Modéliser l'Aléa

TP1 - Chaînes de Markov cachées

Tonf ZHAO & Yonatan DELORO

Pour le 30 avril 2017

1 Crabes de Weldon

Question 1

En figure 1, on représente sur la même figure :

— la loi empirique des données, c'est-à-dire les $P(r = r_i) = \frac{1}{0.004} \frac{N(r \in [r_i - 0.002, r_i + 0.002])}{N_{total}}$ pour tout ratio $r_i = 0.582 + i * 0.04 \in [0.582, 0.694]$

— la loi gaussienne la plus proche de la loi empirique des données, c'est-à-dire $\mathcal{N}(\mu_{obs} = 0.642, \sigma_{obs} = 0.019)$ où μ_{obs} et σ_{obs} correspondent respectivement à la moyenne et à l'écart-type des ratios des données.

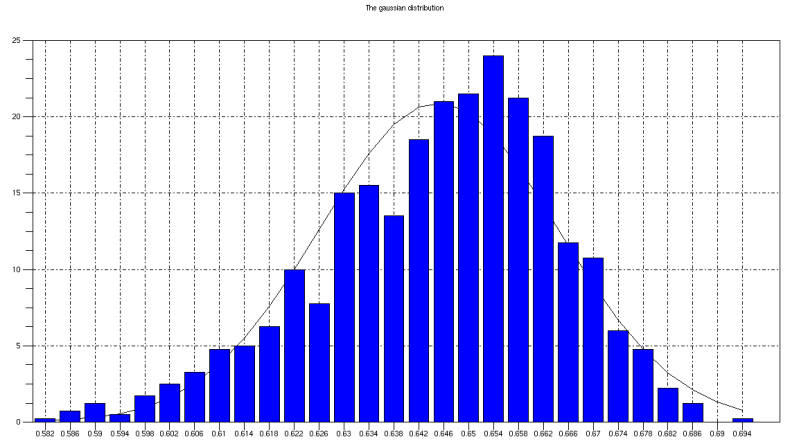


FIGURE 1 – Loi empirique des données (ratios) et loi gaussienne la plus proche

Ainsi on observe qu'une simple loi normale ne suffit par expliquer les données ($P(r = 0.638)$ ou $P(r = 0.654)$ sont par exemple très mal prédits).

Question subsidiaire 1

Testons l'hypothèse de normalité à l'aide du test du χ^2. On pose les hypothèses suivantes :

— $H_0 = \{p = p^0\}$
— $H_1 = \{p \neq p^0\}$

La statistique du test est :

$$D = \sum_{j=1}^{k} \frac{(p_j - q_j)^2}{q_j}$$

Soit Z est de loi $\chi^2(k-1)$, la p-valeur asymptotique est donnée par :

$$p = \mathbb{P}(Z > D)$$

Pour l'hypothèse nulle "les données sont observées à partir d'une loi normale $\mathcal{N}(\mu_{obs} = 0.642, \sigma_{obs} = 0.019)$", on obtient une p-valeur de 0.0000098. On a donc une forte présomption contre l'hypothèse nulle et on la rejeterait.

Question 2

L'étape de maximisation de l'algorithme EM consiste, en reprenant les notations de l'énoncé, à maximiser en θ $Q(\theta, \theta')$, qui se décompose comme :

$$Q(\theta, \theta') = NA_0 + \sum_{j \in I} A_j$$

où $A_0 = \sum_{i \in I} \pi_i \log \pi_i$ et $A_j = \sum_{k=1}^{N} \rho'_{j,k} \log f_{\mu_j, \sigma_j}(y_k)$

— Maximisons d'une part $A_0 = \sum_{i \in I} \pi_i^* \log \pi_i$ sous la contrainte $\pi \in \mathcal{P}_I \Leftrightarrow \sum_I \pi_i = 1$. A_0 est concave, en tant que combinaison linéaire à coefficients positifs de fonctions concaves. Aussi la contrainte d'égalité est affine. Par conséquent, les conditions de Kuhn et Tucker sont des conditions nécessaires et suffisantes d'optimalité.

Le lagrangien du problème de minimisation s'écrit :

$$\mathcal{L}(\pi, \lambda) = -\sum_{i \in I} \pi_i^* \log \pi_i + \lambda \left(\sum_I \pi_i - 1 \right)$$

On a donc :

$$\forall j \in I, \quad \frac{\partial \mathcal{L}(\pi, \lambda)}{\partial \pi_j} = -\frac{\pi_j^*}{\pi_j} + \lambda$$

Les points d'annulation du lagrangien vérifient donc : $\forall j \in I, \quad \pi_j^* = \lambda \pi_j$. En sommant ces égalités, on obtient car $\pi^*, \pi \in I$, $\lambda = 1$. D'où $\forall j \in I, \quad \pi_j^* = \pi_j$.

Ainsi, $\pi = \pi^*$ est l'unique maximiseur de A_0.

— Par ailleurs,

$$A_j = \sum_{k=1}^{N} \rho'_{j,k} \log \left[\frac{1}{\sqrt{2\pi\sigma_j^2}} \exp - \frac{(y_k - \mu_j)^2}{2\sigma_j^2} \right] = \sum_{k=1}^{N} \rho'_{j,k} \left[-\frac{1}{2} \log (2\pi\sigma_j^2) + \frac{(y_k - \mu_j)^2}{2\sigma_j^2} \right]$$

$$= -\frac{1}{2} \log (2\pi\sigma_j^2) \sum_{k=1}^{N} \rho'_{j,k} + \sum_{k=1}^{N} \rho'_{j,k} \frac{(y_k - \mu_j)^2}{2\sigma_j^2}$$

On a :

$$\left\{\begin{array}{rcl}\frac{\partial A_j}{\partial \mu_j} & = & \sum_{k=1}^{N} \rho'_{j,k} \frac{(y_k-\mu_j)}{\sigma_j^2} \\ \frac{\partial A_j}{\partial \sigma_j^2} & = & \sum_{k=1}^{N} \rho'_{j,k} \left[\frac{-1}{2\sigma_j^2} + \frac{(y_k-\mu_j)^2}{2\sigma_j^4}\right]\end{array}\right.$$

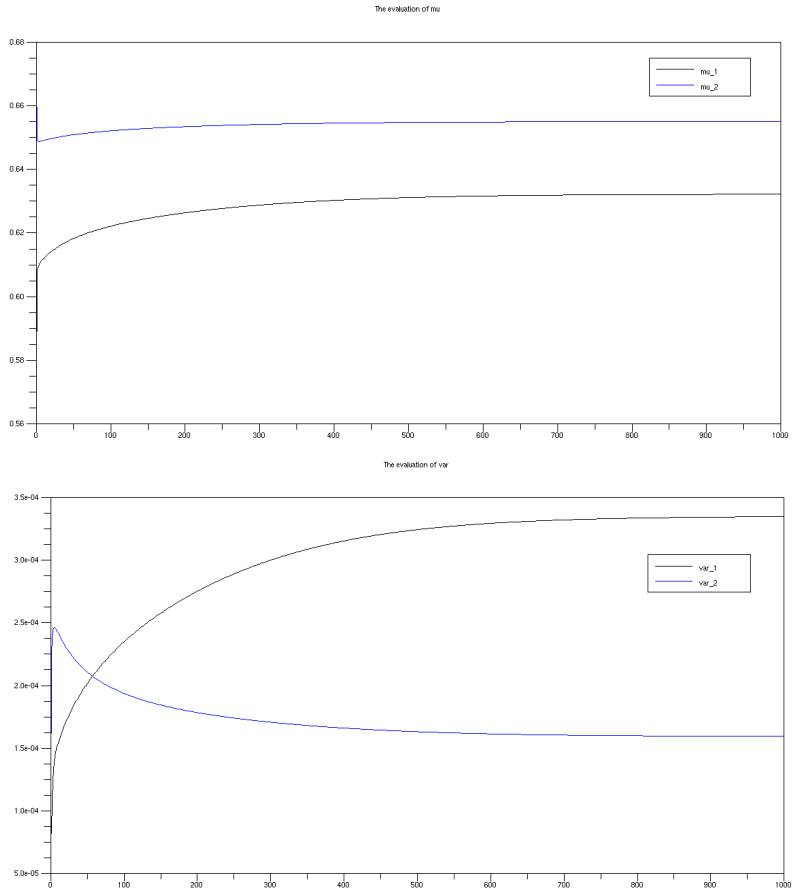
$$\left\{\begin{array}{rcl}\frac{\partial A_j}{\partial \mu_j} & = & 0 \\ \frac{\partial A_j}{\partial \sigma_j^2} & = & 0\end{array}\right. \Leftrightarrow \left\{\begin{array}{rcl}\mu_j \sum_{k=1}^{N} \rho'_{j,k} & = & \sum_{k=1}^{N} \rho'_{j,k} y_k \\ \sum_{k=1}^{N} \rho'_{j,k}(y_k-\mu_j)^2 & = & \sigma_j^2 \sum_{k=1}^{N} \rho'_{j,k}\end{array}\right. \Leftrightarrow \left\{\begin{array}{rcl}\mu_j & = & \frac{\sum_{k=1}^{N} \rho'_{j,k} y_k}{\sum_{k=1}^{N} \rho'_{j,k}} \\ \sigma_j^2 & = & \frac{\sum_{k=1}^{N} \rho'_{j,k}(y_k-\mu_j)^2}{\sum_{k=1}^{N} \rho'_{j,k}}\end{array}\right.$$

Question 3

On applique l'algorithme EM comme précisé dans l'énoncé afin d'obtenir la loi de mélange en supposant l'existence de deux populations.

On obtient après 1000 itérations les lois normales suivantes $\mathcal{N}(\mu_1 = 0.632, \sigma_1 = 0.018)$ et $\mathcal{N}(\mu_2 = 0.655, \sigma_2 = 0.013)$ dans les proportions respectives $\pi_1 = 0.432$ et $\pi_2 = 1 - \pi_1 = 0.568$.

En figure 2, on représente les différents paramètres des lois normales et de π_1 au fil des itérations de l'algorithme EM. Les valeurs indiquées ci-dessus sont donc "signifiantes" puisqu'il y a convergence.



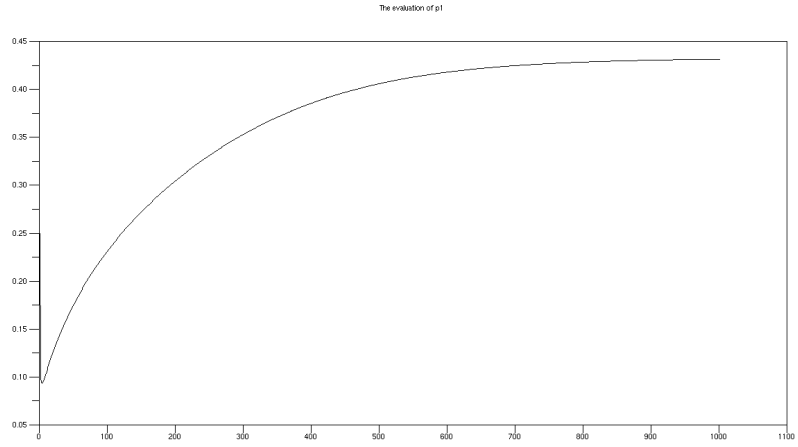


FIGURE 2 – Evolution des paramètres des gaussiennes des 2 populations

En figure 3, on représente la loi empirique et la loi de mélange des deux populations, donnée par $\pi_1\mathcal{N}(\mu_1, \sigma_1) + \pi_2\mathcal{N}(\mu_2, \sigma_2)$ (on y fait également figurer les distributions pondérées des deux populations $\pi_1\mathcal{N}(\mu_1, \sigma_1)$ et $\pi_2\mathcal{N}(\mu_2, \sigma_2)$)

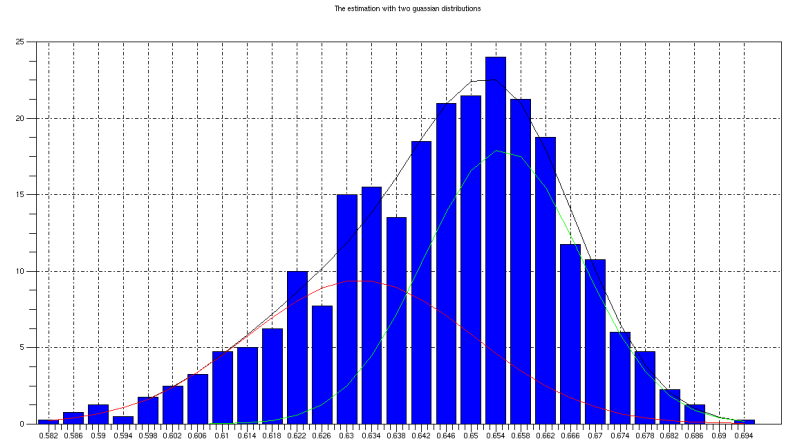


FIGURE 3 – Loi empirique et loi de mélange des 2 populations

On observe visuellement que supposer l'existence de deux populations améliore significativement la modélisation des données empiriques. Pour confirmer cette conjecture, on peut effectuer le test du χ_2 qui nous donne une p-valeur de 0.85. Dans ce cas, on ne rejeterait l'hypothèse nulle.

Question subsidiaire 3

On suppose maintenant que l'on a affaire à des données issues de trois populations de crabes. On applique à nouveau l'algorithme EM.

On obtient après 1000 itérations les lois normales suivantes $\mathcal{N}(\mu_1 = 0.599, \sigma_1 = 0.009)$, $\mathcal{N}(\mu_2 = 0.633, \sigma_2 = 0.016)$ et $\mathcal{N}(\mu_3 = 0.656, \sigma_3 = 0.012)$ dans les proportions respectives $\pi_1 = 0.021$,$\pi_2 = 0.424$ et $\pi_3 = 1 - \pi_1 - \pi_2 = 0.554$.

En figure 4, on observe la convergence des différents paramètres des lois normales et de π_1 au fil des itérations de l'algorithme EM.

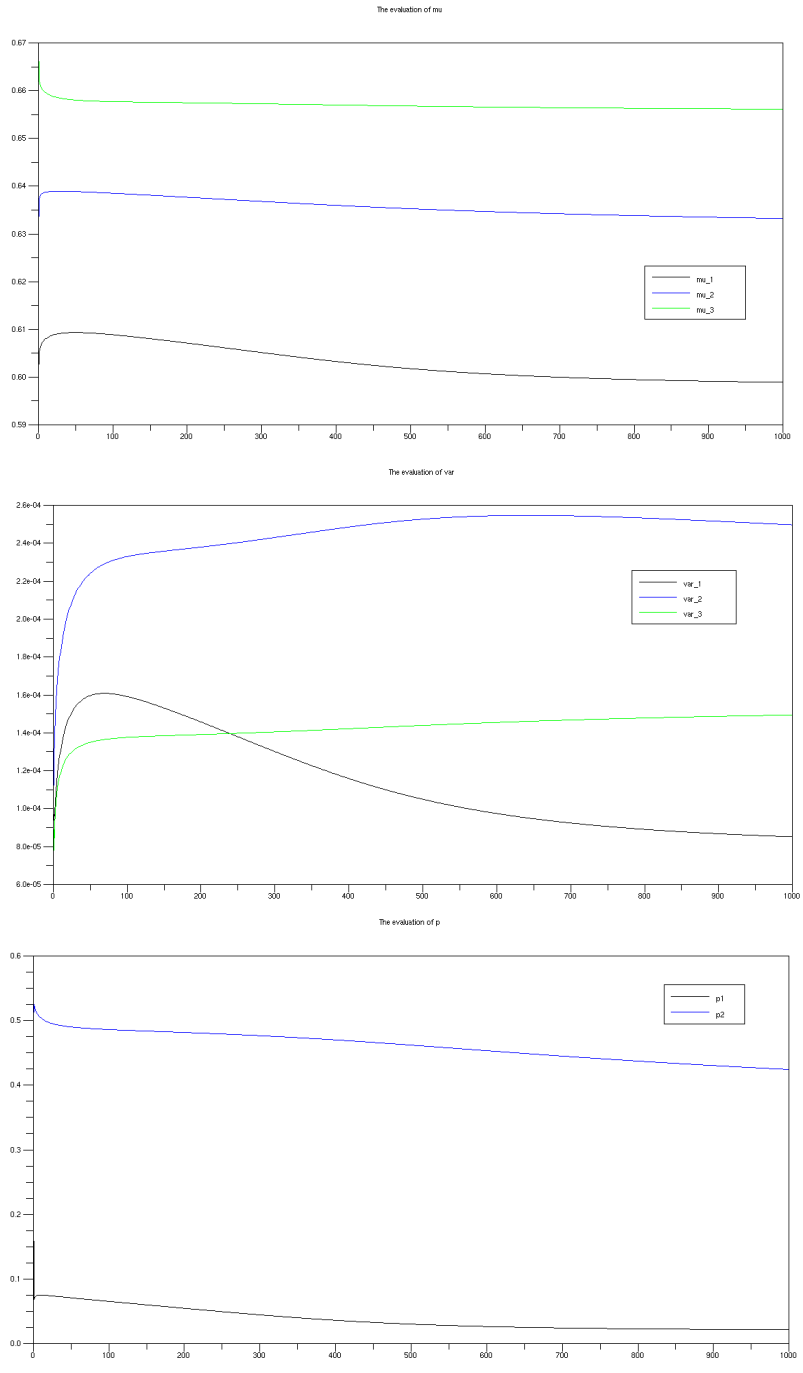


FIGURE 4 – Evolution des paramètres des gaussiennes des 3 populations

En figure 5, on représente toujours la loi empirique, ainsi que la loi de mélange des trois populations, donnée par $\pi_1\mathcal{N}(\mu_1, \sigma_1) + \pi_2\mathcal{N}(\mu_2, \sigma_2) + \pi_3\mathcal{N}(\mu_3, \sigma_3)$.

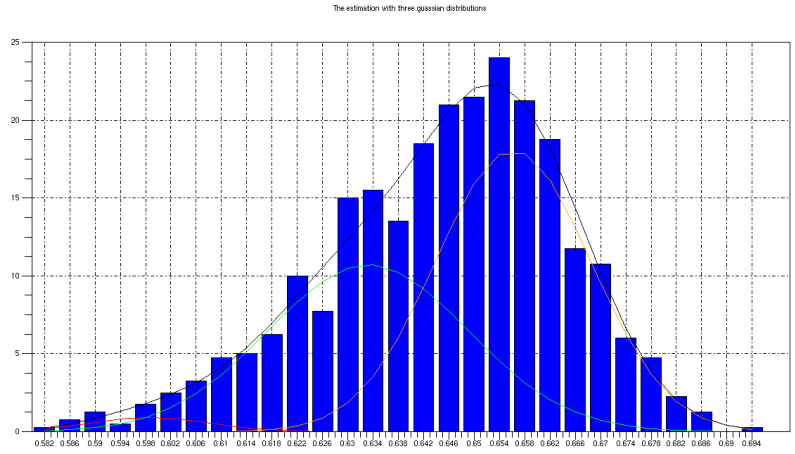


FIGURE 5 – Loi empirique et loi de mélange des 3 populations

On observe visuellement que supposer l'existence d'une troisième population ne change pas énormément le modèle mais il corresponde mieux à la loi empirique. Sa p-valeur pour le test du χ^2 est de 0.9090217, ce qui nous donne un meilleur résultat.

2 Recherche de zones homogènes dans l'ADN

Question 4

Sur le petit fichier

Après 1000 itérations de l'algorithme EM, on obtient comme loi initiale, matrice de transition de S, et matrice de probabilité de $Y|S$:

$$\pi_0 = \begin{pmatrix} 0.290 \\ 0.710 \end{pmatrix}$$

$$a = \begin{pmatrix} 0.657 & 0.343 \\ 0.139 & 0.861 \end{pmatrix}$$

$$b = \begin{pmatrix} 0.292 & 0.206 & 0.020 & 0.482 \\ 0.229 & 0.215 & 0.334 & 0.222 \end{pmatrix}$$

En figure 6 et 7, on observera bien la convergence des termes de a et de b au fil des itérations de l'algorithme EM vers ces valeurs, à condition, comme on l'a fait, de prendre une initialisation pas trop éloignée de ces valeurs (sinon, on risque de "tomber" dans un maximum local)

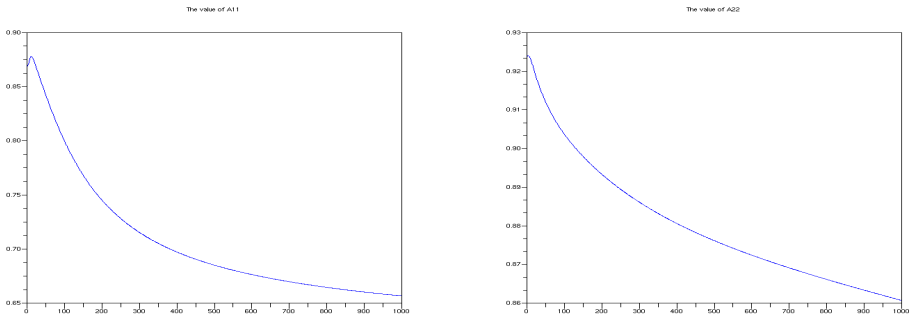


FIGURE 6 – Evolution des termes diagonaux de la matrice a au fil des itérations de l'algorithme EM pour le petit fichier

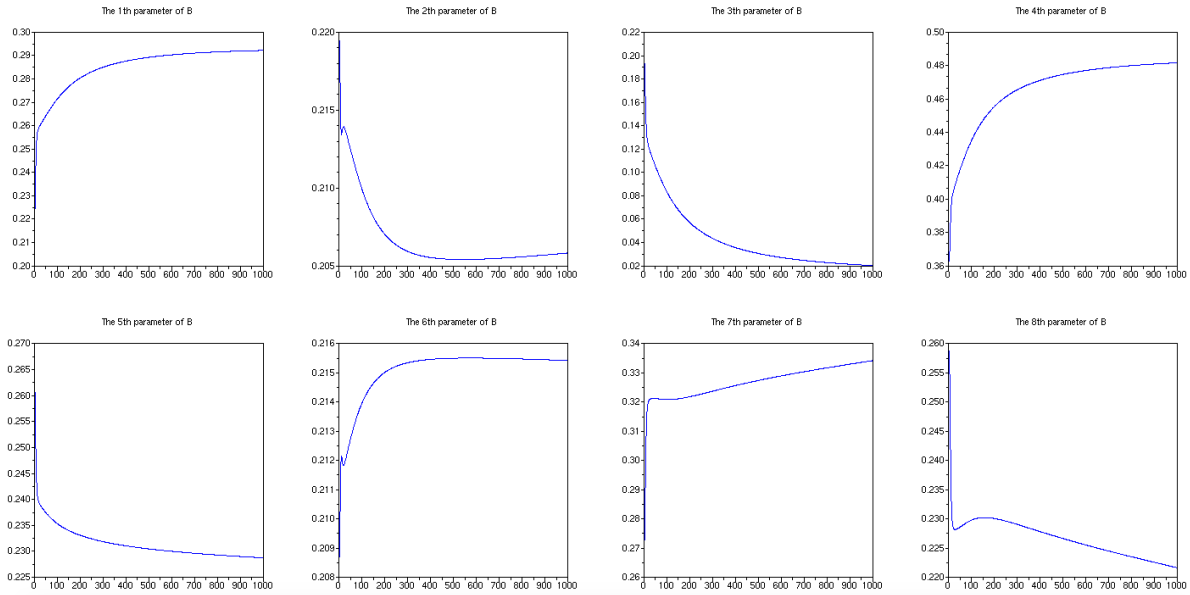


FIGURE 7 – Evolution des termes de la matrice b au fil des itérations de l'algorithme EM pour le petit fichier

Enfin, en figure 8, en représente donc la probabilité d'appartenance des nucléotides à une zone transcrite d'où $P(S_n = +1|Y_1^{N_0} = y_1^{N_0}) > P(S_n = -1|Y_1^{N_0} = y_1^{N_0})$ le long de toute la chaîne (obtenues grâce au lissage). On observe ainsi l'existence de zones homogènes, dont certaines sur plus de 30 nucléotides.

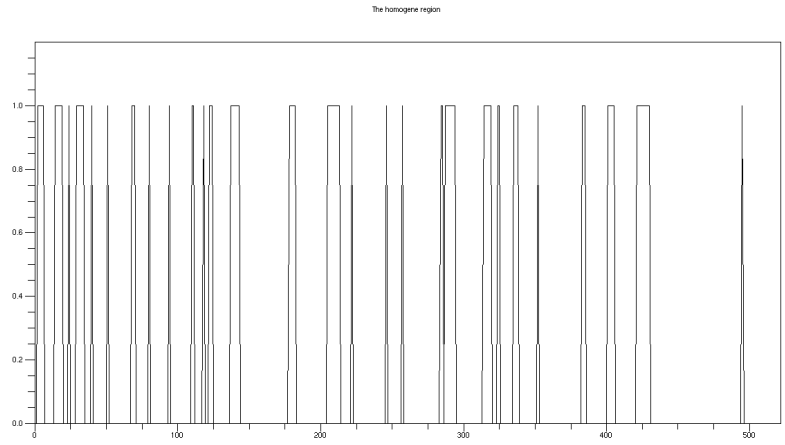


FIGURE 8 – L'appartenance des nucléotides à une zone transcrite +1 pour le petit fichier

Sur le grand fichier

La taille du fichier étant trop importante, on limite le nombre d'itérations de l'algorithme EM à 100. On trouve, avec des conditions initiales pas trop éloignées des valeurs de convergence :

$$\pi_0 = \begin{pmatrix} 0.340 \\ 0.660 \end{pmatrix}$$

$$a = \begin{pmatrix} 0.999772 & 0.000228 \\ 0.0001215 & 0.999875 \end{pmatrix}$$

$$b = \begin{pmatrix} 0.2697 & 0.206 & 0.020 & 0.482 \\ 0.229 & 0.215 & 0.334 & 0.222 \end{pmatrix}$$

En figure 9 et 10, on observera l'évolution des termes de a et b au fil des itérations de l'algorithme EM, pour une initialisation assez proche des valeurs de convergence.

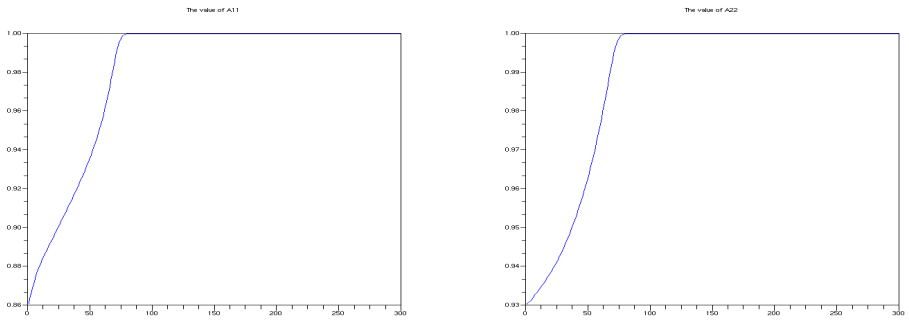


FIGURE 9 – Evolution des termes diagonaux de la matrice a au fil des itérations de l'algorithme EM pour le grand fichier

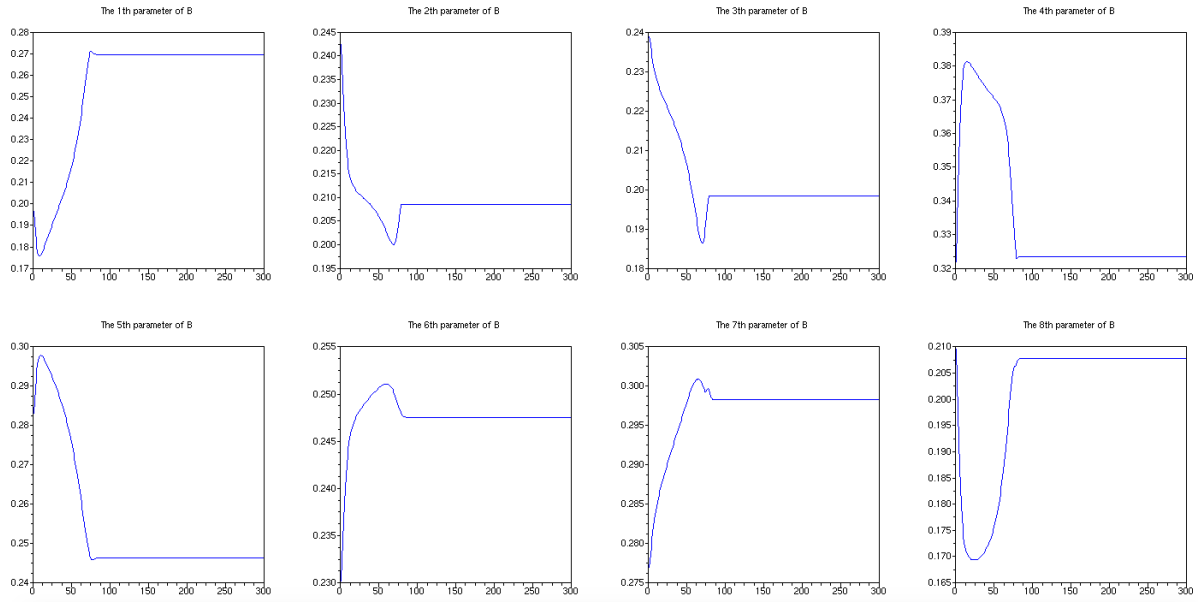


FIGURE 10 – Evolution des termes de la matrice b au fil des itérations de l'algorithme EM pour le grand fichier

Enfin, en figure 11, en représente la probabilité d'appartenance des nucléotides à une zone transcrite $P(S_n = +1|Y_1^{N_0} = y_1^{N_0})$ le long de la chaîne (obtenues grâce au lissage). On observe ainsi clairement l'existence de 6 grandes zones homogènes, trois codantes et trois non-codantes.

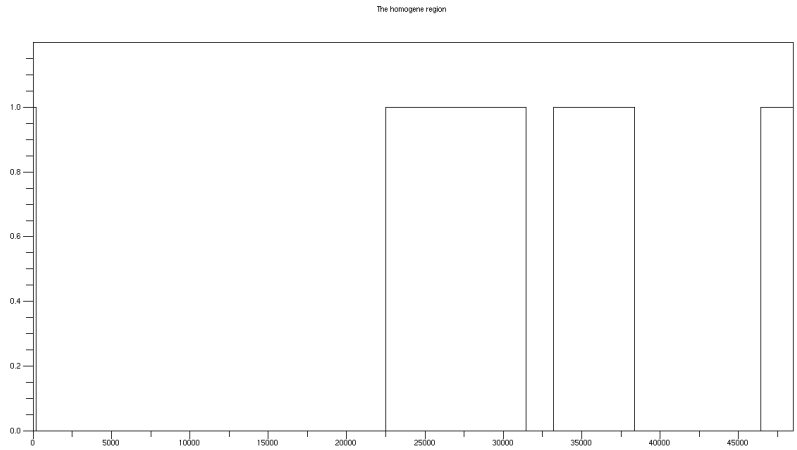


FIGURE 11 – L'appartenance des nucléotides à une zone transcrite +1 pour le grand fichier