Lab Worksheet 12 Solutions

Problem 1: Use Biopython to download 10 influenza hemagglutinin sequences like we did in the Class 21 worksheet. Print the list of genbank identifiers, then fetch and save all of the records to a file called "influenza_HA.gb".

In [1]:
```
# You will need Entrez and Medline to solve this problem
from Bio import Entrez, SeqIO

Entrez.email = "dariya.k.sydykova@utexas.edu"

# let's do a search for influenza H1N1 viruses from Texas
handle = Entrez.esearch(db="nucleotide",  # database to search
                        term="influenza a virus texas h1n1 hemagglutinin comple
te cds",  # search term
                        retmax=10  # number of results that are returned
                        )
record = Entrez.read(handle)
handle.close()

gi_list = record["IdList"] # list of genbank identifiers found

print(gi_list)

# Fetch records from the database
handle = Entrez.efetch(db="nucleotide", id=gi_list, rettype="gb", retmode="text
")
data = handle.read()
handle.close() # close the handle

# Write data to a file
with open("influenza_HA.gb", "w") as outfile:
    outfile.write(data)
```

```
['1609540883', '1609540874', '1609540845', '1609540826', '1609540807', '1609540
788', '1608710259', '1608708049', '1608708047', '1608708045']
```

Problem 2: Restriction enzymes cut DNA by recognizing specific motifs (patterns in the DNA sequence usually less than 10 nucleotides). Some restriction enzymes recognize degenerate motifs. That is, they recognize multiple motifs that differ by only 1 or 2 nucleotides.

Using your sequence file from Problem 1 and **regular expressions**, determine if any of the influenza sequences contain the following restriction sites:

- EcoRI: GAATTC
- BisI: GCNGC, where N represents any nucleotide

In [2]:
```
import re

# Your code goes here

# Open the genbank file with flu sequences
in_handle = open("influenza_HA.gb", "r")
records = SeqIO.parse(in_handle, "genbank")

# Start counters for EcoRI and BisI restriction sites
eco_count = 0
bis_count = 0

# Loop over each sequence
for record in records:
    # Match
    match_eco = re.search(r"GAATTC", str(record.seq))
    match_bis = re.search(r"GC[ATCG]GC", str(record.seq))
    if match_eco:
        eco_count += 1
    if match_bis:
        bis_count += 1

# Remember to close the file handle!
in_handle.close()
print("There are EcoRI restriction sites in", eco_count, "sequences.")
print("There are BisI restriction sites in", bis_count, "sequences.")
```

```
There are EcoRI restriction sites in 6 sequences.
There are BisI restriction sites in 10 sequences.
```