Université Mohamed Khider-Biskra
Département d'informatique

$2^{ème}$ année Master SIOD — ADS
Série d'exercices N°1 — Novembre 2018

Fouille des séquences de données

Exercice 1 Soit les séquences d'ADN suivantes :

TTGTCAAGT, **ATTGCAGTA** et **AATCGACCG**.

1. Représenter les i-plot, f-plot et d-plot de ces séquences,
2. Calculer la séquence des états modaux,
3. Calculer, pour chacune des séquences :
   - Les durées moyennes passées dans chaque état,
   - Les taux de transition,
   - L'entropie longitudinale.
4. Calculer les distances suivantes entre les séquences :
   - Sac de caractères
   - Longest Common Prefix (LCP)

Exercice 2 Soit les séquences d'ADN suivantes :

TCAA, **TGCG** et **CATT**, **CGCC**.

1. Calculer la turbulence de chacune de ces séquences,
2. Dites quelle sont les séquences les plus similaires en utilisant :
   - 2-spectrum
   - LCSS

Exercice 3 Soit la base de données séquentielle suivante représentant l'historique des clients d'une banque :

Client	Séquence	Classe
C1	VVVRCCCVRV	N
C2	RRCCVVVCVR	F
C3	CCCCVRRRCC	F
C4	VRCRCRCCRC	N
C5	CCCCCRVVRR	N
C6	VVVCVVCVVR	F

Avec : V : virement, R : retrait, C : consultation, F :fraude, N :normal.
En utilisant la méthode KPPV (k=3) et la distance sac de caractères, trouver la classe du client ayant la séquence : CVRCVRCCVR

Exercice 4 En utilisant l'algorithme ApprioriAll avec un support minimum de 30%, trouver les séquences fréquentes dans la base séquentielle suivante :

SID	Sequence
10	<a(ac)(adc)>
20	<(ba)(fb)a>
30	<(ab)bfb(ae)>
40	<a(af)d>
50	<d(fac) >
60	<(adf)(ae)>

En déduire les séquences fréquentes maximales.

Responsable de la matière :
Dr A.Djeffal