Предмет: Прикладная биоинформатика
Исполнительница: Смолкина Ю.А
Группа: Адбм
Домашнее задание №1

Смолкина Ю. А.	CP007222.1	ATP synthase subunit alpha

ЧАСТЬ 1

После чтения файла :

```
[57] for seq_record in list(SeqIO.parse("/content/CP007222.1.fasta", "fasta"))[:1]:
         print(repr(seq_record.seq))

     my_seq_str = str(seq_record.seq)
     my_seq_str

Seq('GGTAATTGCCTCGCATAACGCGGTGTGAAAATGGATTGAAGCCCGGGCGGTGGA...CAT')
'GGTAATTGCCTCGCATAACGCGGTGTGAAAATGGATTGAAGCCCGGGCGGTGGATTCTACTCAACTTTAGCCGATGGAGAAAGCCCCGGGATCCGGGCTATTAAAAAGAAGATCTTTTTATTTAGAGATCTGTTCTATTGTGATCTCTTATTAGGATCGCTCCAGGCTGTGGATAACCCGGATCCTGTAATAA
AGATCAATGCGTTGGAAAGGATCACTAGCTGTGAATGATCGGTGATCGTGGCCCGTATAAGCTGGGATCAAAACGGGTACTTATACACAACTCAAAAAGTGAACAACGGTTATTCTTTGGATAACTACCGGTTGATCCAAGCTTTCCACCAGATTTATCCACAATGGATCGCACGATCTTTACACTTATTTGAG
TAAATTAATCCAGGATCCGAGCCAAATCTCCGCTGGATCTTCCGGAATCTCATGTTCAAGGATGTTGATCTTCAGTGTTTCCCCAACCTGTTTTGCGCCAGCGCCTTTCAGTTCTGCTTCTATTTTCTCAATCGCGCCACAAAACGTGTCGTATTCTCGACTACCAATGCCAATTGCGCCGAAACGTACCGCGG
AAAGATCGGGTTTCTGCGTCTGAAGGTCTTCATAGAAAGGGGTCAGGTTGTCCGGAATGTCTCCGGCACCGTGCGTTGAGCTGATTATCAGCCAGATCCCGGAAGTTGACAGATCCTCTAATAACGGACCGTGCACCGTTTCGGTTGAAAAACCGGCAGCTTCCAGCTTTTCCGCCAGATGTTCCGCGACGTAT
TCGGCGCCGCCCAGGGTGCTGCCGCTGATAAGAGTAATGTCTGCCATTGATCGCTCGCTCAGGTAAAGAGGCTGCATTGTACTCTGTGAACGAGCTGGGATCTACCTGTGGAAAAATGTGGGATTAAAAAAGCCGATCATGGCTTGATGGTGCGCATGATCGGGTTCTGCAAGACGATCAGTGTCTCAGTGGAC
TGAATTTCATCAATTGTTTGGATCTTGTTGA...'

[56] my_seq = Seq(my_seq_str)
     type(my_seq)
     my_seq

Seq('GGTAATTGCCTCGCATAACGCGGTGTGAAAATGGATTGAAGCCCGGGCGGTGGA...CAT')
```

1.1. Разрезать геном на 10 частей. (команда **splitter**)
1.2. Посчитать число слов AAAA, ATAT, ATTA, AATT (команда **compseq**)
1.3. Перемешать каждую из 10 частей (команда **shuffleseq)**
1.4. Посчитать число слов AAAA, ATAT, ATTA, AATT в перемешанных последовательностях

Команды выполнялись с помощью : https://www.bioinformatics.nl/cgi-bin/emboss
Проверим, действительно исходный файл распарсился, используя **splitter**

```
[102] #SeqIO.read('/content/outseq_split.txt', "fasta")
      zero = list(SeqIO.parse('/content/outseq_split.txt', "fasta"))[0]#.__dict__
      repr(zero.seq)

'Seq('GGTAATTGCCTCGCATAACGCGGTGTGAAAATGGATTGAAGCCCGGGCGGTGGA...TGA')'

      second = list(SeqIO.parse('/content/outseq_split.txt', "fasta"))[2]#.__dict__
      repr(second.seq)

'Seq('CAGCTGGCTGCAACTGTTTATTAAAAACACAGCACTGTGCAAACACGAAAGTGG...CAG')'
```

Посчитаем в новом файле `outseq_split.txt` число слов AAAA, ATAT, ATTA, AATT с помощью **compseq**

AAAA 9018
ATAT 4725
ATTA 4746
AATT 4303

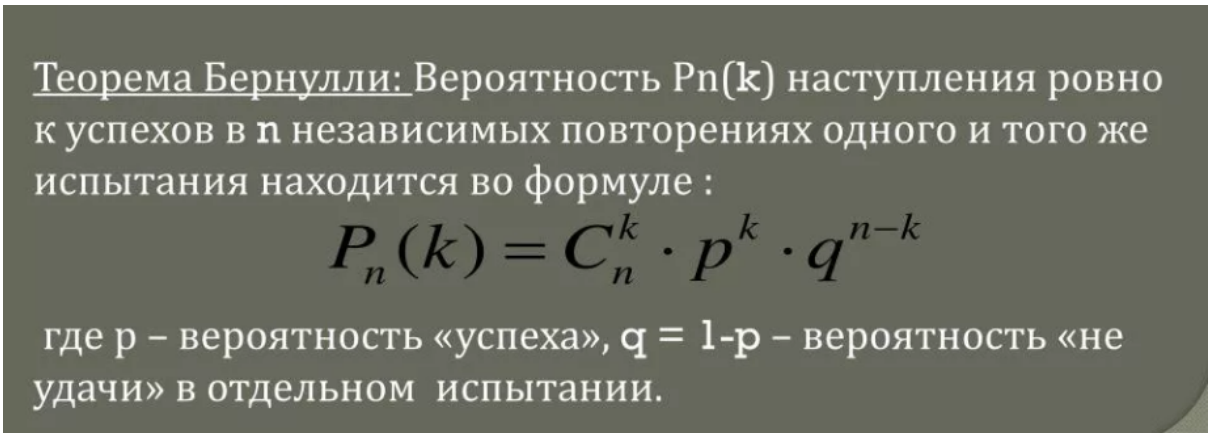
Теперь этот файл compseq_all перемешиваем с помощью **shuffleseq** и опять используем **compseq**

AAAA 3787

ATAT 3650
ATTA 3810
AATT 3757
Вид выгружаемого файла

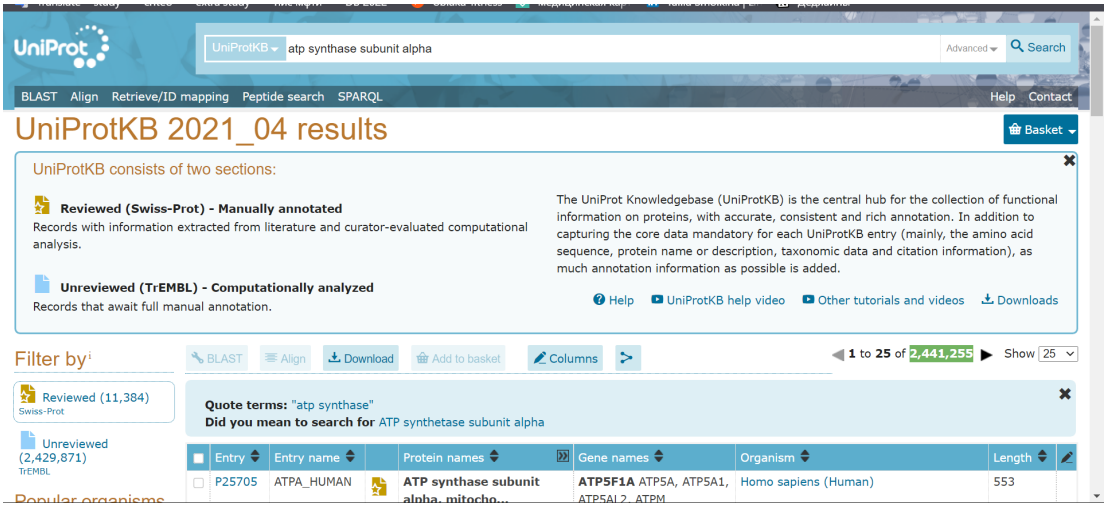
```
Word size       4
Total count     1140908

#
# Word  Obs Count       Obs Frequency   Exp Frequency   Obs/Exp Frequency
#
AAAA    3787            0.0033193       0.0039062       0.8497372
AAAC    4025            0.0035279       0.0039062       0.9031403
AAAG    4120            0.0036112       0.0039062       0.9244567
AAAT    3791            0.0033228       0.0039062       0.8506348
AACA    4106            0.0035989       0.0039062       0.9213153
AACC    4424            0.0038776       0.0039062       0.9926690
AACG    4385            0.0038434       0.0039062       0.9839181
AACT    4103            0.0035963       0.0039062       0.9206422
AAGA    4107            0.0035998       0.0039062       0.9215397
AAGC    4473            0.0039206       0.0039062       1.0036637
AAGG    4403            0.0038592       0.0039062       0.9879570
AAGT    4120            0.0036112       0.0039062       0.9244567
```

слово	**до**	**после**
AAAA	9018	3787
ATAT	4725	3650
ATTA	4746	3810
AATT	4303	3757



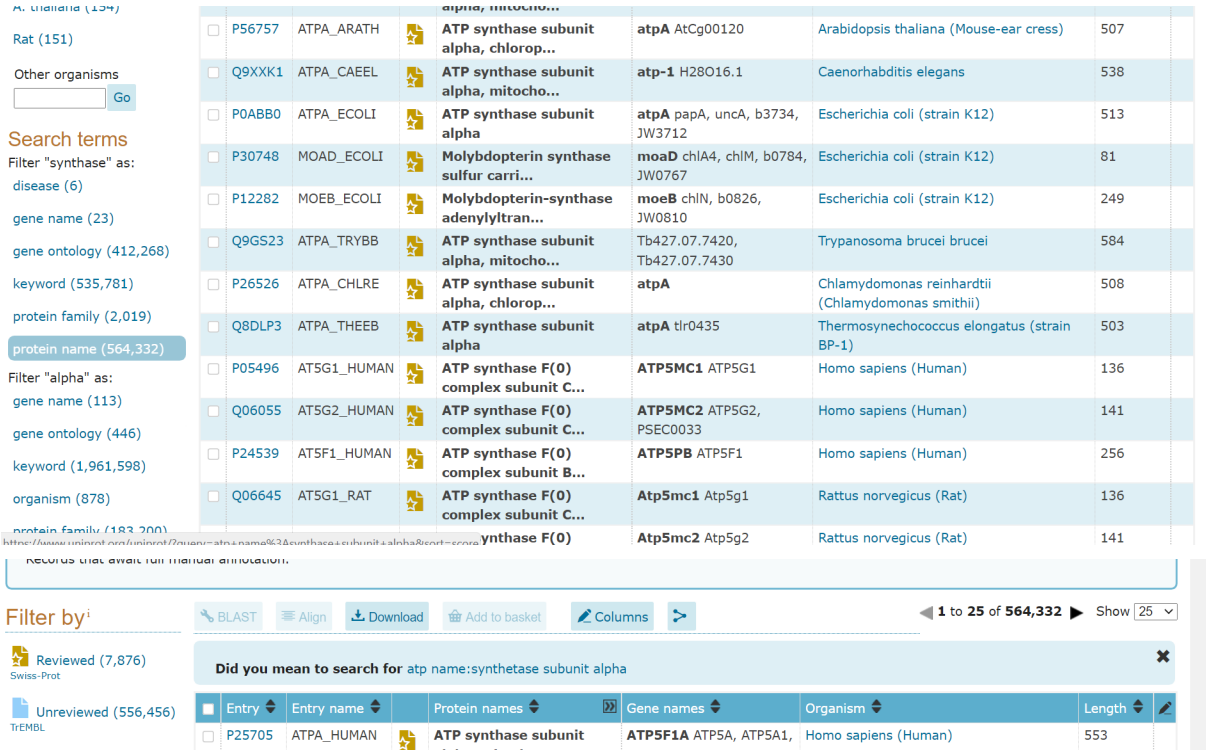
ЧАСТЬ 2

Сколько результатов найдено в UniprotKB по запросу в виде названия белка ATP synthase subunit alpha?В Swiss-Prot? В TrEMBL?



общее кол-во для
UniprotKB - 2441255
Swiss-Prot - 11384
TrEMBL (unreviewed) - 2429871

Сколько результатов найдено при использовании расширенного поиска в поле protein name? В Swiss-Prot? В TrEMBL?

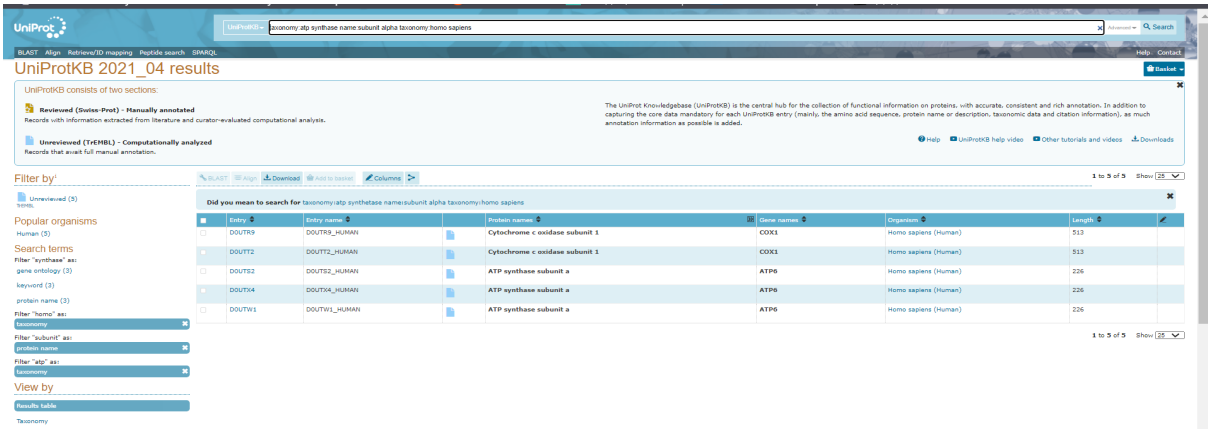


то есть уже

UniprotKB - 564322
Swiss-Prot - 7876

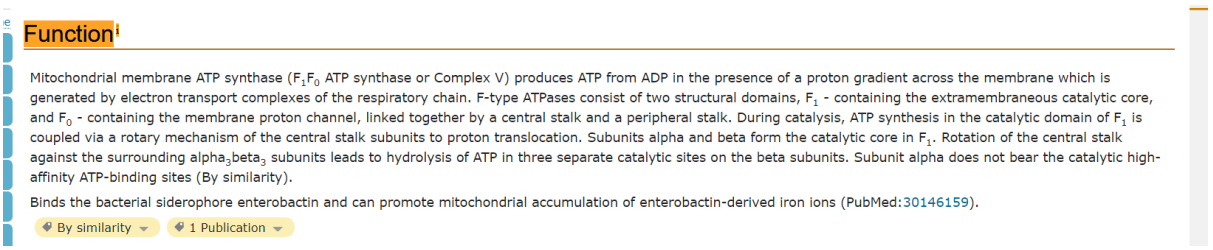
TrEMBL (unreviewed) - 556456

Сколько результатов остается при добавлении фильтра Homo sapiens в поле Taxonomy? В Swiss-Prot? В TrEMBL?

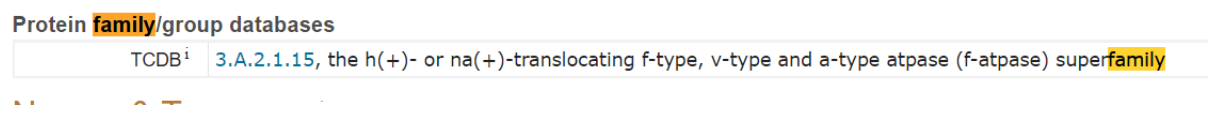


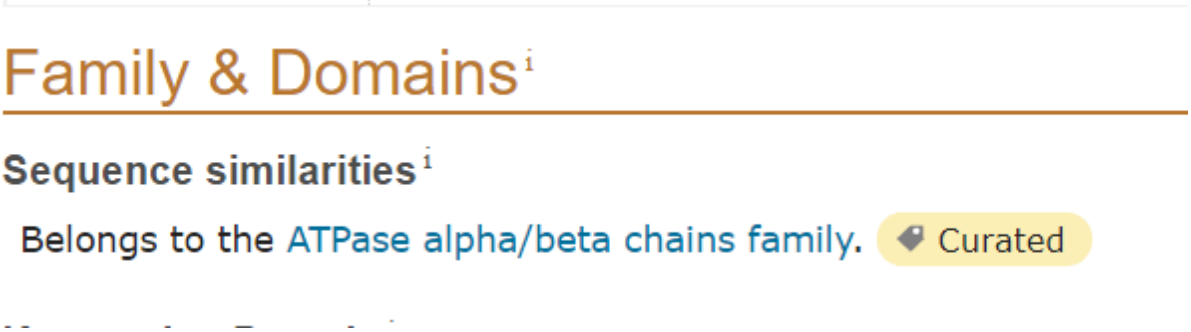
UniprotKB - 5
Swiss-Prot - 1
TrEMBL (unreviewed) - 3

Откройте запись о вашем белке и ответьте на следующие вопросы:
Какова функция белка?



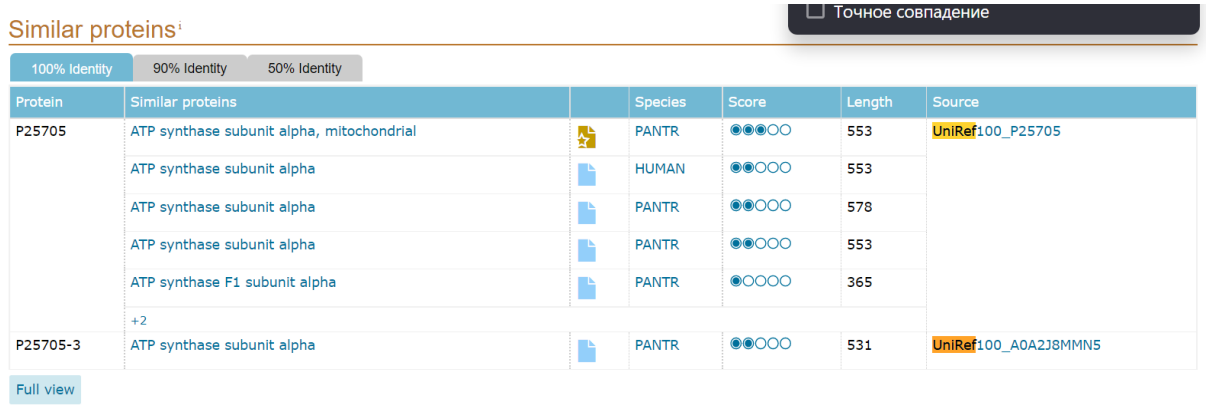
К какому семейству он принадлежит?



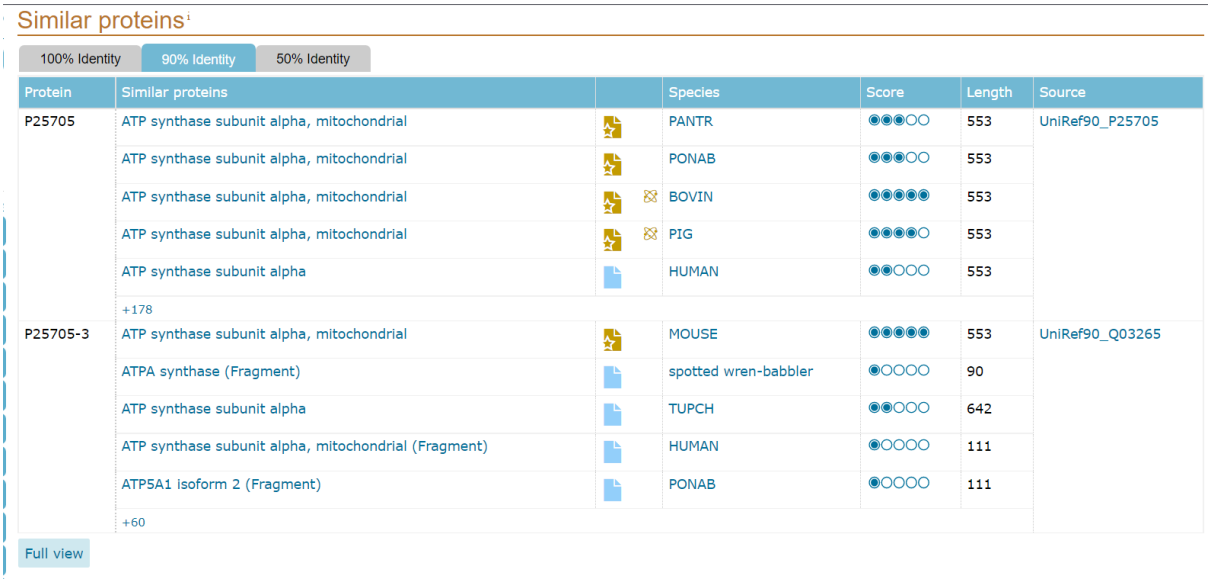


К скольки кластерам UniRef с идентичностью 1.0, 0.9, 0.5 принадлежит белок (включая изоформы)?

Для 1.0 **7**



Для 0.9 **188**



Для 0.5 **417**

BLAST | Align | Download | Add to basket | Columns

1 to 25 of 417 ▶ Show 25

Retrieve P25705

Entry	Entry name		Protein names	Gene names	Organism	Length
Q5R546	ATPA_PONAB		**ATP synthase subunit alpha, mitocho...**	**ATP5F1A** ATP5A1	Pongo abelii (Sumatran orangutan) (Pongo pygmaeus abelii)	553
Q03265	ATPA_MOUSE		**ATP synthase subunit alpha, mitocho...**	**Atp5f1a** Atp5a1	Mus musculus (Mouse)	553
A5A6H5	ATPA_PANTR		**ATP synthase subunit alpha, mitocho...**	**ATP5F1A** ATP5A1	Pan troglodytes (Chimpanzee)	553
P80021	ATPA_PIG		**ATP synthase subunit alpha, mitocho...**	**ATP5F1A** ATP5A1, ATP5A2	Sus scrofa (Pig)	553