

**Maîtrise d'informatique - Bioinformatique**  
**Examen de juin 2002**

Durée : 3 heures.  
 Documents et calculatrices autorisés.

---

**Questions de cours (2 points par question)**

**Question 1.** Construire l'arbre des suffixes du mot TTATTGCAGC.

**Question 2.** 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  $C \leftrightarrow G$ .

**Question 3.** L'algorithme d'alignement global des deux séquences TTGTCAAGT et ATTGCAGTAGC donne la table de programmation dynamique suivante :

		T	T	G	T	C	A	A	G	T
	0	-3	-6	-9	-12	-15	-18	-21	-24	-27
A	-3	-1	-4	-7	-10	-13	-13	-16	-19	-22
T	-6	-1	1	-2	-5	-8	-11	-14	-17	-17
T	-9	-4	1	0	0	-3	-6	-9	-12	-15
G	-12	-7	-2	3	0	-1	-4	-7	-7	-10
C	-15	-10	-5	0	2	2	-1	-4	-7	-8
A	-18	-13	-8	-3	-1	1	4	1	-2	-5
G	-21	-16	-11	-6	-4	-2	1	3	3	0
T	-24	-19	-14	-9	-4	-5	-2	0	2	5
A	-27	-22	-17	-12	-7	-5	-3	0	-1	2
G	-30	-25	-20	-15	-10	-8	-6	-3	2	-1
C	-33	-28	-23	-18	-13	-8	-9	-6	-1	1

Quelles sont les valeurs des paramètres pour une identité, une substitution, une insertion et une délétion qui ont été utilisées ? Construire un alignement optimal.

**Question 4.** Construire un jeu de scores pour l'alignement deux à deux tel que les zones de score positif correspondent à 60% de similarité.

**Exercice 1 : La plus courte super-séquence (7 points)**

Soient  $U$  et  $V$ , deux séquences. Une *super-séquence* de  $U$  et  $V$  est un mot  $S$  qui contient les lettres de  $U$  et  $V$  dans le même ordre. Autrement dit,  $U$  peut être obtenu à partir de  $S$  en appliquant une suite de délétions, et  $V$  également. Nous nous intéressons dans cet exercice à la plus courte super-séquence. Par exemple, si  $U = ATCGCC$  et  $V = ATTGAC$ , la plus courte super-séquence est  $S = ATTCGACC$ .

**Question 5.** Montrer que le meilleur alignement global de  $U$  et  $V$  sans substitutions fournit la plus courte super-séquence.

On note  $\alpha$  le score d'un match (deux lettres identiques),  $\beta$  le coût d'une insertion ou d'une délétion et  $\gamma$  le coût d'une substitution.

**Question 6.** Quelles valeurs de paramètres choisir pour  $\alpha$ ,  $\beta$  et  $\gamma$  afin que l'alignement global optimal soit toujours un alignement sans substitutions ?

**Question 7.** Soient  $m$  la longueur de  $U$ ,  $n$  la longueur de  $V$  et  $l$  la longueur de la plus courte super-séquence de  $U$  et  $V$  ( $m, n \leq l$ ). Soit également  $S$  le score final de l'alignement global entre  $U$  et  $V$ . Montrer que

$$l = n + m - \frac{S - (n + m)\beta}{\alpha - 2\beta}$$

*Indications : considérer les variables  $a$ ,  $b$  et  $c$  correspondant respectivement au nombre de matchs, de délétions et d'insertions dans l'alignement optimal, exprimer  $S$ ,  $n$ ,  $m$  et  $l$  en fonction de  $a$ ,  $b$ ,  $c$ ,  $\alpha$ ,  $\beta$ , puis résoudre.*

**Question 8.** En déduire un algorithme quadratique en temps et linéaire en espace qui permette de calculer la longueur de la plus courte super-séquence. Détailler cet algorithme en pseudo-code.

On considère maintenant le problème de la plus courte super-séquence pour un nombre quelconque de séquences.

**Question 9.** Montrer que le jeu de score utilisé pour deux séquences (trouvé en question 6) permet également de construire la plus courte super-séquence quand on l'applique à l'alignement multiple. Pour l'alignement multiple, on utilise le score SP (*sum of pairs*).

## Exercice 2 : La boîte TATA (5 points)

Les promoteurs des organismes eucaryotes - les organismes évolués - n'ont pas la même forme que chez les procaryotes (voir question 2). On trouve ainsi un motif conservé, appelé la TATA-box, situé 25 à 30 bases en amont du site d'initiation de la transcription. La TATA-box peut être représentée par le motif TATA★A★, où ★ est un symbole joker qui peut remplacer n'importe quel nucléotide. Le problème de recherche de motifs avec jokers se formule ainsi:

*Entrée :* Un motif  $M$  comprenant éventuellement des symboles jokers, un texte  $T$  sans jokers

*Sortie :* Trouver toutes les occurrences de  $M$  dans  $T$

**Question 10.** Où sont les occurrences du motifs TATA★A★ dans le texte TATAGAATATATACAG ?

**Question 11.** Pour un texte de longueur 1000, combien peut-on s'attendre à trouver d'occurrences de TATA★A★ ? Pour simplifier le problème, on supposera que les quatre bases A, C, G et T ont la même fréquence d'apparition (25%) et que les occurrences sont indépendantes l'une de l'autre.

**Question 12.** Comment adapter l'algorithme de Knuth-Morris et Pratt de recherche de motifs exacts au problème de recherche de motifs avec jokers ? Construire la table **Next** pour le motif  $M = \text{TATA} \star \text{A} \star$ .