(a) Writing a small Python code to identify the unique sequence?

from Bio import SeqIO

import re

fasta\_filename='Protein\_seq.fasta'

fasta\_seq=[]

fasta\_des=[]

#loop to get sequence id,seq and description

for record in SeqIO.parse(fasta\_filename, "fasta"):

fasta\_seq.append(str(record.seq))

fasta\_des.append(repr(record.description))

#create dataframe

df = pd.DataFrame(data={'fasta\_des':fasta\_des,'fasta\_seq':fasta\_seq}, index = fasta\_seq)

#remove duplicates and give unique sequences

df = df[~df.index.duplicated()]

#list of sequence

unique\_seq=list(df.fasta\_seq)

print("Total sequences are",len(fasta\_seq))

print("Total unique sequences are",len(unique\_seq))

#print(\*unique\_seq,sep = "\n")

(b) Using Python code how you separate the Species from that unique sequences?

#create list of unique sequences description

seq\_desc=list(df['fasta\_des'])

def ExtractSpecies(seq\_desc):

species=list()

for i in range(len(seq\_desc)):

species.append(re.findall(r"\[(.\*?)\]", seq\_desc[i])[0])

s=set(species)

return(s)

#call function

species=ExtractSpecies(seq\_desc)

print(species)