

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('GGTAATTGCTCGCATAACGCGGTGTGAAAATGGATTGAAGCCCGGGCGGTGGA...CAT')
AGATCAATGGCTTGGAAAGGATCACTAGCTGTGAATGATCGGTGATCGTGGCCGTATAAGCTGGGATCAAAACGGGTACTTATACACAACCTAAAAAGTGAACAACGGTTATTCTTTGGATACTACCGTTGATCCAAAGCTTTCACACAGATTTATCCACAATGGATCGCACGATCTTACACTTATTGGAG
TAAATTAATCCAGATCCGAGCCAAATCTCCGCTGGATCTTCCGGAATCTCATGTTCAAGGATGTTGATCTTCAAGTGTTCCTCCCAACTGTTTTCGCGCAGCGCTTTCAGTTCTGCTTCTATTTTCAATCGCGCCACAAAACGTCGTGATCTCGACTACCAATGCCAATTGCGCCGAAACGTACCGCGG
AAAGATCGGGTTTCTGCTCTGAAGGTTTCATAGAAAAGGGGTAGGTTGTCGGAATGTCCTCGGCACTGTCGCTGAGCTGATTTACAGCCAGATCCCAGAAAGTTGACAGATCTCTAATAACGGACCGTGCACCTTCGGTTGAAAAACCGGCAGCTTCCAGCTTTTCCGCCAGATGTTCCGCCAGCGAT
TCGGCGCCGCGCCAGGGTGTCTCCGCTGATAGAGTAATGTCCTGCAATGATCGCTCAAGTAAAGAGGCTGCATTGTAATCTGTGAACGAGCTGGGATCTACTGTGGAAAAATGTGGGATTAAAAAGCCGATCATGGCTTGATGGTGCAGATGATCGGGTTCTGCAAGACGATCAGTGTCTCAGTGGAC
TGAATTTTCATCAATTGTTGGATCTGTTGA...
```

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

Seq('GGTAATTGCTCGCATAACGCGGTGTGAAAATGGATTGAAGCCCGGGCGGTGGA...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('GGTAATTGCTCGCATAACGCGGTGTGAAAATGGATTGAAGCCCGGGCGGTGGA...TGA')'

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

'Seq('CAGCTGGCTGCAACTGTTTATTAACACAGCACTGTGCAACACGAAAGTGG...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

Теорема Бернулли: Вероятность $P_n(k)$ наступления ровно k успехов в n независимых повторениях одного и того же испытания находится во формуле :

$$P_n(k) = C_n^k \cdot p^k \cdot q^{n-k}$$

где p – вероятность «успеха», $q = 1-p$ – вероятность «неудачи» в отдельном испытании.

ЧАСТЬ 2

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

UniProtKB consists of two sections:

- Reviewed (Swiss-Prot) - Manually annotated**
Records with information extracted from literature and curator-evaluated computational analysis.
- Unreviewed (TrEMBL) - Computationally analyzed**
Records that await full manual annotation.

The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added.

Filter by:

- Reviewed (11,384) Swiss-Prot
- Unreviewed (2,429,871) TrEMBL

Quote terms: "atp synthase"
Did you mean to search for ATP synthetase subunit alpha

Entry	Entry name	Protein names	Gene names	Organism	Length
P25705	ATPA_HUMAN	ATP synthase subunit alpha, mitocho...	ATP5F1A ATP5A, ATP5A1, ATP5A1.2, ATPM	Homo sapiens (Human)	553

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?

Search terms

Filter "synthase" as:
disease (6)
gene name (23)
gene ontology (412,268)
keyword (535,781)
protein family (2,019)
protein name (564,332)

Filter "alpha" as:
gene name (113)
gene ontology (446)
keyword (1,961,598)
organism (878)

Entry	Entry name	Protein names	Gene names	Organism	Length
P56757	ATPA_ARATH	ATP synthase subunit alpha, chlorop...	atpA AtCg00120	Arabidopsis thaliana (Mouse-ear cress)	507
Q9XXK1	ATPA_CAEEL	ATP synthase subunit alpha, mitocho...	atp-1 H28016.1	Caenorhabditis elegans	538
P0AB80	ATPA_ECOLI	ATP synthase subunit alpha	atpA papA, uncA, b3734, JW3712	Escherichia coli (strain K12)	513
P30748	MOAD_ECOLI	Molybdopter synthase sulfur carri...	moaD chlA4, chlM, b0784, JW0767	Escherichia coli (strain K12)	81
P12282	MOEB_ECOLI	Molybdopter synthase adenyltran...	moeB chlN, b0826, JW0810	Escherichia coli (strain K12)	249
Q9GS23	ATPA_TRYBB	ATP synthase subunit alpha, mitocho...	Tb427.07.7420, Tb427.07.7430	Trypanosoma brucei brucei	584
P26526	ATPA_CHLRE	ATP synthase subunit alpha, chlorop...	atpA	Chlamydomonas reinhardtii (Chlamydomonas smithii)	508
Q8DLP3	ATPA_THEEB	ATP synthase subunit alpha	atpA tlr0435	Thermosynechococcus elongatus (strain BP-1)	503
P05496	AT5G1_HUMAN	ATP synthase F(0) complex subunit C...	ATP5MC1 ATP5G1	Homo sapiens (Human)	136
Q06055	AT5G2_HUMAN	ATP synthase F(0) complex subunit C...	ATP5MC2 ATP5G2, PSEC0033	Homo sapiens (Human)	141
P24539	AT5F1_HUMAN	ATP synthase F(0) complex subunit B...	ATP5PB ATP5F1	Homo sapiens (Human)	256
Q06645	AT5G1_RAT	ATP synthase F(0) complex subunit C...	Atp5mc1 Atp5g1	Rattus norvegicus (Rat)	136
		ATP synthase F(0) complex subunit C...	Atp5mc2 Atp5g2	Rattus norvegicus (Rat)	141

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?

Entry	Entry name	Protein name	Gene name	Organism	Length
DOUTS9	DOUTS9_HUMAN	Cytochrome c oxidase subunit 1	COX1	Homo sapiens (Human)	513
DOUTS2	DOUTS2_HUMAN	Cytochrome c oxidase subunit 1	COX1	Homo sapiens (Human)	513
DOUTS2	DOUTS2_HUMAN	ATP synthase subunit a	ATP6	Homo sapiens (Human)	226
DOUTX4	DOUTX4_HUMAN	ATP synthase subunit a	ATP6	Homo sapiens (Human)	226
DOUTW1	DOUTW1_HUMAN	ATP synthase subunit a	ATP6	Homo sapiens (Human)	226

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

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

Function

Mitochondrial membrane ATP synthase (F_1F_0 ATP synthase or Complex V) produces ATP from ADP in the presence of a proton gradient across the membrane which is generated by electron transport complexes of the respiratory chain. F-type ATPases consist of two structural domains, F_1 - containing the extramembraneous catalytic core, and F_0 - containing the membrane proton channel, linked together by a central stalk and a peripheral stalk. During catalysis, ATP synthesis in the catalytic domain of F_1 is coupled via a rotary mechanism of the central stalk subunits to proton translocation. Subunits alpha and beta form the catalytic core in F_1 . Rotation of the central stalk against the surrounding $\alpha_3\beta_3$ subunits leads to hydrolysis of ATP in three separate catalytic sites on the β subunits. Subunit alpha does not bear the catalytic high-affinity ATP-binding sites (By similarity).

Binds the bacterial siderophore enterobactin and can promote mitochondrial accumulation of enterobactin-derived iron ions (PubMed:30146159).

К какому семейству он принадлежит?

Protein family/group databases

TCDB ¹	3.A.2.1.15, the h(+)- or na(+)-translocating f-type, v-type and a-type atpase (f-atpase) superfamily
-------------------	--

Family & Domainsⁱ

Sequence similaritiesⁱ

Belongs to the **ATPase alpha/beta chains family**. Curated

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

Для 1.0 **7**

Similar proteinsⁱ

Точное совпадение

100% Identity	90% Identity	50% Identity				
Protein	Similar proteins	Species	Score	Length	Source	
P25705	ATP synthase subunit alpha, mitochondrial		PANTR	●●●○○	553	UniRef100_P25705
	ATP synthase subunit alpha		HUMAN	●●○○○	553	
	ATP synthase subunit alpha		PANTR	●●○○○	578	
	ATP synthase subunit alpha		PANTR	●●○○○	553	
	ATP synthase F1 subunit alpha		PANTR	●○○○○	365	
	+2					
P25705-3	ATP synthase subunit alpha		PANTR	●●○○○	531	UniRef100_A0A2J8MMN5
Full view						

Для 0.9 **188**

Similar proteinsⁱ

100% Identity	90% Identity	50% Identity			
Protein	Similar proteins	Species	Score	Length	Source
P25705	ATP synthase subunit alpha, mitochondrial	PANTR	●●●○○	553	UniRef90_P25705
	ATP synthase subunit alpha, mitochondrial	PONAB	●●●○○	553	
	ATP synthase subunit alpha, mitochondrial	BOVIN	●●●●●	553	
	ATP synthase subunit alpha, mitochondrial	PIG	●●●●○	553	
	ATP synthase subunit alpha	HUMAN	●●○○○	553	
	+178				
P25705-3	ATP synthase subunit alpha, mitochondrial	MOUSE	●●●●●	553	UniRef90_Q03265
	ATPA synthase (Fragment)	spotted wren-babbler	●○○○○	90	
	ATP synthase subunit alpha	TUPCH	●●○○○	642	
	ATP synthase subunit alpha, mitochondrial (Fragment)	HUMAN	●○○○○	111	
	ATP5A1 isoform 2 (Fragment)	PONAB	●○○○○	111	
	+60				
Full view					

Для 0.5 **417**

<div> <div>BLAST</div> <div>Align</div> <div>Download</div> <div>Add to basket</div> <div>Columns</div> <div></div> </div> <div>1 to 25 of 417</div> <div>Show 25</div>					
<div>Retrieve P25705</div>					
Entry	Entry name	Protein names	Gene names	Organism	Length
<input type="checkbox"/> Q5R546	ATPA_PONAB	ATP synthase subunit alpha, mitocho...	ATP5F1A ATP5A1	Pongo abelii (Sumatran orangutan) (Pongo pygmaeus abelii)	553
<input type="checkbox"/> Q03265	ATPA_MOUSE	ATP synthase subunit alpha, mitocho...	Atp5f1a Atp5a1	Mus musculus (Mouse)	553
<input type="checkbox"/> A5A6H5	ATPA_PANTR	ATP synthase subunit alpha, mitocho...	ATP5F1A ATP5A1	Pan troglodytes (Chimpanzee)	553
<input type="checkbox"/> P80021	ATPA_PIG	ATP synthase subunit alpha, mitocho...	ATP5F1A ATP5A1, ATP5A2	Sus scrofa (Pig)	553