We ask you to use several approaches to answer the **following questions**:

(**1**) Find the genome of the **cauliflower mosaic virus** and **compare this sequence** with your **PCR fragment**.

The **search query** I have used was:

> https://www.ncbi.nlm.nih.gov/genome/?term=cauliflower+mosaic+virus

The **RefSeq entry** (**Accession Number**: **NC_001497**) can be found at:

> https://www.ncbi.nlm.nih.gov/nuccore/NC_001497.1

The genome is a **circular DNA**, with a total length of **8024 bp**.

The **segment** given in the course was (length: **1021 nucleotides**):

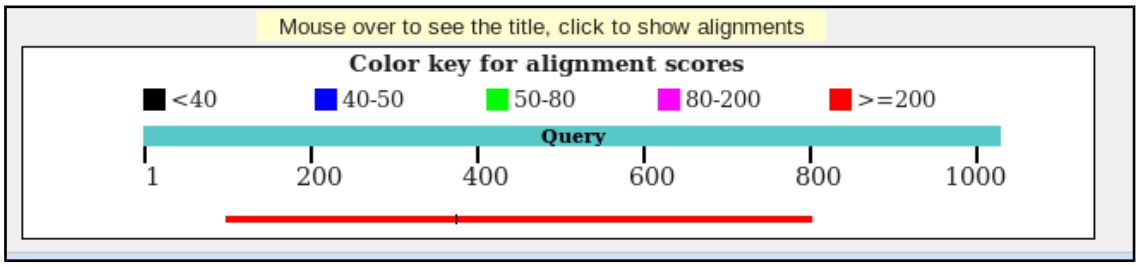
>PCR fragment containing 5x ACGT

```
ATCCCCTCTAAAGAATGGCAGTTTTCCTTTGCATGTAACTATTATGCT
CCCTTCGTTACAAAAATTTTGGACTACTATTGGGCGCGGGGCGCGCCC
GGTCCAAAGACCAGAGGGCTATTGAGACTTTTCAACAAAGGGTAATAT
CGGGAAACCTCCTCGGATTCCATTGCCCAGCTATCTGTCACTTCATCG
AAAGGACAGTAGAAAAGGAAGATGGCTTCTACAAATGCCATCATTGCG
ATAAAGGAAAGGCTATCGTTCAAGATGCCTCTACCGACAGTGGTCCCA
AAGATGGACCCCCACCCACGAGGAACATCGTGGAAAAAGAAGACGTTC
CAACCACGTCTTCAAAGCAAGTGGATTGATGTGATACATGGTGGAGCA
CGACACTCTCGTCTACTCCAAGAATATCAAAGATACAGTCTCAGAAGA
CCAGAGGGCTATTGAGACTTTTCAACAAAGGGTAATATCGGGAAACCT
CCTCGGATTCCATTGCCCAGCTATCTGTCACTTCATCGAAAGGACAGT
AGAAAAGGAAGATGGCTTCTACAAATGCCATCATTGCGATAAAGGAAA
GGCTATCGTTCAAGATGCCTCTACCGACAGTGGTCCCAAAGATGGACC
CCCACCCACGAGGAACATCGTGGAAAAAGAAGACGTTCCAACCACGTC
TTCAAAGCAAGTGGATTGATGTGATATCTCCACTGACGTAAGGGATGA
CGCACAATCCCACTATCCTTCGCAAGACCCTTCCTCTATATAAGGAAG
TTCATTTCATTTGGAGAGGACACGCTGAATTCTCAACACAACATATAC
AAAACAAACGAATCTCAAGCAATCAAGCATTCTACTTCTATTGCAGCA
ATTTAAATCATTTCTTTTAAAGCAAAAGCAATTTTCTGAAAATTTTCA
CCATTTACGAACGATAGCCATGGAGTTGGGACTGAGCTGGATTTTCCT
TTTGGCTATTTTAAAAGGTGTCCAGTGTGAGGTGCAGCTGGTGGAGTC
TGGGGGAGGCCTG
```

Note: The above sequence at **tgagactttt**, marks the beginning of the **35S promoter**.

A **BLAST search** was performed next. The following partial

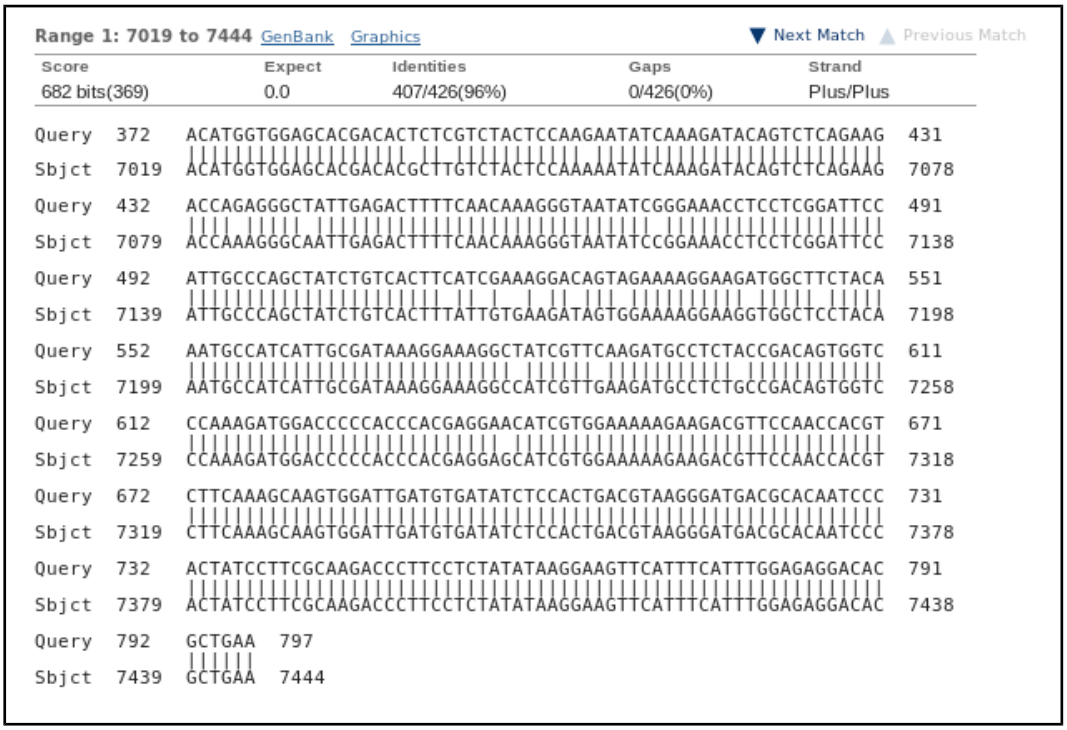
screenshot shows the result:



Evidently, our target sequence can be found to be **contiguous** with our

genome sequence.

The first match is towards nucleotide position **7019** (and including that nucleotide position) in the genome. The following **partial screenshot** shows this:



Range 1: 7019 to 7444 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
682 bits(369)	0.0	407/426(96%)	0/426(0%)	Plus/Plus

```
Query  372   ACATGGTGGAGCACGACACTCTCGTCTACTCCAAGAATATCAAAGATACAGTCTCAGAAG  431
             ||||||||||||||||||| || ||||||||||| |||||||||||||||||||||||
Sbjct  7019  ACATGGTGGAGCACGACACGCTTGTCTACTCCAAAAATATCAAAGATACAGTCTCAGAAG  7078

Query  432   ACCAGAGGGCTATTGAGACTTTTCAACAAAGGGTAATATCGGGAAACCTCCTCGGATTCC  491
             ||||| |||||| |||||||||||||||||||||||||||| |||||||||||||||||||
Sbjct  7079  ACCAAAGGGCAATTGAGACTTTTCAACAAAGGGTAATATCCGGAAACCTCCTCGGATTCC  7138

Query  492   ATTGCCCAGCTATCTGTCACTTCATCGAAAGGACAGTAGAAAAGGAAGATGGCTTCTACA  551
             |||||||||||||||||||||| || |  | || ||| |||||||||| |||||| ||||||
Sbjct  7139  ATTGCCCAGCTATCTGTCACTTTATTGTGAAGATAGTGGAAAAGGAAGGTGGCTCCTACA  7198

Query  552   AATGCCATCATTGCGATAAAGGAAAGGCTATCGTTCAAGATGCCTCTACCGACAGTGGTC  611
             |||||||||||||||||||||||||||| |||||| |||||||||| ||||||||||||
Sbjct  7199  AATGCCATCATTGCGATAAAGGAAAGGCCATCGTTGAAGATGCCTCTGCCGACAGTGGTC  7258

Query  612   CCAAAGATGGACCCCCACCCACGAGGAACATCGTGGAAAAAGAAGACGTTCCAACCACGT  671
             |||||||||||||||||||||||||||| |||||||||||||||||||||||||||||||
Sbjct  7259  CCAAAGATGGACCCCCACCCACGAGGAGCATCGTGGAAAAAGAAGACGTTCCAACCACGT  7318

Query  672   CTTCAAAGCAAGTGGATTGATGTGATATCTCCACTGACGTAAGGGATGACGCACAATCCC  731
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  7319  CTTCAAAGCAAGTGGATTGATGTGATATCTCCACTGACGTAAGGGATGACGCACAATCCC  7378

Query  732   ACTATCCTTCGCAAGACCCTTCCTCTATATAAGGAAGTTCATTTCATTTGGAGAGGACAC  791
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  7379  ACTATCCTTCGCAAGACCCTTCCTCTATATAAGGAAGTTCATTTCATTTGGAGAGGACAC  7438

Query  792   GCTGAA  797
             ||||||
Sbjct  7439  GCTGAA  7444
```

The sequence is not exactly identical though; the identity percentage is **96%** for this contiguous sequence.

(**2**) What is the **accession number** of the **cauliflower mosaic virus genome**?

NC_001497 at → https://www.ncbi.nlm.nih.gov/nuccore/NC_001497

(**3**) What is the **size** of the **cauliflower mosaic virus genome**?

8024 bp

(**4**) How **many proteins** are encoded by the **cauliflower mosaic virus** genome?

7 CDS sequences are shown, as per the above link to NCBI, so we conclude that they are at the least 7 proteins in this genome.

Wikipedia already has an entry for these **7 proteins**, at:

https://en.wikipedia.org/wiki/Cauliflower_mosaic_virus#Genome

The following table shows these **7 proteins**:

ORF I	**Movement Protein**
ORF II	**Insect Transmission Factor**
ORF III	**Structural Protein, DNA-Binding Capabilities**
ORF IV	**Capsid Protein**
ORF V	**Protease, Reverse Transcriptase and RNaseH**
ORF VI	**Translational Activator, Inclusion Body Formation/Trafficking; Possibly more functions (see below)**
ORF VII	**Unknown (Appears to not be required for infection)**

(**5**) How many **publications**, dealing with **cauliflower mosaic virus** are currently listed in **Pubmed**?

I picked Title as search form, leading to this query:

> https://www.ncbi.nlm.nih.gov/pubmed?term=cauliflower%20mosaic%20virus[Title]

The result shows **Items: 1 to 20 of 375**, so we can conclude that at the least a minimum of 375 articles deal with CMV directly in the title. If we extend this to the abstract, using this query:

> https://www.ncbi.nlm.nih.gov/pubmed?term=cauliflower%20mosaic%20virus[Title%2FAbstract]

Then we can obtain **1824** results. So this should technically be the correct answer to the given question.

(**6**) Which is the **most recent publication** mentioning the cauliflower mosaic virus?

We can sort by "**Most Recent**", which appears to be the default already as-is.

The **most recent** one then appears to be:

> https://www.ncbi.nlm.nih.gov/pubmed/29163571

Which is:

> "**Setting Up Shop: The Formation and Function of the Viral Factories of Cauliflower mosaic virus.**", published in **Front Plant Sci. 2017 Oct 30**;8:1832. **doi: 10.3389**/fpls.2017.01832. eCollection 2017.

(7) Give **information** about the **relative position** (base pairs) of the **35S promoter** within the **CaMV genome**?

According to http://www.bios.net/daisy/promoters/242/g1/250.html, the 35S promoter begins with 5'-**TGAGACTTTTCAACAAAGGG**. There are also **CAAT** sequences, in particular **CCACT** and **CACAAT**.

Full **promoter sequence** is:

5'-TGAGACTTTTCAACAAAGGGTAATATCCGGAAACCTCCTCGGATTCCATT
GCCCAGCTATCTGTCACTTTATTGTGAAGATAGTGGAAAAGGAAGGTGGCTCCTACA
AATGCCATCATTGCGATAAAGGAAAGGCCATCGTTGAAGATGCCTCTGCCGACAGTG
GTCCCAAAGATGGACCCCCACCCCACGAGGAGCATCGTGGAAAAAGAAGACGTTCCA
ACCACGTCTTCAAAGCAAGTGGATTGATGTGATATCTCCACTGACGTAAGGGATGAC
GCACAATCCCACTATCCTTCGCAAGACCCTTCCTCTATATAAGGAAGTTCATTTCAT
TTGGAGAGGA

The **biggest overlap** starts at nucleotide position **7092** and ends at nucleotide position 7435. The **identities** are **99%** there. Only difference is at position **7276**, where there appears to be an insertion of C of the PCR fragment. (The Cauliflower genome does not have a C at that position). Note that this overlaps with the gene coding for the "**inclusion body matrix protein**" protein.

(**8**) Which **fragments** of the **CaMV genome** are homologous to the promoter query sequence (starting at base pair __ until base pair __)? Which fragments of the promoter query sequence are homologous to the CaMV genome (starting at base pair __ until base pair __)?

Starting at nucleotide position **7092** and ending at nucleotide position **7435**.

For the **promoter query sequence**, starting at base pair **1** and ending at base pair **345**.

Range 1: 1 to 345 Graphics

Score	Expect	Identities	Gaps	Strand
630 bits(341)	0.0	344/345(99%)	1/345(0%)	Plus/Plus

```
Query  7092  TGAGACTTTTCAACAAAGGGTAATATCCGGAAACCTCCTCGGATTCCATTGCCCAGCTAT  7151
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     TGAGACTTTTCAACAAAGGGTAATATCCGGAAACCTCCTCGGATTCCATTGCCCAGCTAT  60

Query  7152  CTGTCACTTTATTGTGAAGATAGTGGAAAAGGAAGGTGGCTCCTACAAATGCCATCATTG  7211
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    CTGTCACTTTATTGTGAAGATAGTGGAAAAGGAAGGTGGCTCCTACAAATGCCATCATTG  120

Query  7212  CGATAAAGGAAAGGCCATCGTTGAAGATGCCTCTGCCGACAGTGGTCCCAAAGATGGACC  7271
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121   CGATAAAGGAAAGGCCATCGTTGAAGATGCCTCTGCCGACAGTGGTCCCAAAGATGGACC  180

Query  7272  CCCA-CCCACGAGGAGCATCGTGGAAAAAGAAGACGTTCCAACCACGTCTTCAAAGCAAG  7330
              |||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181   CCCACCCCACGAGGAGCATCGTGGAAAAAGAAGACGTTCCAACCACGTCTTCAAAGCAAG  240

Query  7331  TGGATTGATGTGATATCTCCACTGACGTAAGGGATGACGCACAATCCCACTATCCTTCGC  7390
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241   TGGATTGATGTGATATCTCCACTGACGTAAGGGATGACGCACAATCCCACTATCCTTCGC  300

Query  7391  AAGACCCTTCCTCTATATAAGGAAGTTCATTTCATTTGGAGAGGA  7435
              |||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301   AAGACCCTTCCTCTATATAAGGAAGTTCATTTCATTTGGAGAGGA  345
```

(**9**) How many **significant hits** can be found in **BLAST homology searches** performed with the PCR fragment sequence?

I restricted the query towards “cauliflower mosaic virus **CaMV** (taxid:10641)”.

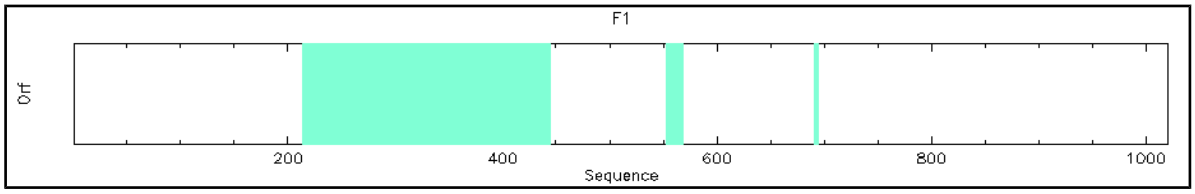
The by far best hit was towards “**Synthetic construct for plastid to nuclear gene transfer**”, with a **Query Cover of 98%**. The **accession ID** is: **AM235741.1**. → https://www.ncbi.nlm.nih.gov/nuccore/AM235741

There were more than **100 significant hits**. Worst result had a **Query Cover** of **87%**, “Cauliflower mosaic virus DNA, complete genome, isolate: **TUR81**”.

(**10**) How many **copies** of the **35S promoter** are used in the **transgenic construct**?

We can find the **TGAGACTTTT** sequence **twice**.

(**11**) Are there any parts within the **35S promoter region** that are **translated**? If yes, which **protein**(s) is/are encoded?



We can see one **large ORF** starting at about nucleotide position **205** and ending at about nucleotide position **444**.

The **protein sequence** for this would be:

> [**205 - 444**]
KRKMASTNAIIAIKERLSFKMPLPTVVPKMDPHPRGTSWKKKT
FQPRLQSKWIDVIHGGARHSRLLQEYQRYSLRRPEGY

I also had a look at the length of the **Antibody light chain proteins**: they range from **211** to **217 amino acids**. For some reason I expected to find an antibody-related sequence, but I was unable to find any homolog in the PCR fragment given.

BLASTx showed the best match being towards “RecName: Full=Transactivator/viroplasmin protein; Short=Tav; AltName: Full=**Inclusion body matrix protein**“.