

Exercices sur la recherche de motifs et les modèles de Markov

1. Motifs exacts : algorithme de Knuth-Morris-Pratt

On désire utiliser l'algorithme de Knuth-Morris-Pratt pour trouver les occurrences du motif GCAGAGAG dans la séquence GCATCGCAGAGAGTATACAGTACG.

Question 1. Calculer le tableau *Next*.

question 2. Exécuter la phase de recherche. Combien il y-a-t-il de comparaisons de caractère à caractère en tout ? Quel aurait été le nombre de comparaisons avec l'algorithme par force brute ?

2. TATTAT

Dans le génome, les sites d'initiation de la transcription de l'ADN en ARN sont précédés d'un site *promoteur* : c'est une région de composition particulière sur laquelle peut se fixer l'ARN polymérase. Dans les organismes procaryotes, ce promoteur est représenté par le motif approché TATTAT.

Le tableau ci-dessous donne les fréquences à chaque position pour TATTAT relevées dans le génome de la bactérie *E. coli*:

	T	A	T	T	A	T
A	0.04	0.88	0.26	0.59	0.49	0.03
C	0.09	0.03	0.11	0.13	0.22	0.05
G	0.07	0.01	0.12	0.16	0.12	0.02
T	0.80	0.08	0.51	0.13	0.18	0.89

Donner un modèle de Markov caché (sans pseudo-comptes) pour représenter un promoteur sur le *brin complémentaire* de l'ADN. À titre de rappel, la correspondance entre les bases est $A \leftrightarrow T$ et $C \leftrightarrow G$.

3. BLAST

On considère les deux séquences ATTCAATTCATTCATTCATTCATTCATTCATTC et ATTGATTGATTGATTGATTGATTGATTGATTG. Quel est, à première vue, leur pourcentage d'identité ? Quand on fait un alignement avec l'algorithme de BLAST, aucune similarité n'est trouvée. Pourquoi ?

4. Le génome de *Bizarrus Examinus*

On considère des séquences nucléotidiques d'un organisme étrange *Bizarrus Examinus*:

1. Le nucléotide actuel est un A, un C, un G ou un T avec une probabilité de 25% si les deux nucléotides précédents sont identiques.
2. Le nucléotide actuel a deux fois plus de chance d'être un C ou un G, plutôt qu'un A ou un T, si les deux nucléotides précédents sont différents. De plus, dans le choix entre C contre G, ou A contre T, les purines (A ou G) vont être préférées dans 60% des cas aux pyrimidines (C ou T).

Montrez que l'on peut modéliser les séquences de *Bizarrus Examinus* par un modèle de Markov, dont vous donnerez les états et les probabilités de transition.