Instituto Politécnico Nacional

Escuela Superior de Cómputo

Bioinformatics

Lab Session 5. Uniprot

González Bocio Erik Alexander 2020630163

Jorge Luis Rosas Trigueros

Fecha de realización: 28 de marzo de 2022

Fecha de entrega: 4 de abril de 2022

Marco teórico:

Para proporcionar a la comunidad científica un recurso único, centralizado y autorizado para secuencias de proteínas e información funcional, las actividades de la base de datos de proteínas Swiss-Prot, TrEMBL y PIR se han unido para formar el consorcio Universal Protein Knowledgebase (UniProt).

El Recurso Universal de Proteínas (UniProt) es un recurso integral para datos de secuencia y anotación de proteínas. Las bases de datos de UniProt son UniProt Knowledgebase (UniProtKB), UniProt Reference Clusters (UniRef) y UniProt Archive (UniParc). El consorcio UniProt y las instituciones anfitrionas EMBL-EBI, SIB y PIR están comprometidas con la preservación a largo plazo de las bases de datos UniProt.

Su misión es proporcionar una base de conocimientos de secuencias de proteínas completa, totalmente clasificada, rica y precisamente anotada, con extensas referencias cruzadas e interfaces de consulta. La base de datos central tendrá dos secciones, correspondientes a las conocidas Swiss-Prot (entradas totalmente seleccionadas manualmente) y TrEMBL (enriquecidas con clasificación automatizada, anotación y extensas referencias cruzadas). Para búsquedas de secuencias convenientes, UniProt también proporciona varias bases de datos de secuencias no redundantes. Las bases de datos UniProt NREF (UniRef) proporcionan subconjuntos representativos de la base de conocimientos adecuados para una búsqueda eficiente. El completo Archivo UniProt (UniParc) se actualiza diariamente desde muchas bases de datos de fuentes públicas. Se puede acceder a las bases de datos de UniProt en línea (http://www.uniprot.org) o descargarlas en varios formatos (ftp://ftp.uniprot.org/pub). Se alienta a la comunidad científica a enviar datos para su inclusión en UniProt.

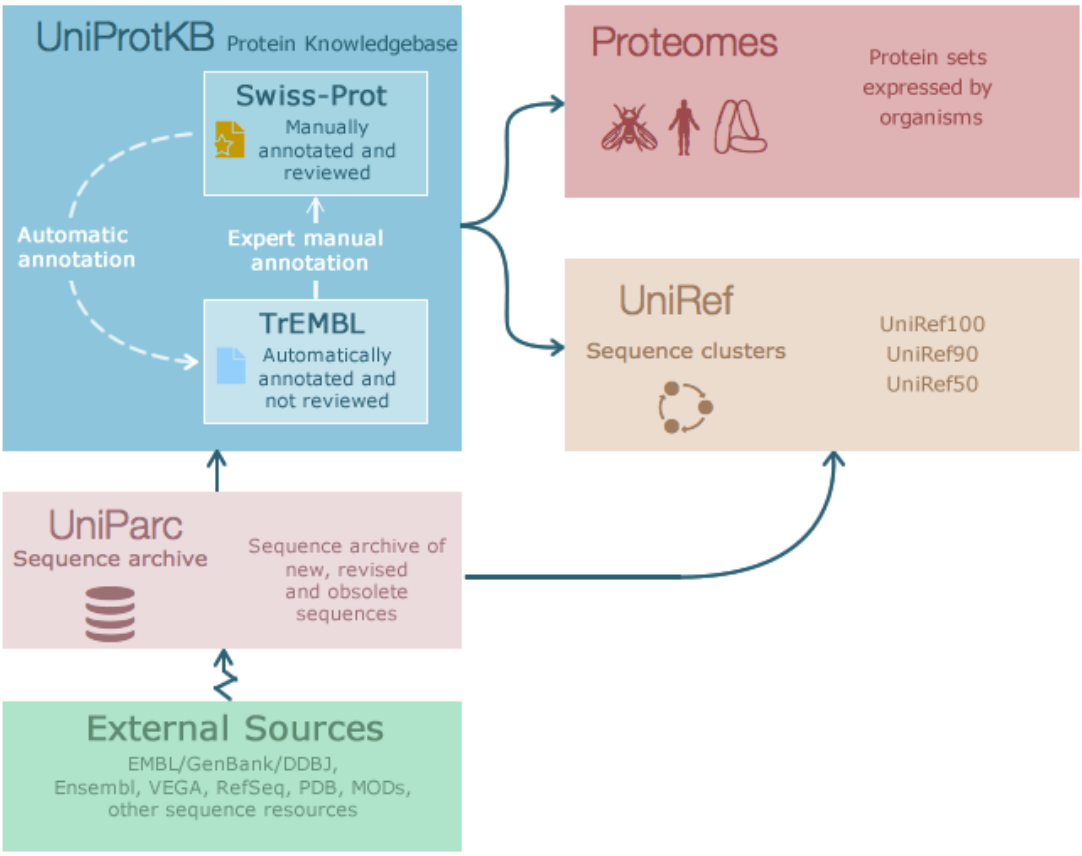

Fig. 1. Diagrama de elementos contenidos en UniProt

UniProt es una colaboración entre el Instituto Europeo de Bioinformática (EMBL-EBI), el Instituto Suizo de Bioinformática SIB y el Recurso de Información de Proteínas (PIR). En los tres institutos, más de 100 personas están involucradas a través de diferentes tareas, como la curación de bases de datos, el desarrollo de software y el soporte.

EMBL-EBI y SIB juntos solían producir Swiss-Prot y TrEMBL, mientras que PIR produjo la Base de Datos de Secuencias de Proteínas (PIR-PSD). Estos dos conjuntos de datos coexistieron con diferentes prioridades de cobertura de secuencias de proteínas y anotaciones. TrEMBL (Translated EMBL Nucleotide Sequence Data Library) se creó originalmente porque los datos de secuencia se generaban a un ritmo que excedía la capacidad de Swiss-Prot para mantenerse al día. Mientras tanto, PIR mantuvo el PIR-PSD y bases de datos relacionadas, incluyendo iProClass, una base de datos de secuencias de proteínas y familias curadas. En 2002, los tres institutos decidieron aunar sus recursos y experiencia y formaron el consorcio UniProt.

El consorcio UniProt está encabezado por Alex Bateman, Alan Bridge y Cathy Wu, con el apoyo de personal clave, y recibe valiosos aportes de un Consejo Asesor Científico independiente.

Material y Equipo:

- Página de UniProt
- Plataforma Teams

- Computadora
- Red de Internet

Desarrollo de la práctica:

En esta práctica se llevó el primer acercamiento a lo que es la herramienta UniProt, herramienta que como lo es PDB es una base de datos, con la diferencia que esta es para datos de secuencia y anotación de proteínas.

El primer ejercicio que se llevó a cabo fue recopilar secuencias para HBA1 de diferentes especies, más específico el de 10 especies diferentes, en las cuales en este caso se usaron las siguientes especies:

1. Human (P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapiens)
2. Dog (P60529|HBA_CANLF Hemoglobin subunit alpha OS=Canis lupus familiaris)
3. Axolotl (P02015|HBA_AMBME Hemoglobin subunit alpha OS=Ambystoma mexicanum)
4. Horse (P01958|HBA_HORSE Hemoglobin subunit alpha OS=Equus caballus)
5. Mouse (P06467|HBAZ_MOUSE Hemoglobin subunit zeta OS=Mus musculus)
6. Pig (P01965|HBA_PIG Hemoglobin subunit alpha OS=Sus scrofa)
7. Lion (P18975|HBA_PANLE Hemoglobin subunit alpha OS=Panthera leo)
8. Jaguar (P63109|HBA_PANON Hemoglobin subunit alpha OS=Panthera onca)
9. Northern persian leopard (P18976|HBA_PANPS Hemoglobin subunit alpha OS=Panthera pardus saxicolor)
10. Red Fox (P21200|HBA_VULVU Hemoglobin subunit alpha OS=Vulpes vulpes)
11. American Aligator (P01999|HBA_ALLMI Hemoglobin subunit alpha OS=Alligator mississippiensis)

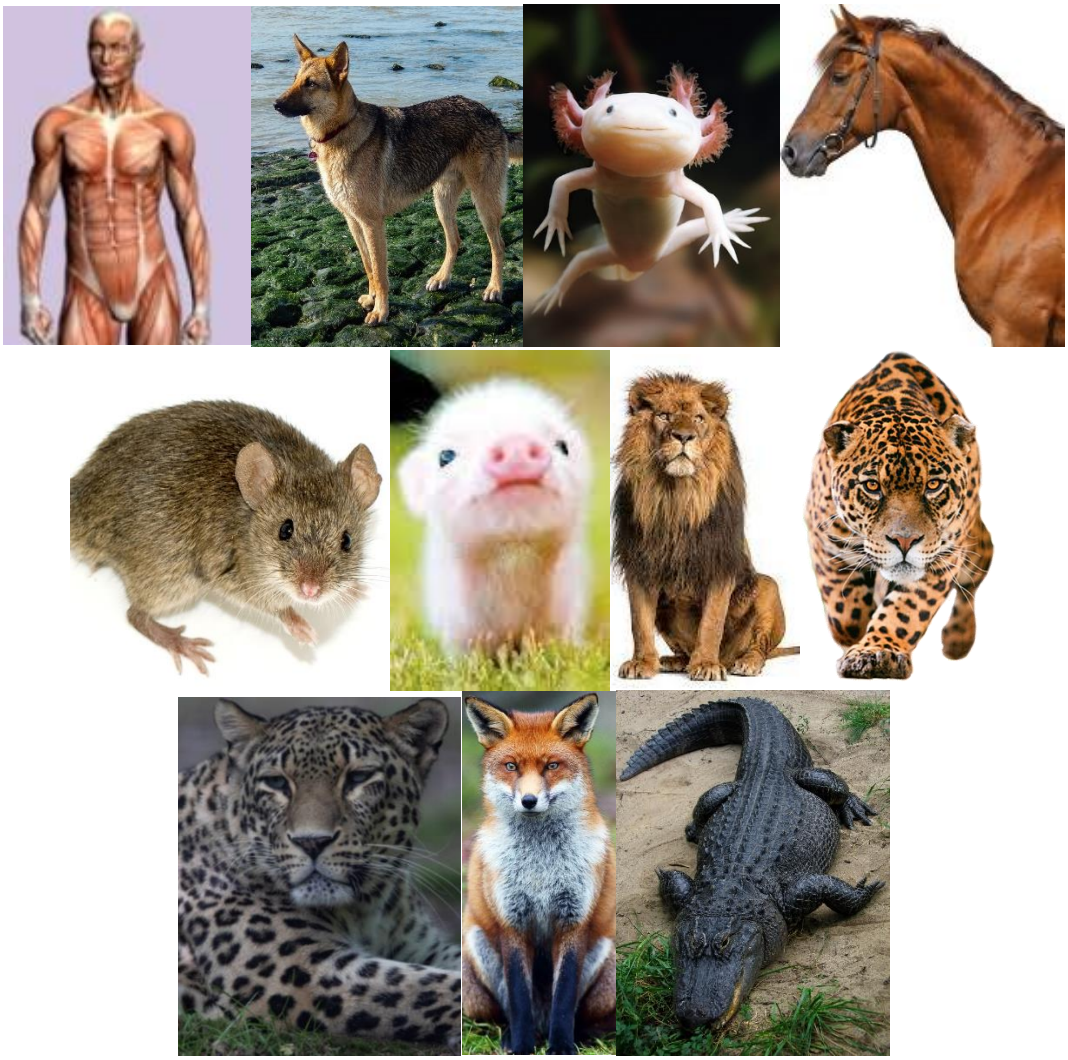

Fig. 2. Imágenes de las especies usadas para hacer la alineación, podemos ver como empieza desde el humano hasta el caimán americano

Luego de eso se llevó a cabo el alineamiento o “Align” de todas las hemoglobinas Alpha de estas especies, por lo cual el resultado fue el siguiente:

```
P69905 HBA_HUMAN          1  -MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGH      59
P60529 HBA_CANLF          1  --VLSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHFDLSPGSAQVKAH      58
P02015 HBA_AMBME          1  MFKLSGEDKANVKAVWDHVKGHEDAFGHEALGRMFTGIEQTHTYFPDKDLNEGSFALHSH      60
P01958 HBA_HORSE          1  -MVLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAH      59
P06467 HBAZ_MOUSE         1  -MSLMKNERAIIMSMWEKMAAQAEPIGTETLERLFCSYPQTKTYFPHFDLHHGSQQLRAH      59
P01965 HBA_PIG            1  --VLSAADKANVKAAWGKVGGQAGAHGAEALERMFLGFPTTKTYFPHFNLSHGSDQVKAH      58
P18975 HBA_PANLE          1  -MVLSSADKNNVKACWGKIGSHAGEYGAEALERTFCSFPTTKTYFPHFDLSHGSAQVQAH      59
P63109 HBA_PANON          1  -MVLSSADKNNVKACWGKIGSHAGEYGAEALERTFCSFPTTKTYFPHFDLSHGSAQVQAH      59
P18976 HBA_PANPS          1  --VLSSADKNNVKACWGKIGSHAGEYGAEALERTFCSFPTTKTYFPHFDLSHGSAQVQTH      58
P21200 HBA_VULVU          1  --VLSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHFDLSPGSAQVKAH      58
                              *   ::  : : * :: .:    * *:* * * .   *:****. :*  **  :: *
```

```
P69905 HBA_HUMAN        60  GKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFT  119
P60529 HBA_CANLF        59  GKKVADALTTAVAHLDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPTEFT  118
P02015 HBA_AMBME        61  GKKVMGALSNAVAHIDDLEATLVKLSDKHAHDLMVDPAEFPRLAEDILVVLGFHLPAKFT  120
P01958 HBA_HORSE        60  GKKVGDALTLAVGHLDDLPGALSNLSDLHAHKLRVDPVNFKLLSHCLLSTLAVHLPNDFT  119
P06467 HBAZ_MOUSE       60  GFKIMTAVGDAVKSIDNLSSALTKLSELHAYILRVDPVNFKLLSHCLLVTMAARFPADFT  119
P01965 HBA_PIG          59  GQKVADALTKAVGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHHPDDFN  118
P18975 HBA_PANLE        60  GQKVADALTKAVVHINDLPNALSDLSDLHAYKLRVDPVNFKFLSHCLLVTLACHHPEEFT  119
P63109 HBA_PANON        60  GQKVADALTKAVAHINDLPNALSDLSDLHAYKLRVDPVNFKFLSHCLLVTLACHHPEEFT  119
P18976 HBA_PANPS        59  GQKVADALTKAVAHINDLPNALSDLSDLHAYKLRVDPVNFKFLSHCLLVTLACHHPEEFT  118
P21200 HBA_VULVU        59  GKKVADALTTAVAHLDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPNEFT  118
                                * *:  *:  **  ::::  :*  **: **: * ***.:*  *:. :* .:. : * .*.

P69905 HBA_HUMAN       120  PAVHASLDKFLASVSTVLTSKYR                                      142
P60529 HBA_CANLF       119  PAVHASLDKFFAAVSTVLTSKYR                                      141
P02015 HBA_AMBME       121  YAVQCSIDKFLHVTMRLCISKYR                                      143
P01958 HBA_HORSE       120  PAVHASLDKFLSSVSTVLTSKYR                                      142
P06467 HBAZ_MOUSE      120  PEVHEAWDKFMSILSSILTEKYR                                      142
P01965 HBA_PIG         119  PSVHASLDKFLANVSTVLTSKYR                                      141
P18975 HBA_PANLE       120  PAVHASLDKFFSAVSTVLTSKYR                                      142
P63109 HBA_PANON       120  PAVHASLDKFFSAVSTVLTSKYR                                      142
P18976 HBA_PANPS       119  PAVHASLDKFFSAVSTVLTSKYR                                      141
P21200 HBA_VULVU       119  PAVHASLDKFFTAVSTVLTSKYR                                      141
                              *: : ***:     :  .***
```

Fig. 3. Alineamiento de las 11 especies elegidas, podemos ver en los resultados

Podemos ver que en algunas se usa el asterisco que como recordamos significa que es un aminoácido idéntico y también una gran cantidad de puntos que son aminoácidos semejantes, más cantidad de dos puntos, haciendo que sean muy parecidos entre ellos.

Además de mostrarnos el alineamiento nos muestra un árbol, el cual nos ayuda a agrupar a las especies en una rama y se muestra en la siguiente imagen:

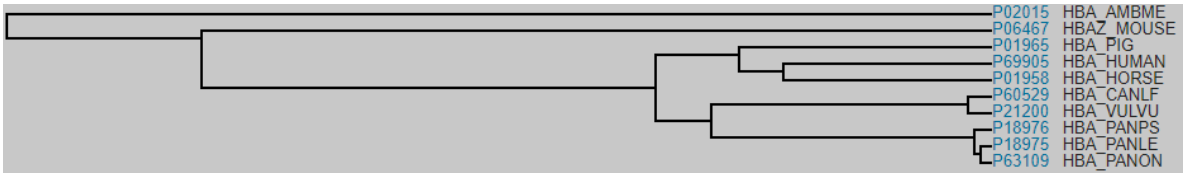

Fig. 4. Imagen del árbol generado luego de la alineación

El siguiente punto de la práctica era realizar un Blast usando la entrada P01979, el blast nos sirve para buscar en todos los datos que hay en la DB y ver cuales coinciden con la entrada, dándonos un porcentaje de coincidencia, lo cual nos ayuda a ver que organismos pueden tener cosas en común. Al meter la entrada nos muestra los siguientes resultados:

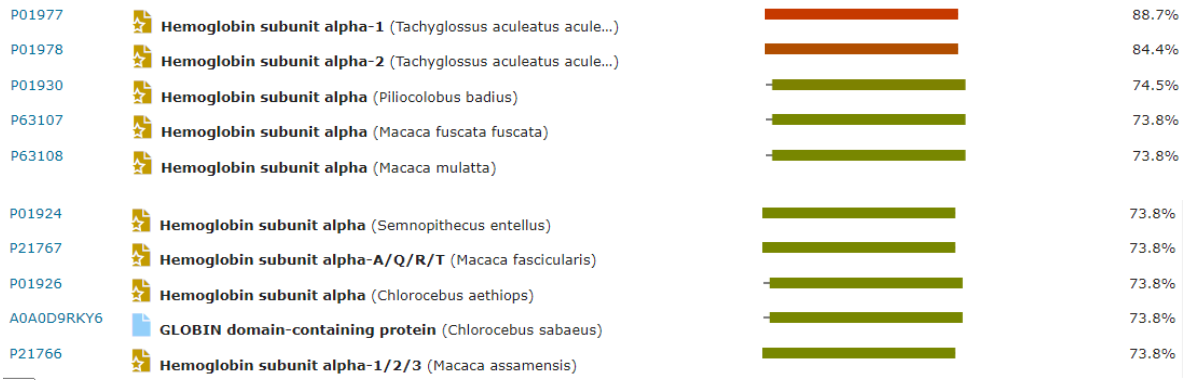

Fig. 5. Resultados generados al llevar a cabo el blast de P01979

El último punto llevado de manera individual fue obtener una secuencia de proteína no caracterizada de un organismo desconocido e investigarla. Por lo que al buscar entre la lista de

estas proteínas se terminó eligiendo la entrada A0A0F7YUR9, por lo que al llevar a cabo el blast se obtuvieron los siguientes resultados:

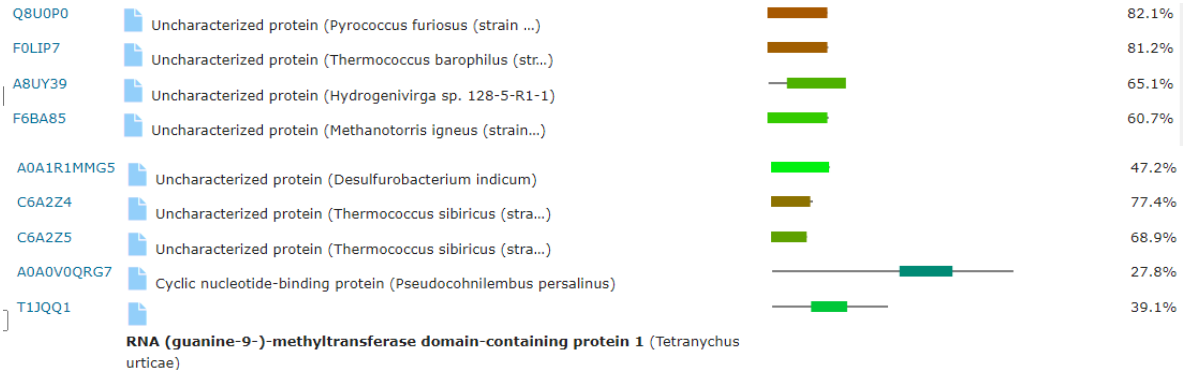

Fig. 6. Resultado generado luego de llevar a cabo el blast de A0A0F7YUR9

Así podemos ver que tiene más parecido a la arquea Pyrococcus furiosus, teniendo un parecido de 82.1% y después la arquea Thermococcus barophilus con un 81.2%

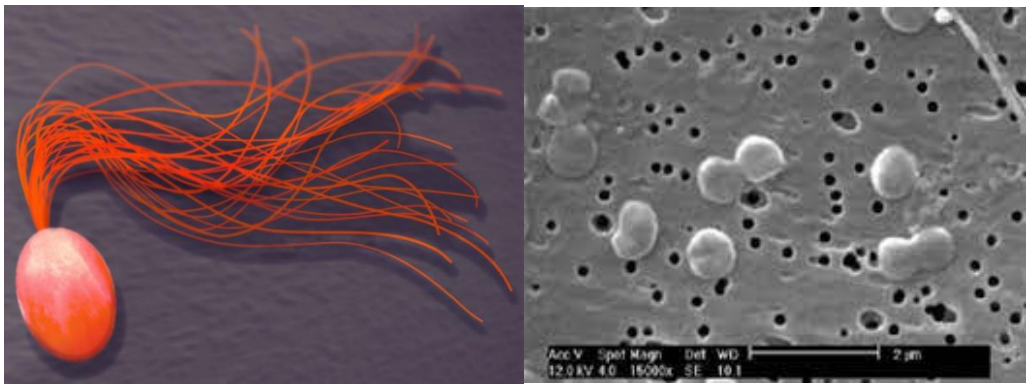

Fig. 7. Imágenes de Pyrococcus furiosus y Thermococcus barophilus, respectivamente

Por lo que podríamos concluir que puede pertenecer a algun tipo de arquea, pareciéndose más al del Pyrococcus furiosus.

La última parte de la práctica fue hecha en equipo, en esta parte se necesitaba reunir 100 secuencias de poli ubiquitina de diferentes organismos, realizar una alineación y usar la cladograma resultante para ilustrar el "árbol de la vida" agregando imágenes de algunos de los organismos.

Luego de elegir los organismos que íbamos a usar para llevar a cabo el alineamiento, se obtuvo el siguiente resultado:

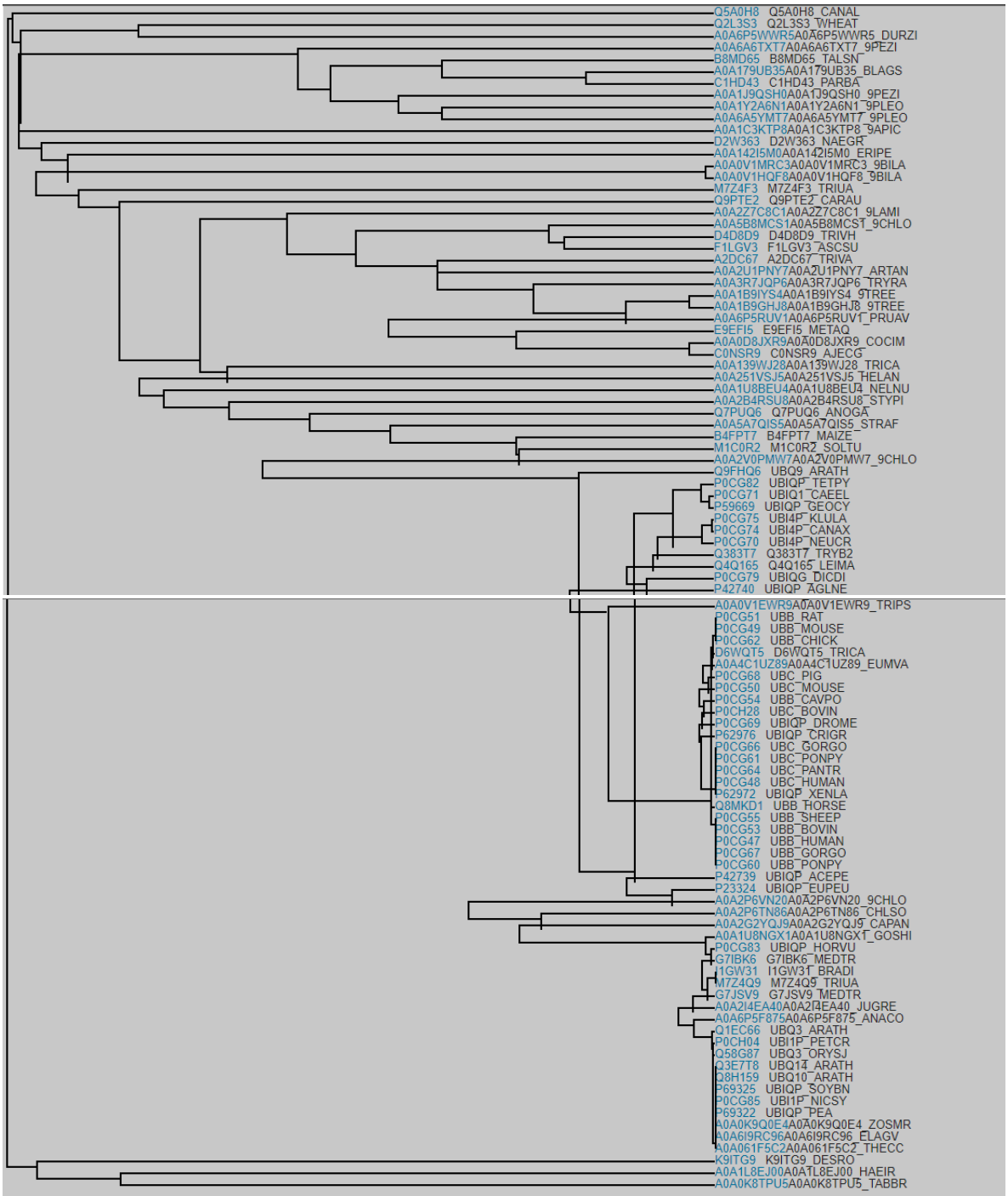

Fig. 8. Árbol resultante del alineamiento entre los 100 organismos con respecto a su poli ubiquitina

Algunos de los organismos usados son:

A0A0V1EWR9(Trichinella pseudospiralis||gusano redondo parasitario)

P0CG51 (Rattus norvegicus||Rata Parda)

P0CG49 (Mus musculus||Raton casera)

P0CG62(Gallus gallus||Gallo Asiático)

D6WQT5(Tribolium castaneum||Gorgojo castaño de la harina)

A0A4C1UZ89(Eumeta variegata||gusano de bolsa de paulownia)

P0CG68(Sus scrofa||Jabalí)

P0CG54(Cavia porcellus||Cuy)

P0CH28(Bos taurus||Vaca/Toro)

P0CG69(Drosophila melanogaster||Mosca de la fruta)

P62976(Cricetulus griseus||Hamster chino)

P0CG66(Gorilla gorilla gorilla||gorila occidental de llanura)

P0CG61(Pongo pygmaeus||orangután de Borneo)

P0CG64(Pan troglodytes||Chimpancé)

P0CG48(Homo sapiens||Humano)

P62972(Xenopus laevis|| rana de uñas africana)

Q8MKD1(Equus caballus||Caballo)

P0CG55(Ovis aries||Oveja)

P0CG53(Bos taurus||Ganado Vacuno)

P42739(Acetabularia peniculus||Acetabularia)

P23324(Euplotes eurystomus||Ciliado)

Aquí podemos ver en el árbol de la vida que los organismos que están primero se agrupan en mayor cantidad que los que están hasta bajo, perteneciendo en su mayoría a hongos y/o plantas y mamíferos, respectivamente, por lo que podemos ver que entre estos hongos y plantas hay mayor agrupación que con mamíferos, bacterias y parásitos, la hay, pero no en gran medida como en el otro caso.

Conclusiones y recomendaciones:

En esta práctica pudimos llevar a cabo el uso de la herramienta uniprot, una herramienta que nos sirve como base de datos para una serie de secuencias de proteínas y en ella pudimos ver un poco de lo que es el alineamiento de estas secuencias, para ver que tenían en común entre varias, además de usar el blast, que es una herramienta para buscar en toda la base de datos y tener una comparación y ver cual era la mas similar entre las registradas y la entrada dada. Me pareció una buena práctica ya que pudimos ver con los árboles que organismos compartían características, en estas podíamos ver como se asociaban organismos agrupándolos en ramas, como en el caso de los ratones, eran de diferentes especies, pero se asociaban, pero lo increíble era cuando lo mismo pasaba con otros organismos o incluso gusanos parasitarios y/o bacterias.

Bibliografía:

- *About UniProt*. Uniprot.org. (2022). Retrieved 2 April 2022, from https://www.uniprot.org/help/about.

- Apweiler, R. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research*, *32*(90001), 115D-119. https://doi.org/10.1093/nar/gkh131