TD Algorithmique du texte

1 Modèles de Markov cachés

On donne le modèle de Markov caché suivant :
- Un alphabet $\mathcal{A} = \{A, C, G, T\}$,
- Un ensemble d'états $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$,
- Un vecteur de probabilités de départ Π,
- Une matrice de transition T,
- Une matrice d'émission E.

avec :

$$\Pi = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad T = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0.8 & 0.1 & 0 & 0 \\ 0 & 0 & 0.5 & 0.4 & 0.1 & 0 \\ 0 & 0 & 0 & 0 & 0.9 & 0.1 \\ 0 & 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad E = \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 & 0 \\ 0.2 & 0.1 & 0.4 & 0.3 \\ 0.8 & 0.2 & 0 & 0 \\ 0.2 & 0.3 & 0.2 & 0.3 \end{bmatrix} \end{array}.$$

Exercice 1
Dessiner le modèle de Markov caché correspondant à ces informations.

Exercice 2
Dites si les séquences suivantes ont pu être produites par ce HMM, et si c'est le cas, donnez une marche ayant permis de la produire, avec la probabilité de cette marche.

1. *ACCGAT*
2. *TACGTTCG*
3. *AAAAA*
4. *ATGGAT*
5. *AGTCAAAAG*

Pour un HMM donné et une séquence w de taille l, on rappelle la méthode d'évaluation de la vraisemblance de w dans le HMM. On remplit la matrice de programmation dynamique α définie par : pour tout $j \in \{1, \ldots, |\mathcal{S}|\}$

$$\begin{cases} \alpha_1(j) &= \pi_j \times e_{j,w_1} \\ \alpha_{i+1}(j) &= e_{j,w_{i+1}} \times \sum_{k \in \mathcal{S}} \alpha_i(k) \times t_{k,j} \ \ \forall i \in \{1, \ldots l-1\} \end{cases}$$

La vraisemblance de la séquence est alors donnée par :

$$P(w|H) = \sum_{j \in \mathcal{S}} \alpha_l(j).$$

Exercice 3
Écrire l'algorithme en pseudocode.

Exercice 4
Calculer la vraisemblance de la séquence AAGACAT.