

Задание 1

Код:

```
x=input()
t=[]
d={}
with open(x) as f:
    g=f.read().split('>')
    t=[g.split('\n',1)[1:] for g in g]
    for i in t:
        for j in i:
            o=j[0:3]
            d[o]=d.get(o,0)+1
l=sorted(d.items())
l=dict(l)
with open("out.txt","w") as o:
    for m,n in l.items():
        print(m,n, file=o)
```

Результаты:

1. Escherichia coli str. K-12 substr. MG1655:

ATG 3890

ATT 4

CTG 2

GTG 338

TTC 1

TTG 80

2. Candidatus Gracilibacteria bacterium 28_42_T64:

ACA 1

ATG 1129

GTG 41

TCA 1

TCT 1

TTG 23

3. *Mycoplasma pneumoniae* M29:

AAA 1

ACA 1

ACT 1

ATA 3

ATC 1

ATG 627

ATT 7

CAA 1

CAC 1

CTA 1

CTC 3

CTG 2

GAA 1

GTG 60

GTT 1

TCC 2

TCT 1

TGA 1

TTA 1

TTC 1

TTG 49

Наиболее частый кодон – ATG. Можно предположить, что другие варианты, которые отличаются от канонического кодона на одну букву – случайные мутации, которые могут не влиять на способность распознавания полимеразой. Какие-то из кодонов являются старт кодонами псевдогенов (например TTC, TCT итд). Какие-то старт кодоны входят в состав белка и не являются, по сути, началом последовательности (возможно есть другие последовательности нуклеотидов, которые влияют на экспрессию этих белков) (например, CTG, TTA).

Задание 2

Код:

```
x=input()
```

```
d = { 'TGA' : 0,
```

```

'TAA' : 0,
'TAG' : 0}

t=[]
u=""

with open(x) as f:

    g=f.readline()

    for i in f:

        if not i.startswith('>'):

            u += ''.join((i.split()))

        if i.startswith('>'):

            for j in range(0, len(u) - 3, 3):

                if u[j:j + 3] in d:

                    print(g)

                    break

            g=i

        u=""

```

Результаты:

1. Escherichia coli str. K-12 substr. MG1655:

```

>lcl|U00096.3_cds_b4587_250 [gene=insN] [locus_tag=b4587]
[db_xref=ASAP:ABE-0285253,ECOCYC:G6130] [protein=IS911A regulator fragment] [pseudo=true]
[location=join(270278..270540,271764..272190)] [gbkey=CDS]

>lcl|U00096.3_cds_AAD13438.1_1459 [gene=fdnG] [locus_tag=b1474]
[db_xref=UniProtKB/Swiss-Prot:P24183] [protein=formate dehydrogenase N subunit alpha]
[transl_except=(pos:586..588,aa:Sec)] [protein_id=AAD13438.1] [location=1547401..1550448] [gbkey=CDS]

>lcl|U00096.3_cds_AAD13456.1_3824 [gene=fdoG] [locus_tag=b3894]
[db_xref=UniProtKB/Swiss-Prot:P32176] [protein=formate dehydrogenase O subunit alpha]
[transl_except=(pos:586..588,aa:Sec)] [protein_id=AAD13456.1] [location=complement(4082772..4085822)]
[gbkey=CDS]

>lcl|U00096.3_cds_AAD13462.1_3997 [gene=fdhF] [locus_tag=b4079]
[db_xref=UniProtKB/Swiss-Prot:P07658] [protein=formate dehydrogenase H]
[transl_except=(pos:418..420,aa:Sec)] [protein_id=AAD13462.1] [location=complement(4297219..4299366)]
[gbkey=CDS]

```

1 последовательность: псевдоген => не важно, где стоит стоп кодон, потому что ген не экспрессируется

2,3,4: кодируют формиат дегидрогеназу. В ней кодон TGA кодирует аминокислоту селеноцистеин (если организм живет в среде без селена – TGA -обычный стоп кодон, но в другом случае – работает последовательность после него, которая заставляет считывать в том числе этот «стоп кодон»).

Задание 3

Код:

```
x=input()
d={"TGA": 0,
  "TAA": 0,
  "TAG": 0}
u=""
with open(x) as f:
    g=f.readline()
    for i in f:
        if not i.startswith('>'):
            u+=''.join((i.split()))
        if i.startswith('>'):
            if u[len(u)-3:len(u)] in d:
                d[u[len(u)-3:len(u)]]+=1
            u=""
            g=i
with open("out.txt","w") as o:
    for m,n in d.items():
        print(m,n, file=o)
```

Результаты:

1. Escherichia coli str. K-12 substr. MG1655:
TGA 1246
TAA 2760
TAG 306

2. Candidatus Gracilibacteria bacterium 28_42_T64:

TGA 1

TAA 999

TAG 188

3. Mycoplasma pneumoniae M29:

TGA 0

TAA 533

TAG 221

Можно увидеть, что во второй и третьей бактерии практически нет стоп кодонов TAG.

По данным NCBI, вторая бактерия имеет альтернативное понимание стоп-кодона TAG, а именно он кодирует аминокислоту глицин.

(<https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=cgencodes#SG25>)

По данным некоторых исследований, третья бактерия имеет вместо стоп кодона – кодон, кодирующий триптофан. (<https://journals.asm.org/doi/10.1128/jb.172.1.504-506.1990>)

Задание 4

Код:

```
x=input()
```

```
e= {
```

```
    "ТТА": 0,
```

```
    "ТТГ": 0,
```

```
    "СТА": 0,
```

```
    "СТС": 0,
```

```
    "СТГ": 0,
```

```
    "СТТ": 0, }
```

```
t=[]
```

```
with open(x) as f:
```

```
    g=f.read().split('\n')
```

```
    for i in g:
```

```
        if not i.startswith('>'):
```

```
t.append(i)
u=''.join(t)
for i in range(0,len(u),3):
    if u[i:i+3] in e:
        e[u[i:i+3]]+=1
l=sorted(e.items())
l=dict(l)
with open("out.txt","w") as o:
    for m,n in l.items():
        print(m,n, file=o)
```

Результаты:

1. Escherichia coli str. K-12 substr. MG1655:

CTA 5203

CTC 14952

CTG 71305

CTT 14728

TTA 18505

TTG 18301

2. Candidatus Gracilibacteria bacterium 28_42_T64:

CTA 4861

CTC 4491

CTG 4147

CTT 8053

TTA 15077

TTG 8048

3. Mycoplasma pneumoniae M29:

CTA 2715

CTC 2362

CTG 2298

СТТ 5771

ТТА 7950

ТТГ 4487

Как и внутри одной бактерии так и между ними – частота разных кодонов, отвечающих за лейцин, разное. Для второй и третьей бактерии более характерны кодоны, не содержащие С или G (например ТТА), что может быть связано с не слишком большим процентным содержанием GC в их последовательностях, в то время как наибольшее число лейцина у первой бактерии кодируется СТG кодоном. Также такое разнообразие в вариантах кодирования возможно из-за специфических транспортных РНК (тРНК), которые являются одним из интересных аспектов в изучении различных бактерий.