# Using Protein Sequence Analysis to Identify Possible Cross-Species Hosts

By R. Urdiales Muñoz, M. A. Grela Hernandez, J. R. Alonso Fernandez.

## Index:

Introduction ..... 2
Abstract ..... 2
SARS-CoV-2 ..... 3
Context and information ..... 3
Analysis workflow ..... 3
Listing similar species ..... 3
Finding evidence for cross-species infection ..... 3
Comparing both the groups of proteins ..... 4
Results comprehension ..... 5
MERS-CoV ..... 6
Context and information ..... 6
Work and study to find reservoirs of the Mers-CoV virus ..... 6
Mers-CoV and and its relationship with animals and humans ..... 7
Analysis workflow ..... 7
Scores of similitud between species and human protein DPP4 ..... 7
Possible reservoirs and interpretation ..... 9
Analysis results ..... 9
Human coronavirus 229E ..... 11
Context and information ..... 11
Analysis workflow ..... 11
Analysis results ..... 12
Conclusions ..... 14

## Introduction

With the last pandemic the concern about animal reservoirs of viruses has increased. For that purpose, several studies have been made, from studying the interaction between the spike protein of those viruses to trying to identify the possible hosts one by one.

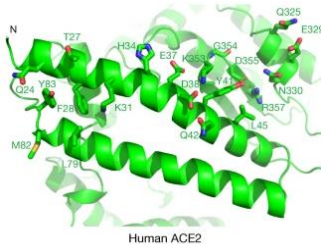
## Abstract

In this homework we propose another way to tighten the possible reservoir by analysing the similitudes between the host receptor proteins for the virus between species where cross-species transmission has already happened and their differences to host receptor proteins for species known to evade that transmission.

We have chosen 3 different human coronaviruses and made three independent implementations of the same idea for their human receptor proteins, so we can compare the results and accuracy for this type of analysis.

# SARS-CoV-2

## Context and information




SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2), which causes the last pandemic's respiratory syndrome known as COVID19. Said virus mainly spreads by air, infecting cells with the ACE2 (angiotensin converting enzyme 2) membrane protein which acts as a receptor for the virus's S (spike) protein. The residues starting the interaction are known to be mainly at the start of the protein, in range (28,42) as shown in the below article.

[Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor]

Additionally, it is known that human cells infected by covid synthetise a variant of the protein without the first 325 residues, outpointing the importance of this domain.

## Analysis workflow

### Listing similar species

The first step to obtain the list of candidate species was to make a protein BLAST against all similar sequences from other species. To accomplish that the sequence for the human protein and for other similar ones were downloaded from UniProt, and loaded into galaxy to perform the BLAST.

As this is a trial for future cases, an approach on how to optimize the runtime was taken into consideration – instead of directly blasting against all possible protein sequences, the representatives for already known clusters of proteins at UniProt were used beforehand to prefilter the rest of the sequences.

The results include these species as the most similar to human regarding ACE2 sequence:

| Species name | Common name | Best BLAST score |
|---|---|---|
| *Gorilla gorilla gorilla* | western lowland gorilla | 1671 |
| *Pan troglodytes* | chimpanzee | 1670 |
| *Pan paniscus* | bonobo | 1669 |
| *Pongo abelii* | sumatran orangutan | 1660 |
| *Nomascus leucogenys* | northern white-cheeked gibbon | 1651 |
| *Piliocolobus tephrosceles* | ugandan red colobus | 1615 |

This and related data are available as flow1 at Galaxy Europe and GitHub (which includes download scripts for automation).

### Finding evidence for cross-species infection

Because the data is mainly today still not linked and interposable, this step had to be done manually. For each species with high enough score in the previous step evidence for infection with human SARS-CoV-2 was searched in scientific literature to classify between 'infection has ocurred' named "yes" in datasets, 'is proved infection cannot occur' named "non", and 'no evidence whatsoever' named "unk".

The following table lists the species for which their ACE2 proteins were in our dataset.

| “yes” (15 sequences) | “non” (3 sequences) |
|---|---|
| *Bos taurus*<br>*Mustela putorius furo*<br>*Neovison vison*<br>*Canis lupus dingo*<br>*Canis lupus familiaris*<br>*Nyctereutes procyonoides*<br>*Felis catus*<br>*Panthera leo*<br>*Gorilla gorilla gorilla*<br>*Chlorocebus aethiops*<br>*Macaca fascicularis*<br>*Macaca mulatta*<br>*Oryctolagus cuniculus*<br>*Mesocricetus auratus*<br>*Peromyscus maniculatus bairdii* | *Eptesicus brasiliensis*<br>*Sus scrofa*<br>*Mus musculus* |

The rest of species whose protein’s sequences were not in any database and more details available as flow2 on GitHub.

## Comparing both the groups of proteins

Proteins from groups “yes” and “non” were aligned altogether using MUSCLE alignment tool to provide a base for the Sequence Logo generator tool. Then, the sequence logo was generated for each group separately.

Finally, with a custom script, the difference between both was named “diff” providing the following with default parameters:


```
  “yes”:
MxxSxWLLxSxxAxTxAQSxxExxxxxFLxKFNxEAExLxYQxxLASWxYNxNITxENxQxMNxAxxKWSAFxxExSxxAxxYxxxxxxxx
xxKxQLxALQxxGxSxLSxxKxxxLNTILxxMSTIxSTGKxxxxxxxQECxxLxxPGLxxIMxxSxDYxxRLWAWExWRxxVGKQLRPLYE
EYVxLxNEMAxxxxYxDYGDYWRxxYExxxxxxYxYxxxQLxxDVExTFxxIxPLYxxLHAYVRxKLMxxYPSxIxPxGCLPAHLLGDMWG
RFWTNLYxLxVPFxxKPxIDVTxxMxxQxWxAxxIFxEAExFFVSxxLPxMTxxFWxNSMLxxxxxxxKxVCHPTAWDLGxxDFRIxMCTx
VTMDxFLTAHHEMGHIQYDMAYAxQPxLLRNGANEGFHEAVGEIMSLSAATPxxLKxxGLLxxxFxEDxETxINFLLKQALTIVGTLPFTY
MLEKWRWMVFxxxIPKxQWMxxWWEMKRxIVGVVEPxPHDETYCDPAxLFHVxxDYSFIRYYTRTxYQFQFxEALCxxAxHxGxLxKCDIS
NSxEAGxxLxxMLxLGxSxPWTxALExxVGxxxMxVxPLLxYFEPLxxWLKxQNxNSxVGWxTxWxPYxDQSIKVRISLKxALGxxAYxWx
xxEMYxFxxSxAYAMRxYFxxxxxxxxxFxxxxxxVxxxKPRxSFxFxVTxxxNxxDxIPRxxVExAxxxxxxRxRxNDxFxLxDNSLEFx
GIxxTLxPPxxxPxxxWLxxFGVVMxxxVxGIxxLIxxGIxxRxKxxxxxxEENPYxxxDxxKGExNxGFxxxDDxQTSxx
  “non”:
MSxSxWLxLSLxxVTxAQSxTExxAxxFLxxFNxEAEDLxxxSxLASWNYNTNITxENxQKMxxAxxKWSAFYExQSxxAxxxxLxEIQxx
xxKRQLQxLQQxGxSxLxADKxKxLxTILxTMSTIYSxGKVxxPxNPQECLxLxxxGLxxIMxxSxDYxxRLWxWExWRxEVGKQLRPLYE
EYVVLxNEMARxNNYxDYGDYWRGDYExxGxxxYxYxRxQLxEDVxRxFxEIKPLYEHLHAYVRxKLMDxYPSxISPTGCLPAHLLGDMWG
RFWTNLYxLTVPFxxKPxIDVTxAMxxQxWDAxxIFxEAEKFxxSxGLPxMTxGFWxNSMLTEPxDGRKVVCHPTAWDLGxxDFRIKMCTK
VTMDxFLTAHHEMGHIQYDMAYAxQPxLLRNGANEGFHEAVGExMSLSxATPxxLKxxGLLPxDFxEDxETEINFLLKQALxIVGTLPFTY
MLEKWRWMVFxGEIPKEQWMxKWWEMKREIVGVVEPLPHDETYCDPAxLFHVxxDYSFIRYxTRTIxxFQFxEALCxxAKxxGxLxKCDIS
NSTEAGxKLLxMLxLGxSxPWTxALExxxGxxxMDxKPLLxYFxPLxxWLKxQNxNSxxGWxxxWxPYADQSIKVRISLKSALGxxAYEWx
xNEMxLFRSSxAYAMRxYFSxxKNxTxPFxxEDVxVxDxKPRxSFxFFVTSPxNxSDxIPRSxVExAIxxxMSRxRINxxFxLxDNxLEFL
GIxPTLxPPxxPPVTxWLIxFGVVMxxVVxGIxxLIxTGIxxRxKxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
  “diff”:
xSxxxxxxLxLxxVxxxxxxTxxxAxxxxxxxxxxxxDxxxxSxxxxxNxxTxxxxxxxxKxxxxxxxxxxxYExQxxxxxxxxLxEIQxx
xxxRxxQxxxQxxxxxxxADxxKxxxxxxxTxxxxYxxxxVxxPxNPxxxLxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxExxxxxxxxxx
xxxVxxxxxxRxNNxxxxxxxxxGDxxxxGxxxxxxxRxxxxExxxRxxxExKxxxEHxxxxxxxxxxDxxxxxxSxTxxxxxxxxxxxxx
xxxxxxxxxTxxxxxxxxxxxxxAxxxxxxDxxxxxxxxxKxxxxxGxxxxxxGxxxxxxxTEPxDGRxVxxxxxxxxxxxxxxxxKxxxK
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxGExxxExxxxKxxxxxxExxxxxxxLxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxIxxxxxxxxxxxxxKxxxxxxxxxx
xxTxxxxKxLxxxxxxxxxxxxxxxxxxxxxxxxxxDxKxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxAxxxxxxxxxxxSxxxxxxxExx
xNxxxLxRSxxxxxxxxxxSxxKNxTxPxxxEDVxxxDxxxxxxxxxFxxSPxxxSxxxxxSxxxxxIxxxMSxxxIxxxxxxxxxxxxxL
xxxPxxxxxxxPxVTxxxIxxxxxxxxVxxxxxxxxxxTxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
```

Finally, the diff pattern was used to identify the most and least probable hosts from the “unk” sequences:

| Most probable transmission | Least probable transmission |
|---|---|
| *Panthera pardus*<br>*Rhinopithecus bieti*<br>*Loxodonta africana*<br>*Ursus maritimus*<br>*Acinonyx jubatus*<br>*Ursus americanus*<br>*Mustela lutreola biedermanni*<br>*Tursiops truncatus*<br>*Cebus imitator*<br>*Lynx canadensis*<br>*Sapajus apella*<br>*Sciurus vulgaris*<br>*Neomonachus schauinslandi*<br>*Panthera tigris altaica*<br>*Saimiri boliviensis boliviensis*<br>*Procyon lotor*<br>*Aotus nancymaae*<br>*Callithrix jacchus*<br>*Rhinopithecus roxellana*<br>*Puma concolor*<br>*Rhinolophus sinicus*<br>*Manis pentadactyla* | *Sus scrofa domesticus*<br>*Mus caroli*<br>*Catagonus wagneri* |

The results are available as flow3 at Galaxy Europe and the scripts and data in GitHub.

## Results comprehension

The low availability of data is reflected in the results as the species in each group reflect the number of sequences in each dataset. We could optimize the parameters of the aligner and the comparer to get a better pattern, but we would require some more data to substantially improve it.

Regarding the already known interactions, it is noticeable that only the asparagine 38 shows up in the diff pattern, which could be explained if the sequences in the dataset happen to have conserved that region while the key regions preventing the infection from the virus occur at some pther point which significally rearranges the 3D structure of the protein.

Still, this has proven adequate to identify the most probable species and a few more worth checking on in future studies apart from the close relatives to the already known ones, like some species of elephant or seal. This is especially interesting because of the papers that suggested the spread of this virus in some marine species.

Even though the idea looks promising, future data will determine how adequate this implementation is and how parameters should be set.

# MERS-CoV

## Context and information

MERS-CoV (Middle East Respiratory Syndrome Coronavirus) is a beta coronavirus that causes severe respiratory illnesses in humans. It is believed to have originated in animals, possibly camels, and is primarily transmitted to humans through contact with infected animals or their respiratory secretions. It can also spread among people through close contact, such as caring for infected patients.

Structure of MERS-CoV spike receptor-binding domain complexed with human receptor DPP4

In this study, the structure of the MERS-CoV S protein and its interaction with the DPP4 protein on the surface of host cells was investigated. It was found that the S protein of the virus binds to the DPP4 protein to enter host cells. Identification of the critical residues involved in the binding between these proteins will help in the development of antiviral treatments for MERS-CoV.

The paper is focused on the identification of critical residues in the S protein of MERS-CoV that are important for viral binding and entry into host cells. Mutations in certain residues were found to significantly reduce binding to the DPP4 protein and the efficiency of viral entry into host cells. The results suggest that the interaction between the virus S protein and the DPP4 protein is crucial for viral entry.

The molecular structure of the MERS-CoV S protein and its interaction with the cellular DPP4 receptor. The researchers identified specific residues in the S protein of the virus that are critical to its ability to bind to the DPP4 receptor and enter host cells.

The MERS-CoV S protein consists of two main subdomains: the core and the receptor-binding subdomain. The core is structurally like that of the SARS-CoV S protein, while the receptor-binding subdomain is markedly different. This helps explain the differences in receptor specificity between MERS-CoV and SARS-CoV.

Furthermore, the enzymatic site of DPP4 was found to be far from the binding site of the MERS-CoV S protein, which explains why DPP4 enzyme inhibitors do not block virus entry into cells. Finally, the importance of structural differences between DPP4 and the ACE2 receptor in determining cell tropism and pathogenesis of MERS-CoV and SARS-CoV in living hosts is discussed.

### Work and study to find reservoirs of the Mers-CoV virus

MERS-CoV can cause a wide range of respiratory symptoms, ranging from mild cold-like symptoms to severe pneumonia with respiratory failure, and in some cases, death. Patients with chronic illnesses or weakened immune systems are at higher risk for severe complications.

Pathological studies in animals infected with MERS-CoV have shown changes in the lungs, such as thickening of alveolar septa, infiltration of inflammatory cells such as lymphocytes and histiocytes, and presence of exudates in alveolar and bronchial spaces. These pathological changes can vary from acute lesions in the early days of infection to chronic or reparative lesions in later stages.

We have a file with the sequence DPP4, dipeptidyl peptidase 4, or also called CD26 is an enzyme found on the surface of many cells in the human body and plays an important role in the regulation of metabolic function and the immune system. DPP4 is a protein that belongs to the family of peptidase enzymes, which are responsible for the hydrolysis of peptides and proteins in the body.

Once the MERS-CoV Spike protein binds to the DPP4 protein on the cell surface, it triggers a series of events that allow entry of the virus into the host cell, including fusion of the viral envelope with the cell membrane and release of the virus' genetic material into the cell. This leads to virus replication and the spread of infection. The interaction between the MERS-CoV Spike protein and the DPP4 protein is a potential target for the development of therapeutic and preventive strategies against MERS-CoV infection. For example, inhibitors of the Spike protein and the DPP4 protein have been investigated as possible approaches to block viral entry into cells and prevent MERS-CoV infection. However, more research is still needed to fully understand the interaction between these proteins and their potential application in the treatment or prevention of MERS-CoV infection.

### Mers-CoV and and its relationship with animals and humans

MERS-CoV is a zoonotic virus (transmitted between animals and humans). It is believed that humans can be infected by direct contact but at present it has only been shown to occur with infected dromedary camels in the Middle East or their milk, urine or other body fluids. Bats are a likely original reservoir; coronaviruses like MERS-CoV have been identified in bats (*Rousettus aegyptiacus*), but epidemiologic evidence of their role in transmission is lacking.

### Visual description of how a dromedary acts as a reservoir.

Several animal species have been evaluated to model human MERS-CoV infection, including rabbits, ferrets, Syrian hamsters and mice, but have been found not to be susceptible to the virus. This was due to the function and distribution of the DPP4 receptor. The structure of the receptor is diverse, but is conserved in humans and non-human primates, making non-human primates an ideal model for infection. The DPP4 receptor of a small non-human primate such as the *Rhesus macaque* and *Callithrix jacchus* have a high similitud and this indicates that they may be future reservoirs due to the high sequence overlap of the DPP4 protein.

### Transmission between animals and humans

In this section we will try to discover possible reservoirs of this virus from the DPP4 protein, a reservoir being a place where an organism or substance can remain latent, persistent or active, and from where it can be released or transmitted to other organisms.

For this, we will use the sequence of the DPP4 protein present in Homo sapiens. We will use the `Blastp` tool present on the Galaxy server to align with the different species whose sequences are similar or intrinsically related to the spread of the virus. These sequences to be compared were obtained from the Uniprot protein database.

## Analysis workflow

Scores of similitud between species and human protein DPP4
This and related data are available in Galaxy Europe and GitHub.

The sequences of animals most closely related to the human DPP4 protein and those sequences belonging to animal species known to have been infected have been chosen.

We have found different types of animal species which could be a possible reservoir and have a high relation with the human protein:

| Uniprot | Species | Result | Paper |
|---|---|---|---|
| A0A0D9RSY5 | *Chlorocebus sabaeus* | No | Article |
| F6VRB0 | *Rhesus macaque* | No | Article |
| A0A2K6UUT9 | *Saimiri boliviensis* | No | ~ |
| A0A2K5F6W0 | *Aotus nancymaae* | No | ~ |
| A0A2K5LB19 | *Cercocebus atys* | No | ~ |
| F7IHU4 | *Callithrix jacchus* | No | Article |
| A0A6D2X1C5 | *Pan troglodytes* | No | ~ |
| G3SI68 | *Gorilla gorilla* | No | ~ |
| A0A2Z5CWD9 | *Carollia perspicillata* | Yes | Article |
| A0A2Z5CWD4 | *Epomops buettikoferi* | Yes | Article |
| A0A2R9CCY4 | *Pan paniscus* | No | ~ |
| H2P7N3 | *Pongo abelii* | No | ~ |
| A0A8I3Q958 | *Canis lupus familiaris* | No | Article |
| A0A075T9L1 | *Camelus dromedarius* | Yes | Article |
| A0A075T9L7 | *Ovis orientalis* | No | Article |
| A0A4W2EX30 | *Bos taurus indicus* | No | Article |
| A0A452FGS0 | *Capra aegagrus hircus* | No | Article |
| A0A2Z5CWB8 | *Rousettus aegyptiacus* | Yes | Article |
| M3XN99 | *Mustela putorius furo* | No | Article |
| A0A075TJ59 | *Mesocricetus auratus* | No | Article |
| P28843 | *Mus musculus* | No | Article |

The result section indicates whether the species can act right now with the data obtained as a reservoir, seeing that only the dromedary and the African bat act as such and species that do not have a paper associated with them are those for which no study has been carried out that can confirm their capacity as a reservoir of Mers-Cov virus.

After reading several studies confirming that most of the animals studied so far do not act as reservoirs of the Mers-Cov virus, but only suffer from it. We will analyse the conserved regions of the DPP4 of the most characteristic species and try to predict whether they can act as reservoirs, as no studies have been carried out on the species we will be dealing with.

## Cross-species transmission, evolution and zoonotic potential of coronaviruses

It should be noted that the case of the spread of alpha and beta variants viruses in marine mammals has also been studied, but no conclusive data have been found that they act as a reservoir. What has been concluded is that coronavirus-contaminated discharges into water bodies should be properly treated to prevent the spread of the virus to marine fauna.

We will take the sequence of the species we are studying that are reservoirs and those that are not, compare with the results using a `sequenceLogo` after aligning the sequences using Galaxy's `ClustalW` tool and interpret the results to see which regions are common to which species and guess whether they may act as reservoirs.

## Possible reservoirs and interpretation

We have aligned the sequences of species that can act as reservoirs and the sequences of species that cannot. We will then compare which regions of the protein are conserved between the two cases by using a sequenceLogo.

Therefore, if we see that the unstudied sequences have certain regions in common with the positive species, we could venture to say that they could act as reservoirs, pending an experimental study to confirm this.

This sequenceLogo comes from WebLogo3, the colours blue, green and black are used to represent different ranges of amino acid or nucleotide conservation in the sequence, which can be measured in terms of bits of information.

- Blue: used to highlight the most conserved amino acids or nucleotides in the sequence, i.e., those that have the most information in common with the reference sequence used.
- Green: used to highlight amino acids or nucleotides that have intermediate conservation in the sequence.
- Black: used to highlight the least conserved amino acids or nucleotides in the sequence, i.e., those that have less information in common with the reference sequence used. Sometimes we have different symbols stacked on top of each other, this is because these amino acids appear in different sequences in the same position.

Furthermore, we must consider that the logo could be larger and thicker, thats indicate a higher frequency of occurrence of that amino acid or nucleotide in the sequence, while smaller and thinner letters indicate a lower frequency of occurrence.

SequenceLogo of positive cases

SequenceLogo of negative cases

sequences of Species which have been shown to be reservoir species

Sequences of Species not proven to be reservoirs

# Analysis results

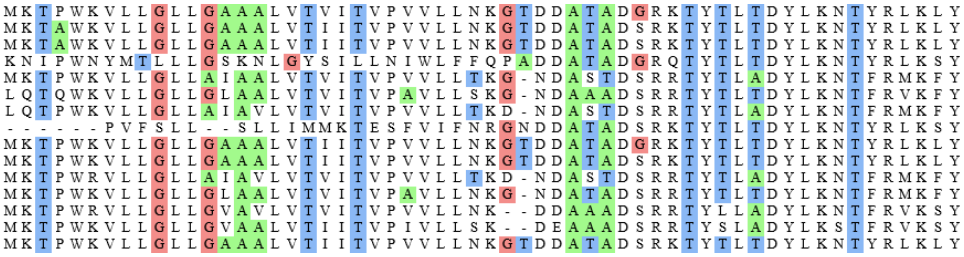
With these results we could venture to say there are some regions that could be the ones that do that the species that can act as a reservoir.

If we analyse the sequenceLogo we see that in the negative cases we have some gaps at the beginning, while in the positive cases there are hardly any gaps at all. This could be due to missing sequence strings in these areas, but as we can see that the length of both positive and negative cases is the same, it could be due to an error.

Even so, this form of representation does not give us much information, so we are going to try looking only at the shared nucleotides.

**Positive cases**

**Negative cases**



As we can see, both cases present a high similarity, so it is not possible to differentiate the reason why they act as a reservoir or not. Only one part has been shown, but in the following points the rest can be accessed.

The **Nucleotide display** (colour scheme-Nucleotide) scale shall be used to view the sequences in a more informative way in the page ngphylogeny.fr.

Conserved domains in positive cases

Conserved domains in negative cases

With all that we have seen, the papers and the results, we can conclude that both Mers-CoV and Sars-CoV are betacoronaviruses that are transmitted in a similar way, and can affect both humans and animals, even affecting marine species.

In the case of Mers-CoV, no relationship has been found between whether the host can act as a reservoir or not in terms of the conserved regions of the protein in positive and negative cases.

No difference can be seen, as in both cases there are conserved areas with a great similarity, therefore, we cannot say for sure that we see a clear difference between sequences. Perhaps with the study of molecular dynamics or with further experiments we can discern between the two cases.

Therefore, more experiments and further research in this field would be required to provide conclusive data.

# Human coronavirus 229E

## Context and information

The first human coronavirus was isolated from the B814 strain in 1965. It was discovered in the respiratory tract of patients with the common cold during studies on nasal samples. The 229E species was isolated from cultured standard tissues. This virus is generally associated with upper respiratory tract diseases, accounting for a total of 15-30% of common cold cases in humans. This coronavirus uses an S protein (called spike protein) that mediates the binding to the receptor and subsequent fusion of the virus to the host cell. The receptor, which we will study below, has a binding domain (RBD) but the system by which the S protein maintains its function once it binds is unknown. This protein has hydrophilic subunits that, in its study, reveal part of its helical core in the solvent. The binding between this protein and its receptor is very specific and structurally conserved within members of the Alphacoronavirus family, even when they bind to different receptors in different locations. This, along with interspecies jumping, could suggest a mechanism for acquiring new receptor interactions. The functional domain of the S protein is composed of an N-terminal S1 domain that contains the receptor binding domain (RBD) and a C-terminal S2 region that mediates membrane fusion. This is a trimer that maintains this structure before and after fusion with the cell. It is the binding domains that change their conformation before and after fusion.

### Analysis workflow

We have taken the primary sequence source from UniProtKB/P15144 . With these files we began this study about the infection of coronviruses in other potentially infective species. P15144 is the ID code of human aminopeptidase (h(APN)) in UniProt database. The h(APN) is the well know receptor of Human-Cov-229E which is a virus knows to cause a common cold. There are another species that have this receptor with their genetic sequences.

Initially we used the `blastp` function on the initial fasta document to allign all the sequences. Once we got this we are going to past the format of the document from tabular data to `.fasta` for other operation. In the second step, we will search for species that have a closer relationship with the virus and those that do not. In this case, we have divided the dataset into three parts. We will take into account those sequences that have between 70% and 99.99% similarity to our human aminopeptidase sequence. Less than 70-60% would lose a lot of genetic information. The 100% sequence would represent the h (APN) sequence. We will divide this dataset into three fasta files, one for 70-80%, 80-90%, and 90-99.99%. When searching the literature, we see how there is no evidence of cases in which these animals have been infected with the 229E coronavirus from the 80-90% dataset. Therefore, we will choose to study only the 70-80% and 90-99.99% datasets.

After selecting the datasets, we are going to use, based on the literature, we proceed to carry out a deeper study of the genetic sequences of the receptors of each animal species. In this case, we have selected three animal species from the dataset that have a 70-80% similarity. In these three animal species, according to scientific literature, cases of 229E infection have been observed, in fact, they are the animals where more pathological cases have been confirmed. On the other hand, we have decided to use three species that are closer to 100% similarity with the original sequence. Those species are: *Macaca mulatta*, *Pan troglodytes*, *Mandrillus leucophaeus*, *Camelus dromedarius*, *Vicugna pacos* y *Mus musculus.*

After finishing the literature search, we decided to use these three species to conduct the study of the 229E receptor sequence. To do this, we first performed an analysis of the sequences with the usegalaxy.com application using the `sequence` Logo tool. Next, we used the NGPhylogeny application that performs multiple alignments and phylogenetic trees. From this, we will identify the

sequence differences between the different receptors. For this, we will create two datasets, one with all the species that are infected, and another with those that are not infected.

## Analysis results

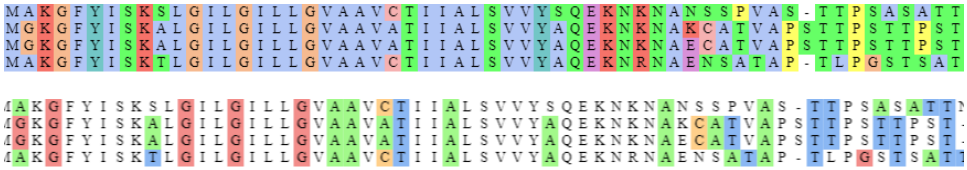
After reviewing the literature, we will study the sequences of those species that are more closely related and those that are not. First, we will perform a general scan of the sequence using Sequence Logo, as we mentioned earlier. We will do this for both the species that are infected and those that are not infected. For space reasons, we will provide a link to our Github repository where these images can be found.

Sequence Logo for infected species

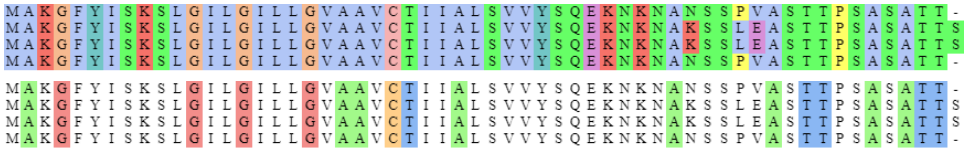
Sequence Logo for non infected species

This study is very general, because the layout of the image that this tool provides does not facilitate the interpretation of specific nucleotides. However, we can see that there are nucleotides that appear strongly conserved. The letters that appear alone are those that are repeated a lot among sequences. As we advance, the size and thickness of the letter varies according to how it is repeated among the sequences. In the case of animals that can be infected by 229E, we see that there are many letters grouped in the final sections. This means that there is a lot of variation in nucleotides in these areas. However, we see that in the first part, the variation is lower. In the second case, we see more blue letters than green and black ones. The color scheme follows the following pattern: blue letters are related to the most conserved nucleotides in the sequence. We see that the sequences that do not get infected have more of these types of nucleotides because they have a greater similarity to the original sequence. This is different for the infected sequence.

To finish, we will analyze the sequences in a specific way. Using the NGPhylogeny application, we can see the sequences in depth, from the point of view of the most conserved nucleotides. Later, we can see a generated phylogenetic tree for the different datasets.



In the image, we see a grid of different colors, one for each nucleotide. If we color it with the clustal2 color system, we can see in more detail (in blue) the nucleotides that are the same in the sequences. With the application, we can also see each of the nucleotides that are different at the same level in the 4 compared sequences. In the case of the species that are seen to be infected, we have more different nucleotides, especially in the final part of the sequences. This makes their similarity to the protein different. We could say that the infectivity potential of human coronavirus 229E is given by some very conserved areas that are more involved in the functionality of the protein.





In the case of the species that do not get infected, we have many more conserved sequences. As we can see in clustal2, there are many more areas of blue color, indicating repeated nucleotides. If we look at the nucleotides at the specific level, there are fewer different nucleotides. Although we can see that at the end of the sequence, there are areas with different nucleotides. We also observe that the conserved areas mostly coincide with those sequences that do get infected, with the exception of a few. We could say that it is these differences in conserved areas that generate a change in the functional structure of the protein, allowing infection. In both data sets, we can see that there are more conserved sequences in the anterior part of the sequences than in the posterior part. This may indicate that one area may be structural nucleotides and others functional nucleotides, with more similar sequences in the functional ones.

## Conclusions

As an overview of the three implementations, we can see that the current amount of data available does not make for very detailed comparisons. Even though, we could still some conclusions. To summarize:

- For SARS-CoV-2 the data provided for a decent pattern that can reasonably distinguish hosts and even predict potential reservoirs.
- As for MERS-CoV spreading is not very documented, we need more information for that virus.
- Regarding HCoV 229E, a noticeable difference was seen at the end of the sequences to make a pattern. In conclusion, we can say that the infectivity capacity of the S protein of coronavirus 229E in other animals is not determined by its phylogenetic relationship with humans or genetic similarity. It is characterized by the conserved sequences that, according to the literature, are more related to the functional part of the receptor, while the rest of the sequence is related to the structural part.

For next steps, in silico models could be used to increase the amount of data and potentially the accuracy of conclusions. Another compelling approach would be comparing 3D structures on top of the bare sequences, as we have seen there are important regions for the pattern across the proteins, which would make for proteins whose interaction rely on glycosylation, for some other viruses. Finally, it is to mention that there also exist viruses which bind to more than one protein, so the comparison should be made for all proteins in such cases.

## External data

GitHub repository: https://github.com/jricardo-um/analisis-datos-omicos-trabajo-2