

# MATH0488 Projet Geurts (rapport)

Di MatteoAlyssa, DucheneLouis, GerardManon

TOTAL POINTS

**38.6 / 55**

QUESTION 1

## Partie 1: 1.1 Chaînes de Markov 9 pts

1.1 Q1.1.1 2 / 2

✓ - 0 pts Correct

1.2 Q1.1.2 0.95 / 2

✓ - 0.3 pts La présentation des graphes doit être améliorée

✓ - 0.25 pts Erreur légère dans la justification/Manquement dans la justification

✓ - 0.5 pts Discussion assez pauvre

1 C'est plutôt  $\$P(X_{t=i}) = P(X_{\{t-1\}}) Q_{\{:, i\}}$

2 Cela aurait été mieux de les mettre côté à côté pour voir qu'ils convergent vers les mêmes distributions

3 Quelles conditions la chaîne respecte-t-elle pour être stationnaire?

1.3 Q1.1.3 1 / 1

✓ - 0 pts Correct

1.4 Q1.1.4 1.5 / 3

✓ - 1.5 pts Justification théorique incorrecte

4 C'est parce que la chaîne est irréductible et apériodique (et donc ergodique) qu'elle admet une distribution stationnaire (unique). C'est vers

cette distribution que l'on converge avec une séquence suffisamment longue.

Avec une chaîne ergodique, les statistiques temporelles convergent vers les statistiques obtenues sur plusieurs réalisations (ici, la statistique est la fréquence d'apparition).

1.5 Q1.1.5 0.75 / 1

✓ - 0.25 pts Faute légère/Manquement léger

5 On peut voir également que ce n'est pas parce que les fréquences d'apparition sont proches que les séquences ont été générées par la même matrice de transition

QUESTION 2

## Partie 1: 1.2 Echantillonnage de Gibbs 10 pts

2.1 Q1.2.1: Balance détaillée -->

Distribution invariante 2 / 2

✓ - 0 pts Démonstration correcte

✓ - 0 pts Condition pour unicité de la distribution invariante correcte

2.2 Q1.2.2: Gibbs --> Balance détaillée 2.4 / 3

✓ - 0 pts Cas où on reste dans le même état: correct

✓ - 0.1 pts Cas où 2+ variables changent: incorrect

ou manquant

✓ - 0.5 pts Cas où une seule variable change:  
démonstration globalement correcte mais  
comportant une/des petite.s erreur.s.

✓ - 0 pts Conditions pour que l'algorithme de Gibbs  
fonctionne: réponse complète et correcte (=dire que  
la chaîne doit être apériodique et irréductible).

6 La probabilité de choisir une variable parmi  
\$N variables selon une distribution uniforme  
est égale à  $\frac{1}{N}$  et non pas  
 $\frac{1}{N-1}$ .

### 2.3 Q1.2.3: Gibbs avec 2 variables 4 / 5

✓ - 0 pts Choix de la distribution conjointe:  
justification correcte (= chaîne irréductible + la  
somme de tous les éléments de la matrice doit valoir  
1)

✓ - 1 pts Calcul des distributions conditionnelles:  
incorrect ou manquant

✓ - 0 pts Explications concernant l'implémentation  
de l'algorithme de Gibbs (nombre d'itérations, etc.):  
ok

✓ - 0 pts Comparaison entre les fréquences  
d'apparition des états et la matrice  $Q_Y$ : analyse  
ok

✓ - 0 pts Alternance de  $Y_1$  et  $Y_2$ : ok

### QUESTION 3

## Partie 2 36 pts

### 3.1 Q2.2.1 Dérivations mathématiques 7 / 10

✓ - 0 pts Discussion ordres supérieurs présente.

✓ - 1 pts Quelques petites erreurs

✓ - 1 pts Certains passages pas suffisamment  
explicatifs.

✓ - 1 pts Quelques problèmes de formes (notations  
pas claires, pas expliquées, etc.).

7 Non, pas du tout. S dépend de A.

8 Problème de notations. Le produit sur  $k$   
doit porter aussi sur cette partie.

9 Ce terme n'est pas suffisant. Il faut un terme  
comme ça pour chaque vecteur de paramètres de  
multinomiale. Il était plus simple de passer à un  
 $\text{proto}$  directement.

10 C'est surtout lié à l'utilisation d'une approche  
bayésienne.

11 Le passage ici aurait pu être mieux expliqué.

12 Pas bien formulé. C'est le comptage des  
nucléotides dans les régions hors motif de toutes  
les séquences + le comptage dans le motif de la  
séquence  $k$ .

13 Ce n'est pas correct. C'est le nombre de  
nucléotide dans le motif de la séquence  $k$ ,  
pas dans la séquence.

14 Non, c'est une propriété de la fonction  
 $\Gamma$ , pas de Dirichlet.

15 "taille différente" ? C'est mal formulé.

16 Ce n'est pas correct. Il manque  $\alpha_i$ .  
On obtient ça par la même propriété de la  
fonction  $\Gamma$ .

17 Parce que  $\tilde{n}^{-k} = \sum_j j^{k}$

18

Manque toujours \$\$\alpha\$\$.

- 19 C'est le dénominateur
- 20 Non, la formule n'est pas correcte. On ne doit pas multiplier pour chaque valeur \$\$v\$\$ possible. Il faut sélectionner le \$\$v\$\$ par position \$\$j\$\$ qui correspondant au nucléotide à la position \$\$j-1\$\$.

### 3.2 Q2.2.2 Implémentation 8 / 10

Description de l'implémentation

✓ - 0 pts Correct

Sophistication de l'implémentation

✓ - 2 pts Bien

- Points forts: choix des positions renvoyées selon KL, shift implémenté (mais pas fonctionnel). Code ok. Description suffisante.

- 21 TB

### 3.3 Q2.2.3 Expérimentations 3 / 10

+ 3 Point adjustment

- Aucune expérimentation et aucun résultat montré, même lié à la compétition. Il n'y a que des courtes discussions (correctes cependant).

- 22 Etonnant. Ca devrait changer pas mal d'un run à l'autre.

### 3.4 Q2.2.4 Compétition 6 / 6

✓ + 2 pts Participation données artificielles

✓ + 1 pts Participation données réelles

✓ + 2 pts Résultats excellents données artificielles

✓ + 1 pts Résultats excellents données réelles

# 1 Première partie : chaînes de Markov et échantillonnage de Gibbs

## 1.1 Chaînes de Markov

### 1.1.1 Matrice de transition $Q$ avec $[Q]_{i,j} = \mathbb{P}(X_{(t+1)} = j | X_t = i)$

La matrice de transition, reprenant les probabilités de passer d'un nucléotide à l'autre, peut être déterminée en identifiant simplement, dans la séquence, le nombre de transitions d'un type de nucléotide vers un autre. Il vient alors 16 valeurs correspondantes aux différentes paires possibles (AA, AC, ..., TG, TT) qu'il suffit de diviser par le nombre de paires commençant par la même base pour obtenir les probabilités correspondantes à ces transitions. Il en découle ainsi la matrice  $Q$  suivante :

$$Q = \begin{pmatrix} 0.09722222 & 0.50520833 & 0.27777778 & 0.11979167 \\ 0.25321888 & 0.24606581 & 0.26180258 & 0.23891273 \\ 0.31100478 & 0.49282297 & 0.1076555 & 0.08851675 \\ 0.69607843 & 0.09803922 & 0.09803922 & 0.10784314 \end{pmatrix}$$

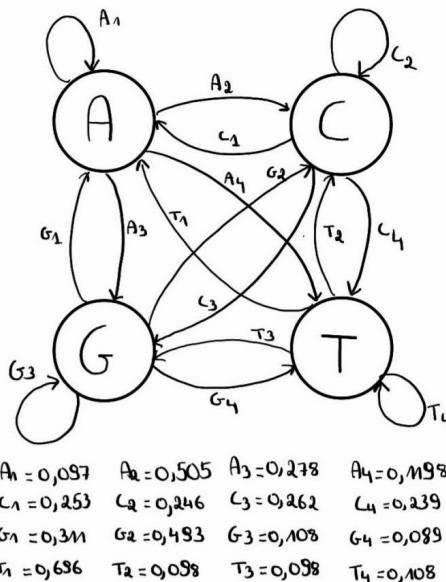


FIGURE 1. Diagramme d'états de la chaîne

### 1.1.2 $\mathbb{P}(X_t = i)$ et $Q^t$

La matrice de transition permettant de passer de l'état " $t$ " à l'état " $t+1$ ", il est possible d'exprimer  $\mathbb{P}(X_t = i)$  comme :

$$\mathbb{P}(X_t = i) = \mathbb{P}(X_{t-1} = i) Q = \mathbb{P}(X_0 = i) Q^t.$$

$\mathbb{P}(X_t = i)$  peut donc être exprimé, pour tout  $t$ , par l'unique expression de  $\mathbb{P}(X_0 = i)$  et, plus généralement, la distribution de probabilité  $\pi_t$  sur base de  $\pi_0$ .

Dans le premier cas, le nucléotide de départ de la séquence est choisi au hasard, la distribution de probabilité initiale est donc tirée d'une loi uniforme.  $\pi_0 = \mathbb{P}(X_0 = i) = (0.25, 0.25, 0.25, 0.25)$ .

Dans le second cas, la séquence débute avec un nucléotide C, dès lors, la probabilité en  $t = 0$  d'avoir C sera de 1, les probabilités des autres nucléotides étant nulles.  $\pi_0 = \mathbb{P}(X_0 = i) = (0, 1, 0, 0)$ .

1.1 Q1.1.1 2 / 2

✓ - 0 pts Correct

# 1 Première partie : chaînes de Markov et échantillonnage de Gibbs

## 1.1 Chaînes de Markov

### 1.1.1 Matrice de transition $Q$ avec $[Q]_{i,j} = \mathbb{P}(X_{(t+1)} = j | X_t = i)$

La matrice de transition, reprenant les probabilités de passer d'un nucléotide à l'autre, peut être déterminée en identifiant simplement, dans la séquence, le nombre de transitions d'un type de nucléotide vers un autre. Il vient alors 16 valeurs correspondantes aux différentes paires possibles (AA, AC, ..., TG, TT) qu'il suffit de diviser par le nombre de paires commençant par la même base pour obtenir les probabilités correspondantes à ces transitions. Il en découle ainsi la matrice  $Q$  suivante :

$$Q = \begin{pmatrix} 0.09722222 & 0.50520833 & 0.27777778 & 0.11979167 \\ 0.25321888 & 0.24606581 & 0.26180258 & 0.23891273 \\ 0.31100478 & 0.49282297 & 0.1076555 & 0.08851675 \\ 0.69607843 & 0.09803922 & 0.09803922 & 0.10784314 \end{pmatrix}$$

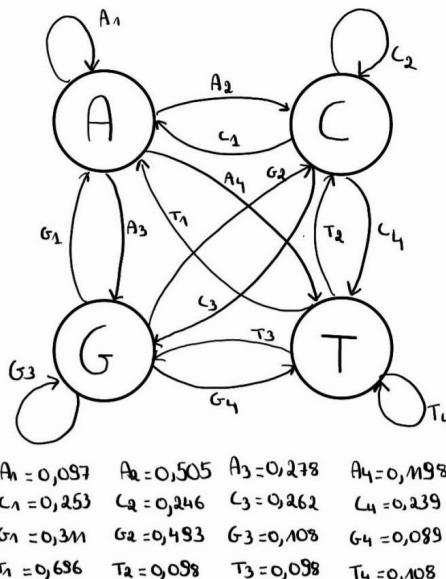


FIGURE 1. Diagramme d'états de la chaîne

### 1.1.2 $\mathbb{P}(X_t = i)$ et $Q^t$

La matrice de transition permettant de passer de l'état " $t$ " à l'état " $t+1$ ", il est possible d'exprimer  $\mathbb{P}(X_t = i)$  comme :

$$\mathbb{P}(X_t = i) = \mathbb{P}(X_{t-1} = i) Q = \mathbb{P}(X_0 = i) Q^t.$$

$\mathbb{P}(X_t = i)$  peut donc être exprimé, pour tout  $t$ , par l'unique expression de  $\mathbb{P}(X_0 = i)$  et, plus généralement, la distribution de probabilité  $\pi_t$  sur base de  $\pi_0$ .

Dans le premier cas, le nucléotide de départ de la séquence est choisi au hasard, la distribution de probabilité initiale est donc tirée d'une loi uniforme.  $\pi_0 = \mathbb{P}(X_0 = i) = (0.25, 0.25, 0.25, 0.25)$ .

Dans le second cas, la séquence débute avec un nucléotide C, dès lors, la probabilité en  $t = 0$  d'avoir C sera de 1, les probabilités des autres nucléotides étant nulles.  $\pi_0 = \mathbb{P}(X_0 = i) = (0, 1, 0, 0)$ .

Grâce à ces deux états initiaux  $\mathbb{P}(X_0 = i)$ , il est possible de générer les deux évolutions de  $\mathbb{P}(X_t = i)$  avec  $t$  représentées ci-dessous.

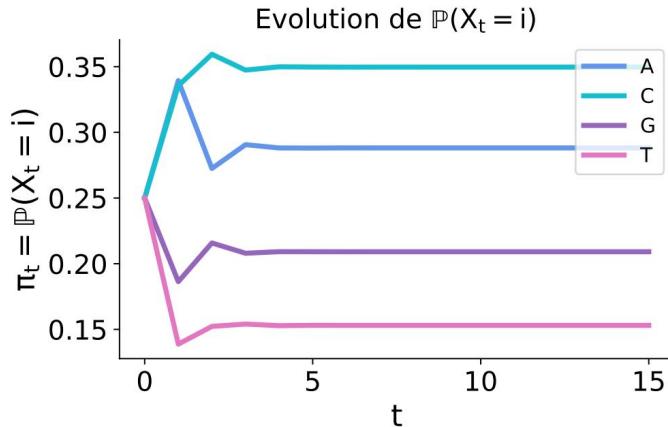


FIGURE 2.  $\mathbb{P}(X_t = i)$  avec 1<sup>e</sup> lettre aléatoire

2

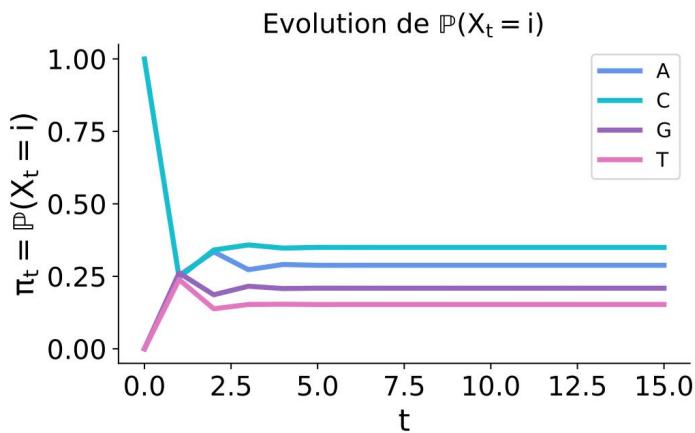


FIGURE 3.  $\mathbb{P}(X_t = i)$  avec 1<sup>e</sup> lettre étant le C

Les graphes montrent que l'évolution des distributions de probabilité tend, dans les deux cas, vers un état stationnaire au bout de plusieurs étapes. Cet état stationnaire étant commun aux deux cas illustrés, la distribution initiale n'a donc pas d'influence sur la convergence du système.

Pour illustrer ce phénomène, il convient de s'intéresser à la matrice de transition  $Q$  en différents états "t" :

$$Q^5 = \begin{pmatrix} 0.2885758 & 0.34929747 & 0.20893048 & 0.15319625 \\ 0.28801639 & 0.34994209 & 0.20920564 & 0.15283589 \\ 0.28867458 & 0.34915719 & 0.20882761 & 0.15334062 \\ 0.28689841 & 0.3504818 & 0.2095796 & 0.15304019 \end{pmatrix}$$

$$Q^{14} = \begin{pmatrix} 0.28814407 & 0.34967484 & 0.20910455 & 0.15307654 \\ 0.28814407 & 0.34967484 & 0.20910455 & 0.15307654 \\ 0.28814407 & 0.34967484 & 0.20910455 & 0.15307654 \\ 0.28814407 & 0.34967484 & 0.20910455 & 0.15307654 \end{pmatrix}$$

$$Q^{15} = \begin{pmatrix} 0.28814407 & 0.34967484 & 0.20910455 & 0.15307654 \\ 0.28814407 & 0.34967484 & 0.20910455 & 0.15307654 \\ 0.28814407 & 0.34967484 & 0.20910455 & 0.15307654 \\ 0.28814407 & 0.34967484 & 0.20910455 & 0.15307654 \end{pmatrix}$$

3

La matrice de transition étant conservée entre les états 14 et 15, il y a bien convergence vers une distribution stationnaire, qui n'était pas encore atteinte l'état 5. Ainsi,  $\mathbb{P}(X_t = i) = \mathbb{P}(X_0 = i) \times Q^t$  reste constant pour des valeurs de  $t$  suffisamment grandes.

### 1.1.3 $\mathbb{P}(X_t = \infty)$

La distribution de probabilité en  $t \rightarrow \infty$  étant stationnaire, et ce peu importe la condition initiale, elle peut être déterminée dès l'état  $t$  où se fait la convergence. Ainsi en prenant, par exemple,  $t = 15$  dans  $\mathbb{P}(X_t = i) = \mathbb{P}(X_0 = i) \times Q^t$ .

Il vient :

$$\mathbb{P}(X_\infty = i) = \mathbb{P}(X_{15} = i) = (0.28814407, 0.34967484, 0.20910455, 0.15307654)$$

Par ailleurs, du point de vue théorique, une distribution de probabilité invariante est nécessairement un vecteur propre à gauche de la matrice de transition, celui-ci étant associé à un vecteur unitaire. Ce critère mène au vecteur  $(0.28814407, 0.34967484, 0.20910455, 0.15307654)$  qui correspond bien à la distribution stationnaire déterminée précédemment.

### 1.1.4 Proportion d'apparition des nucléotides pour une longueur T croissante de chaîne

La réalisation est générée de la manière suivante :

En partant d'un nucléotide choisi arbitrairement, une nouvelle base est tirée à partir de la matrice de transition Q jusqu'à avoir atteint une chaîne de longueur T. Pour chaque longueur de la chaîne croissante, la proportion d'apparition de chacune des bases est calculée. Ces proportions et leur évolution sont représentées dans la Figure 4.

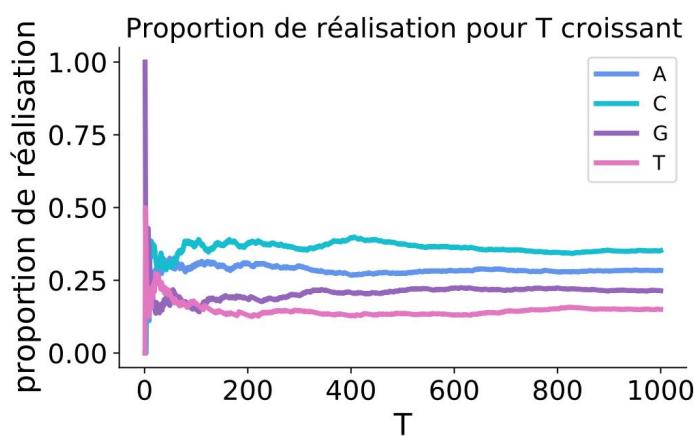


FIGURE 4. Proportion de réalisation des bases pour une longueur T croissante

Il y a convergence vers  $\mathbb{P}(X_\infty = i)$ . Plus l'échantillon considéré est grand, plus les proportions observées lors de la réalisation tendent vers la distribution de probabilité qui caractérise cette réalisation. Les proportions d'apparition des nucléotides respectent donc bien la loi des grands nombres.

4

1.2 Q1.1.2 0.95 / 2

- ✓ - 0.3 pts *La présentation des graphes doit être améliorée*
  - ✓ - 0.25 pts *Erreur légère dans la justification/Manquement dans la justification*
  - ✓ - 0.5 pts *Discussion assez pauvre*
- 1 C'est plutôt  $\$P(X_t=i) = P(X_{\{t-1\}}) Q_{\{:, i\}}\$$
  - 2 Cela aurait été mieux de les mettre côté à côté pour voir qu'ils convergent vers les mêmes distributions
  - 3 Quelles conditions la chaîne respecte-t-elle pour être stationnaire?

$$Q^{15} = \begin{pmatrix} 0.28814407 & 0.34967484 & 0.20910455 & 0.15307654 \\ 0.28814407 & 0.34967484 & 0.20910455 & 0.15307654 \\ 0.28814407 & 0.34967484 & 0.20910455 & 0.15307654 \\ 0.28814407 & 0.34967484 & 0.20910455 & 0.15307654 \end{pmatrix}$$

3

La matrice de transition étant conservée entre les états 14 et 15, il y a bien convergence vers une distribution stationnaire, qui n'était pas encore atteinte l'état 5. Ainsi,  $\mathbb{P}(X_t = i) = \mathbb{P}(X_0 = i) \times Q^t$  reste constant pour des valeurs de  $t$  suffisamment grandes.

### 1.1.3 $\mathbb{P}(X_t = \infty)$

La distribution de probabilité en  $t \rightarrow \infty$  étant stationnaire, et ce peu importe la condition initiale, elle peut être déterminée dès l'état  $t$  où se fait la convergence. Ainsi en prenant, par exemple,  $t = 15$  dans  $\mathbb{P}(X_t = i) = \mathbb{P}(X_0 = i) \times Q^t$ .

Il vient :

$$\mathbb{P}(X_\infty = i) = \mathbb{P}(X_{15} = i) = (0.28814407, 0.34967484, 0.20910455, 0.15307654)$$

Par ailleurs, du point de vue théorique, une distribution de probabilité invariante est nécessairement un vecteur propre à gauche de la matrice de transition, celui-ci étant associé à un vecteur unitaire. Ce critère mène au vecteur  $(0.28814407, 0.34967484, 0.20910455, 0.15307654)$  qui correspond bien à la distribution stationnaire déterminée précédemment.

### 1.1.4 Proportion d'apparition des nucléotides pour une longueur T croissante de chaîne

La réalisation est générée de la manière suivante :

En partant d'un nucléotide choisi arbitrairement, une nouvelle base est tirée à partir de la matrice de transition Q jusqu'à avoir atteint une chaîne de longueur T. Pour chaque longueur de la chaîne croissante, la proportion d'apparition de chacune des bases est calculée. Ces proportions et leur évolution sont représentées dans la Figure 4.

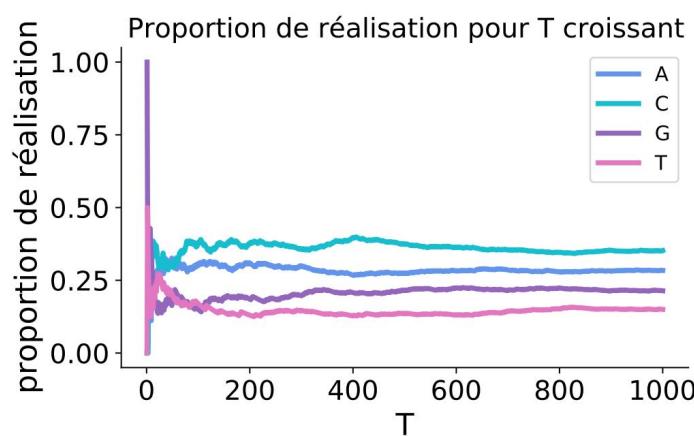


FIGURE 4. Proportion de réalisation des bases pour une longueur T croissante

Il y a convergence vers  $\mathbb{P}(X_\infty = i)$ . Plus l'échantillon considéré est grand, plus les proportions observées lors de la réalisation tendent vers la distribution de probabilité qui caractérise cette réalisation. Les proportions d'apparition des nucléotides respectent donc bien la loi des grands nombres.

4

1.3 Q1.1.3 1 / 1

✓ - 0 pts Correct

$$Q^{15} = \begin{pmatrix} 0.28814407 & 0.34967484 & 0.20910455 & 0.15307654 \\ 0.28814407 & 0.34967484 & 0.20910455 & 0.15307654 \\ 0.28814407 & 0.34967484 & 0.20910455 & 0.15307654 \\ 0.28814407 & 0.34967484 & 0.20910455 & 0.15307654 \end{pmatrix}$$

3

La matrice de transition étant conservée entre les états 14 et 15, il y a bien convergence vers une distribution stationnaire, qui n'était pas encore atteinte l'état 5. Ainsi,  $\mathbb{P}(X_t = i) = \mathbb{P}(X_0 = i) \times Q^t$  reste constant pour des valeurs de  $t$  suffisamment grandes.

### 1.1.3 $\mathbb{P}(X_t = \infty)$

La distribution de probabilité en  $t \rightarrow \infty$  étant stationnaire, et ce peu importe la condition initiale, elle peut être déterminée dès l'état  $t$  où se fait la convergence. Ainsi en prenant, par exemple,  $t = 15$  dans  $\mathbb{P}(X_t = i) = \mathbb{P}(X_0 = i) \times Q^t$ .

Il vient :

$$\mathbb{P}(X_\infty = i) = \mathbb{P}(X_{15} = i) = (0.28814407, 0.34967484, 0.20910455, 0.15307654)$$

Par ailleurs, du point de vue théorique, une distribution de probabilité invariante est nécessairement un vecteur propre à gauche de la matrice de transition, celui-ci étant associé à un vecteur unitaire. Ce critère mène au vecteur  $(0.28814407, 0.34967484, 0.20910455, 0.15307654)$  qui correspond bien à la distribution stationnaire déterminée précédemment.

### 1.1.4 Proportion d'apparition des nucléotides pour une longueur T croissante de chaîne

La réalisation est générée de la manière suivante :

En partant d'un nucléotide choisi arbitrairement, une nouvelle base est tirée à partir de la matrice de transition Q jusqu'à avoir atteint une chaîne de longueur T. Pour chaque longueur de la chaîne croissante, la proportion d'apparition de chacune des bases est calculée. Ces proportions et leur évolution sont représentées dans la Figure 4.

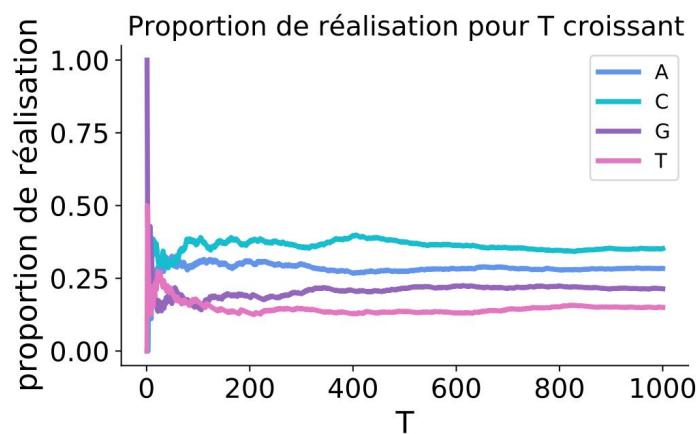


FIGURE 4. Proportion de réalisation des bases pour une longueur T croissante

Il y a convergence vers  $\mathbb{P}(X_\infty = i)$ . Plus l'échantillon considéré est grand, plus les proportions observées lors de la réalisation tendent vers la distribution de probabilité qui caractérise cette réalisation. Les proportions d'apparition des nucléotides respectent donc bien la loi des grands nombres.

4

1.4 Q1.1.4 1.5 / 3

✓ - 1.5 pts *Justification théorique incorrecte*

- 4 C'est parce que la chaîne est irréductible et apériodique (et donc ergodique) qu'elle admet une distribution stationnaire (unique). C'est vers cette distribution que l'on converge avec une séquence suffisamment longue.

Avec une chaîne ergodique, les statistiques temporelles convergent vers les statistiques obtenues sur plusieurs réalisations (ici, la statistique est la fréquence d'apparition).

### 1.1.5 Comparaison à d'autres séquences

#### Fréquences d'apparition

Les fréquences d'apparition des différents nucléotides de chaque séquence peuvent être obtenues en comptant le nombre de chaque nucléotide et en divisant par la longueur de la séquence.

Les fréquences ainsi obtenues sont donc :

fréquences de la séquence 1 = [0.288, 0.3495, 0.209, 0.153]

fréquences de la séquence 2 = [0.306, 0.341, 0.197, 0.1555]

fréquences de la séquence 3 = [0.2845, 0.36, 0.1985, 0.1565]

La fréquence de la séquence 1 correspond à  $\mathbb{P}(X_\infty = i)$ .

Les trois séquences présentent beaucoup de similitudes. Seules les proportions du nucléotide C dans la séquence 3 et du nucléotide A dans la séquence 2 comportent un écart de valeur plus important, l'écart maximal étant relatif à ce second cas.

#### Probabilités

Les probabilités que ces séquences aient été générées par la chaîne de Markov estimée sur base de la première séquence sont calculées au moyen de la formule suivante :

$$P(X_0 = x_i) \times \prod_{i=1}^L P(X_i = x_i | X_{i-1} = x_{i-1})$$

Avec "L" la longueur de la chaîne

Ces probabilités, résultant d'un produit de valeur entre 0 et 1, sont très faibles. Il est donc préférable de les calculer sous forme de logarithme.

$$\begin{aligned} & \log \left( P(X_0 = x_i) \times \prod_{i=1}^L P(X_i = x_i | X_{i-1} = x_{i-1}) \right) \\ &= \log P(X_0 = x_i) + \log \left( \sum_{i=1}^L P(X_i = x_i | X_{i-1} = x_{i-1}) \right) \end{aligned}$$

Ainsi :

$\log_{10}$  probabilités de la séquence 1 = -1054.4897930466145

$\log_{10}$  probabilités de la séquence 2 = -1291.098922487848

$\log_{10}$  probabilités de la séquence 3 = -1042.727684912839

Une fois de plus, la séquence 2 est la plus éloignée du lot. Cela indique qu'elle a une plus faible probabilité d'avoir été générée par la chaîne de Markov.

1.5 Q1.1.5 0.75 / 1

✓ - 0.25 pts *Faute légère/Manquement léger*

- 5 On peut voir également que ce n'est pas parce que les fréquences d'apparition sont proches que les séquences ont été générées par la même matrice de transition

## 1.2 Échantillonnage de Gibbs

### 1.2.1 Équations de balance

Pour que  $\pi_0$  soit invariante, il faut que  $\pi_s = \pi_0 Q$

En appliquant  $\pi_n = \pi_0 Q^n$  pour une composante de  $\pi$ , il en découle :

$$\pi_1(i) = \sum_{j=1}^N \pi_0(j)[Q]_{j,i} \quad \forall (i,j) \in \{1, \dots, N\}^2$$

Si les équations de balance sont respectées, il vient alors :

$$\pi_1(i) = \sum_{j=1}^N \pi_0(i)[Q]_{i,j}$$

$\pi_0(i)$  ne dépendant pas de  $j$ , il peut être sorti de la somme :

$$\pi_1(i) = \pi_0(i) \sum_{j=1}^N [Q]_{i,j}$$

Or, la matrice de transition étant une matrice stochastique, la somme sur une ligne de celle-ci vaut 1. Dès lors, en utilisant la propriété  $\pi_n = \pi_0 Q^n$  :

$$\pi_1 = \pi_0 Q = \pi_0$$

$\pi_0$  est donc une distribution invariante de la chaîne de Markov.

Pour que cette distribution  $\pi_0$  soit unique, il faut que la chaîne soit irréductible.

### 1.2.2 Cas de la matrice de transition de la méthode Gibbs

Montrer que la matrice de transition de la chaîne de Markov définie par la méthode d'échantillonnage de Gibbs satisfait aux équations de balance détaillées pour la distribution invariante  $\mathbb{P}(Y)$  revient à montrer que :

$$\mathbb{P}(y^{(t)}) Q_{y^{(t+1)}, y^{(t)}} = \mathbb{P}(y^{(t+1)}) Q_{y^{(t)}, y^{(t+1)}}$$

Et donc :

$$\mathbb{P}(y^{(t)}) \mathbb{P}(y^{(t+1)}|y^{(t)}) = \mathbb{P}(y^{(t+1)}) \mathbb{P}(y^{(t)}|y^{(t+1)}) \quad (1)$$

où  $\mathbb{P}(y^{(t+1)}|y^{(t)})$  est la probabilité de passer de  $y^{(t)}$  à  $y^{(t+1)}$  en appliquant l'échantillonnage de Gibbs. Ce cas est trivial lorsque  $y^{(t)} = y^{(t+1)}$ .

À l'itération au temps  $t + 1$ , l'algorithme met à jour la composante  $i$  choisie uniformément dans la distribution

$$y_i^{(t+1)} \sim \mathbb{P}\left(Y_i|Y_1 = y_1^{(t)}, \dots, Y_{i-1} = y_{i-1}^{(t)}, Y_{i+1} = y_{i+1}^{(t)}, \dots, Y_N^{(t)}\right)$$

Les notations se simplifient en écrivant :  $y_i^{(t+1)} \sim \mathbb{P}\left(Y_i|Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t)}\right)$

2.1 Q1.2.1: Balance détaillée --> Distribution invariante 2 / 2

✓ - 0 pts *Démonstration correcte*

✓ - 0 pts *Condition pour unicité de la distribution invariante correcte*

## 1.2 Échantillonnage de Gibbs

### 1.2.1 Équations de balance

Pour que  $\pi_0$  soit invariante, il faut que  $\pi_s = \pi_0 Q$

En appliquant  $\pi_n = \pi_0 Q^n$  pour une composante de  $\pi$ , il en découle :

$$\pi_1(i) = \sum_{j=1}^N \pi_0(j)[Q]_{j,i} \quad \forall (i,j) \in \{1, \dots, N\}^2$$

Si les équations de balance sont respectées, il vient alors :

$$\pi_1(i) = \sum_{j=1}^N \pi_0(i)[Q]_{i,j}$$

$\pi_0(i)$  ne dépendant pas de  $j$ , il peut être sorti de la somme :

$$\pi_1(i) = \pi_0(i) \sum_{j=1}^N [Q]_{i,j}$$

Or, la matrice de transition étant une matrice stochastique, la somme sur une ligne de celle-ci vaut 1. Dès lors, en utilisant la propriété  $\pi_n = \pi_0 Q^n$  :

$$\pi_1 = \pi_0 Q = \pi_0$$

$\pi_0$  est donc une distribution invariante de la chaîne de Markov.

Pour que cette distribution  $\pi_0$  soit unique, il faut que la chaîne soit irréductible.

### 1.2.2 Cas de la matrice de transition de la méthode Gibbs

Montrer que la matrice de transition de la chaîne de Markov définie par la méthode d'échantillonnage de Gibbs satisfait aux équations de balance détaillées pour la distribution invariante  $\mathbb{P}(Y)$  revient à montrer que :

$$\mathbb{P}(y^{(t)}) Q_{y^{(t+1)}, y^{(t)}} = \mathbb{P}(y^{(t+1)}) Q_{y^{(t)}, y^{(t+1)}}$$

Et donc :

$$\mathbb{P}(y^{(t)}) \mathbb{P}(y^{(t+1)}|y^{(t)}) = \mathbb{P}(y^{(t+1)}) \mathbb{P}(y^{(t)}|y^{(t+1)}) \quad (1)$$

où  $\mathbb{P}(y^{(t+1)}|y^{(t)})$  est la probabilité de passer de  $y^{(t)}$  à  $y^{(t+1)}$  en appliquant l'échantillonnage de Gibbs. Ce cas est trivial lorsque  $y^{(t)} = y^{(t+1)}$ .

À l'itération au temps  $t + 1$ , l'algorithme met à jour la composante  $i$  choisie uniformément dans la distribution

$$y_i^{(t+1)} \sim \mathbb{P}\left(Y_i|Y_1 = y_1^{(t)}, \dots, Y_{i-1} = y_{i-1}^{(t)}, Y_{i+1} = y_{i+1}^{(t)}, \dots, Y_N^{(t)}\right)$$

Les notations se simplifient en écrivant :  $y_i^{(t+1)} \sim \mathbb{P}\left(Y_i|Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t)}\right)$

En introduisant  $y_i$  dans les termes de l'égalité (1), il vient :

$$Q_{y^{(t+1)}, y^{(t)}} = \mathbb{P}(y^{(t+1)} | y^{(t)}) = \frac{1}{N-1} \times \mathbb{P}\left(Y_i = y_i^{(t+1)} | Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t)}\right)$$

Multiplier par  $\mathbb{P}(y^{(t)})$ , permet d'exprimer  $\mathbb{P}(y^{(t+1)} | y^{(t)}) \mathbb{P}(y^{(t)})$  comme :

$$\begin{aligned} \mathbb{P}(y^{(t+1)} | y^{(t)}) \mathbb{P}(y^{(t)}) &= \frac{1}{N-1} \times \mathbb{P}\left(Y_i = y_i^{(t+1)} | Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t)}\right) \times \mathbb{P}(y^{(t)}) \\ &= \frac{1}{N-1} \times \mathbb{P}\left(Y_i = y_i^{(t+1)} | Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t)}\right) \\ &\quad \times \mathbb{P}\left(Y_i = y_i^{(t)} | Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t)}\right) \times \mathbb{P}\left(Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t)}\right) \end{aligned}$$

En suivant un raisonnement semblable, il vient :

$$\begin{aligned} \mathbb{P}(y^{(t)} | y_i^{(t+1)}) \mathbb{P}(y^{(t+1)}) &= \frac{1}{N-1} \times \mathbb{P}\left(Y_i = y_i^{(t)} | Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t+1)}\right) \times \mathbb{P}(y^{(t+1)}) \\ &= \frac{1}{N-1} \times \mathbb{P}\left(Y_i = y_i^{(t)} | Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t+1)}\right) \\ &\quad \times \mathbb{P}\left(Y_i = y_i^{(t+1)} | Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t+1)}\right) \times \mathbb{P}\left(Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t+1)}\right) \end{aligned}$$

Dans l'algorithme, les composantes autres que celles correspondantes à  $i$  sont inchangées. Ainsi,  $y_x^{(t+1)} = y_x^{(t)}$   $\forall x \in \{1, \dots, N\} \setminus i$ , il en découle alors :

$$\begin{aligned} \mathbb{P}(y^{(t+1)} | y^{(t)}) \mathbb{P}(y^{(t)}) &= \frac{1}{N-1} \times \mathbb{P}\left(Y_i = y_i^{(t+1)} | Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t)}\right) \\ &\quad \times \mathbb{P}\left(Y_i = y_i^{(t)} | Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t)}\right) \quad \times \mathbb{P}\left(Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t+1)}\right) \end{aligned}$$

Et,

$$\begin{aligned} \mathbb{P}(y^{(t)} | y^{(t+1)}) \mathbb{P}(y^{(t+1)}) &= \frac{1}{N-1} \times \mathbb{P}\left(Y_i = y_i^{(t)} | Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t)}\right) \\ &\quad \times \mathbb{P}\left(Y_i = y_i^{(t+1)} | Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t)}\right) \quad \times \mathbb{P}\left(Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t+1)}\right) \end{aligned}$$

Et donc,

$$\mathbb{P}(y^{(t+1)} | y^{(t)}) \mathbb{P}(y^{(t)}) = \mathbb{P}(y^{(t)} | y^{(t+1)}) \mathbb{P}(y^{(t+1)})$$

Pour garantir une convergence de la chaîne vers la distribution stationnaire unique, il faut que celle-ci soit ergodique, donc irréductible et apériodique.

### 1.2.3 Distribution conjointe

- (a) Pour pouvoir appliquer l'algorithme, il faut que les conditions citées dans le point précédent soient respectées.

La période étant le plus grand commun diviseur de l'ensemble des nombres  $n \in \mathbb{N}_0$  tel que  $Q_Y^n(i, i) \neq 0$ . Pour être apériodique, il faut que la chaîne de Markov considérée aie une période de 1, donc que  $Q_Y(i, i) \neq 0$ .

De plus, pour s'assurer d'avoir une chaîne irréductible, une solution consiste à simplement éviter les valeurs nulles dans  $Q_Y$ .

Tous les éléments de  $Q_Y$  doivent sommer à 1 et

La somme des probabilités se trouvant dans  $Q_Y$  doit être unitaire.

## 2.2 Q1.2.2: Gibbs --> Balance détaillée 2.4 / 3

- ✓ - 0 pts Cas où on reste dans le même état: correct
  - ✓ - 0.1 pts Cas où 2+ variables changent: incorrect ou manquant
  - ✓ - 0.5 pts Cas où une seule variable change: démonstration globalement correcte mais comportant une/des petite.s erreur.s.
  - ✓ - 0 pts Conditions pour que l'algorithme de Gibbs fonctionne: réponse complète et correcte (=dire que la chaîne doit être apériodique et irréductible).
- 6 La probabilité de choisir une variable parmi  $\$N\$$  variables selon une distribution uniforme est égale à  $\$frac{1}{N}$  et non pas  $\$frac{1}{N-1}$ .

En introduisant  $y_i$  dans les termes de l'égalité (1), il vient :

$$Q_{y^{(t+1)}, y^{(t)}} = \mathbb{P}(y^{(t+1)} | y^{(t)}) = \frac{1}{N-1} \times \mathbb{P}\left(Y_i = y_i^{(t+1)} | Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t)}\right)$$

Multiplier par  $\mathbb{P}(y^{(t)})$ , permet d'exprimer  $\mathbb{P}(y^{(t+1)} | y^{(t)}) \mathbb{P}(y^{(t)})$  comme :

$$\begin{aligned} \mathbb{P}(y^{(t+1)} | y^{(t)}) \mathbb{P}(y^{(t)}) &= \frac{1}{N-1} \times \mathbb{P}\left(Y_i = y_i^{(t+1)} | Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t)}\right) \times \mathbb{P}(y^{(t)}) \\ &= \frac{1}{N-1} \times \mathbb{P}\left(Y_i = y_i^{(t+1)} | Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t)}\right) \\ &\quad \times \mathbb{P}\left(Y_i = y_i^{(t)} | Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t)}\right) \times \mathbb{P}\left(Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t)}\right) \end{aligned}$$

En suivant un raisonnement semblable, il vient :

$$\begin{aligned} \mathbb{P}(y^{(t)} | y_i^{(t+1)}) \mathbb{P}(y^{(t+1)}) &= \frac{1}{N-1} \times \mathbb{P}\left(Y_i = y_i^{(t)} | Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t+1)}\right) \times \mathbb{P}(y^{(t+1)}) \\ &= \frac{1}{N-1} \times \mathbb{P}\left(Y_i = y_i^{(t)} | Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t+1)}\right) \\ &\quad \times \mathbb{P}\left(Y_i = y_i^{(t+1)} | Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t+1)}\right) \times \mathbb{P}\left(Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t+1)}\right) \end{aligned}$$

Dans l'algorithme, les composantes autres que celles correspondantes à  $i$  sont inchangées. Ainsi,  $y_x^{(t+1)} = y_x^{(t)}$   $\forall x \in \{1, \dots, N\} \setminus i$ , il en découle alors :

$$\begin{aligned} \mathbb{P}(y^{(t+1)} | y^{(t)}) \mathbb{P}(y^{(t)}) &= \frac{1}{N-1} \times \mathbb{P}\left(Y_i = y_i^{(t+1)} | Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t)}\right) \\ &\quad \times \mathbb{P}\left(Y_i = y_i^{(t)} | Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t)}\right) \quad \times \mathbb{P}\left(Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t+1)}\right) \end{aligned}$$

Et,

$$\begin{aligned} \mathbb{P}(y^{(t)} | y^{(t+1)}) \mathbb{P}(y^{(t+1)}) &= \frac{1}{N-1} \times \mathbb{P}\left(Y_i = y_i^{(t)} | Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t)}\right) \\ &\quad \times \mathbb{P}\left(Y_i = y_i^{(t+1)} | Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t)}\right) \quad \times \mathbb{P}\left(Y_{\{1, \dots, N\} \setminus i} = y_{\{1, \dots, N\} \setminus i}^{(t+1)}\right) \end{aligned}$$

Et donc,

$$\mathbb{P}(y^{(t+1)} | y^{(t)}) \mathbb{P}(y^{(t)}) = \mathbb{P}(y^{(t)} | y^{(t+1)}) \mathbb{P}(y^{(t+1)})$$

Pour garantir une convergence de la chaîne vers la distribution stationnaire unique, il faut que celle-ci soit ergodique, donc irréductible et apériodique.

### 1.2.3 Distribution conjointe

- (a) Pour pouvoir appliquer l'algorithme, il faut que les conditions citées dans le point précédent soient respectées.

La période étant le plus grand commun diviseur de l'ensemble des nombres  $n \in \mathbb{N}_0$  tel que  $Q_Y^n(i, i) \neq 0$ . Pour être apériodique, il faut que la chaîne de Markov considérée aie une période de 1, donc que  $Q_Y(i, i) \neq 0$ .

De plus, pour s'assurer d'avoir une chaîne irréductible, une solution consiste à simplement éviter les valeurs nulles dans  $Q_Y$ .

Tous les éléments de  $Q_Y$  doivent sommer à 1 et

La somme des probabilités se trouvant dans  $Q_Y$  doit être unitaire.

$$Q_Y = \begin{pmatrix} 0.1 & 0.05 & 0.02 & 0.08 \\ 0.04 & 0.11 & 0.07 & 0.03 \\ 0.07 & 0.05 & 0.04 & 0.09 \\ 0.05 & 0.06 & 0.08 & 0.06 \end{pmatrix}$$

Par simplicité, plutôt que de s'assurer de la convergence, nous ferons beaucoup d'itérations de l'algorithme et ensuite nous vérifierons si il y a eu convergence.

- (b) Les fréquences d'apparition de chaque état convergent vers les probabilités définies par la matrice de référence  $Q_Y$  avec la longueur de la chaîne de Markov, comme prédit par la loi des grands nombres. La méthode d'échantillonnage de Gibbs donne effectivement une meilleure approximation de  $Q_Y$  pour une chaîne de Markov de longueur 5000 que 100.

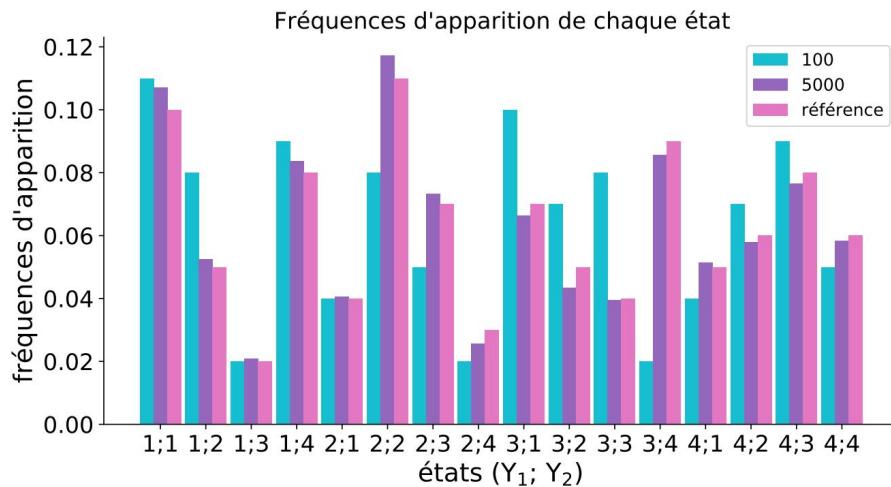


FIGURE 5. Fréquences d'apparition avec sélection uniforme

- (c) Les résultats diffèrent légèrement du cas précédent par leur vitesse de convergence. L'alternance entre la modification de  $Y_1$  et de  $Y_2$  permet en effet d'atteindre plus rapidement  $Q_y$ , l'écart entre les probabilités définies par celle-ci et celles de la réalisation étant globalement plus réduit que dans le cas précédent.

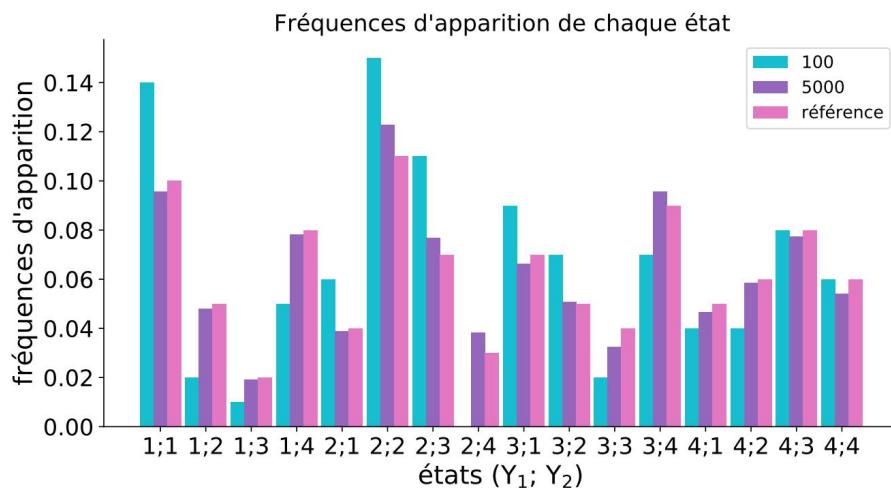


FIGURE 6. Fréquences d'apparition avec sélection alternée

### 2.3 Q1.2.3: Gibbs avec 2 variables 4 / 5

- ✓ - 0 pts Choix de la distribution conjointe: justification correcte (= chaîne irréductible + la somme de tous les éléments de la matrice doit valoir 1)
- ✓ - 1 pts Calcul des distributions conditionnelles: incorrect ou manquant
- ✓ - 0 pts Explications concernant l'implémentation de l'algorithme de Gibbs (nombre d'itérations, etc.): ok
- ✓ - 0 pts Comparaison entre les fréquences d'apparition des états et la matrice  $Q_Y$ : analyse ok
- ✓ - 0 pts Alternance de  $Y_1$  et  $Y_2$ : ok

## 2 Deuxième partie : détection de motifs dans des séquences d'ADN

### 2.1 Dérivations mathématiques

Développement menant à l'expression  $\mathbb{P}(A_k|A_{-k}, S)$

Pour calculer  $\mathbb{P}(S, A|\Theta, \Phi)$ , nous avons tout d'abord appliqué la formule des probabilités composées (chain rule). Ensuite,  $\mathbb{P}(A)$  a été obtenu à partir de  $\mathbb{P}(A|\Theta, \Phi)$  en remarquant que  $A$  était indépendant de  $\Phi$  et  $\Theta$ . Cette probabilité peut ensuite être exprimée en sachant qu'une position  $a_k$  est tirée de façon uniforme dans  $\{1, \dots, |S_k| - W + 1\}$ . Cette probabilité doit être exprimée sous la forme d'un produit sur  $N$  afin de considérer tous les  $a_k$ . Afin d'identifier des termes indépendants,  $S$  a été décomposé en trois parties différentes. Ces parties correspondent à la séquence avant le motif générée selon  $\Phi$ , la séquence du motif générée selon  $\Theta$  et la séquence après le motif générée selon  $\Phi$ . Décomposer  $S$  permet d'avoir des termes indépendants dans les probabilités, en effet, par exemple si une partie de la séquence est générée selon  $\Phi$  elle ne dépend pas de  $\Theta$ .  $S$  et  $A$  sont aussi indépendants.<sup>7</sup> Ensuite, l'expression peut être exprimée en prenant en compte les différentes distributions. Le motif de taille  $W$  est tiré d'une distribution multinomiale de paramètre  $\theta_{i,j}$ . Dans le cas d'un modèle de Markov d'ordre 0, les nucléotides en dehors du motif sont indépendants les uns des autres et ceux-ci sont tirés d'une distribution multinomiale définie par  $K$  paramètres  $\Phi$ .

$$\begin{aligned}
 \mathbb{P}(S, A|\Theta, \Phi) &= \mathbb{P}(A|\Theta, \Phi) \cdot \mathbb{P}(S|A, \Theta, \Phi) \\
 &= \mathbb{P}(A) \cdot \mathbb{P}(S|A, \Theta, \Phi) \\
 &= \left( \prod_{k=1}^N \frac{1}{|S_k| - W + 1} \right) \cdot \mathbb{P}(1, \dots, a_{k-1}, a_k, \dots, a_k + W - 1, a_k + W, \dots, |S_k| |A, \Theta, \Phi) \\
 &= \left( \prod_{k=1}^N \frac{1}{|S_k| - W + 1} \right) \cdot \mathbb{P}(1, \dots, a_{k-1} | \Phi) \cdot \mathbb{P}(a_k, \dots, a_k + W - 1 | \Theta) \cdot \mathbb{P}(a_k + W, \dots, |S_k| | \Phi) \\
 &= \left( \prod_{k=1}^N \frac{1}{|S_k| - W + 1} \right) \cdot \frac{M!}{M_1! M_2! M_3! M_4!} \prod_{i=1}^K \prod_{j=1}^W \theta_{ij}^{n_{ij}} \phi_i^{\tilde{n}_{i0}}
 \end{aligned}$$

où  $M$  est le nombre total de nucléotides,  $M_1$  le nombre total de A,  $M_2$  le nombre total de C,  $M_3$  le nombre total de G et  $M_4$  le nombre total de T.

Dans l'expression ci-dessus,  $n_{ij}$  correspond au nombre total de  $i$  dans la position  $j$  des motifs et  $\tilde{n}_{i0}$  correspond au nombre de fois que  $i$  apparaît dans toutes les séquences en dehors du motif.

L'expression obtenue devient :

$$\mathbb{P}(S, A|\Theta, \Phi) \propto \prod_{i=1}^K \prod_{j=1}^W \theta_{ij}^{n_{ij}} \phi_i^{\tilde{n}_{i0}}$$

où les constantes ont disparu grâce à la proportionnalité.

Ensuite, il faut calculer  $\mathbb{P}(A_k|A_{-k}, S)$ . Pour cela, il faut commencer par exprimer  $P(A|S)$  qui par la loi de Bayes est proportionnel à  $P(S, A)$

$$\mathbb{P}(A|S) = \frac{\mathbb{P}(S, A)}{\mathbb{P}(S)} \propto \mathbb{P}(S, A)$$

$\mathbb{P}(S, A)$  peut être exprimé comme :

$$\mathbb{P}(S, A) \propto \iint \mathbb{P}(S, A | \Theta, \Phi) p(\Theta) p(\Phi) d\Phi d\Theta$$

où  $p(\Phi)$  suit une distribution de Dirichlet  $\text{Dir}(\alpha)$  où  $\alpha = (\alpha_1, \dots, \alpha_K)$  et où  $p(\Theta)$  suit une distribution du produit de Dirichlet  $\text{PD}(\mathbf{B})$  avec  $\mathbf{B} = (\beta_1, \dots, \beta_W)$  et  $\beta_j = (\beta_{1j}, \dots, \beta_{Kj})^T$ .

En exprimant  $\mathbb{P}(S, A | \Theta, \Phi)$  par l'expression trouvée ci-dessus et en exprimant les propriétés de la distribution de Dirichlet, il vient :

$$\begin{aligned} \mathbb{P}(S, A) &\propto \iint \mathbb{P}(S, A | \Theta, \Phi) p(\Theta) p(\Phi) d\Phi d\Theta \\ &\propto \iint \prod_{i=1}^K \prod_{j=1}^W \theta_{ij}^{n_{ij}} \phi_i^{\tilde{n}_{i0}} p(\Theta) p(\Phi) d\Phi d\Theta \\ &\propto \prod_{i=1}^K \prod_{j=1}^W \left[ \Gamma(\tilde{n}_{i0} + \alpha_i) \cdot \Gamma(n_{ij} + \beta_{ij}) \right] \end{aligned}$$
11

Les termes  $\alpha_i$  et  $\beta_{ij}$  sont introduits afin d'éviter d'avoir des probabilités nulles.<sup>10</sup> En effet, dans le cas de séquences de petites tailles, si un nucléotide n'apparaît pas au départ, le comptage relatif sera nul et la probabilité qui en découle aussi. Il n'y aura donc aucune chance de le tirer. Cela biaiserait la convergence de l'algorithme vers une séquence qui ne contiendrait jamais un nucléotide dû à la non-apparition dans l'échantillon initial.

$n_{ij}$  et  $\tilde{n}_{i0}$  peuvent être exprimés comme :

$$\tilde{n}_{i0} = \tilde{n}_{i0}^{-k} - \tilde{n}_{i0}^k \quad \text{et} \quad n_{ij} = n_{ij}^{-k} + j_{ij}^k$$

Où

- $\tilde{n}_{i0}^{-k}$  = nombre total de nucléotide  $i$  dans la région hors motif sauf le  $k^{i\text{ème}}$ <sup>12</sup>
- $\tilde{n}_{i0}^k$  = nombre total de nucléotide  $i$  dans la séquence  $k$ <sup>13</sup>
- $n_{ij}^{-k}$  = nombre de  $i$  dans la position  $j$  de tous les motifs sauf le  $k^{i\text{ème}}$
- $j_{ij}^k$  = 1 si la base à la position  $j$  du motif  $k$  est  $i$

Ensuite, pour trouver l'expression de  $\mathbb{P}(A_k | A_{-k}, S)$ , il faut utiliser le fait que ce terme est proportionnel à  $\mathbb{P}(A | S)$  et considérer les fonctions  $A_{-k}$  comme des constantes. En exprimant  $n_{ij}$  et  $\tilde{n}_{i0}$  par leur expression et en divisant l'expression obtenue ci-dessus pour  $\mathbb{P}(S, A)$  par deux constantes  $\Gamma(\tilde{n}_{i0}^{-k} + \alpha_i)$  et  $\Gamma(n_{ij}^{-k} + \beta_{ij})$ , il vient :

$$\begin{aligned} \mathbb{P}(A_k | A_{-k}, S) &\propto \prod_{i=1}^K \prod_{j=1}^W \left[ \frac{\Gamma(\tilde{n}_{i0}^{-k} - \tilde{n}_{i0}^k + \alpha_i) \Gamma(n_{ij}^{-k} + j_{ij}^k + \beta_{ij})}{\Gamma(\tilde{n}_{i0}^{-k} + \alpha_i) \Gamma(n_{ij}^{-k} + \beta_{ij})} \right] \\ &\propto \prod_{i=1}^K \prod_{j=1}^W \frac{\Gamma(\tilde{n}_{i0}^{-k} - \tilde{n}_{i0}^k + \alpha_i)}{\Gamma(\tilde{n}_{i0}^{-k} + \alpha_i)} (n_{ij}^{-k} + \beta_{ij})^{j_{ij}^k} \end{aligned}$$

Où le terme  $(n_{ij}^{-k} + \beta_{ij})^{j_{ij}^k}$  a été obtenu en utilisant la propriété de la distribution de Dirichlet statuant :

$$\frac{\Gamma(n_{ij}^{-k} + j_{ij}^k + \beta_{ij})}{\Gamma(n_{ij}^{-k} + \beta_{ij})} = (n_{ij}^{-k} + \beta_{ij})^{j_{ij}^k}$$
14

L'expression de  $\mathbb{P}(A_k|A_{-k}, S)$  peut encore être simplifiée. En effet,  $\tilde{n}_{i0}^k$  et  $\tilde{n}_{i0}^{-k}$  étant de taille différente, l'expression devient :

$$\frac{\Gamma(\tilde{n}_{i0}^{-k} - \tilde{n}_{i0}^k + \alpha_i)}{\Gamma(\tilde{n}_{i0}^{-k} + \alpha_i)} \approx \frac{1}{\tilde{n}_{i0}^k} \quad \text{16}$$

Il vient donc :

$$\mathbb{P}(A_k|A_{-k}, S) \propto \prod_{i=1}^K \prod_{j=1}^W \frac{(n_{ij}^{-k} + \beta_{ij})^{j_{ij}^k}}{\tilde{n}_{i0}^k} \quad \text{18}$$

Qui peut être réécrit comme :

$$\mathbb{P}(A_k|A_{-k}, S) \propto \prod_{i=1}^K \prod_{j=1}^W \frac{(n_{ij}^{-k} + \beta_{ij})^{j_{ij}^k}}{(\tilde{n}_{i0}^{-k} + \alpha_i)^{j_{ij}^k}} \quad \text{17}$$

En utilisant la propriété de Dirichlet disant que :

$$\mathbb{E}(\theta_{ij}|n_{ij}^{-k}) \propto n_{ij}^{-k} + \beta_{ij} \quad \text{et} \quad \mathbb{E}(\phi_i|\tilde{n}_{i0}^{-k}) \propto \tilde{n}_{i0}^{-k} + \alpha_i$$

la probabilité finale est :

$$\mathbb{P}(A_k|A_{-k}, S) \propto \prod_{i=1}^K \prod_{j=1}^W \left( \frac{\mathbb{E}(\theta_{ij}|n_{ij}^{-k})}{\mathbb{E}(\phi_i|\tilde{n}_{i0}^{-k})} \right)^{j_{ij}^k}$$

Où

$$\mathbb{E}(\theta_{ij}|n_{ij}^{-k}) = \frac{n_{ij}^{-k} + \beta_{ij}}{\sum_{l=1}^K n_{lj}^{-k} + \sum_{l=1}^K \beta_{lj}} = \frac{n_{ij}^{-k} + \beta_{ij}}{N - 1 + \sum_{l=1}^K \beta_{lj}}$$

et

$$\mathbb{E}(\phi_i|\tilde{n}_{i0}^{-k}) = \frac{\tilde{n}_{i0}^{-k} + \alpha_i}{\sum_{l=1}^K \tilde{n}_{l0}^{-k} + \sum_{l=1}^K \alpha_l} = \frac{\tilde{n}_{i0}^{-k} + \alpha_i}{|S| - (N - 1) W + \sum_{l=1}^K \alpha_l}$$

### Estimation des paramètres $\Theta$ et $\Phi$

L'estimation des paramètres  $\Theta$  du modèle de motif, lorsque les positions A des motifs sont connues, peut être faite à partir des comptages des nucléotides qui apparaissent à chaque position dans les motifs. Il en est de même pour  $\Phi$  en considérant les nucléotides qui apparaissent à chaque position en dehors des motifs.

Ainsi :

$$\theta_{ij} = \mathbb{E}(\theta_{ij}|n_{ij}) = \frac{n_{ij} + \beta_{ij}}{\sum_{l=1}^K n_{lj} + \sum_{l=1}^K \beta_{lj}} = \frac{n_{ij} + \beta_{ij}}{N + \sum_{l=1}^K \beta_{lj}}$$

et

$$\phi_i = \mathbb{E}(\phi_i|\tilde{n}_{i0}) = \frac{\tilde{n}_{i0} + \alpha_i}{\sum_{l=1}^K \tilde{n}_{l0} + \sum_{l=1}^K \alpha_l} = \frac{\tilde{n}_{i0} + \alpha_i}{|S| - N W + \sum_{l=1}^K \alpha_l}$$

## Généralisation aux ordres supérieurs

Pour généraliser à l'ordre supérieur, il faut adapter les probabilités définissant les nucléotides hors motif, soit le numerateur  $\mathbb{E}(\phi_i|\tilde{n}_{i0}^{-k})$  dans l'expression de  $\mathbb{P}(A_k|A_{-k}, S)$ . Un modèle de Markov d'ordre supérieur fait intervenir, dans la distribution de probabilité, les transitions entre les nucléotides. La distribution  $\Phi = (\phi_1, \dots, \phi_K) \in \mathbb{R}^K$  relative aux nucléotides hors motif pour l'ordre 0 devient alors, à l'ordre  $n$  :

$$\Phi = (\phi_1, \dots, \phi_K) \in \mathbb{R}^{K \times K^n} \quad \text{où} \quad \phi_i = (\phi_1, \dots, \phi_K)^T \in \mathbb{R}^{K \times K^{n-1}}$$

Pour l'ordre 1, la distribution de probabilité des nucléotides hors motif correspond donc à une matrice  $K \times K$  et il vient alors :

$$\mathbb{P}(A_k|A_{-k}, S) \propto \prod_{i=1}^K \prod_{v=1}^K \prod_{j=1}^W \left( \frac{\mathbb{E}(\theta_{ij}|n_{ij}^{-k})}{\mathbb{E}(\phi_{iv}|\tilde{n}_{iv}^{-k})} \right)^{j_{ij}^k} \quad 20$$

Où :

$$\mathbb{E}(\theta_{ij}|n_{ij}^{-k}) = \frac{n_{ij}^{-k} + \beta_{ij}}{\sum_{l=1}^K n_{lj}^{-k} + \sum_{l=1}^K \beta_{lj}} \quad \text{et} \quad \mathbb{E}(\phi_{iv}|\tilde{n}_{iv}^{-k}) = \frac{\tilde{n}_{iv}^{-k} + \alpha_{iv}}{\sum_{l=1}^K \tilde{n}_{lv}^{-k} + \sum_{l=1}^K \alpha_{lv}}$$

## 2.2 Implémentation

L'algorithme de détection de motifs a été élaboré en suivant l'échantillonnage de Gibbs. Initialement, aucune information à propos des positions de motifs n'est connue. L'algorithme commence donc avec des positions  $a_k$  de motifs tirés uniformément selon celles possibles, soit entre 1 et  $|S_l| - W + 1$  où  $S_l$  est la longueur de la séquence  $l$  et  $W$  la longueur du motif. Ensuite, l'algorithme crée une copie des positions trouvées afin de conserver tous les vecteurs A trouvés à chaque itération. En effet, ce n'est pas uniquement la dernière valeur trouvée qui nous intéresse. Ensuite, il faut choisir une séquence pour laquelle la probabilité des positions du motif sera calculée. La position du nouveau motif de la séquence est donc tiré selon la probabilité conditionnelle  $\mathbb{P}(A_k|A_{-k}, S)$ . Juste avant cette étape peut intervenir une étape de décalage. Enfin, l'algorithme est censé se répéter jusqu'à l'obtention d'une convergence vers les positions des motifs. Cela étant difficile en pratique, notre approche a plutôt consisté en un choix d'un nombre d'itérations maximales.

Pour le choix de la séquence dont le motif va être modifié, nous avons opté pour un balayage systématique de la variable  $i$ , car, comme déterminé dans la partie 1.2.3, l'algorithme converge plus rapidement. Toutefois, un balayage non systématique, en tirant le variable  $i$  uniformément, aurait également fonctionné. Cela aurait simplement demander plus d'itérations pour converger.

La valeur de  $\hat{A}$  retournée par notre algorithme est la valeur de A trouvée ayant le meilleur score. C'est donc pour cela qu'il fallait conserver les différents vecteurs A. Le score que nous avons choisi de privilégié est la distance de Kullback-Leiber. En effet, cette dernière mesure à quel point les paramètres  $\Theta$  et  $\Phi$  sont différents. Cela permet donc de quantifier la proximité du motif trouvé avec le motif réel.

En ce qui concerne la valeur de  $\hat{\Theta}$ , la formule établie au point 2.1 peut être utilisée. Celle-ci correspond à l'estimation des paramètres  $\Theta$  pour  $\hat{A}$ . Pour l'implémentation du score selon la distance de Kullback-Leiber,  $\Phi$  était aussi nécessaire. Il a aussi été trouvé grâce au développement de la même section.

Pour permettre à l'algorithme de converger sans rester bloqué dans une forme décalée du modèle optimal, il convient de faire subir aux motifs identifiés un décalage de quelques nucléotides toutes les  $M$  itérations. Remarquons cependant qu'il n'est pas toujours possible de procéder au décalage. En effet, les positions du motif sont limitées aux valeurs entre 1 et  $|S_l| - W + 1$  où  $S_l$  est la longueur de la séquence  $l$  et  $W$  la longueur du motif.

Une méthode permettant de déterminer la nécessité d'un décalage des positions des motifs consiste à comparer  $\mathbb{P}(A'|S)$  et  $\mathbb{P}(A|S)$  où  $A$  désigne les positions des motifs avant le décalage et où  $A'$  désigne la position des motifs après le décalage.

Ces probabilités s'exprimant selon des termes discrets, il convient de faire leur comparaison sous la forme d'un rapport. De fait, la fonction gamma s'exprime comme une factorielle pour les éléments discrets. Or, les comptages qui apparaissent dans l'argument des fonctions gammes de  $\mathbb{P}(A'|S)$  et  $\mathbb{P}(A|S)$  sont très proches, le rapport de  $\mathbb{P}(A'|S)$  et  $\mathbb{P}(A|S)$  peut donc se simplifier en un produit de quelques facteurs plus simple et efficace à calculer.

$$\frac{\mathbb{P}(A'|S)}{\mathbb{P}(A|S)} = \frac{\prod_{i=1}^K \Gamma(\tilde{n}_{i0}' + \alpha_i) \prod_{j=1}^W \Gamma(n_{ij}' + \beta_{ij})}{\prod_{i=1}^K \Gamma(\tilde{n}_{i0} + \alpha_i) \prod_{j=1}^W \Gamma(n_{ij} + \beta_{ij})}$$

Dès lors, si le rapport est plus grand que 1, le décalage est avantageux.

De plus, afin de sortir d'un optimum local, il est intéressant de permettre à l'algorithme de procéder au décalage dans certains cas où le rapport est inférieur à 1. Pour implémenter cela, nous faisons appel à une probabilité proportionnelle au rapport sous la forme d'un nombre aléatoire  $\in [0, 1]$ . Si celui-ci est inférieur au rapport, nous procédons au décalage. 21

Cependant, notre implémentation n'a pas mené à un code de décalage fonctionnel, l'algorithme générant des résultats erronés après l'ajout de celui-ci. Pour palier à ce problème, nous avons écrit un code de décalage simplifié n'intervenant qu'après la convergence de l'algorithme. Une fois l'optimum déterminé, nous procédons à un décalage d'un nucléotide vers la gauche et la droite afin de vérifier que les scores de ces motifs décalés soient bien inférieurs à celui initialement sélectionné. Dans le cas contraire, le motif décalé de plus haut score est sélectionné.

---

**Algorithm 1** Échantillonnage de Gibbs.

```
function ALGOGIBBS(S, T)          Où S sont les N séquences, T le nombre d'itération
    Choisir des valeurs pour commencer :
    for i = 1 to N do
         $A_i^{(0)} \sim U\{1, \dots, |S_i| - W + 1\}$ 
         $A^{(0)} = (A_1^{(0)}, \dots, A_N^{(0)})$ 
        Remplir  $\Theta$  avec  $\theta_{ij} = \mathbb{E}(\theta_{ij}|n_{ij}) \quad \forall i \in \{1, \dots, K\}$  et  $\forall j \in \{1, \dots, W\}$ 
        Remplir  $\Phi$  avec  $\phi_i = \mathbb{E}(\phi_i|\tilde{n}_{i0}) \quad \forall i \in \{1, \dots, K\}$ 
        Calcul du scoreKL( $\Theta, \Phi$ )
    for t = 0 to T do
         $A^{(t+1)} = A^{(t)}$ 
        i = (i + 1) % N                                 $i \sim U\{1, \dots, N\}$  aurait été possible
        // endroit où l'étape du décalage aurait dû avoir lieu toutes les M itérations
         $A_i^{(t+1)} \sim \mathbb{P}(A_i|A_{-k}, S)$ 
        Remplir  $\Theta$  avec  $\theta_{ij} = \mathbb{E}(\theta_{ij}|n_{ij}) \quad \forall i \in \{1, \dots, K\}$  et  $\forall j \in \{1, \dots, W\}$ 
        Remplir  $\Phi$  avec  $\phi_i = \mathbb{E}(\phi_i|\tilde{n}_{i0}) \quad \forall i \in \{1, \dots, K\}$ 
        Calcul du scoreKL( $\Theta, \Phi$ )
    tMax = indice de la composante de A ayant le plus grand scoreKL
    return  $A^{(tMax)}, \Theta^{(tMax)}$ 
```

---

**Algorithm 2** Code du décalage

```
for s in {-1, 1} do :
    Décaler de s le contenu de  $A^{(t+1)}$  en conservant  $A'_l \in \{1, \dots, |S_l| - W + 1\} \quad \forall l \in \{1, \dots, N\}$ 
    Si cette condition ne peut pas être remplie, passé à l'itération de s suivante
    if  $\frac{P(A'|S)}{P(A^{(t+1)}|S)} > 1$  then
         $A_{choisi} = A'$ 
    else
        aléatoire  $\sim U\{0, 1\}$ 
        if aléatoire  $< \frac{P(A'|S)}{P(A^{(t+1)}|S)}$  then
             $A_{choisi} = A'$ 
     $A = A_{choisi}$ 
```

---

## 2.3 Expériences et résultats

### 2.3.1 Obtenez-vous les mêmes résultats quand vous relancez l'algorithme ?

Les résultats de l'algorithme sont semblables d'un test à l'autre lorsque les paramètres sont les mêmes. En effet, en relançant plusieurs fois l'algorithme pour les mêmes paramètres de départ, celui-ci permet de converger vers les mêmes positions de motif pour la majorité des séquences étudiées. Les différences n'apparaissent que pour les séquences pour lesquelles le code n'a pas convergé vers la bonne position  $a_k$ .

### **2.3.2 Quel impact ont les paramètres de l'algorithme (nombre d'itérations, gestion du décalage, degré du modèle de séquence hors motif, paramètres des distributions a priori, etc.) sur ses performances ?**

Le nombre d'itérations permet d'améliorer les chances de trouver les bonnes positions de motifs. Celui-ci a donc un impact majeur dans l'algorithme. Lorsqu'il y a trop peu d'itérations, la probabilité que l'algorithme converge vers le résultat optimal est très faible, cette probabilité étant directement proportionnelle au nombre de modifications réalisées.

L'augmentation du nombre d'itération implique en revanche une plus longue durée de calcul. L'algorithme ne permettra donc peut être pas d'obtenir un niveau de précision suffisant dans un laps de temps raisonnable, limitant donc l'usage de celui-ci.

L'algorithme pouvant être bloqué dans un optimum local, la gestion du décalage est importante. En effet, sans cette dernière, le résultat obtenu correspond à celui attendu décalé d'une ou plusieurs positions.

### **2.3.3 Quel est l'impact des paramètres du problème (nombre et longueur des séquences de S, longueur du motif) sur les performances ?**

Plus il y a de séquences, plus les probabilités obtenues sont générales et peuvent permettre de trouver les motifs. Un set de séquence de petite taille mènerait à des distributions de probabilité très spécifiques qui ne permettraient pas d'identifier des motifs présentant plus de variabilité que ceux des autres séquences.

Similairement, plus les séquences et motifs sont longs, moins les probabilités qui en découlent seront biaisées.

Cependant, une augmentation de la longueur des séquences et des motifs engendre une très forte augmentation de la durée de calcul. L'algorithme n'est donc potentiellement pas déployable pour des jeux de données de très grandes tailles qu'impliquent souvent les travaux de recherches en bio-informatique.

### **2.3.4 Quel est l'effet de l'ordre du modèle de Markov pour les séquences hors motif ?**

N'ayant pas implémenté cela dans le code, nous ne pouvons pas analyser de vrais résultats. Cependant, selon la théorie, prendre en compte un ordre supérieur améliore la détection des motifs étant donné que les dépendances spatiales entre les nucléotides hors motif sont également considérées.

### **2.3.5 Que se passe-t-il si la longueur W utilisée par l'algorithme ne correspond pas à la longueur du vrai motif ? Pouvez-vous trouver une stratégie pour déterminer la longueur du motif dans le cas où elle serait inconnue ?**

Dans le cas où la longueur du motif est inconnue, il faut tester plusieurs valeurs possibles et décider laquelle a le plus de probabilité d'être la vraie taille. Il faut faire cela de manière jusqu'à avoir testé les différentes valeurs qui auraient pu être choisies. Notons que la probabilité convergera vers la bonne valeur, il sera donc inutile de tester des valeurs s'éloignant de la probabilité. Le critère de probabilité se base sur  $\mathbb{P}(S, A)$ , mais il est plus évident de calculer  $\mathbb{P}(S, A|\Theta, \Phi)$ .

$$\mathbb{P}(S, A|\Theta, \Phi) \propto \prod_{i=1}^K \phi_i^{\tilde{n}_{i0}} \prod_{j=1}^W \theta_{ij}^{n_{ij}}$$

Les résultats étant très faibles, il est plus évident de comparer leur logarithme, soit :

$$\log \mathbb{P}(S, A | \Theta, \Phi) \propto \sum_{i=1}^K \tilde{n}_{i0} \log(\phi_i) \sum_{j=1}^W n_{ij} \log(\theta_{ij})$$

Parmi ces résultats obtenus, celui ayant la plus haute probabilité est celui ayant la valeur absolue la plus faible.

Notons que la distance de Kullback-Leiber utilisée pour déterminer  $\hat{A}$  ne peut plus l'être ici. En effet, plus la longueur de motif prédite est faible, plus ce score sera grand.

### 3 Codes

Le code concernant la partie 1 se trouve dans `projet_part1.py` et celui concernant la partie 2 se trouve dans `projet_part2.py`.

### 4 Contribution de chacun des membres

- Alyssa Di Matteo : 33%
- Louis Duchêne : 33%
- Manon Gerard : 33%

### 3.1 Q2.2.1 Dérivations mathématiques 7 / 10

- ✓ - 0 pts Discussion ordres supérieurs présente.
  - ✓ - 1 pts Quelques petites erreurs
  - ✓ - 1 pts Certains passages pas suffisamment explicités.
  - ✓ - 1 pts Quelques problèmes de formes (notations pas claires, pas expliquées, etc.).
- 7 Non, pas du tout. S dépend de A.
- 8 Problème de notations. Le produit sur \$\$k\$\$ doit porter aussi sur cette partie.
- 9 Ce terme n'est pas suffisant. Il faut un terme comme ça pour chaque vecteur de paramètres de multinomiale. Il était plus simple de passer à un \$\$\propto\$\$ directement.
- 10 C'est surtout lié à l'utilisation d'une approche bayesienne.
- 11 Le passage ici aurait pu être mieux expliqué.
- 12 Pas bien formulé. C'est le comptage des nucléotides dans les régions hors motif de toutes les séquences + le comptage dans le motif de la séquence \$\$k\$\$.
- 13 Ce n'est pas correct. C'est le nombre de nucléotide dans le motif de la séquence \$\$k\$\$, pas dans la séquence.
- 14 Non, c'est une propriété de la fonction \$\$\Gamma\$\$, pas de Dirichlet.
- 15 "taille différente" ? C'est mal formulé.
- 16 Ce n'est pas correct. Il manque \$\$\alpha\_i\$\$. On obtient ça par la même propriété de la fonction \$\$\Gamma\$\$.
- 17 Parce que \$\$\tilde{n}^{-k\_{i0}} = \sum\_j j^{ij}^k
- 18 Manque toujours \$\$\alpha\$\$.
- 19 C'est le dénominateur
- 20 Non, la formule n'est pas correcte. On ne doit pas multiplier pour chaque valeur \$\$v\$\$ possible. Il faut sélectionner le \$\$v\$\$ par position \$\$j\$\$ qui correspondant au nucléotide à la position \$\$j-1\$\$.

### 3.2 Q2.2.2 Implémentation 8 / 10

Description de l'implémentation

✓ - 0 pts *Correct*

Sophistication de l'implémentation

✓ - 2 pts *Bien*

Points forts: choix des positions renvoyées selon KL, shift implémenté (mais pas fonctionnel).

Code ok. Description suffisante.

21 TB

### 3.3 Q2.2.3 Expérimentations 3 / 10

#### + 3 Point adjustment

💬 Aucune expérimentation et aucun résultat montré, même lié à la compétition. Il n'y a que des courtes discussions (correctes cependant).

22 Etonnant. Ca devrait changer pas mal d'un run à l'autre.

### 3.4 Q2.2.4 Compétition 6 / 6

- ✓ + 2 pts *Participation données artificielles*
- ✓ + 1 pts *Participation données réelles*
- ✓ + 2 pts *Résultats excellents données artificielles*
- ✓ + 1 pts *Résultats excellents données réelles*



UNIVERSITÉ DE LIÈGE  
FACULTÉ DES SCIENCES APPLIQUÉES

---

# Méthodes de Monte Carlo par chaînes de Markov pour la détection de motifs en bioinformatique

---

MATH0488-1 : Éléments de processus stochastiques

*Auteurs :*  
Alyssa DI MATTEO s201486  
Louis DUCHÈNE S202097  
Manon GERARD s201354

*Professeur :*  
Pierre GEURTS

*Encadrants :*  
Vân Anh HUYNH-THU  
Yann CLAES

3<sup>e</sup> année de Bachelier Ingénieur Civil  
Année académique 2022 - 2023

## Références

- [1] C.E. LAWRENCE, J.S. LIU et A.F NEUWALD. « Bayesian Models for Multiple Local Sequence Alignment and Gibbs Sampling Strategies ». *Journal of the American Statistical Association* (Dec 1995). URL : <https://www.jstor.org/stable/2291508>.
- [2] C.E. LAWRENCE et al. « Detecting subtle sequence signals : a gibbs sampling strategy for multiple alignment ». *Science*. 262(5131) : 208-214. Oct 1993.
- [3] Maneesh SAHANI. « Probabilistic Unsupervised Learning Sampling Methods ». *University College London* (2015). URL : [http://www.gatsby.ucl.ac.uk/teaching/courses/ml1-2015/lect12-handout.pdf?fbclid=IwAR3Ecd-WwnYLIrrcom2TWjGv6-o7m51\\_iM1BMQCIpn7MBQuV8KkyJcxPbGk](http://www.gatsby.ucl.ac.uk/teaching/courses/ml1-2015/lect12-handout.pdf?fbclid=IwAR3Ecd-WwnYLIrrcom2TWjGv6-o7m51_iM1BMQCIpn7MBQuV8KkyJcxPbGk).
- [4] G THIJS et al. « A gibbs sampling method to detect over-represented motifs in the upstream regions of co-expressed genes ». *Proceedings of the Fifth Annual International Conference on Computational Biology*. RECOMB '01, page 305-312, New York, NY, USA, 2001. Association for Computing Machinery.