

Отчет.

Часть 1

1) Для анализа были выбраны следующие протеины человека:

HELZ_HUMAN	CCS_HUMAN	JAK1_HUMAN	IRS4_HUMAN
UBC_HUMAN	NFIB_HUMAN	CHD6_HUMAN	DDB1_HUMAN
ZBP1_HUMAN	ACTB_HUMAN		

Далее при помощи продвинутого поиска в uniprot ищем ортологи для этих протеинов (протеины специально выбирались так, чтобы у каждого были ортологи).

H2QDQ4_PANTR	H2Q473_PANTR	H2PZ66_PANTR
A0A2I3SS11_PANTR	UBC_PANTR	H2RCU0_PANTR
A0A2I3SY12_PANTR	H2QKN2_PANTR	ACTB_PANTR

Далее выравниваем при помощи мегги попарно все эти протеины с их ортологами.

```
1. sp|P42694|HELZ_HUMAN Probable helicase with zinc finger domain OS=Homo sapiens OX=9606 GN=HELZ PE=1 SV=2
2. tr|H2QDQ4|H2QDQ4_PANTR Helicase with zinc finger OS=Pan troglodytes OX=9598 GN=HELZ PE=2 SV=1
M D R A K K C E A C E L K R D D Y E M A L K H C E A L L L G Y I M A D F G P C P L E I R I K I E L L Y R I A F L D L K N Y V A D E D C R H V L G E G L A I
H E D R A K K C E A C E L K R D D Y E M A L K H C E A L L L G Y I M A D F G P C P L E I R I K I E L L Y R I A F L D L K N Y V A D E D C R H V L G E G L A I

1. sp|Q14618|CCS_HUMAN Copper chaperone for superoxide dismutase OS=Homo sapiens OX=9606 GN=CCS PE=1 SV=1
2. tr|H2Q473|H2Q473_PANTR Superoxide dismutase copper chaperone OS=Pan troglodytes OX=9598 GN=CCS PE=2 SV=2
M A D G D G D L C L E F A V A N H C D C C V A V K K L U G V A V L S E V H L E D Q M V L V H L P F Q V V A L L G G R A V L K C H R D L G N L E A K
M A D G D G D L C L E F A V A N H C D C C V A V K K L U G V A V L S E V H L E D Q M V L V H L P F Q V V A L L G G R A V L K C H R D L G N L E A K

1. sp|P23458|JAK1_HUMAN Tyrosine-protein kinase JAK1 OS=Homo sapiens OX=9606 GN=JAK1 PE=1 SV=2
2. tr|H2PZ66|H2PZ66_PANTR Tyrosine-protein kinase OS=Pan troglodytes OX=9598 GN=JAK1 PE=2 SV=1
M Q Y L N I K E D C N A M A F C A K M R S K K T E V N L E A P E F G V E V I F Y L S D R E P L R L G S G E T A E E L C I R A A A C R I S P L C H N L F A L Y D E N T K L W Y A P R T I I
M Q Y L N I K E D C N A M A F C A K M R S K K T E V N L E A P E F G V E V I F Y L S D R E P L R L G S G E T A E E L C I R A A A C R I S P L C H N L F A L Y D E N T K L W Y A P R T I I

1. sp|Q14654|IRS4_HUMAN Insulin receptor substrate 4 OS=Homo sapiens OX=9606 GN=IRS4 PE=1 SV=1
2. tr|A0A2I3SS11|A0A2I3SS11_PANTR Insulin receptor substrate 4 OS=Pan troglodytes OX=9598 GN=IRS4 PE=4 SV=1
M A S C F R D Q A T R R L R G A A A A A A A A A A V V T P L L S G P T A L I G T G S S C P G A M M L S A T G S R S D S E S E E D L P V G E E V C K R G Y L R K K H G
M A S C F R D Q A T R R L R G A A A A A A A A A A V V T P L L S G P T A L I G T G S S C P G A M M L S A T G S R S D S E S E E D L P V G E E V C K R G Y L R K K H G

1. sp|P0CC48|UBC_HUMAN Polyubiquitin-C OS=Homo sapiens OX=9606 GN=UBC PE=1 SV=3
2. sp|P0CC64|UBC_PANTR Polyubiquitin-C OS=Pan troglodytes OX=9598 GN=UBC PE=1 SV=1
M I F V K K L U G K I I L E V E P S D T I E V K A K I D D K E G P P D Q R L I F A K K L E D G R L E D Y N I K E S T L H L V L R L R G G M I F V K I L G K T I I L E V E P S D T I E N
M I F V K K L U G K I I L E V E P S D T I E V K A K I D D K E G P P D Q R L I F A K K L E D G R L E D Y N I K E S T L H L V L R L R G G M I F V K I L G K T I I L E V E P S D T I E N

1. sp|Q00712|NFIB_HUMAN Nuclear factor 1 B-type OS=Homo sapiens OX=9606 GN=NFIB PE=1 SV=2
2. tr|H2RCU0|H2RCU0_PANTR Nuclear factor 1 OS=Pan troglodytes OX=9598 GN=NFIB PE=3 SV=2
M M Y S P I C L T D D F H P F I E A L L P H R A I A Y T F L A R K R K Y F K K H K R M K D E E R A V K D E L L E K P E I K K M A R L L A K R K D I R D E Y R E D F V L V T C
M M Y S P I C L T D D F H P F I E A L L P H R A I A Y T F L A R K R K Y F K K H K R M K D E E R A V K D E L L E K P E I K K M A R L L A K R K D I R D E Y R E D F V L V T C

1. sp|Q8TD26|CHD6_HUMAN Chromodomain-helicase-DNA-binding protein 6 OS=Homo sapiens OX=9606 GN=CHD6 PE=1 SV=4
2. tr|H2QKD4|H2QKD4_PANTR DNA helicase OS=Pan troglodytes OX=9598 GN=CHD6 PE=2 SV=1
M K K I Q K K E K L S N L V L N H S P H S D A S V F Y K S P F F D C S T D D E E K I E D V A H C L P K D L Y A E E E A A T L F P R K M T H N G M E D S G C
M K K I Q K K E K L S N L V L N H S P H S D A S V F Y K S P F F D C S T D D E E K I E D V A H C L P K D L Y A E E E A A T L F P R K M T H N G M E D S G C

1. sp|Q16531|DDB1_HUMAN DNA damage-binding protein 1 OS=Homo sapiens OX=9606 GN=DDB1 PE=1 SV=1
2. tr|A0A2I3SY12|A0A2I3SY12_PANTR DNA damage-binding protein 1 OS=Pan troglodytes OX=9598 GN=DDB1 PE=3 SV=1
D M N R L K V I K Y S G K I E H F P W R F H T E R K K E P A T G F I D D L I E F L D I R P K M E V V A N L Y D D G G M K R E A A D D L I K V V E I L T R I H
D M N R L K V I K Y S G K I E H F P W R F H T E R K K E P A T G F I D D L I E F L D I R P K M E V V A N L Y D D G G M K R E A A D D L I K V V E I L T R I H

1. sp|Q9H171|ZBP1_HUMAN Z-DNA-binding protein 1 OS=Homo sapiens OX=9606 GN=ZBP1 PE=1 SV=2
2. tr|H2QKN2|H2QKN2_PANTR Z-DNA binding protein 1 OS=Pan troglodytes OX=9598 GN=ZBP1 PE=4 SV=2
N S N K M I P C V A P G G V A S G E G E P G E D A R R P A D T Q R H F F R I G O F I T P H K K L P K L E T M T L G R R H K A A G H Y V D E A S H E G S W W G G G I
N S N K M I P C V A P G G V A S G E G E P G E D A R R P A D T Q R H F F R I G O F I T P H K K L P K L E T M T L G R R H K A A G H Y V D E A S H E G S W W G G G I

1. sp|P60709|ACTB_HUMAN Actin cytoplasmic 1 OS=Homo sapiens OX=9606 GN=ACTB PE=1 SV=1
2. sp|Q5R1X3|ACTB_PANTR Actin cytoplasmic 1 OS=Pan troglodytes OX=9598 GN=ACTB PE=2 SV=1
F S I N K C V D I R K D L Y A T V L G G T M Y P G I A D R H K E I T A L A P S T M K I I A P P E R K Y V N I G G I L A L S L T F D M M I K E Y D E S G P I V H R K C F
F S I N K C V D I R K D L Y A T V L G G T M Y P G I A D R H K E I T A L A P S T M K I I A P P E R K Y V N I G G I L A L S L T F D M M I K E Y D E S G P I V H R K C F
```

И при помощи BLAST ищем среднее identity и similarity:

NAME	IDENTITIES	SIMILARITIES
H2QDQ4_PANTR	99%	99%
H2Q473_PANTR	99%	98%
H2PZ66_PANTR	99%	99%
A0A2I3SS11_PANTR	97%	98%
UBC_PANTR	100%	100%
H2RCU0_PANTR	93%	92%
H2QKD4_PANTR	98%	98%
A0A2I3SY12_PANTR	98%	98%
H2QKN2_PANTR	99%	98%
ACTB_PANTR	99%	99%

Таким образом, среднее identity равно 98.1%, среднее similarity 97.9

Часть 2

Задание 1

0) Просто считаем по формуле. После 10 циклов ПЦР:

$$(2 * 2^{10}) / (2 * 2^{10} + 3 * 2^{10}) = 0.4, \text{ т.е. } 40\%$$

После 40 циклов ПЦР:

$$(2 * 2^{40}) / (2 * 2^{40} + 3 * 2^{40}) = 0.4, \text{ т.е. те же самые } 40\%$$

1) Загружаем файл 80.fasta, подгружаем его в blast. Смотрим на все ряды, выбираем в каждом лучшую находку. Ставим ограничение, чтобы identity было не менее 99%. Результаты сохраняем в файл reads.txt (сохраняем только названия видов у которых лучшая находка для каждого read). Если в read нет ничего, оставляем пустую строчку. Результаты сохранены в файле read.txt. Как мы можем наблюдать, хотя несколько чтений (а именно 2) и принадлежат *Felis catus*, все же чтений, которые принадлежат *Canis lupus familiaris* больше, а именно 6. Кроме того есть 3 чтения, которые принадлежат кошачьим (*Acinonyx jubatus*, *Lynx canadensis*, *Panthera pardus*) и 8 которые принадлежат *Canis lupus* (волку). Опираясь на этот результат, можно сделать вывод, что скорее всего виновник загрязнения это владелец собаки, т.е. Иванов (т.к. чтений, которые принадлежат *Canis lupus familiaris* больше, чем чтений, которые принадлежат *Felis catus*. Если считать все чтения, которые принадлежат кошачьим и собачьим, то тоже складывается впечатление, что больше загрязнения пришло от владельца собаки.

2) Теперь давайте воспользуемся файлом, который мы получили в пункте 1. Вот сводка по организмам:

Вид	Проценты
Не определен	20%
<i>Felis catus</i>	2%
<i>Canis lupus familiaris</i>	6%
<i>Canis lupus</i>	8%
<i>Acinonyx jubatus</i>	1%
<i>Lynx canadensis</i>	1%
<i>Panthera pardus</i>	1%
<i>Sciurus carolinensis</i>	1%
<i>Homo sapiens</i>	11%

Mus musculus	48%
Mastomys coucha	1%

Итак, мы видим, что среди источников загрязнения есть 3 основных группы. Первая – это человек (возможно лаборант занес свое днк или что-нибудь в этом роде). Также есть кошачьи и собачьи, причем процент собачьих значительно больше. Основные выводы по этому поводу сделаны в пункте 1, тут можно лишь добавить, что возможно не один Иванов виноват в загрязнении, вина Петрова тоже есть, т.к. кошачьи откуда-то взялись. Все остальное – это 48% Mus musculus днк которой исследовали и по 1% на 2 других грызунов.

Задание 2

- а) Воспользуемся следующим скриптом, чтобы сгенерировать требуемый fasta файл:

```
with open('seq.fasta', 'w') as file:
    s =
    "CCGGCAGCCCTGTTATTGTTTGGCTCCACATTTACATTTCTGCCTCTTGCAGCAGCATT
    TCCGGTTTCTTTTTGCCGGAGCAGCTCACTATTCACCCGAT"
    print(len(s))
    for i in range(100):
        file.write('>fragment_')
        file.write(str(100-i))
        file.write('\n')
        file.write(s[:100-i] + '\n')
```

Далее загружаем этот файл в blast, пока что не ограничиваем поиск только человеком. Получаем, что последняя длина последовательности, когда E-value ещё нормальное (меньше 0.05) это 25

blast.ncbi.nlm.nih.gov/Blast.cgi

NIH National Library of Medicine
National Center for Biotechnology Information

BLAST® » blastn suite » results for RID-63GE2Z91016

Home Recent Results Saved Strategies Help

< Edit Search Save Search Search Summary ▾

Job Title **fragment_100**

RID **63GE2Z91016** Search expires on 05-16 13:23 pm [Download All ▾](#)

Results for **76-icjQuery_45787 fragment_25(25bp)** ▾

Program **BLASTN** [Citation ▾](#)

Database **nt** [See details ▾](#)

Query ID **icjQuery_45787**

Description **fragment_25**

Molecule type **dna**

Query Length **25**

Other reports [Distance tree of results](#) [MSA viewer](#) [?](#)

Filter Results

Organism only top 20 will appear ☐ exclude

Type common name, binomial, taxid or group name

[+ Add organism](#)

Percent Identity to E value to Query Coverage to

[Filter](#) [Reset](#)

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments Download ▾ Select columns ▾ Show 100 ▾ [?](#)

☒ select all 40 sequences selected [GenBank](#) [Graphics](#) [Distance tree of results](#) [MSA Viewer](#)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	Eukaryotic synthetic construct chromosome 17	eukaryotic synt...	46.4	92.7	100%	0.038	100.00%	88299790	CP034495.1
<input checked="" type="checkbox"/>	PREDICTED: Pan paniscus WD repeat containing antisense to TP53 (WRAP53), transcript variant X1...	Pan paniscus	46.4	46.4	100%	0.038	100.00%	2600	XM_055102123.1
<input checked="" type="checkbox"/>	Homo sapiens BRD4-independent group 4 enhancer GRCh37_chr17:7589699-7590898 (LOC12686248)...	Homo sapiens	46.4	46.4	100%	0.038	100.00%	1421	NG_086981.2
<input checked="" type="checkbox"/>	PREDICTED: Macaca thibetana thibetana tumor protein p53 (TP53), transcript variant X1, mRNA	Macaca thibeta...	46.4	46.4	100%	0.038	100.00%	3463	XM_050765346.1
<input checked="" type="checkbox"/>	Homo sapiens isolate CHM13 chromosome 17	Homo sapiens	46.4	46.4	100%	0.038	100.00%	84276897	CP068261.2

b) А теперь добавим в организмы человек. Получаем, что последняя длина последовательности, когда E-value ещё нормальное (меньше 0.05) это 21, т.е. она уменьшилась.

blast.ncbi.nlm.nih.gov/Blast.cgi

Your search is limited to records that include: human (taxid:9606)

Job Title **fragment_100**

RID **63GF6PRF013** Search expires on 05-16 13:23 pm [Download All ▾](#)

Results for **80-icjQuery_90423 fragment_21(21bp)** ▾

Program **BLASTN** [Citation ▾](#)

Database **nt** [See details ▾](#)

Query ID **icjQuery_90423**

Description **fragment_21**

Molecule type **nucleic acid**

Query Length **21**

Other reports [Distance tree of results](#) [MSA viewer](#) [?](#)

Filter Results

Organism only top 20 will appear ☐ exclude

Type common name, binomial, taxid or group name

[+ Add organism](#)

Percent Identity to E value to Query Coverage to

[Filter](#) [Reset](#)

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments Download ▾ Select columns ▾ Show 100 ▾ [?](#)

☒ select all 37 sequences selected [GenBank](#) [Graphics](#) [Distance tree of results](#) [MSA Viewer](#)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	Homo sapiens BRD4-independent group 4 enhancer GRCh37_chr17:7589699-7590898 (LOC126862483) on c...	Homo sapiens	39.2	39.2	100%	0.047	100.00%	1421	NG_086981.2
<input checked="" type="checkbox"/>	Homo sapiens isolate CHM13 chromosome 17	Homo sapiens	39.2	39.2	100%	0.047	100.00%	84276897	CP068261.2
<input checked="" type="checkbox"/>	Homo sapiens tumor protein p53 (TP53), RefSeqGene (LRG_321) on chromosome 17	Homo sapiens	39.2	39.2	100%	0.047	100.00%	32772	NG_017013.2
<input checked="" type="checkbox"/>	Homo sapiens DNA, chromosome 17, nearly complete genome	Homo sapiens	39.2	39.2	100%	0.047	100.00%	80688777	AP023477.1
<input checked="" type="checkbox"/>	Homo sapiens WD repeat containing antisense to TP53 (WRAP53), RefSeqGene (LRG_375) on chromosome 17	Homo sapiens	39.2	39.2	100%	0.047	100.00%	24432	NG_028245.1
<input checked="" type="checkbox"/>	Homo sapiens isolate SE3a WDR79 (WDR79) gene, exons 1 through 3 and partial cds, and TP53 (TP53) gene...	Homo sapiens	39.2	39.2	100%	0.047	100.00%	4016	EU877060.1
<input checked="" type="checkbox"/>	Homo sapiens isolate SETa WDR79 (WDR79) gene, exons 1 through 3 and partial cds, and TP53 (TP53) gene...	Homo sapiens	39.2	39.2	100%	0.047	100.00%	4016	EU877056.1
<input checked="" type="checkbox"/>	Homo sapiens isolate SE5a WDR79 (WDR79) gene, exons 1 through 3 and partial cds, and TP53 (TP53) gene...	Homo sapiens	39.2	39.2	100%	0.047	100.00%	4016	EU877052.1
<input checked="" type="checkbox"/>	Homo sapiens isolate CH3b WDR79 (WDR79) gene, exons 1 through 3 and partial cds, and TP53 (TP53) gene...	Homo sapiens	39.2	39.2	100%	0.047	100.00%	4016	EU876951.1
<input checked="" type="checkbox"/>	Homo sapiens isolate SE3a WDR79 (WDR79) gene, exons 1 through 3 and partial cds, and TP53 (TP53) gene...	Homo sapiens	39.2	39.2	100%	0.047	100.00%	4016	EU877048.1
<input checked="" type="checkbox"/>	Homo sapiens isolate SE2b WDR79 (WDR79) gene, exons 1 through 3 and partial cds, and TP53 (TP53) gene...	Homo sapiens	39.2	39.2	100%	0.047	100.00%	4016	EU877047.1

Это так, потому что E-Value показывает вероятность случайного совпадения. Соответственно, если мы ограничим поиск только одним лишь человеком она будет меньше и нам нужен будет меньший кусок последовательности.