

## Отчет.

Для анализа был выбран tetraodon, а именно его 21 хромосома. Был скачан файл в формате .fasta. Далее был загружен файл zhunt3-alan.c  
Далее были выполнены следующие команды:

```
!gcc zhunt3-alan.c -lm -o zhunt3
```

```
!chmod a+x zhunt3
```

```
!./zhunt3 12 8 12 chr21.fa
```

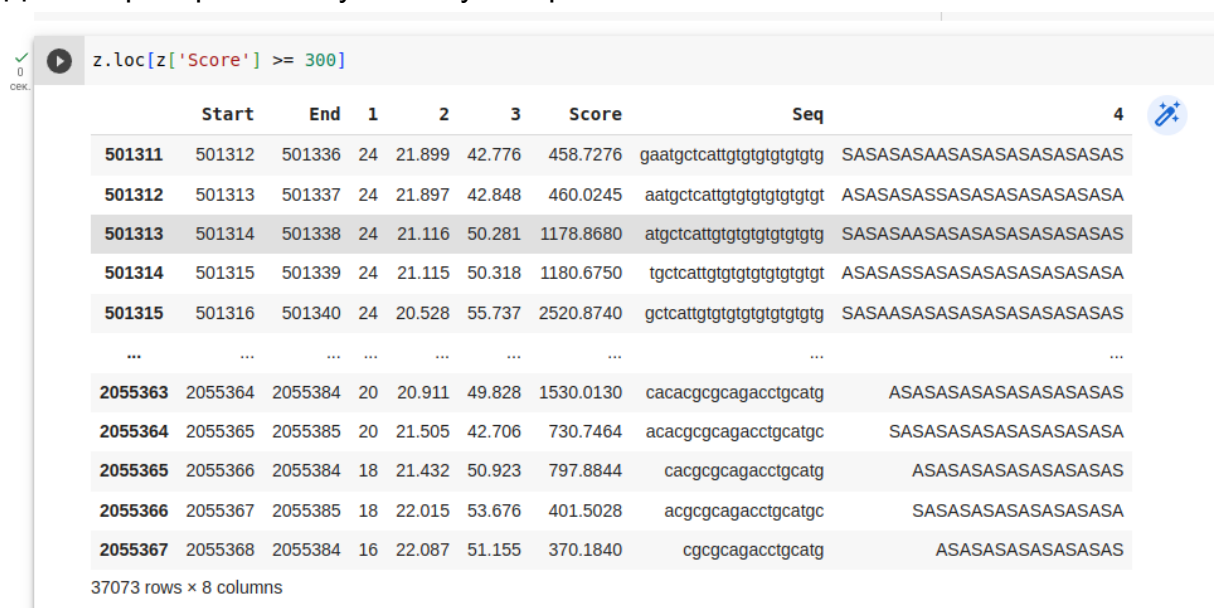
Получен файл chr21.fa.Z-SCORE

1) Для анализа будем использовать pandas:

```
import pandas as pd
```

```
z=pd.read_csv("chr21.fa.Z-SCORE", skiprows=1,  
names=["Start","End","1","2","3","Score","Seq","4"], delim_whitespace=True)
```

Далее проверяем все участки у которых z-score



	Start	End	1	2	3	Score	Seq	4
501311	501312	501336	24	21.899	42.776	458.7276	gaatgctcattgtgtgtgtgtgtg	SASASASAASASASASASASASAS
501312	501313	501337	24	21.897	42.848	460.0245	aatgctcattgtgtgtgtgtgtgt	ASASASASSASASASASASASASAS
501313	501314	501338	24	21.116	50.281	1178.8680	atgctcattgtgtgtgtgtgtgtg	SASASAAASASASASASASASASAS
501314	501315	501339	24	21.115	50.318	1180.6750	tgctcattgtgtgtgtgtgtgtgt	ASASASSASASASASASASASASAS
501315	501316	501340	24	20.528	55.737	2520.8740	gctcattgtgtgtgtgtgtgtgtg	SASAASASASASASASASASASAS
...	...	...	...	...	...	...	...	...
2055363	2055364	2055384	20	20.911	49.828	1530.0130	cacacgcgcagacctgcatg	ASASASASASASASASASASAS
2055364	2055365	2055385	20	21.505	42.706	730.7464	acacgcgcagacctgcatgc	SASASASASASASASASASASAS
2055365	2055366	2055384	18	21.432	50.923	797.8844	cacgcgcagacctgcatg	ASASASASASASASASASASAS
2055366	2055367	2055385	18	22.015	53.676	401.5028	acgcgcagacctgcatgc	SASASASASASASASASASASAS
2055367	2055368	2055384	16	22.087	51.155	370.1840	cgcgcagacctgcatg	ASASASASASASASASASASAS

37073 rows x 8 columns

Получается, что участков z-ДНК 37073

2) Теперь приступаем к следующему заданию. Устанавливаем необходимую библиотеку:

```
!pip install biopython
```

```
from Bio import SeqIO
```

Теперь запускаем следующий код, чтобы узнать количество квадруплексов:

```
inp = "chr21.fa"
for r in SeqIO.parse(inp, "fasta"):
    name, seq = r.id, str(r.seq)
    pattern="(?:G{3,}[ATGC]{1,7}){3,}G{3,}"
    ans=[m.start(),m.end(),m.group(0)] for m in re.finditer(pattern,seq)
len(ans)
```

77

Итого, их всего 77.

- 3) Теперь 3 задание. Определим, куда попадают z-ДНК и квадруплексы. Качаем файл tetNig2.ensGene.gtf.gz распаковываем, подгружаем в колаб. Также качаем bedtools. Далее запускаем следующий код:

```
with open('quad.bed', 'w') as the_file:
    for i in range(len(ans)):
        the_file.write("chr21\t"+str(ans[i][0])+"\t"+str(ans[i][1])+"\n")
!./bedtools2/bin/bedtools intersect -a quad.bed -b tetNig2.ensGene.gtf -wao | tail
```

chr21	5648036	5648070	chr21	ensGene.v101	exon	5647660	5648185	.	.	.	gene_id "ENSTNIG00000014037.1"; transcript_id "ENSTNIT00000017260.1"; exo
chr21	5648036	5648070	chr21	ensGene.v101	CDS	5647660	5648185	.	2	.	gene_id "ENSTNIG00000014037.1"; transcript_id "ENSTNIT00000017260.1"; exo
chr21	5652716	5652737	.	.	-1	-1	.	.	.	0	.
chr21	5658950	5658982	.	.	-1	-1	.	.	.	0	.
chr21	5665955	5665988	.	.	-1	-1	.	.	.	0	.
chr21	5665373	5665407	.	.	-1	-1	.	.	.	0	.
chr21	5666107	5666140	.	.	-1	-1	.	.	.	0	.
chr21	5674814	5674838	chr21	ensGene.v101	transcript	5673691	5676578	.	.	.	gene_id "ENSTNIG00000014039.1"; transcript_id "ENSTNIT00000017262
chr21	5675430	5675461	chr21	ensGene.v101	transcript	5673691	5676578	.	.	.	gene_id "ENSTNIG00000014039.1"; transcript_id "ENSTNIT00000017262
chr21	5780057	5780093	.	.	-1	-1	.	.	.	0	.

Т.е. все квадруплексы лежат и в межгенном пространстве, и в генах, и на пересечении. Т.е. все z-ДНК лежат и в межгенном пространстве, и в генах, и на пересечении.

Разберемся теперь с z-ДНК. Запускаем похожий код:

```
[79] with open('z.bed', 'w') as z_file:
    for i, r in z.iterrows():
        z_file.write("chr21\t" + str(r['Start']) + "\t" + str(r['End']) + "\n")
```

Посмотрим последние 10 строк:

```
!./bedtools2/bin/bedtools intersect -a z.bed -b tetNig2.ensGene.gtf -wao | tail
```

chr21	2056017	2056033	.	.	.	-1	-1	.	.	.	0
chr21	2056018	2056034	.	.	.	-1	-1	.	.	.	0
chr21	2056019	2056035	.	.	.	-1	-1	.	.	.	0
chr21	2056020	2056036	.	.	.	-1	-1	.	.	.	0
chr21	2056021	2056037	.	.	.	-1	-1	.	.	.	0
chr21	2056022	2056038	.	.	.	-1	-1	.	.	.	0
chr21	2056023	2056039	.	.	.	-1	-1	.	.	.	0
chr21	2056024	2056040	.	.	.	-1	-1	.	.	.	0
chr21	2056025	2056041	.	.	.	-1	-1	.	.	.	0
chr21	2056026	2056042	.	.	.	-1	-1	.	.	.	0

Поищем по ключевому слову gene:

```
./bedtools/bin/bedtools intersect -a z.bed -b tetNig2.ensGene.gtf -wao | grep "gene" | tail
```

chr	start	end	chr	start	end	feature	start	end	gene_id	transcript_id
chr21	2047169	2047185	chr21	2047170	2047172	CDS	2047170	2047172	ENSTNIG00000001441.1	ENSTNIT00000003639.1
chr21	2047169	2047185	chr21	2047170	2047172	start_codon	2047170	2047172	ENSTNIG00000001441.1	ENSTNIT00000003639.1
chr21	2047170	2047186	chr21	2047170	2047172	transcript	2047170	2047172	ENSTNIG00000001441.1	ENSTNIT00000003639.1
chr21	2047170	2047186	chr21	2047170	2047172	exon	2047170	2047172	ENSTNIG00000001441.1	ENSTNIT00000003639.1
chr21	2047170	2047186	chr21	2047170	2047172	CDS	2047170	2047172	ENSTNIG00000001441.1	ENSTNIT00000003639.1
chr21	2047170	2047186	chr21	2047170	2047172	start_codon	2047170	2047172	ENSTNIG00000001441.1	ENSTNIT00000003639.1
chr21	2047171	2047187	chr21	2047171	2047172	transcript	2047171	2047172	ENSTNIG00000001441.1	ENSTNIT00000003639.1
chr21	2047171	2047187	chr21	2047171	2047172	exon	2047171	2047172	ENSTNIG00000001441.1	ENSTNIT00000003639.1
chr21	2047171	2047187	chr21	2047171	2047172	CDS	2047171	2047172	ENSTNIG00000001441.1	ENSTNIT00000003639.1
chr21	2047171	2047187	chr21	2047171	2047172	start_codon	2047171	2047172	ENSTNIG00000001441.1	ENSTNIT00000003639.1

Т.е. все z-ДНК лежат и в межгенном пространстве, и в генах, и на пересечении.

4) Давайте напишем 4. Находим промоторы, записываем их в promoters.txt

```
!head promoters.txt
```

Запустить код в ячейке (Ctrl+Enter)  
Код в ячейке выполнялся с момента последнего изменения

Выполнил пользователь Максим Куваев  
01:48 (0 минут назад)  
Время выполнения: 0.315 сек.

```
chr21 536676 537676 "ENSTNIG00000005728.1";  
chr21 537499 538499 "ENSTNIG00000005727.1";  
chr21 541931 542931 "ENSTNIG00000005726.1";  
chr21 547662 548662 "ENSTNIG00000005725.1";  
chr21 555809 556809 "ENSTNIG00000005725.1";
```

Далее делаем gtf:

```
with open("promoters.txt", "r") as file:  
    content = file.read()  
  
content = content.replace(" ", "\t")  
  
with open("promoters.gtf", "w") as file:  
    file.write(content)  
!head promoters.gtf
```

```
chr21 505321 506321 "ENSTNIG00000003754.1";  
chr21 513501 514501 "ENSTNIG00000002616.1";  
chr21 518066 519066 "ENSTNIG00000002617.1";  
chr21 522990 523990 "ENSTNIG00000002618.1";  
chr21 526176 527176 "ENSTNIG00000002619.1";  
chr21 536676 537676 "ENSTNIG00000005728.1";  
chr21 537499 538499 "ENSTNIG00000005727.1";  
chr21 541931 542931 "ENSTNIG00000005726.1";  
chr21 547662 548662 "ENSTNIG00000005725.1";  
chr21 555809 556809 "ENSTNIG00000005725.1";
```

И находим пересечение:

```
✓ 4 OK. !./bedtools2/bin/bedtools intersect -a z.bed -b promoters.gtf -wao | awk '{print $7}' > z.txt
genes = set()
with open('z.txt', 'r') as file:
    for l in file:
        genes.add(l.strip().replace('"', '').replace(';', ''))
genes.remove('.')
genes

'ENSTNIG00000009162.1',
'ENSTNIG00000009163.1',
'ENSTNIG00000009164.1',
'ENSTNIG00000009165.1',
'ENSTNIG00000009166.1',
'ENSTNIG00000009167.1',
'ENSTNIG00000009169.1',
'ENSTNIG00000009170.1',
'ENSTNIG00000009171.1',
'ENSTNIG00000009172.1',
'ENSTNIG00000009173.1',
'ENSTNIG00000009174.1',
'ENSTNIG00000009175.1',
'ENSTNIG00000009176.1',
'ENSTNIG00000009177.1'
```

5) По сути, работаем аналогично 4 заданию. Промотеры уже есть.  
Остается:

```
✓ 0 OK. !./bedtools2/bin/bedtools intersect -a quad.bed -b promoters.gtf -wao | awk '{print $7}' > quad.txt
genes = set()
with open('quad.txt', 'r') as file:
    for l in file:
        genes.add(l.strip().replace('"', '').replace(';', ''))
genes.remove('.')
genes

{'ENSTNIG00000002904.1',
'ENSTNIG00000005634.1',
'ENSTNIG00000009170.1',
'ENSTNIG00000019799.1'}
```