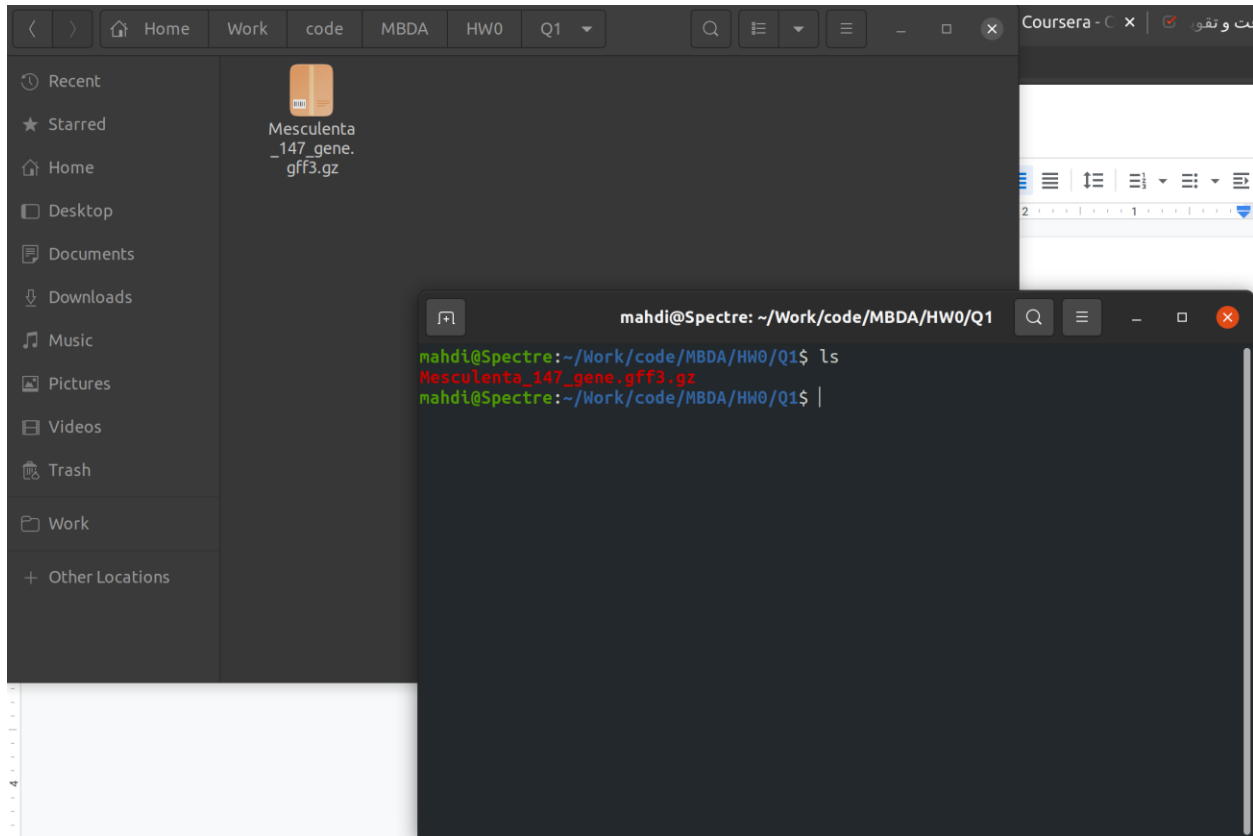


به نام خدا

مهدی کافی - تمرین شماره ۰

## سوال اول:

در ابتدا فایل Mesculenta\_147\_gene.gff3.gz را دانلود کرده و در فولدر سوال اول میریزیم.



۱. سپس با دستور gunzip، به صورت زیر فایل را از حالت فشرده خارج کرده و نامش را به Cassavagenes.txt تغییر می‌دهیم.

```
mahdi@Spectre: ~/Work/code/MBDA/HW0/Q1
mahdi@Spectre:~/Work/code/MBDA/HW0/Q1$ ls
Mesculenta_147_gene.gff3.gz
mahdi@Spectre:~/Work/code/MBDA/HW0/Q1$ gunzip -c Mesculenta_147_gene.gff3.gz > C
assavagenes.txt
mahdi@Spectre:~/Work/code/MBDA/HW0/Q1$ ls
Cassavagenes.txt  Mesculenta_147_gene.gff3.gz
mahdi@Spectre:~/Work/code/MBDA/HW0/Q1$ |
```

سپس از فایل head میگیریم تا با ساختار آن آشنا شویم.

```
mahdi@Spectre: ~/Work/code/MBDA/HW0/Q1
Cassavagenes.txt Mesculenta_147_gene.gff3.gz
mahdi@Spectre:~/Work/code/MBDA/HW0/Q1$ head Cassavagenes.txt
##gff-version 3
scaffold00005 phytozome9_0 gene 34993 37484 . - . I
D=cassava4.1_003195m.g;Name=cassava4.1_003195m.g
scaffold00005 phytozome9_0 mRNA 34993 37484 . - . I
D=PAC:17967872;Name=cassava4.1_003195m.g;pacid=17967872;longest=1;Parent=cassava4.1_003195m.g
scaffold00005 phytozome9_0 CDS 35367 37166 . - 0 I
D=PAC:17967872.CDS.1;Parent=PAC:17967872;pacid=17967872
scaffold00005 phytozome9_0 five_prime_UTR 37167 37484 . - .
ID=PAC:17967872.five_prime_UTR.1;Parent=PAC:17967872;pacid=17967872
scaffold00005 phytozome9_0 CDS 35230 35275 . - 0 I
D=PAC:17967872.CDS.2;Parent=PAC:17967872;pacid=17967872
scaffold00005 phytozome9_0 CDS 34993 35126 . - 2 I
D=PAC:17967872.CDS.3;Parent=PAC:17967872;pacid=17967872
scaffold00005 phytozome9_0 gene 2612 5455 . + . I
D=cassava4.1_022599m.g;Name=cassava4.1_022599m.g
scaffold00005 phytozome9_0 mRNA 2612 5455 . + . I
D=PAC:17967873;Name=cassava4.1_022599m.g;pacid=17967873;longest=1;Parent=cassava4.1_022599m.g
scaffold00005 phytozome9_0 CDS 2612 2928 . + 0 I
D=PAC:17967873.CDS.1;Parent=PAC:17967873;pacid=17967873
mahdi@Spectre:~/Work/code/MBDA/HW0/Q1$
```

۲. به نظر می‌رسد که در هر سطر نوع داده آن سطر را مشخص کرده‌است. حال کفایت خطوطی که نوع آن‌ها gene است را با دستور grep پیدا کنیم و سپس با دستور wc تعداد این خطوط، تعداد کلمات و سائز را به دست آوریم و در نهایت با دستور cut تعداد خطوط را از خروجی دستور wc مشخص کنیم.

```

mahdi@Spectre: ~/Work/code/MBDA/HW0/Q1
mahdi@Spectre:~/Work/code/MBDA/HW0/Q1$ ls
Mesculenta_147_gene.gff3.gz
mahdi@Spectre:~/Work/code/MBDA/HW0/Q1$ gunzip -c Mesculenta_147_gene.gff3.gz > Cassavagenes.txt
mahdi@Spectre:~/Work/code/MBDA/HW0/Q1$ ls
Cassavagenes.txt Mesculenta_147_gene.gff3.gz
mahdi@Spectre:~/Work/code/MBDA/HW0/Q1$ head Cassavagenes.txt
##gff-version 3
scaffold00005 phytozome9_0 gene 34993 37484 . - . ID=cassava4.1_003195m.g;Name=cassava4.1_003195m.g
scaffold00005 phytozome9_0 mRNA 34993 37484 . - . ID=PAC:17967872;Name=cassava4.1_003195m.g;pacid=17967872;longest=1;Parent=cassava4.1_003195m.g
scaffold00005 phytozome9_0 CDS 35367 37166 . - 0 ID=PAC:17967872.CDS.1;Parent=PAC:17967872;pacid=17967872
scaffold00005 phytozome9_0 five_prime_UTR 37167 37484 . - . ID=PAC:17967872.five_prime_UTR.1;Parent=PAC:17967872;pacid=17967872
scaffold00005 phytozome9_0 CDS 35230 35275 . - 0 ID=PAC:17967872.CDS.2;Parent=PAC:17967872;pacid=17967872
scaffold00005 phytozome9_0 CDS 34993 35126 . - 2 ID=PAC:17967872.CDS.3;Parent=PAC:17967872;pacid=17967872
scaffold00005 phytozome9_0 gene 2612 5455 . + . ID=cassava4.1_022599m.g;Name=cassava4.1_022599m.g
scaffold00005 phytozome9_0 mRNA 2612 5455 . + . ID=PAC:17967873;Name=cassava4.1_022599m.g;pacid=17967873;longest=1;Parent=cassava4.1_022599m.g
scaffold00005 phytozome9_0 CDS 2612 2920 . + 0 ID=PAC:17967873.CDS.1;Parent=PAC:17967873;pacid=17967873
mahdi@Spectre:~/Work/code/MBDA/HW0/Q1$ cat Cassavagenes.txt | grep gene | wc | cut -d" " -f 3
30666
mahdi@Spectre:~/Work/code/MBDA/HW0/Q1$ |

```

۳. برای ریختن سطرهایی که دارای کلمه **gene** هستند در یک فایل کافیست که خروجی **grep gene** محتوای این فایل را با **>** درون فایل **Genes.txt** بریزیم.

```

mahdi@Spectre: ~/Work/code/MBDA/HW0/Q1
mahdi@Spectre:~/Work/code/MBDA/HW0/Q1$ cat Cassavagenes.txt | grep gene > Genes.txt
mahdi@Spectre:~/Work/code/MBDA/HW0/Q1$ ls
Cassavagenes.txt Genes.txt Mesculenta_147_gene.gff3.gz
mahdi@Spectre:~/Work/code/MBDA/HW0/Q1$ head Genes.txt
scaffold00005 phytozome9_0 gene 34993 37484 . - . ID=cassava4.1_003195m.g;Name=cassava4.1_003195m.g
scaffold00005 phytozome9_0 gene 2612 5455 . + . ID=cassava4.1_022599m.g;Name=cassava4.1_022599m.g
scaffold00005 phytozome9_0 gene 25077 26923 . + . ID=cassava4.1_028923m.g;Name=cassava4.1_028923m.g
scaffold00007 phytozome9_0 gene 3279 4340 . + . ID=cassava4.1_023044m.g;Name=cassava4.1_023044m.g
scaffold00008 phytozome9_0 gene 49334 55454 . + . ID=cassava4.1_000805m.g;Name=cassava4.1_000805m.g
scaffold00008 phytozome9_0 gene 55741 58420 . - . ID=cassava4.1_030173m.g;Name=cassava4.1_030173m.g
scaffold00009 phytozome9_0 gene 37732 37935 . - . ID=cassava4.1_028373m.g;Name=cassava4.1_028373m.g
scaffold00009 phytozome9_0 gene 28757 33544 . + . ID=cassava4.1_028576m.g;Name=cassava4.1_028576m.g
scaffold00010 phytozome9_0 gene 233998 234939 . + . ID=cassava4.1_015436m.g;Name=cassava4.1_015436m.g
scaffold00010 phytozome9_0 gene 144624 150646 . - . ID=cassava4.1_016957m.g;Name=cassava4.1_016957m.g
mahdi@Spectre:~/Work/code/MBDA/HW0/Q1$ |

```

۴. برای مشخص کردن تعداد ژن‌های فایل Genes.txt کافیهست که محتوای این فایل را به دستور WC بدهیم و با دستور cut تعداد خطوط خروجی دستور WC را مشخص کنیم.

```
mahdi@Spectre: ~/Work/code/MBDA/HW0/Q1
mahdi@Spectre:~/Work/code/MBDA/HW0/Q1$ cat Genes.txt | wc | cut -d " " -f 3
30666
mahdi@Spectre:~/Work/code/MBDA/HW0/Q1$ |
```

## سوال دوم:

در ابتدا با دستور wget فایل را دانلود می‌کنیم.

```
mahdi@Spectre: ~/Work/code/MBDA/HW0/Q2
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ ls
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ wget http://ce.sharif.edu/~smohseni/hw0_2.tar.gz
--2021-02-20 17:01:51-- http://ce.sharif.edu/~smohseni/hw0_2.tar.gz
Resolving ce.sharif.edu (ce.sharif.edu)... 81.31.168.124
Connecting to ce.sharif.edu (ce.sharif.edu)|81.31.168.124|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 519 [application/x-gzip]
Saving to: 'hw0_2.tar.gz'

hw0_2.tar.gz          100%[=====]          519  --.-KB/s    in 0s

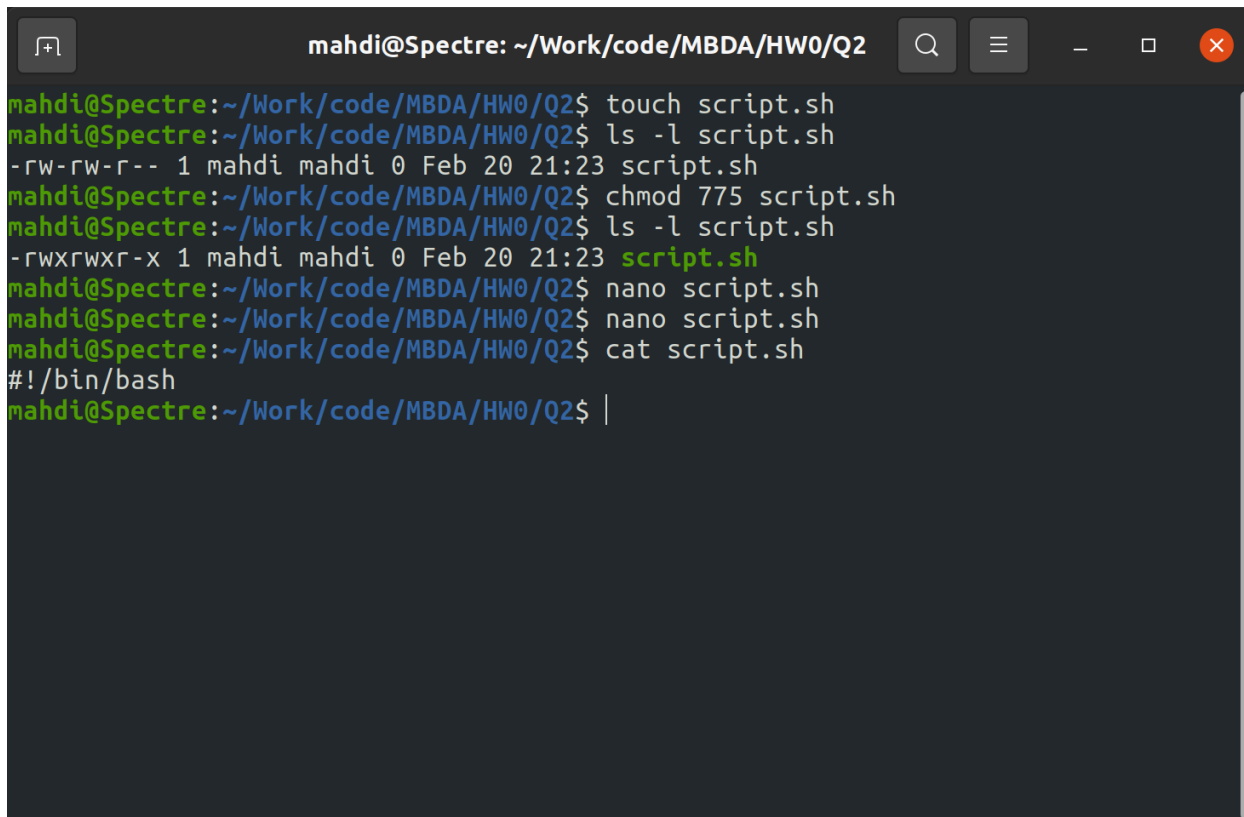
2021-02-20 17:01:51 (16.5 MB/s) - 'hw0_2.tar.gz' saved [519/519]

mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ ls
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ |
```

با دستور **tar** فایل را از حالت فشرده خارج می‌کنیم.

```
mahdi@Spectre: ~/Work/code/MBDA/HW0/Q2
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ tar -xzf hw0_2.tar.gz
hw0_2/
hw0_2/gene9/
hw0_2/gene9/gene9.txt
hw0_2/gene4/
hw0_2/gene4/gene4.txt
hw0_2/gene10/
hw0_2/gene10/gene10.txt
hw0_2/gene6/
hw0_2/gene6/gene6.txt
hw0_2/gene7/
hw0_2/gene7/gene7.txt
hw0_2/gene2/
hw0_2/gene2/gene2.txt
hw0_2/gene3/
hw0_2/gene3/gene3.txt
hw0_2/gene5/
hw0_2/gene5/gene5.txt
hw0_2/gene1/
hw0_2/gene1/gene1.txt
hw0_2/gene8/
hw0_2/gene8/gene8.txt
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ ls
hw0_2  hw0_2.tar.gz
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ |
```

حال برای ایجاد `bash script` یک فایل به نام `script.sh` می‌سازیم و با دستور `chmod` دسترسی اجرا به آن می‌دهیم. سپس فایل را باز کرده و در خط اول عبارت `#!/bin/bash` را می‌نویسیم.



```
mahdi@Spectre: ~/Work/code/MBDA/HW0/Q2
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ touch script.sh
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ ls -l script.sh
-rw-rw-r-- 1 mahdi mahdi 0 Feb 20 21:23 script.sh
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ chmod 775 script.sh
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ ls -l script.sh
-rwxrwxr-x 1 mahdi mahdi 0 Feb 20 21:23 script.sh
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ nano script.sh
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ nano script.sh
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ cat script.sh
#!/bin/bash
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ |
```

حال می‌خواهیم که محتوای تمامی فایل‌های ژن‌ها را در یک فایل که نام آن را از کاربر می‌گیریم بریزیم برای این کار از wildcardها استفاده می‌کنیم.

```
maahdi@Spectre: ~/Work/code/MBDA/HW0/Q2
maahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ cat script.sh
#!/bin/bash

echo "Enter the file name: "
read file_name
cat /home/maahdi/Work/code/MBDA/HW0/Q2/hw0_2/gene*/gene*.txt > $file_name
maahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ ls
hw0_2  hw0_2.tar.gz  script.sh
maahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ sh script.sh
Enter the file name:
gene.out
maahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ cat gene.out
a1 + 10
pixrs1fg - 7
dfg - 12
tvsa - 5
ldt4 - 6
xbk + 9
lpm + 11
xgo + 5
xti - 10
spvrtyv + 15
maahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ |
```

در ادامه با دستور sort خطوط فایل gene.out را بر حسب طول ژن مرتب می‌کنیم. در دستور sort از optionهای k, V استفاده کردیم که به ترتیب برای انتخاب مرتب سازی بر حسب ستون داده شده به دستور و مرتب سازی بر اساس کل عدد داخل ستون و نه فقط رقم اول آن هستند. خروجی دستور sort را در یک فایل موقت می‌ریزیم و سپس دوباره این محتوای مرتب شده را به فایل genes.out برمی‌گردانیم.



```
mahdi@Spectre: ~/Work/code/MBDA/HW0/Q2
GNU nano 4.8 script.b.sh
#!/bin/bash

echo "Enter the file name: "
read file_name
cat /home/mahdi/Work/code/MBDA/HW0/Q2/hw0_2/gene*/gene*.txt > $file_name
sort -Vk 3 $file_name > tmp_file
cat tmp_file > $file_name
rm -f tmp_file

^G Get Help      ^O Write Out     ^W Where Is      [ Read 8 lines ]
^X Exit          ^R Read File     ^\ Replace       ^K Cut Text
                  ^J Justify       ^U Paste Text    ^J To Spell
                  ^C Cur Pos      ^- Go To Line    M-U Undo
                  M-E Redo
```

```
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ ls
hw0_2  hw0_2.tar.gz  script.b.sh  script.sh
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ sh script.b.sh
Enter the file name:
gene.out
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ ls
gene.out  hw0_2  hw0_2.tar.gz  script.b.sh  script.sh
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ cat gene.out
tvsa - 5
xgo + 5
ldt4 - 6
pixrs1fg - 7
xbk + 9
a1 + 10
xti - 10
lpm + 11
dfg - 12
spvrtyv + 15
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ |
```

در مرحله بعدی کافیت که محتوای فایل `gene.out` را خط به خط بخوانیم، در هر خط ۳ عبارت را در ۳ متغیر مجزای `name`، `pos`، `len` بریزیم و حالا کافیت که با داشتن طول رشته تصادفی، یک رشته تصادفی از حروف `ACGT` و به طول محتوای متغیر `len` با ترکیب دو دستور `tr` و `head` بسازیم در دستور `head -C` استفاده می‌کنیم و به آن متغیر `len` را می‌دهیم تا رشته‌ای تصادفی به طول `len` کاراکتر داشته‌باشیم، سپس خطوطی می‌سازیم که ترکیب نام، موقعیت و رشته تصادفی مربوط به هر خط هستند و سپس این خط را به فایلی موقت اضافه می‌کنیم. در نهایت محتوای مورد نظر ما در فایل موقت است و این محتوا را در فایل `gene.out` که نامش از کاربر گرفته شده‌است، میریزیم.

```
mahdi@Spectre: ~/Work/code/MBDA/HW0/Q2
GNU nano 4.8 script.sh Modified
#!/bin/bash

echo "Enter the file name: "
read file_name
cat /home/mahdi/Work/code/MBDA/HW0/Q2/hw0_2/gene*/gene*.txt > $file_name
sort -Vk 3 $file_name > tmp_file
cat tmp_file > $file_name
rm -f tmp_file

while IFS=" " read -r name pos len
do
    rand_str=$(tr -dc "ACGT" </dev/urandom | head -c $len)
    printf '%s %s %s\n' "$name" "$pos" "$rand_str" >> tmp_file
done < "$file_name"
cat tmp_file > $file_name
rm -f tmp_file

[ Read 15 lines ]
^G Get Help ^O Write Out ^W Where Is [ Read 15 lines ]
^X Exit ^R Read File ^_ Replace ^K Cut Text ^J Justify ^C Cur Pos M-U Undo
^U Paste Text ^T To Spell ^_ Go To Line M-E Redo

mahdi@Spectre: ~/Work/code/MBDA/HW0/Q2
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ ls
hw0_2 hw0_2.tar.gz script.sh
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ sh script.sh
Enter the file name:
gene.out
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ ls
gene.out hw0_2 hw0_2.tar.gz script.sh
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ cat gene.out
tvsa - TTCCT
xgo + GCCAT
ldt4 - AGGGGG
pixrs1fg - ATTATCG
xbk + AACTCTCCG
a1 + GCTCGGACCG
xti - GCAGTTTCAT
lpm + CCATATCCGAG
dfg - TCCTTCGATTTC
spvrtyv + ACCTCGTAGAGTCCC
mahdi@Spectre:~/Work/code/MBDA/HW0/Q2$ |
```

در نهایت محتوای فایل gene.out را مشاهده می کنیم.