAST use](image)

4. Discussion

That is, the consensus sequence has a PIN=98.36% with respect to the Feline Pan leukopenia virus, referring to several viral isolates and strains described in the Genbank ® database.

Hmmm... But only 10 results are shown...(Figure 1)
Only for reasons of space, the remaining 90 results are not shown, but the last 10 results indicate a PIN of 98.36%. Undoubtedly... it is the virus already mentioned...!!!(Figure 2)

![Table of results](image)

**Figure 2** Continuation of results from BLAST use

Thus, with the use of these biotools in our laboratory we have detected and identified several pathogens of veterinary interest [6, 7, 8, 9, 10].

5. Conclusion

Both programs can significantly help us to scrutinize the nucleotide sequence of any pathogen, in our case, and compare it with the official data stored in Genbank® (https://www.youtube.com/watch?v=LcvNpwS8hu8).

Compliance with ethical standards

Acknowledgments

We thank the FAVET students (for achieving part of their dreams with us) and the Wolf Foundation, Illinois, USA (since 2020).

References


[10] Vera C, Jara MA, Navarro C. A preliminary study for the F gen of Canine Distemper Virus as target for phylogenetic analysis. GSC Advanced Research and Reviews. 2022, 10(01); 105–118