# Exercises on finding ori in a bacterial genome

## Finding frequent patterns

We are interested in finding a pattern that appears most frequently than others in a DNA string called ori, which corresponds to the origin of replication of DNA.

The genomic region of *Vibrio cholerae* is given as a text file containing a single string (v\_cholerae\_oric.txt).

### PatternCount

Using python, define a function PatternCount taking a pattern and a text in argument, returning the number of occurrences of this pattern in the text. Apply your function on pattern ‘TGATCA’ and the ori of *Vibrio cholerae*.

### FrequentWords

Notion : k-mer = a string of length k

Define a function FrequentWords taking a text and an integer k in argument and returning a list of the most frequent k-mer(s) in this text. Hint : you can use a frequency map of all k-mers in the text.

### PatternMatching

Define a function that returns a list of all the starting positions of a pattern in a text.

Apply your function on the assumed sequence of the DnaA box and its reverse complement.

### Other genomes

Look for the DnaA sequence you found in the ori region of Thermotoga petrophila.

Find all the 9-mers appearing 3 times or more in this region.

## Some optimization

### PatternToNumber

Define a function PatternToNumber which transforms a pattern of DNA letters into an index.

### NumberToPattern

Define a function NumberToPattern which returns a pattern of length k for a given index.

### ComputingFrequencies

Define ComputingFrequencies that returns the frequency array of the most frequent k-mers in a text (inputs : k and the text).

### FasterFrequentWords

Write a second FrequentWords function which is faster…

## Finding Clumps of patterns

### ClumpFinding

Notion :  a k-mer forms an (L, t)-clump inside a string Genome if there is an interval of Genome of length L in which this k-mer appears at least t times.

Define ClumpFinding(genome, k, L, t) : for a given genome, returns all patterns that form a (L, t)-clump in an interval L.

## Skew diagram

### Skew

Define a function Skew(i, sequence) that returns the difference between total count of G and C (G - C) from nucleotide 0 to nucleotide i of the sequence.

## Study of assumed E.coli ori

Find the most frequent 9-mers in eColi\_supOri.txt. How frequent are they ?

### HammingDistance

Notion : Hamming distance between 2 sequences is the number of nucleotides that mismatch. Two nucleotides mismatch if they are at the same position in their respective sequence but are different.

Define HammingDistance(seqA, seqB) that returns the Hamming distance between seqA and seqB.

### ApproximatePatternMatching

Define ApproximatePatternMatching(pattern, text, d) that returns all starting positions where pattern appears as a substring of text with at most d mismatches.

### ApproximatePatternCount

Define ApproximatePatternCount(pattern, text, d) that returns the number of occurrences of pattern and all of its cognate sequences with at most d mismatches, in text.