**TALLER 1 MAESTRÍA BIOINFORMATICA Y BIOLOGIA COMPUTACIONAL**

**BASES DE DATOS BIOLÓGICAS**

Profesor : Luis F. Rivera

Tutor: Andres F. Londoño

Fecha Máxima Entrega: 21 Sep 2022

Grupo: A

Diego Mauricio Gómez Londoño

Juan Diego Polanco Franco , Nota:

1) Desde una máquina windows realice la descarga (ISO) del sistema operativo Ubuntu 20.4 Desktop e y proceder con la instalación apoyándose en lo practicado en clase.

- Una vez cargado el sistema abrir la terminal e instalar el visor interactivo de procesos “htop” con el comando :

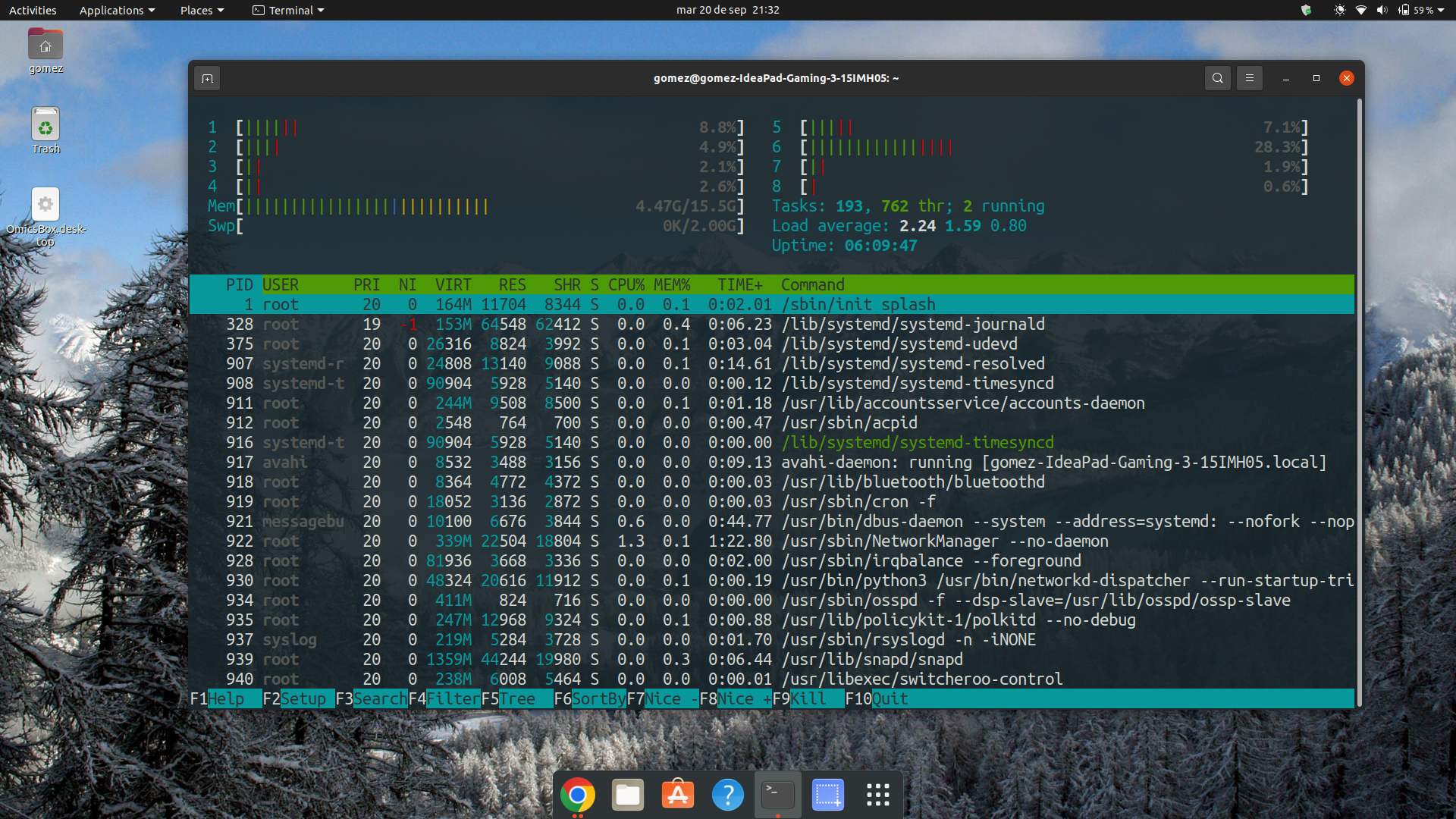
*sudo apt-get install htop*

- Ejecutar en la terminal el comando htop y adjuntar screenshot .

- Usar el navegador firefox para descargar el instalador de chrome para ubuntu y proceder con su instalación con el fin de tener acceso al drive y guardar los puntos siguientes de este taller.

- Tener debidamente instalado el Ubuntu 20.4 para la proxima sesion.

Ubuntu 20.04 se ejecuta de manera local en las computadoras y se han instalado tanto el navegador Google Chrome como el visor de procesos Htop:



2) Sobre el sistema anteriormente Instalado, realizar la instalación de Mongodb, generando un usuario admin protegido con contraseña para garantizar la protección de sus datos.

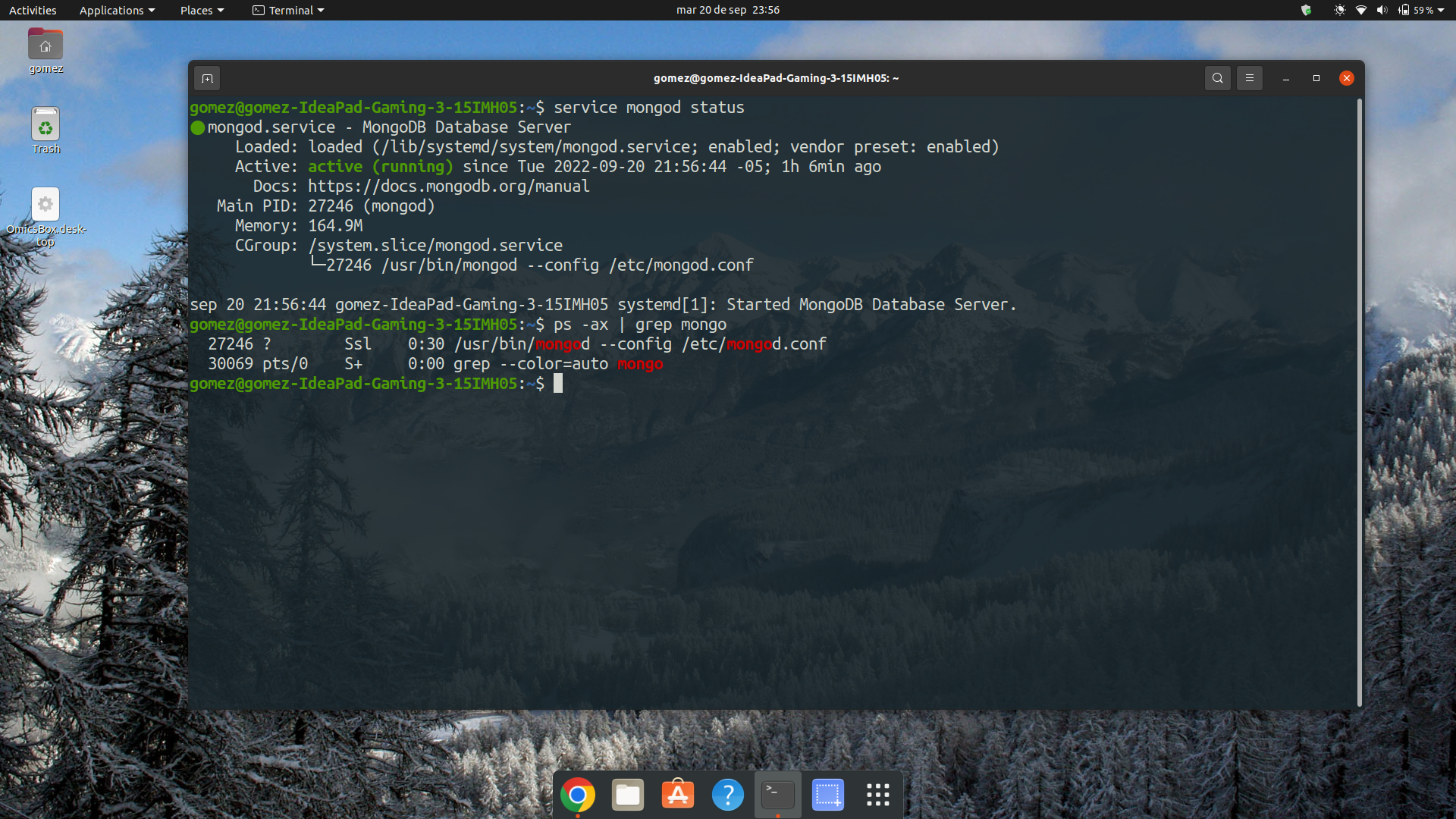
Pueden usar el siguiente link como guia.

<https://websiteforstudents.com/install-mongodb-on-ubuntu-18-04-lts-beta-server/>

* Después de la instalación respectiva abra una nueva terminal y ejecute los siguientes comandos y adjunte screenshots respectivos :

*service mongod status*

*ps -ax | grep mongo*



3) Se debe crear una Base de datos especializada en proteínas anticancerígenas cuya fuente no provenga de ratas y ratones(ver opciones avanzadas de busqueda ), descargando de uniprot (<https://www.uniprot.org/>) TODA la información correspondiente que incluya específicamente los siguientes campos:

**Entry Entry\_name Status Organism Length Sequence EC\_number PubMed\_ID**

Después de extraer la información respectiva responda las siguientes preguntas adjuntando el comando usado.

1. Número de proteínas encontradas.

7732

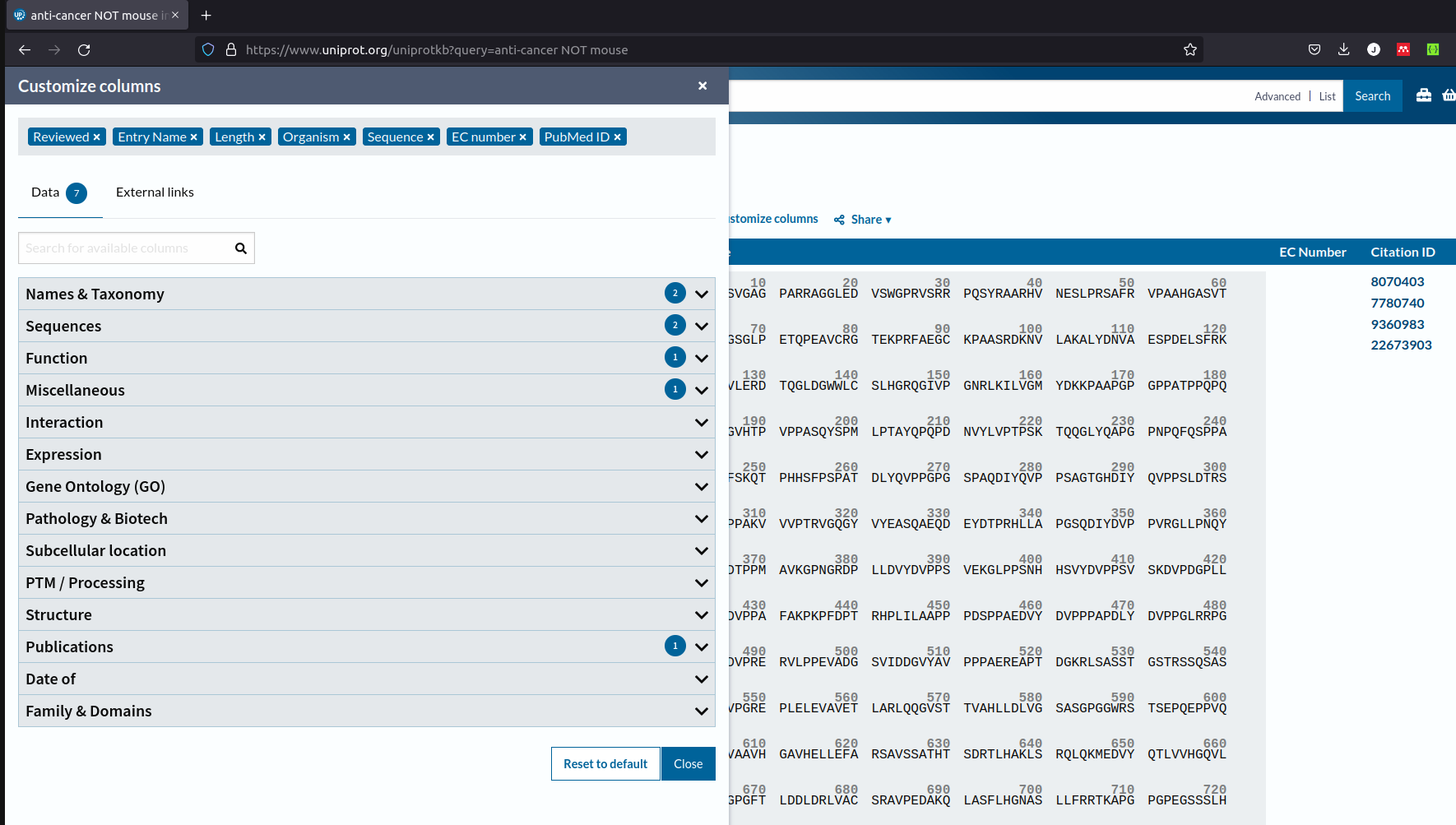
1. Número de proteínas con algún soporte o revisión experimental.

109

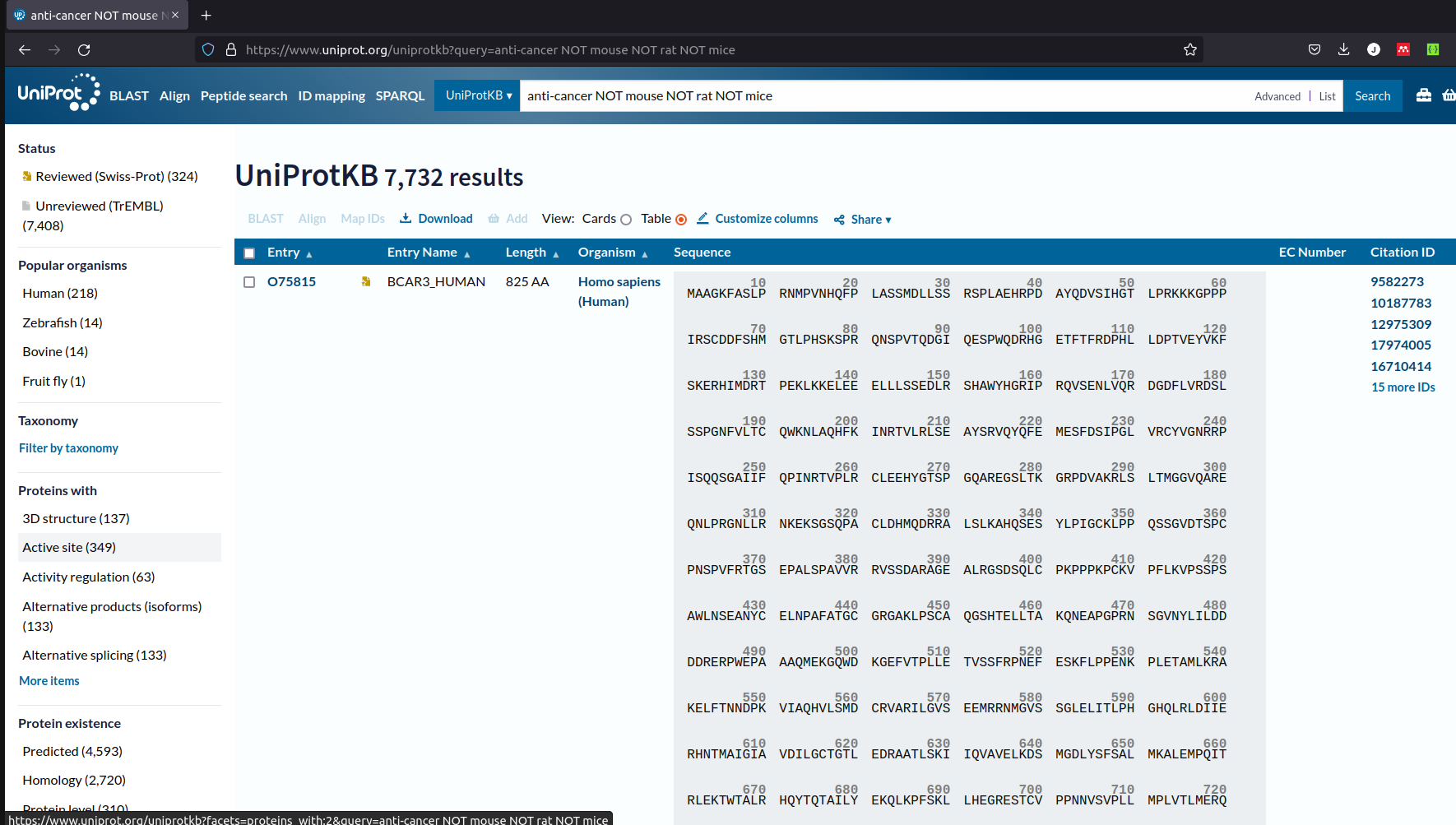
1. A partir del archivo descargado genere un archivo fasta con las secuencias de péptidos menores o iguales a 50 residuos de aminoácidos.(puede usar comandos como awk, sed, atom, y/o editor de preferencia ) , Cuantas secuencias se generaron?

176

4) Genere un archivo JSON a partir del archivo descargado en el punto 3, realice la importación o carga a mongoDB y realice tres diferentes búsquedas que considere relevantes. (Adjuntar archivo JSON , screenshots que considere necesarios y los comandos empleados en todo el proceso).

**Búsqueda y selección de proteínas anticancerígenas** 

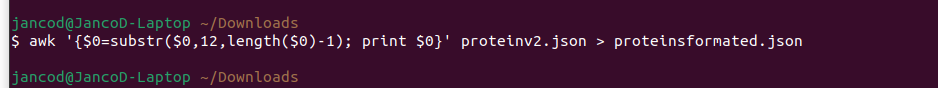
7732 Resultados



Descarga del archivo en formato JSON, se descomprime

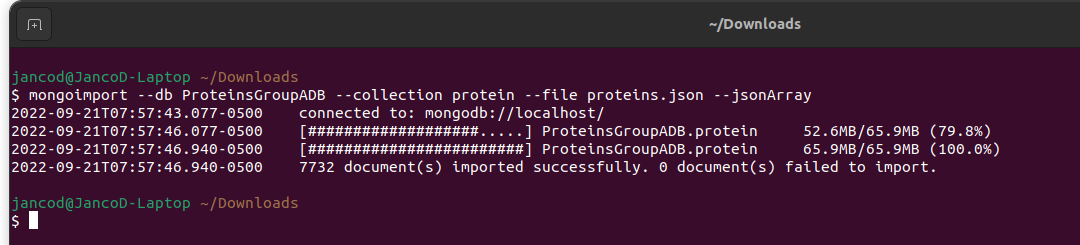
Formateo del Archivo JSON

awk '{$0=substr($0,12,length($0)-1); print $0}' proteinv2.json > proteinsformated.json

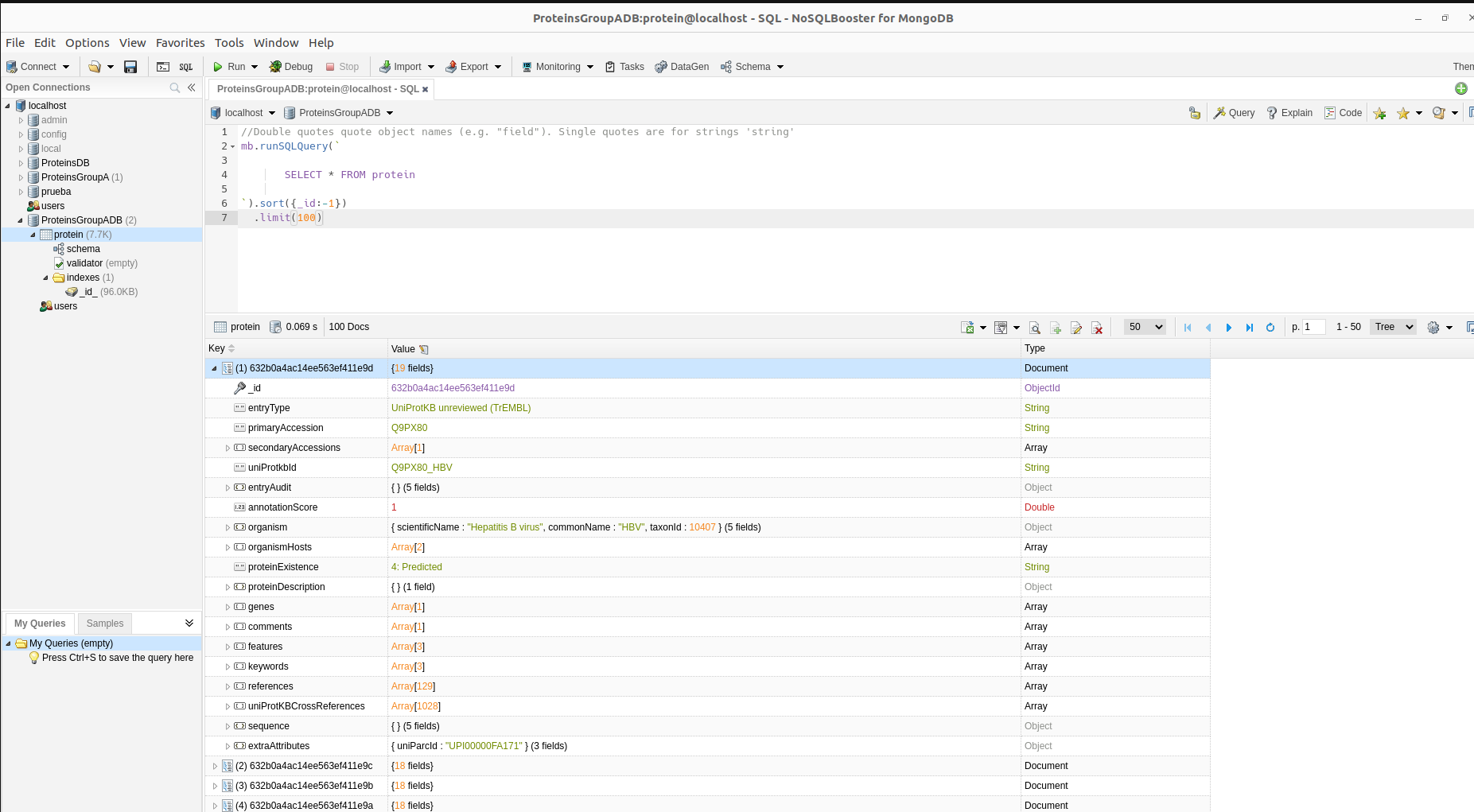


Importacion del archivo Formateado a MongoDB

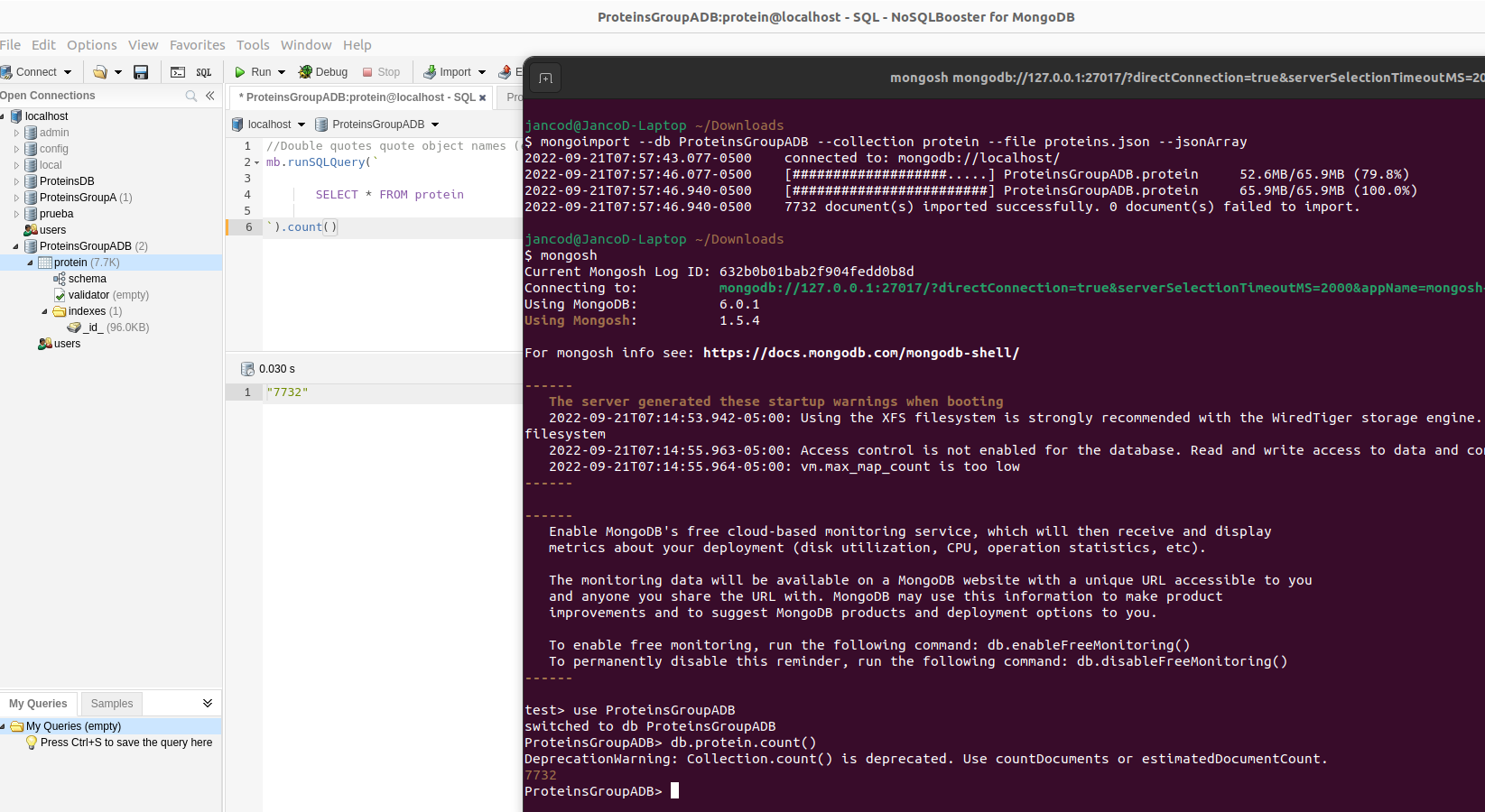
mongoimport --db ProteinsGroupADB --collection protein --file proteins.json --jsonArray



Se puede observar en un cliente MongoDB la información importada

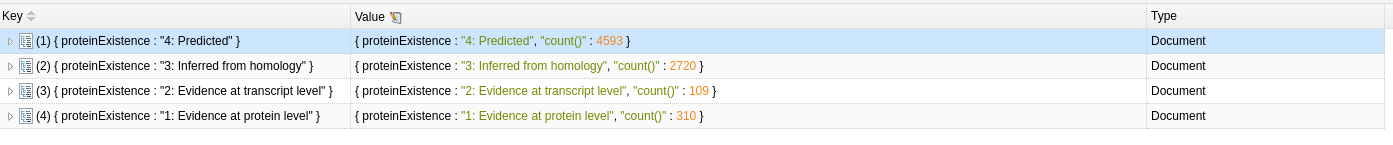


Se cuentan proteinas



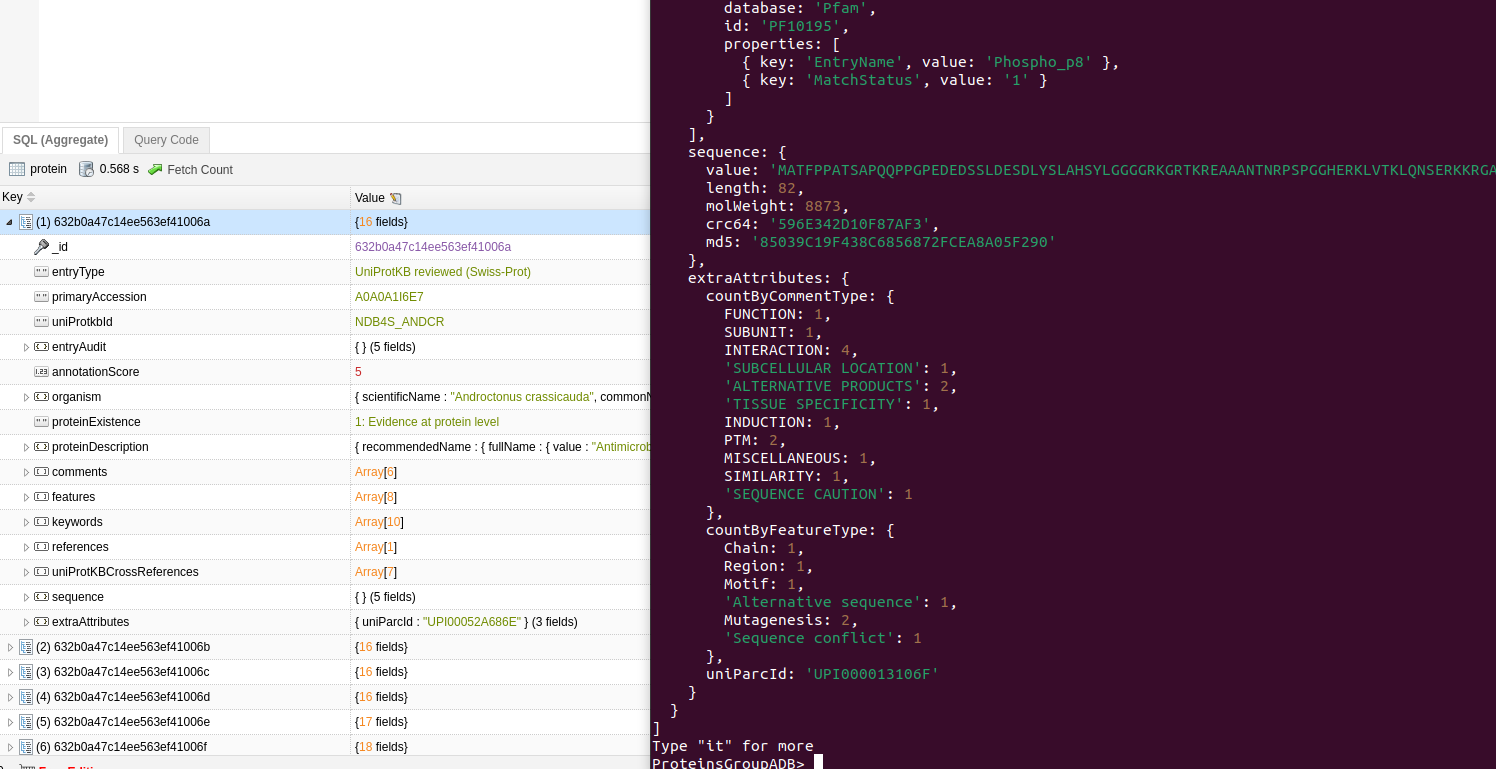
Tipo de identificación del de las proteínas y cuantas hay de acuerdo a su tipo





Proteinas con mejor Score





**Revisión de secuencias y filtro por número de aminoácidos (menor o igual a 50)**



