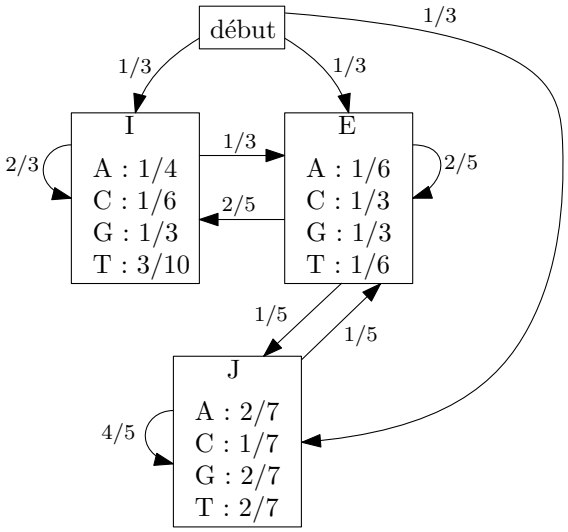
BIN702 - Série d'exercices #5 : le modélisation de séquences avec modèles de Markov

Manuel Lafond

Exercice 1: Le modèle de Markov ci-bas tente, de façon complètement irréaliste, de modéliser une séquence où chaque nucléotide peut être dans un intron (I), un exon (E), ou une région inutile (J, pour "junk").



a. Calculez la probabilité de générer $ACCT$ avec la séquence d'états $JEEI$.

Solution. Le calcul va comme suit :

$$\frac{1}{3} \cdot \frac{2}{7} \cdot \frac{1}{5} \cdot \frac{1}{3} \cdot \frac{2}{5} \cdot \frac{1}{3} \cdot \frac{2}{5} \cdot \frac{3}{10} = \frac{4}{39375}$$

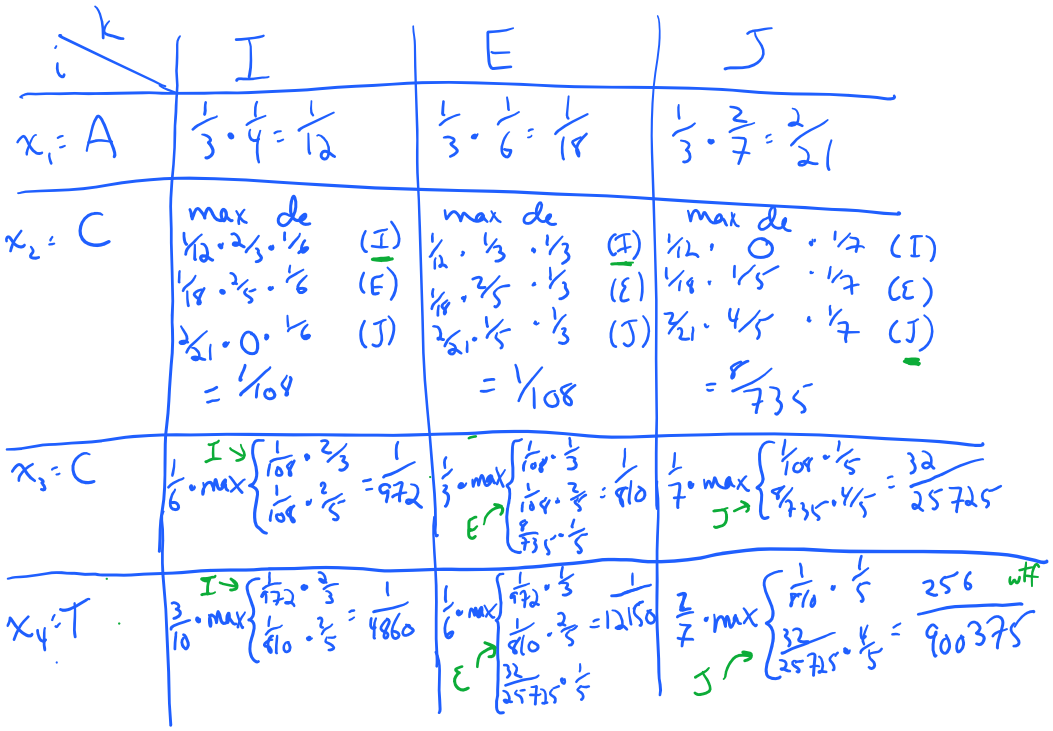
(le tout calculé mentalement, bien sûr !)

□

b. En utilisant l'algorithme de Viterbi, calculez le chemin le plus probable pour la séquence $ACCT$. La façon la plus simple est de remplir une matrice de programmation dynamique.

Solution. Il est certain que je ne demanderais jamais un calcul aussi laborieux en examen. Je regrette d'avoir donné cet exercice, car je suis maintenant obligé de fournir une solution. Par contre, si vous passez à travers cet exercice, ça devrait vous permettre de bien comprendre l'algorithme.

Rappelons que la récurrence calcule $V[i, q]$, la probabilité du meilleur chemin qui génère $s_1 \dots s_i$ en terminant à l'état q. La table suivante illustre tous les calculs (j'ai fait cet exercice en utilisant une autre notation : j'ai mis un k au lieu d'un q, et les x_i devraient être des s_i). En vert, l'état qui a mené à la valeur optimale observée.



Au final, la meilleure probabilité possible est $\frac{256}{900375}$ (par calcul mental, bien sûr), et le chemin le plus probable donnant cette séquence est simplement $JJJJ$.

Pour fins pédagogique, quelques explications. La rangée 1 utilise les conditions initiales

$$V[1, q] = Pr[\text{premier état} = q] \cdot e(q, s_1)$$

où $e(q, s_i)$ est la probabilité que l'état q émette s_i. Les autres rangées

utilisent la récurrence

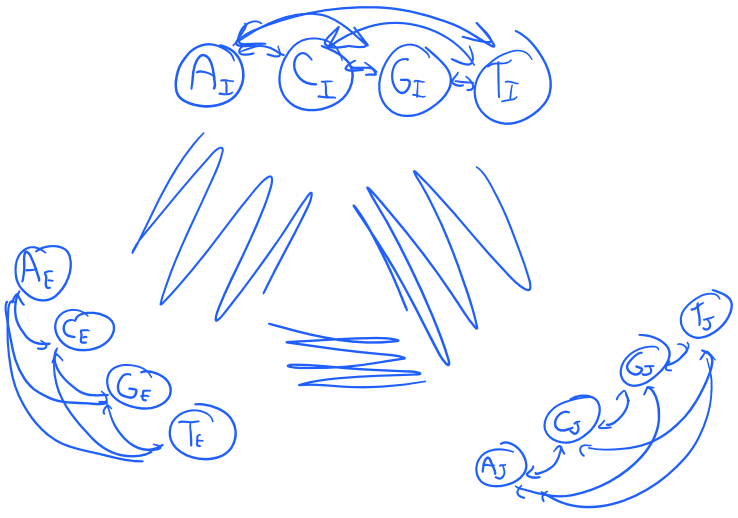
$$V[i,q] = \max_{h \in Q}(V[i-1,h]\alpha(h,q)e(q,s_i))$$

Ceci fonctionne car pour obtenir un chemin C optimal terminant à q et générant $s_1 \dots s_i$, on doit d'abord prendre un chemin C' terminant dans un certain état h et générant $s_1 \dots s_{i-1}$. La probabilité de C' est donnée par $V[i-1,h]$. On obtient C en ajoutant q à C' et en générant s_i, ce qui demande de multiplier la probabilité de $V[i-1,h]$ par la probabilité de passer de h à q ($\alpha(h,q)$) et la probabilité de générer s_i ($e(q,s_i)$). Bien sûr, on ne connaît pas l'état précédent h, alors on les essaie tous et on prend le maximum, ce qui trouvera forcément le bon état précédent optimal.

□

c. Ce modèle ne tient pas compte des dépendances entre les caractères consécutifs. Par exemple, on pourrait modéliser le fait que CG est plus rare dans l'état J que dans l'état E. Proposez un modèle de Markov caché alternatif pour pallier ce manque (ne donnez que les états - inutile de tenter d'inférer les probabilités de transitions et d'émissions).

Solution. La dépendance entre les paires de caractères peut être modélisée en ayant 4 états par classe (comme les ilôts CpG). Une proposition serait la suivante.



□