Badji Mokhtar Annaba University **2020/2021**
Master 1 BDMA
Programming in R language
Project N° 1 **Ahmed Boulemden**

Pattern recognition on biological data

Description:

The project aims to implement a biological data analysis process. In this context, you are asked to use the dataset of *Cytokines* sequences. Students have to implement every work from each of the process phases.

The process is composed of the following phases :

1- Data acquisition

The used dataset contains biological sequences of proteins from the *Cytokines* family. In this phase of the process, you have to :

- Download the TAB-seperated file from https://www.uniprot.org/uniprot/?query=cytokines&sort=score.
  The file includes 14909 entry (as of January, 30th 2021). *i:* you have to activate the presence of the sequence in the result table before the download.
- Implement a code which organizes the data from the file into the following structure. Use the import integrated GUI of r-studio (import from Text (base)),

ID	Name	Specie	Sequence
P10145	IL8_HUMAN Interleukin-8	Homo sapiens	MTSKLAVALLA……

2- Feature extraction

Implement a code for n-gram features extraction with n=1 and 2 (as explained).
Features are then saved in the following structure form.

ID	A	B	C	…	…	AA	AB	…	…
P10145	26	15	13			5	3		

- Exemple: For «ABC»
A = 1, B = 1, C = 1, AB = 1, BC = 1, rest = 0

3- Model application

The model application phase is intended for the implementation of the K-means clustering algorithm. Use this implementation to perform a clustering based on the extracted features from the previous phase. Use the **Euclidian distance.**

4- Results and evaluation.

In the last phase, students have to implement codes for visualizing and evaluating the clustering results.