Analyse Statistique des Séquences Biologiques
TD n°5: Motifs

C.Bouyioukos, G. Nuel

November 15, 2017

Exercice 1 (accès internet)**.** Récupérer à partir de `http://rsat.ulb.ac.be/rsat/` les régions *upstream* d'une famille de gènes de la levure (au format fasta); ex: `GAL_up800`, `GCN_up800`, etc. Récupérer ensuite à partir de JASPAR au moins un motif connu chez la levure (*Saccharomyces cerevisiae*) au format `sites` ou bien `pfm`.

Exercice 2 (`random`; `Biopython`: `SeqIO`, `Motif`; `R`; `matplotlib`)**.** Pour un choix seuil donné, compter le nombre de *hits* du motif que vous avez choisi dans le jeu de donnée, puis donner la significativité de cette observation sous la forme: 1) d'une *p-value* empirique; 2) d'un *z-score*. Reprendre cette démarche pour deux autres valeurs de seuil.

Exercice 3 (`biopython`: `SeqIO`, `ExPASy`, `shuffle`, `re.compile`, `find_all`)**.** Récupérer dans SWISSPROT les 16 β-globines suivantes:

Q7M2Y5	Q9YGW1	P02112	P68873
P68871	P68223	P02075	P60524
P02088	P68872	P02062	P07412
P02070	P68226	P02067	P02094

Combien de ces globines contiennent-elles l'expression régulière `G[AG]E[AVF]L` ? Utiliser le modèle de shuffling pour évaluer la significativité de cette observation.