TD Algorithmique du texte

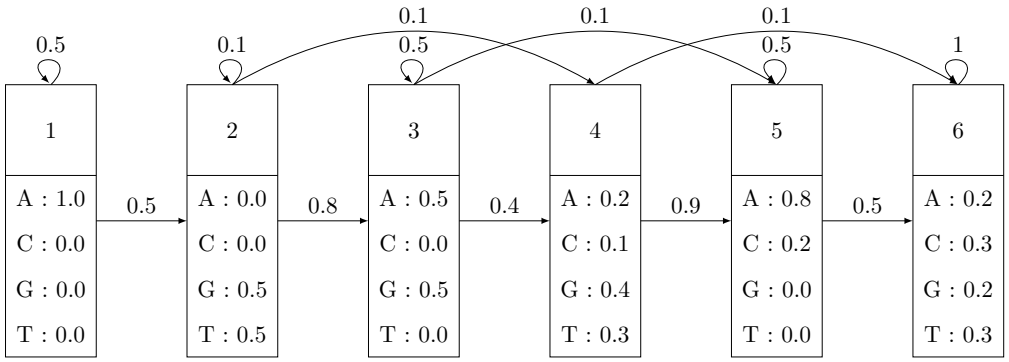
1 Modèles de Markov cachés

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 :

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

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

Remarque : Le vecteur de départ n'apparaît pas sur le dessin.



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*
   On peut produire le premier "A" en l'état 1 (état de départ à coup sûr), mais le deuxième "C" ne peut être produit ni dans l'état 1 ni dans l'état 2, qui sont les seuls états que l'on peut atteindre depuis l'état 1. Cette séquence ne peut donc être produite par le HMM.
2. *TACGTTCG*
   On ne peut pas produire le "T" depuis l'état 1 qui est l'état de départ à coup sûr : cette séquence ne peut donc être produite par le HMM.

3. *AAAAA*
   On peut produire le premier "A" en l'état 1 (état de départ à coup sûr), mais pour produire un autre "A" on doit rester dans l'état 1 (car dans l'état 2 la probabilité d'émettre un "A" est 0 et on ne peut pas atteindre d'autres état que l'état 1 ou l'état 2 depuis l'état 1). Idem pour les autres "A". La seule marche possible ici est donc la marche : 11111. Pour calculer la vraisemblance de cette marche, on utilise le calcul suivant :

$$1 \times (0.5 \times 0.5 \times 0.5 \times 0.5) \times (1 \times 1 \times 1 \times 1 \times 1) = \frac{1}{2^4}$$

   où le premier 1 est la probabilité de démarrer dans l'état 1, le bloc suivant représente les probabilités de transition sur la marche, et le dernier bloc représente les probabilités d'émission des "A" dans chacun des états 1 de la marche.

4. *ATGGAT*
   Avec cette séquence, on a plusieurs possibilités de marches. On va choisir un exemple ici. On démarre à coup sûr dans l'état 1 pour produire le premier "A". Puis on peut produire le "T" dans l'état 2 (ici on n'a pas le choix). Le troisième caractère, "G", peut être produit dans l'état 2 ou l'état 3, disons dans l'état 3. Le quatrième caractère peut être produit dans l'état 4 par exemple, le "A" dans l'état 5 et le "T" dans l'état 6. La marche 123456 convient donc pour produire la séquence dans ce HMM. Pour calculer la vraisemblance de cette marche, on utilise le calcul suivant :

$$1\times(0.5\times0.8\times0.4\times0.9\times0.5)\times(1\times0.5\times0.5\times0.4\times0.8\times0.3) = \frac{27}{5^6} = \frac{1728}{10^6} = 1.728\times10^{-3}$$

5. *AGTCAAAAG*
   Pour cette séquence on a également plusieurs choix possibles, voici une marche qui produit la séquence dans le HMM : 122455556.
   Le calcul de la probabilité de cette marche donne :

$$\frac{9}{2^3 \times 5^9} = \frac{9 \times 2^6}{10^9} = \frac{576}{10^9} = 5.76 \times 10^{-7}.$$

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.

cf. transparents du cours...

Exercice 4
Calculer la vraisemblance de la séquence AAGACAT.

On utilise l'algorithme forward sur l'entrée H et $w = AAGACAT$. Donc on commence par remplir une matrice α :

	1	2	3	4	5	6
A						
A						
G						
A						
C						
A						
T						

On initialise la première ligne avec les probabilités de départ de chaque état, multipliée par la probabilité d'émettre un "A" dans cet état respectivement. Comme la seule probabilité de départ non nulle est celle de l'état 1, on obtient :

	1	2	3	4	5	6
A	1	0	0	0	0	0
A						
G						
A						
C						
A						
T						

Pour chacune des cases de la ligne suivante, on doit faire la somme sur la ligne précédente des contenus des cases multipliés par la probabilité de transition et la probabilité d'émission. Comme certaines de ces probabilités sont nulles, cela simplifie le calcul. Ainsi, pour la deuxième ligne, toutes les cases à 0 de la première ligne ne vont pas contribuer à des probabilités non nulle dans la deuxième. La seule contribution non nulle proviendrait de la première case. Étant donné qu'il y a aussi des probabilités nulles de transition de l'état 1 à l'état 3, 4, 5 et 6, la première case ne peut pas contribuer non plus à une probabilité non nulle pour les cases correspondant à ces états dans la deuxième ligne. Enfin, la probabilité d'émettre un "A" dans l'état

2 étant nulle, on peut directement mettre un zéro dans la case correspondant à l'état 2 dans la deuxième ligne. Il reste un terme non nul, celui de la première case, qui correspond au produit de la probabilité se trouvant dans la case de l'état 1 à la première ligne, multipliée par la probabilité de transition de l'état 1 vers l'état 1, et par la probabilité d'émettre un "A" dans l'état 1 : $1 \times \frac{1}{2} \times 1$. Cela nous donne un remplissage de la deuxième ligne :

	1	2	3	4	5	6
A	1	0	0	0	0	0
A	$\frac{1}{2}$	0	0	0	0	0
G						
A						
C						
A						
T						

De même pour la ligne suivante, on peut déjà mettre des zéros dans les cases des états 3, 4, 5, 6 que l'on ne peut pas atteindre depuis le seul état de la ligne précédente qui n'a pas de probabilité nulle. Et on peut mettre un zéro dans la case correspondant à l'état 1 puisque la probabilité d'émettre un "G" dans l'état 1 est nulle. Il reste à calculer la probabilité de la case correspondant à l'état 2, qui correspond au produit de la probabilité se trouvant dans la case de l'état 1 à la deuxième ligne, multipliée par la probabilité de transition de l'état 1 vers l'état 2, et par la probabilité d'émettre un "G" dans l'état 2 : $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{2^3}$. Cela nous donne un remplissage de la deuxième ligne :

	1	2	3	4	5	6
A	1	0	0	0	0	0
A	$\frac{1}{2}$	0	0	0	0	0
G	0	$\frac{1}{2^3}$	0	0	0	0
A						
C						
A						
T						

Dans la ligne suivante, une fois mises à zéros les cases dans lesquelles la probabilité d'émettre un "A" est nulle et les cases correspondant aux états qui ne sont pas accessibles depuis l'état 2, il reste les cases des états 3 et 4. Pour l'état 3, on multiplie la probabilité de la case correspondant à l'état 2

sur la troisième ligne par la probabilité de transition de l'état 2 à l'état 3 et la probabilité d'émettre un "A" dans l'état 3 : $\frac{1}{2^3} \times \frac{4}{5} \times \frac{1}{2} = \frac{1}{2^2 \times 5}$. De même pour l'état 4 : $\frac{1}{2^3} \times \frac{1}{10} \times \frac{1}{5} = \frac{1}{2^4 \times 5^2}$, ce qui nous donne :

	1	2	3	4	5	6
A	1	0	0	0	0	0
A	$\frac{1}{2}$	0	0	0	0	0
G	0	$\frac{1}{2^3}$	0	0	0	0
A	0	0	$\frac{1}{2^2 \times 5}$	$\frac{1}{2^4 \times 5^2}$	0	0
C						
A						
T						

Dans la ligne suivante, il y a trois termes non nuls : celui de l'état 4, dont le calcul est : $\frac{1}{2^2 \times 5} \times \frac{1}{10} \times \frac{2}{5} = \frac{1}{2^2 \times 5^3}$; celui de l'état 5, dont le calcul contient cette fois deux termes : $\frac{1}{2^{\times} 5} \times \frac{1}{10} \times \frac{1}{5} + \frac{1}{2^4 \times 5^2} \times \frac{9}{10} \times \frac{1}{5} = \frac{29}{2^5 \times 5^4}$; et enfin celui de l'état 6 : $\frac{1}{2^4 \times 5^2} \times \frac{1}{10} \times \frac{3}{10} = \frac{3}{2^6 \times 5^4}$. On obtient le tableau :

	1	2	3	4	5	6
A	1	0	0	0	0	0
A	$\frac{1}{2}$	0	0	0	0	0
G	0	$\frac{1}{2^3}$	0	0	0	0
A	0	0	$\frac{1}{2^2 \times 5}$	$\frac{1}{2^4 \times 5^2}$	0	0
C	0	0	0	$\frac{1}{2^2 \times 5^3}$	$\frac{29}{2^5 \times 5^4}$	$\frac{3}{2^6 \times 5^4}$
A						
T						

La sixième ligne contient deux termes non nuls, celui de l'état 5 et de l'état 6. Pour l'état 5, on a à nouveau deux termes dans le calcul (venant de l'état 4 et de l'état 5 de la ligne supérieure) : $\frac{1}{2^2 \times 5^3} \times \frac{9}{10} \times \frac{4}{5} + \frac{29}{2^5 \times 5^4} \times \frac{1}{2} \times \frac{4}{5} = \frac{101}{2^4 \times 5^5}$. Pour l'état 6 : $\frac{29}{2^5 \times 5^4} \times \frac{1}{2} \times \frac{1}{5} + \frac{3}{2^6 \times 5^4} \times 1 \times \frac{1}{5} = \frac{1}{2 \times 5^5}$. Cela donne :

	1	2	3	4	5	6
A	1	0	0	0	0	0
A	$\frac{1}{2}$	0	0	0	0	0
G	0	$\frac{1}{2^3}$	0	0	0	0
A	0	0	$\frac{1}{2^2\times 5}$	$\frac{1}{2^4\times 5^2}$	0	0
C	0	0	0	$\frac{1}{2^2\times 5^3}$	$\frac{29}{2^5\times 5^4}$	$\frac{3}{2^6\times 5^4}$
A	0	0	0	0	$\frac{101}{2^4\times 5^5}$	$\frac{1}{2\times 5^5}$
T						

Pour la dernière ligne, seule la dernière case contient une probabilité non nulle, car la probabilité d'émettre un "T" est nulle dans l'état 5. Le calcul de la probabilité dans l'état 6 est alors : $\frac{101}{2^4\times 5^5}\times\frac{1}{2}\times\frac{3}{10}+\frac{1}{2\times 5^5}\times 1\times\frac{3}{10}=\frac{351}{2^6\times 5^6}$. On a donc le tableau α rempli :

	1	2	3	4	5	6
A	1	0	0	0	0	0
A	$\frac{1}{2}$	0	0	0	0	0
G	0	$\frac{1}{2^3}$	0	0	0	0
A	0	0	$\frac{1}{2^2\times 5}$	$\frac{1}{2^4\times 5^2}$	0	0
C	0	0	0	$\frac{1}{2^2\times 5^3}$	$\frac{29}{2^5\times 5^4}$	$\frac{3}{2^6\times 5^4}$
A	0	0	0	0	$\frac{101}{2^4\times 5^5}$	$\frac{1}{2\times 5^5}$
T	0	0	0	0	0	$\frac{351}{2^6\times 5^6}$

Pour calculer la vraisemblance de la séquence, on fait la somme des probabilités sur la dernière ligne : $\frac{351}{2^6\times 5^6}=\frac{351}{10^6}=3.51\times 10^{-4}$.