Université de Nantes
L3 - Module S32M130
Année 2008/2009

Ségolen Geffray
segolen.geffray@univ-nantes.fr

TD 3 : Modélisation pour les brins d'ADN et chaînes de Markov

Exercice 1.
La succession des nucléotides d'un brin d'ADN peut se modéliser comme une chaîne de Markov homogène $(X_n)_{n\geq 0}$ à espace d'états $E = \{A, C, G, T\}$ de matrice de transition

$$\mathcal{P} = \begin{pmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{pmatrix}$$

où pour tous i et j dans E, on a $p_{i,j} \in [0, 1]$ et $\sum_{k\in E} p_{i,k} = 1$.

1. Tracer le graphe des états associé à $\mathcal{P}$ et déterminer la nature des différents états.
2. Montrer qu'il existe une unique loi stationnaire notée π.
3. Dans la suite, on fait l'hypothèse que la loi initiale $p^{(0)}$ de la chaîne (X_n) coïncide avec la probabilité stationnaire π. Montrer que la loi de X_n est indépendante de n.
4. Pour tout $i \in E$, déterminer en fonction de π la limite presque-sûre et l'espérance de

$$\widehat{\pi}_{i,n} = \frac{1}{n} \sum_{k=0}^{n-1} I(X_k = i)$$

5. Notons p un sous-ensemble des paramètres de la matrice de transition choisis pour être non linéairement liés, par exemple

$$p = (p_{A,A}, p_{A,C}, p_{A,G}, p_{C,A}, p_{C,C}, p_{C,G}, p_{G,A}, p_{G,C}, p_{G,G}, p_{T,A}, p_{T,C}, p_{T,G})$$

Pour toute suite de réels $(i_0, ..., i_n)$ à valeurs dans E^{n+1}, on appelle fonction de log-vraisemblance la fonction L_p définie par

$$L_p(i_0, ..., i_n) = \log\Big(\mathbb{P}[X_0 = i_0, ..., X_n = i_n]\Big)$$

Exprimer $L_p(i_0, ..., i_n)$ en fonction de p (et de π).
6. Déterminer $\widehat{p}_n(i_0, ..., i_n) = \text{argmax}_p\{L_p(i_0, ..., i_n)\}$. On admettra que la matrice hessienne de L_p est définie négative en la solution annulant le vecteur gradient de L_p.
7. Déterminer la limite presque-sûre de $\widehat{p}_n(X_0, ..., X_n)$. On utilisera le fait que $(Y_n)_{n\in\mathbb{N}}$ où $Y_n = (X_n, X_{n+1})$ est une chaine de Markov à valeurs dans $E \times E$ de matrice de transition $\widetilde{P} = \Big(\widetilde{P}_{(i_1,i_2),(j_i,j_2)} = I(i_2 = j_1)p_{j_1,j_2}\Big)_{i_1,i_2,j_1,j_2\in E}$ et d'unique probabilité invariante $\widetilde{\pi} = (\widetilde{\pi}_{i,j} = \pi_i p_{i,j})_{i,j\in E}$.
8. Considérons le brin d'ADN suivant :

$$CAGCGAGGCAGCCGACCTCGGTAACCTCGG$$

a) Proposer une "bonne" approximation de la loi initiale.

b) Proposer une "bonne" approximation de la matrice de transition.

9. On suppose dorénavant que $p_{i,j} = p_j$ pour tous i et j dans E avec la contrainte $p_j \in [0,1]$ pour tout $j \in E$ et $\sum_{k \in E} p_k = 1$. Notons p' un sous-ensemble des paramètres de la matrice de transition de ce nouveau modèle, choisis pour être non linéairement liés, par exemple

$$p' = (p_A, p_C, p_G)$$

Montrer que la chaine de Markov (X_n) associée à la nouvelle matrice de transition $\mathcal{P}'$ induite par cette paramétrisation admet une unique loi stationnaire π' dont on déterminera l'expression. Dans la suite, on fait l'hypothèse que la loi initiale $p^{(0)}$ de la chaîne (X_n) coïncide avec la probabilité stationnaire π'.

10. Pour toute suite de réels $(i_0, ..., i_n)$ à valeurs dans E^{n+1}, la fonction de log-vraisemblance est à présent la fonction $L_{p'}$ définie par

$$L_{p'}(i_0, ..., i_n) = \log\Big(\mathbb{P}[X_0 = i_0, ..., X_n = i_n]\Big)$$

Déterminer $\widehat{p'}_n = \text{argmax}_{p'}\{L_{p'}(i_0, ..., i_n)\}$.

11. Pour tout i dans E, déterminer la limite presque-sûre et l'espérance de $\widehat{p'}_{i,n}$.

12. Considérons le brin d'ADN suivant :

$$CAGCGAGGCAGCCGACCTCGGTAACCTCGG$$

Proposer une "bonne" approximation de la matrice de transition dans ce nouveau modèle.

Exercice 2.

La détection des séquences codantes est un point important de l'analyse du génôme. Le dinucléotide CG (noté CpG) est relativement rare dans le génôme humain. La raison pour laquelle le dinucléotide CpG est rare tient au fait que lorsque la cytosine C est suivie par la guanine G, elle a tendance à se méthyler et méthyl-C a une forte probabilité de muter en thymine T. Par contre, ce processus de méthylation est inhibé aux abords des promoteurs de gènes et des codons START (Bird, 1987). Ces régions, dites ilôts CpG, contiennent donc une concentration plus importante en dinucléotides CpG. Leur longueur varie de quelques centaines à quelques milliers de bases. La présence d'ilôts CpG peut donc être un indicateur du début d'une séquence codante. Par conséquent, l'identification des ilôts CpG peut aider à localiser les gènes dans l'ADN.

1. Une première question intéressante est la suivante : étant donné une courte séquence d'ADN, provient-t-elle d'un ilôt CpG ou non ? En modélisant la succession des nucléotides d'un brin d'ADN comme une chaîne de Markov homogène $(X_n)_{n \geq 0}$ à espace d'états $E = \{A, C, G, T\}$, Durbin et al. (1999) ont fourni les valeurs suivantes pour la matrice de transition à l'intérieur et à l'extérieur des ilôts CpG respectivement :

$$\mathcal{P}^+ = \begin{pmatrix} 0.180 & 0.274 & 0.426 & 0.120 \\ 0.171 & 0.368 & 0.274 & 0.188 \\ 0.161 & 0.339 & 0.375 & 0.125 \\ 0.079 & 0.355 & 0.384 & 0.182 \end{pmatrix}$$

et

$$\mathcal{P}^- = \begin{pmatrix} 0.300 & 0.205 & 0.285 & 0.210 \\ 0.322 & 0.298 & 0.078 & 0.302 \\ 0.248 & 0.246 & 0.298 & 0.208 \\ 0.177 & 0.239 & 0.292 & 0.292 \end{pmatrix}$$

Pour répondre à la question posée, on peut procéder comme suit. On détermine le log-odds-ratio de la séquence considérée $(i_0, ..., i_n)$ défini par :

$$LOR(i_0, ..., i_n) = \log\left(\frac{\prod_{k=0}^{n-1} p^+_{i_k, i_{k+1}}}{\prod_{k=0}^{n-1} p^-_{i_k, i_{k+1}}}\right)$$

Si $LOR(i_0, ..., i_n) > 0$, alors on décide que la séquence considérée est un ilôt CpG.
Soit la séquence ACGTACG. S'agit-il d'un ilôt CpG ?

2. Une seconde question intéressante est la suivante : étant donné une longue séquence d'ADN, contient-elle des ilôts CpG ? Pour y répondre, introduisons l'espace des états $F = \{0, 1\}$ où 1 code pour l'appartenance à un ilôt CpG et où 0 code pour la non-appartenance à un ilôt CpG. On choisit de modéliser la succession de la nature (CpG ou non-CpG) des nucléotides d'un brin d'ADN par une chaîne de Markov homogène $(Y_n)_{n\geq 0}$ à espace d'états F de matrice de transition $\mathcal{P}$ dont le terme (i, j) est donné par

$$p_{i,j} = \mathbb{P}[Y_1 = j | Y_0 = i]$$

Soit $\nu = (\nu_j)_{j\in F}$ la loi initiale de la chaîne (Y_n) en notant $\nu_j = \mathbb{P}[Y_0 = j]$ pour $j \in F$. La chaîne de Markov (Y_n) est non-observée : on dit qu'elle est cachée. En revanche, on observe la succession des nucléotides modélisée par une suite de variables (X_n) à valeurs dans $E = \{A, C, G, T\}$. On fait les hypothèses suivantes :

• les (X_n) sont mutuellement indépendants conditionnellement aux (Y_n), ce qui peut s'énoncer sous la forme suivante : pour tout $n \in \mathbb{N}$, pour tous $j_0, ..., j_n$ dans F, pour tous $i_0, ..., i_n$ dans E, on a :

$$\mathbb{P}[X_0 = i_0, ..., X_n = i_n | Y_0 = j_0, ..., Y_n = j_n] = \prod_{k=1}^{n} \mathbb{P}[X_k = i_k | Y_0 = j_0, ..., Y_n = j_n]$$

• pour tout $n \in \mathbb{N}$, la loi de X_n ne dépend que de Y_n, ce qui entraine en particulier que l'on a pour tout $N \in \mathbb{N}$, pour tout $n \in \{0, ..., N\}$, pour tout i dans E, pour tous $j_0, ..., j_n$ dans F

$$\mathbb{P}[X_n = i | Y_0 = j_0, ..., Y_N = j_N] = \mathbb{P}[X_n = i | Y_n = j_n]$$

et

$$\mathbb{P}[X_n = i | Y_0 = j_0, ..., Y_n = j_n, X_0 = i_0, ..., X_{n-1} = i_{n-1}] = \mathbb{P}[X_n = i | Y_n = j_n]$$

• pour tout $n \in \mathbb{N}$, la loi de X_n conditionnelle à Y_n est indépendante de n. On appelle alors probabilité d'émission la probabilité suivante définie pour $i \in E$ et $j \in F$:

$$q(j, i) = \mathbb{P}[X_n = i | Y_n = j]$$

• pour tout $n \in \mathbb{N}^*$, la loi de Y_n ne dépend que de Y_{n-1} (et est indépendante de n par la propriété de Markov)

a) Montrer que sous ces hypothèses, le modèle est entièrement déterminé par la donnée de ν la loi initiale de (Y_n), de $\mathcal{P}$ la matrice de transition de (Y_n) et de la matrice des probabilités d'émission $\mathcal{Q} = (q(j,i))_{j\in F, i\in E}$. Pour cela, on montrera que pour tout $n \in \mathbb{N}$, pour tous $j_0, ..., j_n$ dans F, pour tous $i_0, ..., i_n$ dans E, on a :

$$\mathbb{P}[X_0 = i_0, ..., X_n = i_n, Y_0 = j_0, ..., Y_n = j_n] = \nu_{j_0}\, q(j_0, i_0) \left(\prod_{k=1}^{n} p_{j_{k-1}, j_k} q(j_k, i_k) \right)$$

b) Soit $(i_n)_{n\in\mathbb{N}}$ une suite fixée de E. Pour tout $j \in F$, définissons pour $n = 0$

$$V_j^{(0)}(i_0) = \nu_j q(j, i_0)$$

et pour $n \in \mathbb{N}^*$

$$V_j^{(n)}(i_0, ..., i_n) = \max_{j_0,...,j_{n-1}\in F} \Big\{ \mathbb{P}[Y_0 = j_0, ..., Y_{n-1} = j_{n-1}, Y_n = j, X_0 = i_0, ..., X_n = i_n] \Big\}$$

Montrer que la suite $(V_j^{(n)}(i_0, ..., i_n))_{n\in\mathbb{N}}$ vérifie la relation de récurrence suivante :

$$V_j^{(n+1)}(i_0, ..., i_{n+1}) = q(j, i_{n+1}) \max_{k\in F} \Big\{ p_{k,j} V_k^{(n)}(i_0, ..., i_n) \Big\}$$

c) Si l'on admet que l'on connait les paramètres $(\nu, \mathcal{P}, \mathcal{Q})$ du modèle, l'algorithme de Viterbi permet de calculer le chemin $(j_0^*, ..., j_n^*)$ qui maximise la vraisemblance à $i_0, ..., i_n$ fixés dans E, à savoir :

$$(j_0^*, ..., j_n^*) = \text{argmax}_{j_0,...,j_n\in F} \Big\{ \mathbb{P}[Y_0 = j_0, ..., Y_n = j_n, X_0 = i_0, ..., X_n = i_n] \Big\}$$

Pour $i_0, ..., i_n$ fixés dans E, l'algorithme de Viterbi consiste à calculer les $V_j^{(m)}(i_0, ..., i_m)$ pour $m = 0, ..., n$ puis à retrouver le chemin optimal pas à pas dans le sens rétrograde : connaissant j_m^*, on trouve j_{m-1}^* par la formule :

$$j_{m-1}^* = \psi_{j_m^*}(m)$$

avec

$$\psi_j(m) = \text{argmax}_{k\in F} \Big\{ p_{k,j} V_k^{(m-1)}(i_0, ..., i_{m-1}) \Big\}$$

L'algorithme de Viterbi s'écrit comme suit :

1. initialisation : $V_j^{(0)}(i_0) = \nu_j\, q(j, i_0)$
2. récurrence : pour $m = 1, ..., n$

$$V_j^{(m+1)}(i_0, ..., i_{m+1}) = q(j, i_{m+1}) \max_{k\in F} \Big\{ p_{k,j} V_k^{(m)}(i_0, ..., i_m) \Big\}$$

3. étape finale :

$$j_n^* = \text{argmax}_{k\in F} \Big\{ V_k^{(n)}(i_0, ..., i_n) \Big\}$$

4. récurrence rétrograde : pour $m = 0, ..., n-1$

$$j_m^* = \psi_{j_{m+1}^*}(m+1)$$

On fournit les valeurs suivantes des paramètres du modèle :

$$\nu = (9/10, 1/10)$$

$$\mathcal{P} = \begin{pmatrix} 9/10 & 1/10 \\ 2/10 & 8/10 \end{pmatrix}$$

$$\mathcal{Q} = \begin{pmatrix} 1/4 & 1/6 & 1/4 & 1/3 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}$$

Considérons le brin d'ADN suivant :

$$CAGCTGCGCG$$

Mettre en oeuvre l'algorithme de Viterbi pour retrouver la succession de la nature (CpG ou non-CpG) des nucléotides du brin d'ADN fourni.