Descripción de script para la identificación de la proteína E, correspondiente al genoma del viral de la cepa SARS-COV2 **SARS-CoV-2** (*severe acute respiratory syndrome coronavirus)*

## A partir del script adnframes.py ubicado en el repositorio GitHub (<https://github.com/Anulasi/python-1/blob/master/w6/adnframes.py>) se procedió a identificar la proteína E.

## Líneas 5-12: Se establecieron las variables correspondientes al procesamiento del ADN complementario (bases nitrogenadas) universal para cada organismo.

## Líneas 14-19: Se establecieron las variables correspondientes al procesamiento del ADN complementario a RNAm (RNA mensajero), indicando que, adenina = adenina "A":"A", citosina=citosina "C":"C", timina=uracilo "T":"U”, y guanina = guanina "G":"G".

## Líneas 21-86: Para poder traducir las secuencias de ARNm a proteína, los organismos utilizan claves de codones de inicio y paro, los cuales, conforman 64 en total, siendo 61 codificantes para aminoácidos, y únicamente 3 de paro (UAA, UAG, UGA); estos codones se conforman de la combinación de tres bases nitrogenadas (A, C, G y U) para ARNm. Las líneas 21-86 corresponden a las 64 combinaciones posibles para cada codón, y su equivalente en la formación de aminoácidos.

## Líneas 88-153: Se implementó la nomenclatura aminoacidica utilizada en genética molecular y en bases de datos, la cual consta de una sola letra, esto permite la representación de la estructura primaria de una proteína mediante la disposición consecutiva de letras sin espacios ni signos intermedios, disponiendo a la izquierda el aminoácido N-terminal y a la derecha el aminoácido C-terminal.

## Líneas 155-175: Se introdujo la secuencia de la proteína E en el sentido 5’ 3’ para encontrar la cadena complementaria antisentido (-) dentro del archivo covid19.txt en la variable f, metiendo los resultados en u renglón nuevo dentro de la variable x

## Líneas 178-198: Se realizó la búsqueda de la secuencia complementaria en la variable X, y se obtuvo el resultado esperado tanto de la cadena antisentido como de la parte codificante, así mismo, se pidió la longitud de cada una de las salidas.

## Líneas 200-210: Hecho el paso anterior, se realizó la síntesis de RNAm a partir de las cadenas – y +, para crear posteriormente identificar los codones.

## Líneas 212-249: Se realizó la identificación de los marcos de lectura abiertos, y de paro, de la transcripción, mediante la construcción de los codones de inicio y paro identificados en el genoma de SARS-CoV-2, utilizando longitudes de palabra no mayores a 3, utilizando ciclos for y condicionantes if, else.

## Líneas 252-266: Utilizando el juego de codones, se identificaron las combinaciones de aminoácidos disponibles mediante ciclos for y condicionantes (if y else), así como la identificación de codones codificantes para un mismo aminoácido, y codones codificantes para un solo aminoácido, utilizando ciclos for dentro de las variables i y j.

## Líneas 270-282: Se obtuvieron los marcos de lectura abiertos así como su longitud, posteriormente se imprimieron dichos fragmentos, así como sus aminoácidos correspondientes en la nomenclatura de genética molecular, obteniendo como resultado la secuencia de la proteína E, la cual fue corroborada mediante blast en NCBI (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

## <https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2?from=26245&to=26472&report=fasta>

## <https://www.ncbi.nlm.nih.gov/nuccore/1798174254>

## 