Chaînes de Markov régulées pour l'analyse de séquences biologiques

Nicolas Vergne

Journée commune NormaSTIC, Normandie Mathématiques, le 20 mai 2016

Avant-Propos

gggcggcgacctcgcgggttttcgctatttatgaaaattttccggtttaaggcgtttccgttcttcttcgt
cataacttaatgtttttatttaaaataccctctgaaaagaaaggaaacgacaggtgctgaaagcgaggctt
tttggcctctgtcgtttcctttctctgttttgtccgtggaatgaacaatggaagtcaacaaaaagcagct
ggctgacattttcggtgcgagtatccgtaccattcagaactggcaggaacagggaatgcccgttctgcgag
gcggtggcaagggtaatgaggtgctttatgactctgccgccgtcataaaatggtatgccgaaagggatgct
gaaattgagaacgaaaagctgcgccgggaggttgaagaactgcggcaggccagcgaggcagatctccagcc
aggaactattgagtacgaacgccatcgacttacgcgtgcgcaggccgacgcacaggaactgaagaatgcca
gagactccgctgaagtggtggaaaccgcattctgtactttcgtgctgtcgcggatcgcaggtgaaattgcc
agtattctcgacgggctccccctgtcggtgcagcggcgttttccggaactggaaaaccgacatgttgattt
cctgaaacgggatatcatcaaagccatgaacaaagcagccgcgctggatgaactgataccggggttgctga
gtgaatatatcgaacagtcaggttaacaggctgcggcattttgtccgcgccgggcttcgctcactgttcag
gccggagccacagaccgccgttgaatgggcggatgctaattactatctcccgaaagaatccgcataccagg
aagggcgctgggaaacactgccctttcagcgggccatcatgaatgcgatgggcagcgactacatccgtgag
gtgaatgtggtgaagtctgcccgtgtcggttattccaaaatgctgctgggtgtttatgcctactttataga

Chaînes de Markov

- Processus stochastique : $(X_t)_{t\in\mathbb{N}}$
- Propriété de Markov (ordre 1) : $\mathbb{P}(X_t \mid X_u, u < t) = \mathbb{P}(X_t \mid X_{t-1})$
- Chaîne de Markov : loi initiale μ_0 et matrice de transition Π
- Pour un espace d'état $\mathcal{A} = \{a, c, g, t\}$, nous avons à l'ordre 1 :

$$\mu_0 = \begin{pmatrix} \mu_0(a) & \mu_0(c) & \mu_0(g) & \mu_0(t) \end{pmatrix}$$

$$\Pi = \begin{pmatrix} \pi_{aa} & \pi_{ac} & \pi_{ag} & \pi_{at} \\ \pi_{ca} & \pi_{cc} & \pi_{cg} & \pi_{ct} \\ \pi_{ga} & \pi_{gc} & \pi_{gg} & \pi_{gt} \\ \pi_{ta} & \pi_{tc} & \pi_{tg} & \pi_{tt} \end{pmatrix}$$

où $\pi_{uv} = \mathbb{P}(X_t = v \mid X_{t-1} = u)$, avec $u \in \mathcal{A}$ et $v \in \mathcal{A}$ et où nous pouvons choisir par exemple 0.25 pour $\mu_0(u)$, $\forall u \in \mathcal{A}$.

Plan

Plan

Les modèles de Markov

Chaînes de Markov

Homogénéité des séquences. Faux ! Exemple : pourcentage en gc.

$X = (X_t)_{t \in [\![0;n]\!]} \longrightarrow \Pi$

Les modèles de Markov

Chaînes de Markov

Homogénéité des séquences. Faux ! Exemple : pourcentage en gc.
$X = (X_t)_{t \in [\![0;n]\!]} \longrightarrow \Pi$

Chaînes de Markov cachées

Différentes plages homogènes. Modélisation d'un certain nombre de "phénomènes" biologiques. Exemple : composition différente des régions codantes et non-codantes. $X = (X_t)_{t \in [\![0;n]\!]} \longrightarrow \Pi_1, \Pi_2, \ldots$

Les modèles de Markov

Chaînes de Markov

Homogénéité des séquences. Faux ! Exemple : pourcentage en gc.
$X = (X_t)_{t \in [\![0;n]\!]} \longrightarrow \Pi$

Chaînes de Markov cachées

Différentes plages homogènes. Modélisation d'un certain nombre de "phénomènes" biologiques. Exemple : composition différente des régions codantes et non-codantes. $X = (X_t)_{t \in [\![0;n]\!]} \longrightarrow \Pi_1, \Pi_2, \ldots$

Chaînes de Markov régulées

Hétérogénéité continue. Modélisation de phénomènes biologiques continus. Exemple : transition (parfois brutale) entre deux états d'une chaîne de Markov cachée ou pourcentage en gc. $X = (X_t)_{t \in [\![0;n]\!]} \longrightarrow \Pi_{\frac{t}{n}}$

Isochores

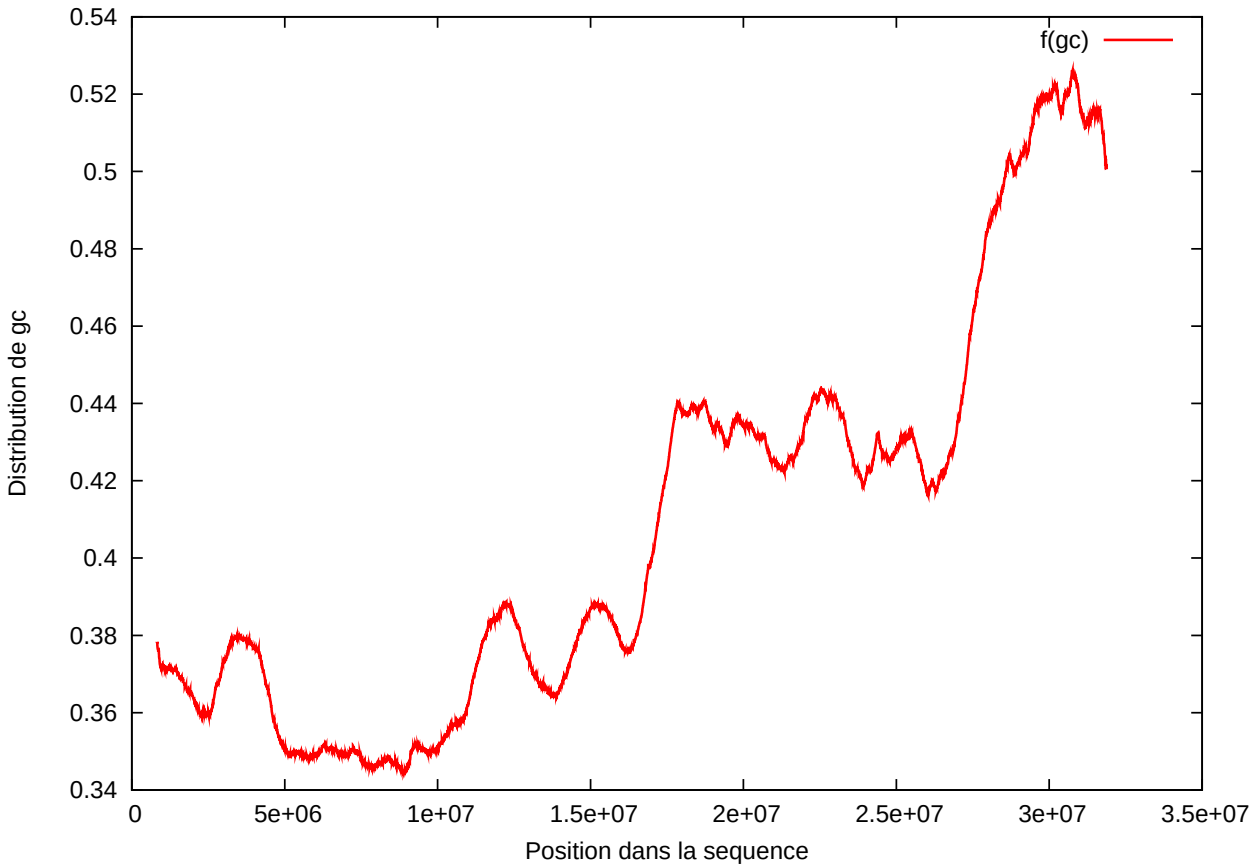
Phénomène biologique

Le pourcentage en gc varie le long d'une séquence.

Isochores

Isochore	Pourcentage en gc
L1	gc $<$ 38 %
L2	38 % $<$ gc $<$ 42 %
H1	42 % $<$ gc $<$ 47 %
H2	47 % $<$ gc $<$ 52 %
H3	52 % $<$ gc

Pourcentage en gc sur le chromosome 21 de l'homme



Définition et notations

$\mathcal{A}$ est l'alphabet (espace d'état) utilisé (par exemple $\mathcal{A}$ ={a,c,g,t}).

Définition (Chaîne de Markov régulée)

Soit $X = (X_t)_{t \in [\![0;n]\!]}$ une suite de variables aléatoires. Une chaîne de Markov régulée d'ordre k est uniquement définie par sa matrice de transition

$$\Pi_{\frac{t}{n}}(u, v) = \mathbb{P}\left(X_t = v | X_{t-k} \dots X_{t-1} = u\right)$$

et une loi initiale μ_0 avec $u = u_1 u_2 \dots u_k$ le passé markovien et $(u_1, u_2, \dots, u_k, v) \in \mathcal{A}^{k+1}$.

Ordre k

$\Pi_0 \rightsquigarrow \mu_0$	$\Pi_{\frac{k}{n}}$	$\Pi_{\frac{k+1}{n}}$	$\cdots$	Π_1
$X_0 \dots X_{k-1}$	X_k	X_{k+1}	$\cdots$	X_n

Plan

Dérive linéaire : le modèle

Les chaînes de Markov régulées polynomiales ont deux paramètres d'ordre :

- l'ordre markovien : k
- le degré de la dérive : d

Dérive linéaire

Deux matrices formant deux points d'appuis : Π_0 au début de la séquence et Π_1 à la fin de la séquence. On varie linéairement de l'une à l'autre :

$$\Pi_{\frac{t}{n}}(u,v) = \left(1 - \frac{t}{n}\right)\Pi_0(u,v) + \left(\frac{t}{n}\right)\Pi_1(u,v)$$

Estimation de Π_0 et Π_1

- Maximum de vraisemblance
- Régression matricielle
- Point par point

Π_0 et Π_1 : Maximum de vraisemblance

- La vraisemblance (à l'ordre 1) :

$$\ell(X, \Pi_0, \Pi_1) \quad = \quad \mu_0(X_0) \prod_{t=1}^{n} \left[\left(1 - \frac{t}{n}\right) \Pi_0(X_{t-1}, X_t) + \left(\frac{t}{n}\right) \Pi_1(X_{t-1}, X_t) \right].$$

- La Log-vraisemblance :

$$L(X, \Pi_0, \Pi_1) \quad = \quad \ln \mu_0(X_0) + \sum_{t=1}^{n} \sum_{u \in \mathcal{A}} \mathbb{1}_{\{X_{t-1}=u\}} \sum_{v \in \mathcal{A}} \mathbb{1}_{\{X_t=v\}} \ln \left(\Pi_{\frac{t}{n}}(u, v) \right).$$

- Nous posons

$$\Pi_{\frac{t}{n}}(u, u) = \left(1 - \frac{t}{n}\right) \left(1 - \sum_{v \in \mathcal{A} \setminus \{u\}} \Pi_0(u, v)\right) + \left(\frac{t}{n}\right) \left(1 - \sum_{v \in \mathcal{A} \setminus \{u\}} \Pi_1(u, v)\right),$$

ainsi $L = \ln \mu_0(X_0) + \sum_{t=1}^{n} \sum_{u \in \mathcal{A}} \mathbb{1}_{\{X_{t-1}=u\}}$

$$\left(\left(\sum_{v \in \mathcal{A} \setminus \{u\}} \mathbb{1}_{\{X_t=v\}} \ln \left(\Pi_{\frac{t}{n}}(u, v) \right) \right) + \mathbb{1}_{\{X_t=u\}} \ln \left(\Pi_{\frac{t}{n}}(u, u) \right) \right).$$

Π_0 et Π_1 : Maximum de vraisemblance

- On annule la dérivée pour obtenir le système suivant :

$$\begin{cases} \displaystyle\sum_{t=1}^{n}\left(1-\frac{t}{n}\right)\frac{\mathbb{1}_{\{X_{t-1}=u,X_t=v\}}}{\Pi_{\frac{t}{n}}(u,v)} = \sum_{t=1}^{n}\left(1-\frac{t}{n}\right)\frac{\mathbb{1}_{\{X_{t-1}=u,X_t=u\}}}{\Pi_{\frac{t}{n}}(u,u)} \\ \displaystyle\sum_{t=1}^{n}\left(\frac{t}{n}\right)\frac{\mathbb{1}_{\{X_{t-1}=u,X_t=v\}}}{\Pi_{\frac{t}{n}}(u,v)} = \sum_{t=1}^{n}\left(\frac{t}{n}\right)\frac{\mathbb{1}_{\{X_{t-1}=u,X_t=u\}}}{\Pi_{\frac{t}{n}}(u,u)} \end{cases}$$

- Un système de $2|\mathcal{A}|(|\mathcal{A}|-1)$ équations à $2|\mathcal{A}|(|\mathcal{A}|-1)$ inconnues.
- En réalité $|\mathcal{A}|$ systèmes $2(|\mathcal{A}|-1)$ équations à $2(|\mathcal{A}|-1)$ inconnues.

Π_0 et Π_1 : Maximum de vraisemblance

Exemple d'un de ces systèmes à l'ordre 1 pour l'alphabet {a, c, g, t}

$$
\begin{cases}
\bullet \sum_{t=1}^{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=c\}} \left(\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)\Pi_0(a,c) + \left(\frac{t}{n}\right)\Pi_1(a,c)} = \sum_{t=1}^{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=a\}} \left(\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)(1-\Pi_0(a,c)-\Pi_0(a,g)-\Pi_0(a,t)) + \left(\frac{t}{n}\right)(1-\Pi_1(a,c)-\Pi_1(a,g)-\Pi_1(a,t))} \\
\bullet \sum_{t=1}^{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=g\}} \left(\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)\Pi_0(a,g) + \left(\frac{t}{n}\right)\Pi_1(a,g)} = \sum_{t=1}^{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=a\}} \left(\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)(1-\Pi_0(a,c)-\Pi_0(a,g)-\Pi_0(a,t)) + \left(\frac{t}{n}\right)(1-\Pi_1(a,c)-\Pi_1(a,g)-\Pi_1(a,t))} \\
\bullet \sum_{t=1}^{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=t\}} \left(\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)\Pi_0(a,t) + \left(\frac{t}{n}\right)\Pi_1(a,t)} = \sum_{t=1}^{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=a\}} \left(\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)(1-\Pi_0(a,c)-\Pi_0(a,g)-\Pi_0(a,t)) + \left(\frac{t}{n}\right)(1-\Pi_1(a,c)-\Pi_1(a,g)-\Pi_1(a,t))} \\
\bullet \sum_{t=1}^{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=c\}} \left(1-\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)\Pi_0(a,c) + \left(\frac{t}{n}\right)\Pi_1(a,c)} = \sum_{t=1}^{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=a\}} \left(1-\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)(1-\Pi_0(a,c)-\Pi_0(a,g)-\Pi_0(a,t)) + \left(\frac{t}{n}\right)(1-\Pi_1(a,c)-\Pi_1(a,g)-\Pi_1(a,t))} \\
\bullet \sum_{t=1}^{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=g\}} \left(1-\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)\Pi_0(a,g) + \left(\frac{t}{n}\right)\Pi_1(a,g)} = \sum_{t=1}^{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=a\}} \left(1-\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)(1-\Pi_0(a,c)-\Pi_0(a,g)-\Pi_0(a,t)) + \left(\frac{t}{n}\right)(1-\Pi_1(a,c)-\Pi_1(a,g)-\Pi_1(a,t))} \\
\bullet \sum_{t=1}^{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=t\}} \left(1-\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)\Pi_0(a,t) + \left(\frac{t}{n}\right)\Pi_1(a,t)} = \sum_{t=1}^{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=a\}} \left(1-\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)(1-\Pi_0(a,c)-\Pi_0(a,g)-\Pi_0(a,t)) + \left(\frac{t}{n}\right)(1-\Pi_1(a,c)-\Pi_1(a,g)-\Pi_1(a,t))}
\end{cases}
$$

Π_0 et Π_1 : Régression matricielle

- On divise la séquence en N segments de tailles m :

$$\underbrace{X_0 \dots X_m}_{S_0} \underbrace{X_{m+1} \dots X_{2m}}_{S_1} \cdots \underbrace{X_{(N-1)m+1} \dots X_n}_{S_{N-1}}$$

- Sur chaque segment, on estime une matrice de transition :

$$\widehat{\Pi_{S_\ell}}(u,v) = \frac{N_{S_\ell}(uv)}{N_{S_\ell}(u+)} = \frac{\sum_{t \in S_\ell^\star} \mathbb{1}\{X_{t-k} \dots X_{t-1} = u, X_t = v\}}{\sum_{j \in S_\ell^\star} \mathbb{1}\{X_{j-k} \dots X_{j-1} = u\}}$$

- On minimise la somme $\sum_{\ell \in [\![0, N-1]\!]} d\left(\widehat{\Pi_{S_\ell}}, (1-\tau_\ell)\Pi_0 + \tau_\ell \Pi_1\right)$.

Π_0 et Π_1 : Régression matricielle

Expression de $\widehat{\Pi_0}$ et $\widehat{\Pi_1}$

$$\begin{cases} \widehat{\Pi_0}(u,v) &= \dfrac{B_1C_2(u,v) - B_2C_1(u,v)}{A_2B_1 - A_1B_2} \\ \widehat{\Pi_1}(u,v) &= \dfrac{A_2C_1(u,v) - A_1C_2(u,v)}{A_2B_1 - A_1B_2} \end{cases} \quad \text{avec}$$

$$A_1 = \sum_{\ell=0}^{N-1} 1-\tau_\ell, \quad B_1 = \sum_{\ell=0}^{N-1} \tau_\ell, \quad C_1(u,v) = \sum_{\ell=0}^{N-1} \widehat{\Pi_{S_\ell}}(u,v),$$

$$A_2 = \sum_{\ell=0}^{N-1} \tau_\ell(1-\tau_\ell), \quad B_2 = \sum_{\ell=0}^{N-1} {\tau_\ell}^2, \quad C_2(u,v) = \sum_{\ell=0}^{N-1} \tau_\ell \widehat{\Pi_{S_\ell}}(u,v).$$

Stochasticité de $\widehat{\Pi_0}$ et $\widehat{\Pi_1}$

- $\sum_{v\in\mathcal{A}} \widehat{\Pi_0}(u,v) = \sum_{v\in\mathcal{A}} \widehat{\Pi_1}(u,v) = 1.$
- Mais tous les termes ne sont pas forcément positifs...

Π_0 et Π_1 : Point par point

- Rappel : $\Pi_{\frac{t}{n}}(u,v) = \left(1-\frac{t}{n}\right)\Pi_0(u,v) + \frac{t}{n}\Pi_1(u,v)$
- On minimise les "erreurs de prédictions" :

$$\sum_{t=1}^{n}\left[\sum_{u\in\mathcal{A}^k}\mathbb{1}\{X_{t-k}\dots X_{t-1}=u\}\left[\sum_{v\in\mathcal{A}}\left(\Pi_{\frac{t}{n}}(u,v)-\mathbb{1}_{\{X_t=v\}}\right)^2\right]\right]$$

Π_0 et Π_1 : Point par point

Expression de $\widehat{\Pi_0}$ et $\widehat{\Pi_1}$

$$\begin{cases} \widehat{\Pi_0}(u,v) &= \dfrac{B_2C_1 - B_1C_2}{A_1B_2 - A_2B_1} \\ \widehat{\Pi_1}(u,v) &= \dfrac{A_1C_2 - A_2C_1}{A_1B_2 - A_2B_1} \end{cases} \quad \text{avec}$$

$$A_1 = A_1(u) = 2\sum_{t=1}^{n} \mathbb{1}_u \left(1 - \frac{t}{n}\right)^2, \qquad A_2 = A_2(u) = 2\sum_{t=1}^{n} \mathbb{1}_u \left(1 - \frac{t}{n}\right)\left(\frac{t}{n}\right),$$

$$B_1 = B_1(u) = 2\sum_{t=1}^{n} \mathbb{1}_u \left(1 - \frac{t}{n}\right)\left(\frac{t}{n}\right), \qquad B_2 = B_2(u) = 2\sum_{t=1}^{n} \mathbb{1}_u \left(\frac{t}{n}\right)^2,$$

$$C_1 = C_1(u,v) = 2\sum_{t=1}^{n} \mathbb{1}_{uv} \left(1 - \frac{t}{n}\right), \qquad C_2 = C_2(u,v) = 2\sum_{t=1}^{n} \mathbb{1}_{uv} \left(\frac{t}{n}\right).$$

Stochasticité de $\widehat{\Pi_0}$ et $\widehat{\Pi_1}$

- $\displaystyle\sum_{v\in\mathcal{A}} \widehat{\Pi_0}(u,v) = \sum_{v\in\mathcal{A}} \widehat{\Pi_1}(u,v) = 1$
- Mais tous les termes ne sont pas forcément positifs...

Dérive polynomiale

Définition

Une dérive polynomiale de degré d nécessite $d+1$ matrices formant $d+1$ points d'appuis $\Pi_{\frac{i}{d}}$:

$$\Pi_{\frac{t}{n}}(u,v) = \sum_{i=0}^{d} p_i(t)\Pi_{\frac{i}{d}}(u,v)$$

- Les $\Pi_{\frac{i}{d}}$ sont réparties uniformément sur la séquence ;
- Les p_i sont des polynômes de degré d tels que

$$\forall (i,j) \in \{0,\dots,d\}^2,\ p_i\left(\frac{nj}{d}\right) = \mathbb{1}_{\{i=j\}};$$

- Pour $t = ni/d$, $\Pi_{\frac{t}{n}} = \Pi_{\frac{i}{d}}$;
- $\sum_{v\in\mathcal{A}} \Pi_{\frac{t}{n}}(u,v) = 1.$

Dérive polynomiale : $\Pi_{\frac{t}{n}}(u,v)$

Chaîne de Markov régulées de degré 2

$$\left(2\frac{t^2}{n^2}-3\frac{t}{n}+1\right)\Pi_0(u,v)+\left(-4\frac{t^2}{n^2}+4\frac{t}{n}\right)\Pi_{\frac{1}{2}}(u,v)+\left(2\frac{t^2}{n^2}-\frac{t}{n}\right)\Pi_1(u,v).$$

Chaîne de Markov régulées de degré 3

$$\begin{aligned}&\left(-\frac{9}{2}\frac{t^3}{n^3}+9\frac{t^2}{n^2}-\frac{11}{2}\frac{t}{n}+1\right)\Pi_0+\left(\frac{27}{2}\frac{t^3}{n^3}-\frac{45}{2}\frac{t^2}{n^2}+9\frac{t}{n}\right)\Pi_{\frac{1}{3}}\\ +\quad&\left(-\frac{27}{2}\frac{t^3}{n^3}+18\frac{t^2}{n^2}-\frac{9}{2}\frac{t}{n}\right)\Pi_{\frac{2}{3}}+\left(\frac{9}{2}\frac{t^3}{n^3}-\frac{9}{2}\frac{t^2}{n^2}+\frac{t}{n}\right)\Pi_1.\end{aligned}$$

$\Pi_{\frac{i}{d}}$: Régression matricielle

Fonction à minimiser

$$\sum_{\ell\in[\![0,N-1]\!]}\sum_{u\in\mathcal{A}^k}\sum_{v\in\mathcal{A}}\left(\widehat{\Pi_{S_\ell}}(u,v)-\sum_{i=0}^{d}p_i(n\tau_\ell)\Pi_{\frac{i}{d}}(u,v)\right)^2$$

Systèmes à résoudre

Pour chaque couple (u,v), un système de $d+1$ équations à $d+1$ inconnues :

$$\begin{pmatrix} \sum_{\ell=0}^{N-1}A_0(n\tau_\ell)A_0(n\tau_\ell) & \cdots & \sum_{\ell=0}^{N-1}A_0(n\tau_\ell)A_d(n\tau_\ell) \\ \vdots & & \vdots \\ \sum_{\ell=0}^{N-1}A_d(n\tau_\ell)A_0(n\tau_\ell) & \cdots & \sum_{\ell=0}^{N-1}A_d(n\tau_\ell)A_d(n\tau_\ell) \end{pmatrix}\begin{pmatrix}\widehat{\Pi_0}(u,v)\\ \vdots \\ \widehat{\Pi_1}(u,v)\end{pmatrix}=\begin{pmatrix}\sum_{\ell=0}^{N-1}A_0(n\tau_\ell)\widehat{\Pi_{S_\ell}}(u,v)\\ \vdots \\ \sum_{\ell=0}^{N-1}A_d(n\tau_\ell)\widehat{\Pi_{S_\ell}}(u,v)\end{pmatrix}$$

$\Pi_{\frac{i}{d}}$: Point par point

Fonction à minimiser

$$\sum_{t=1}^{n}\left[\sum_{u\in\mathcal{A}^k}\mathbb{1}_u\left[\sum_{v\in\mathcal{A}}\left(\Pi_{\frac{t}{n}}(u,v)-\mathbb{1}_{uv}\right)^2\right]\right]$$

Systèmes à résoudre

Pour chaque couple (u,v), un système de $d+1$ équations à $d+1$ inconnues :

$$\begin{pmatrix} \sum_{t=1}^{n}\mathbb{1}_u A_0(t)A_0(t) & \cdots & \sum_{t=1}^{n}\mathbb{1}_u A_0(t)A_d(t) \\ \vdots & & \vdots \\ \sum_{t=1}^{n}\mathbb{1}_u A_d(t)A_0(t) & \cdots & \sum_{t=1}^{n}\mathbb{1}_u A_d(t)A_d(t) \end{pmatrix} \begin{pmatrix} \widehat{\Pi_0}(u,v) \\ \vdots \\ \widehat{\Pi_1}(u,v) \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^{n}A_0(t)\mathbb{1}_{uv} \\ \vdots \\ \sum_{t=1}^{n}A_d(t)\mathbb{1}_{uv} \end{pmatrix}$$

Log-vraisemblance : séquences simulées

Log-vraisemblance

$$L(X, \Pi_0, \Pi_1) = \ln \mu_0(X_0) + \sum_{t=1}^{n} \sum_{u \in \mathcal{A}} \mathbb{1}_{\{X_{t-1}=u\}} \sum_{v \in \mathcal{A}} \mathbb{1}_{\{X_t=v\}} \ln \left(\Pi_{\frac{t}{n}}(u, v) \right).$$

Degré		0	1	2	3	4	5
Ordre	Régression	−67191	−66999	−66962	−66910	−66909	−66907
0	Point	**−67191**	**−66999**	**−66962**	**−66910**	**−66909**	**−66907**
Ordre	Régression	−66718	−66504	−66448	−66382	−66376	−66368
1	Point	**−66710**	**−66501**	**−66445**	**−66380**	**−66374**	**−66366**
Ordre	Régression	−66706	−66482	−66407	−66321	−66295	−66275
2	Point	**−66693**	**−66477**	**−66402**	**−66317**	**−66290**	**−66270**
Ordre	Régression	−66630	−66331	−66186	−66038	−65938	−65883
3	Point	**−66612**	**−66320**	**−66169**	**−66014**	**−65898**	**−65817**

Log-vraisemblance : séquences réelles

Degré		0	1	2	3	4	5
Ordre	Régression	−67191	−66973	−66934	−66873	−66760	−66680
0	Point	**−67191**	**−66973**	**−66934**	**−66873**	**66760**	**−66680**
Ordre	Régression	−66743	−66500	−66439	−66362	−66234	−66146
1	Point	**−66714**	**−66483**	**−66419**	**−66345**	**−66220**	**−66135**
Ordre	Régression	−66052	−65657	−65577	−65438	−65281	−65160
2	Point	**−66005**	**−65631**	**−65544**	**−65410**	**−65255**	**−65139**
Ordre	Régression	−65661	−65168	−65033	−64809	−64597	−64432
3	Point	**−65579**	**−65116**	**−64951**	**−64746**	**−64497**	**−64329**

Conclusion

Par la suite, nous préconisons la méthode d'estimation point par point.

Plan

Splines

Splines polynomiales

- Polynômes par morceaux de degré fixé (en général 3 : splines cubiques)
- Regression polynomiale
- Choix des *noeuds* (répartition uniforme)
- Contraintes de continuité aux noeuds
- Plus flexible que les polynômes

Splines

Splines polynomiales

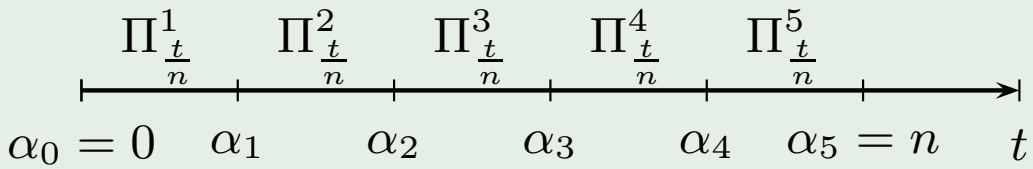
- Polynômes par morceaux de degré fixé (en général 3 : splines cubiques)
- Regression polynomiale
- Choix des *noeuds* (répartition uniforme)
- Contraintes de continuité aux noeuds
- Plus flexible que les polynômes

Splines de matrices : estimation

- Estimation globale : un unique système linéaire
- Aller retour avec fonctions de base : plusieurs petits sytèmes
- Aller retour sans fonctions de base : plusieurs petits sytèmes

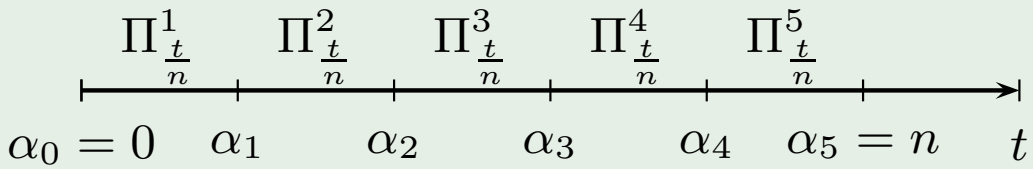
Splines de matrices

Découpage d'une séquence en 5 morceaux

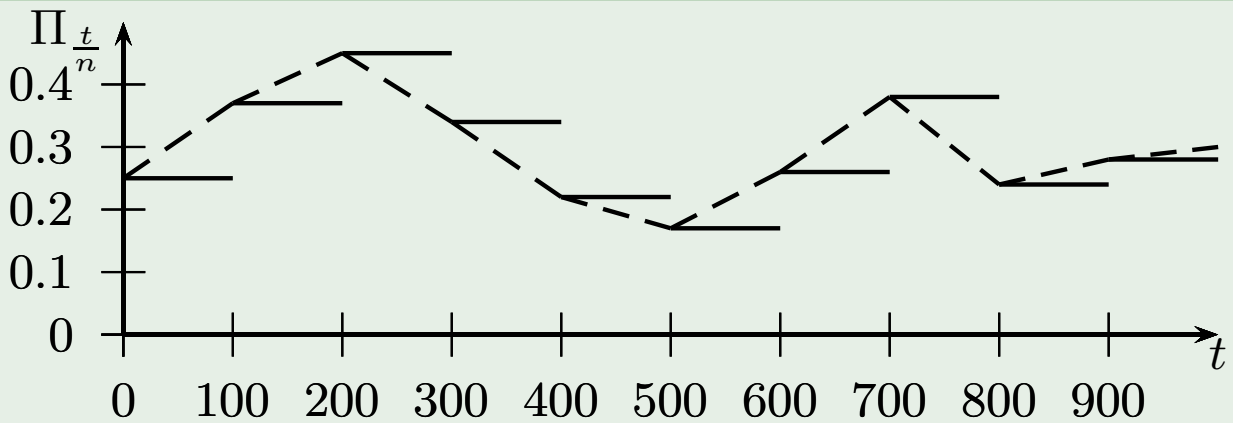


Splines de matrices

Découpage d'une séquence en 5 morceaux



Splines polynomiales de degré 0 et 1



L'axe des ordonnées "représente" l'espace des matrices.

Estimation globale

Le modèle

$$\Pi^i_{\frac{t}{n}}(u,v) \quad = \quad M^i_0(u,v) + \frac{t}{n} M^i_1(u,v) + \cdots + \frac{t^d}{n^d} M^i_d(u,v).$$

Contraintes

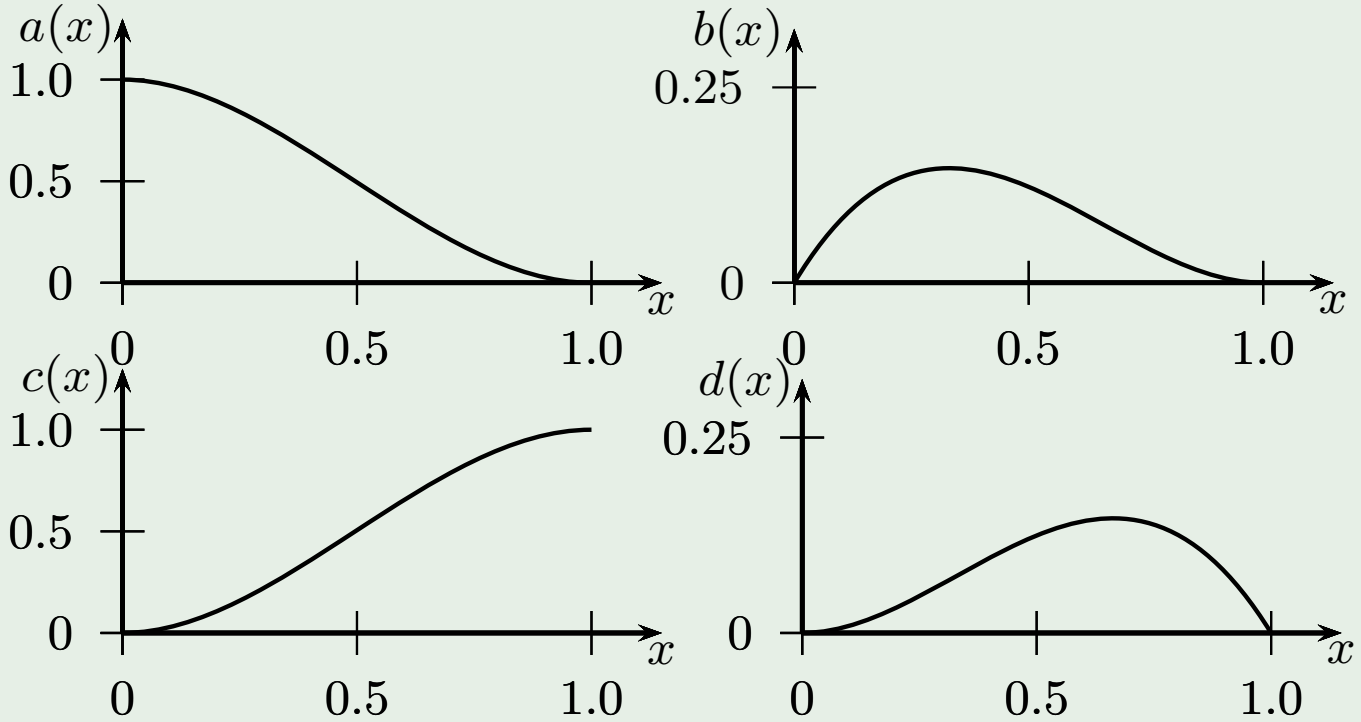
- $\Pi^i_{\frac{\alpha_i}{n}}(u,v) = \Pi^{i+1}_{\frac{\alpha_i}{n}}(u,v)$ pour i allant de 1 à $N-1$ ($N-1$ contraintes)
- $\left(\Pi^i_{\frac{\alpha_i}{n}}\right)^{(j)}(u,v) = \left(\Pi^{i+1}_{\frac{\alpha_i}{n}}\right)^{(j)}(u,v)$ pour i allant de 1 à $N-1$ et j allant de 1 à $d-1$ ($(N-1)(d-1)$ contraintes)

Minimisation

Pour chaque couple (u,v), $(N-1)d$ contraintes et $N(d+1)$ paramètres : méthode lagrangienne.

Aller retour avec fonctions de bases

Les fonctions de base des polynômes de degré 3



$a(x) = 2x^3 - 3x^2 + 1$, $b(x) = x^3 - 2x^2 + x$,
$c(x) = -2x^3 + 3x^2$, $d(x) = -x^3 + x^2$.

Aller retour avec fonctions de bases : modèle

Le modèle

$$\Pi^i_{\frac{t}{n}} = A_i a\left(\frac{t-\alpha_{i-1}}{\alpha_i-\alpha_{i-1}}\right) + B_i b\left(\frac{t-\alpha_{i-1}}{\alpha_i-\alpha_{i-1}}\right) + C_i c\left(\frac{t-\alpha_{i-1}}{\alpha_i-\alpha_{i-1}}\right) + D_i d\left(\frac{t-\alpha_{i-1}}{\alpha_i-\alpha_{i-1}}\right)$$

Aller retour avec fonctions de bases : modèle

Le modèle

$$\Pi^i_{\frac{t}{n}} = A_i a\left(\frac{t-\alpha_{i-1}}{\alpha_i-\alpha_{i-1}}\right) + B_i b\left(\frac{t-\alpha_{i-1}}{\alpha_i-\alpha_{i-1}}\right) + C_i c\left(\frac{t-\alpha_{i-1}}{\alpha_i-\alpha_{i-1}}\right) + D_i d\left(\frac{t-\alpha_{i-1}}{\alpha_i-\alpha_{i-1}}\right)$$

Simplification $\Pi^i_{\frac{t}{n}}$

Les contraintes ainsi que les propriétés des fonctions de bases simplifient le modèle

$$\Pi^i_{\frac{t}{n}} = A_i a\left(\frac{t-\alpha_{i-1}}{\alpha_i-\alpha_{i-1}}\right) + B_i b\left(\frac{t-\alpha_{i-1}}{\alpha_i-\alpha_{i-1}}\right) + A_{i+1} a\left(\frac{\alpha_i-t}{\alpha_i-\alpha_{i-1}}\right) + \frac{\alpha_{i-1}-\alpha_i}{\alpha_{i+1}-\alpha_i} B_{i+1} b\left(\frac{\alpha_i-t}{\alpha_i-\alpha_{i-1}}\right)$$

Minimisation

Les contraintes sont cette fois intégrées au modèle : méthode classique.

Aller retour avec fonctions de bases : algorithme

Le modèle $\Pi^i_{\frac{t}{n}}$

$$\Pi^i_{\frac{t}{n}} = A_i a\left(\frac{t-\alpha_{i-1}}{\alpha_i-\alpha_{i-1}}\right) + B_i b\left(\frac{t-\alpha_{i-1}}{\alpha_i-\alpha_{i-1}}\right) + A_{i+1} a\left(\frac{\alpha_i - t}{\alpha_i-\alpha_{i-1}}\right) + \frac{\alpha_{i-1}-\alpha_i}{\alpha_{i+1}-\alpha_i} B_{i+1} b\left(\frac{\alpha_i - t}{\alpha_i-\alpha_{i-1}}\right)$$

Algorithme

- Initialisation sur les deux premiers segments : A_1, B_1, A_2, B_2
- Aller sur chaque segment i de 2 à N : A_i, B_i
- Retour sur chaque segment i de $N-1$ à 1 : A_i, B_i

Aller retour sans fonctions de base

Le modèle

$$\Pi^i_{\frac{t}{n}} = H^i_0 + \frac{t}{n} H^i_1 + \frac{t^2}{n^2} H^i_2 + \frac{t^3}{n^3} H^i_3.$$

Minimisation

Pour chaque couple (u, v), $2(N-1)$ contraintes et $3N$ paramètres : méthode lagrangienne.

Algorithme

Le principe est le même que précédemment.

Log-vraisemblances des différentes estimations

Phage Lambda

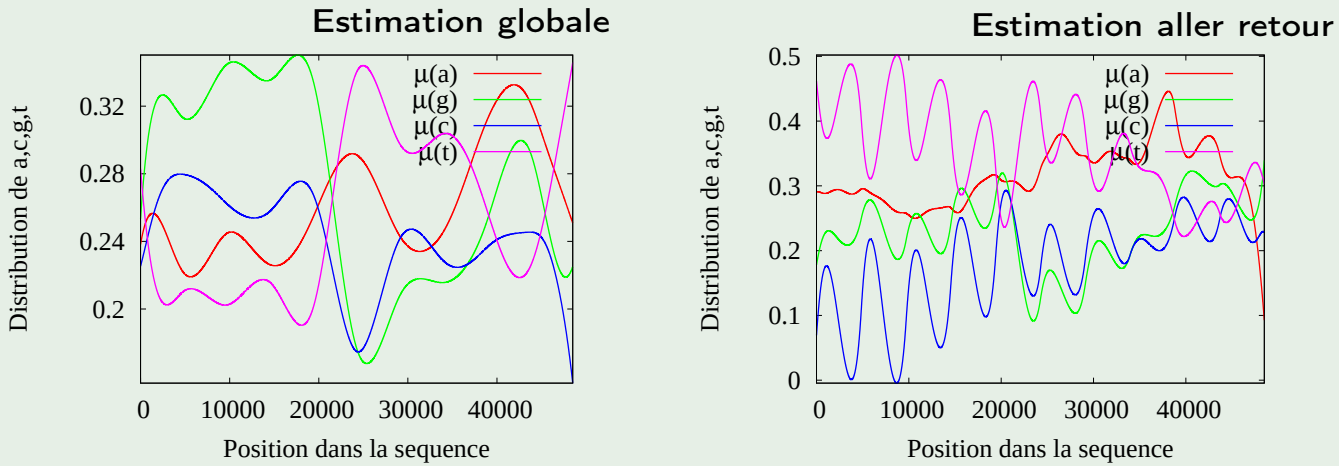
Nombre de	segments	2	3	4	5	10
	Globale	−66753	**−66681**	**−66658**	**−66661**	**−66543**
Ordre 0	Sans base	**−66665**	−67587	−72329	71979	−72680
	Bases	−67439	−68049	−70115	−68473	−83406
	Globale	−66213	**−66136**	**−66110**	**−66101**	**−65930**
Ordre 1	Sans base	**−66119**	−67429	−71213	−72184	−80817
	Bases	−66578	−68377	−76442	−70565	−780727
	Globale	−65244	**−65135**	**−65073**	**−65047**	**−64761**
Ordre 2	Sans base	**−65112**	−66843	−71439	−75769	−82707
	Bases	−67853	−69738	−75917	−72260	−79626
	Globale	−64486	**−64319**	**−64170**	**−64059**	**−63406**
Ordre 3	Sans base	**−64290**	−67234	−72830	−77288	−80974
	Bases	−70331	−70894	−76871	−75301	−80273

Remarque

Ces log-vraisemblances sont meilleures que celles obtenues par dérive polynomiale dans le cas d'une estimation globale.

Globale / Aller retour

Oscillations

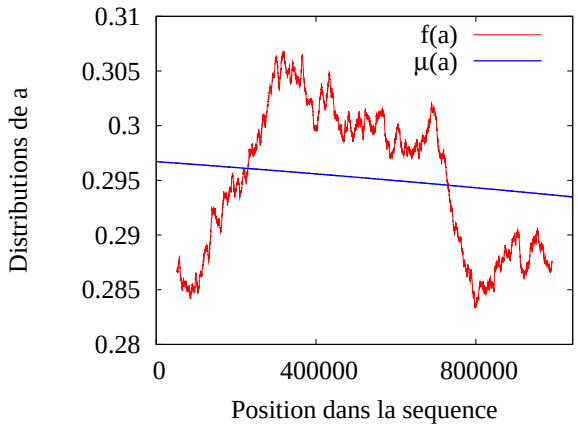


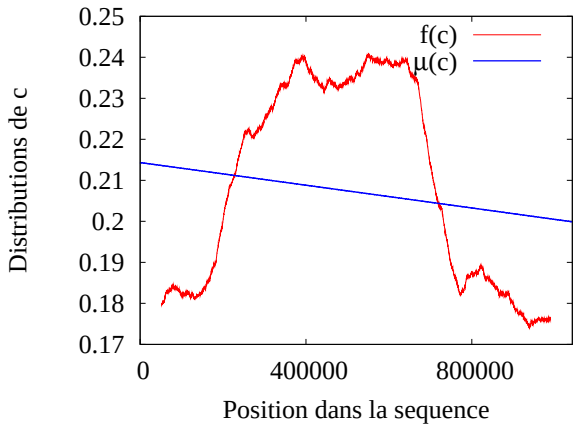
Conclusion

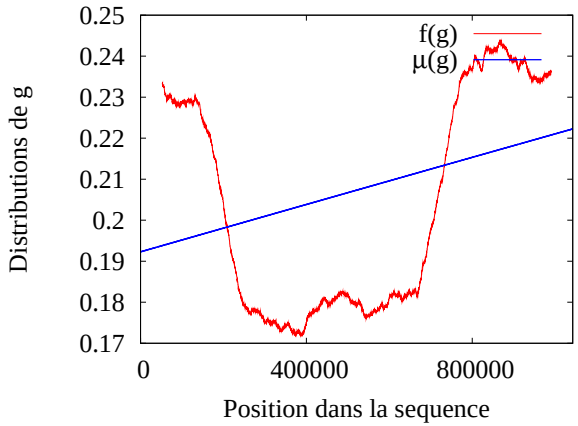
Par la suite, nous préconisons la méthode d'estimation globale.

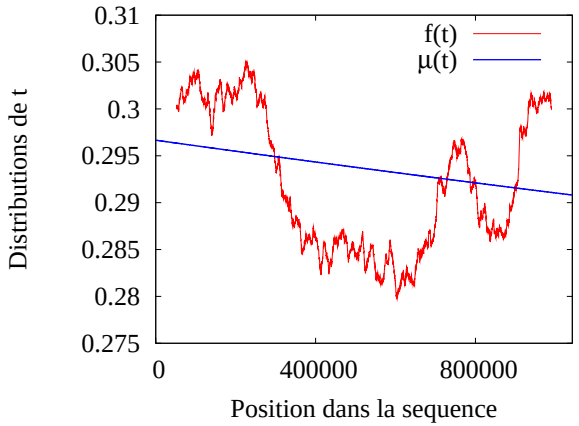
Plan

Fréquences / Lois stationnaires (degré 1) sur *Chlamydia trachomatis*

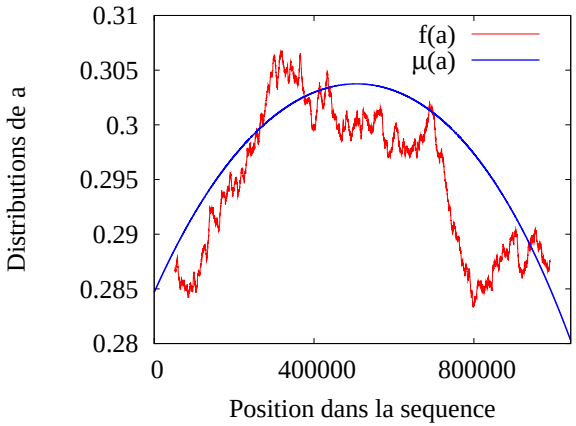


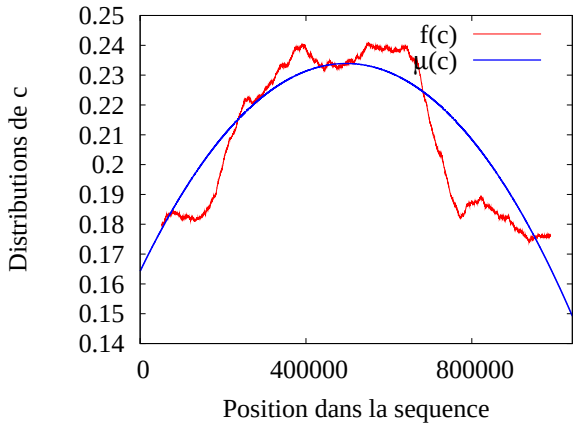


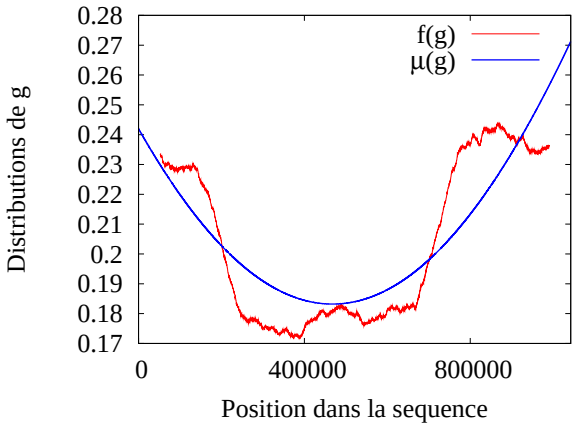


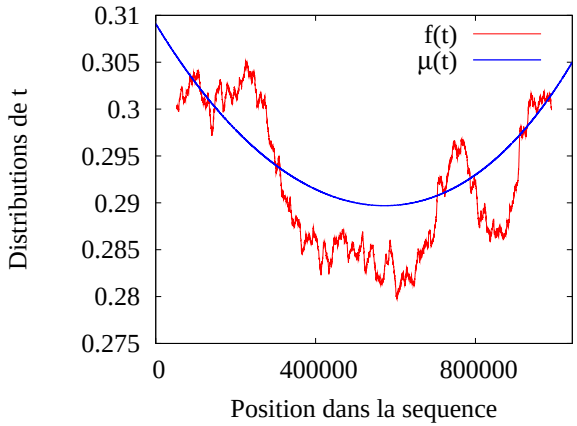


Fréquences / Lois stationnaires (degré 2)

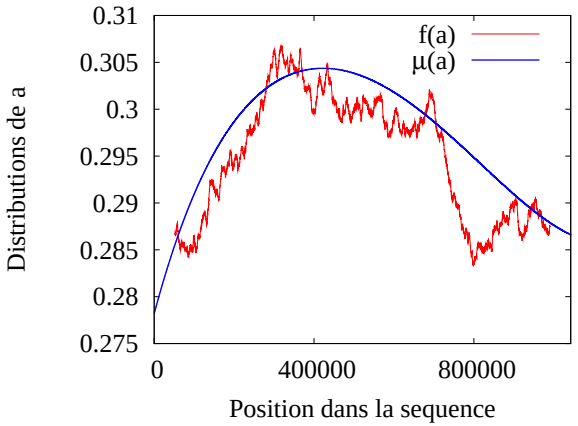


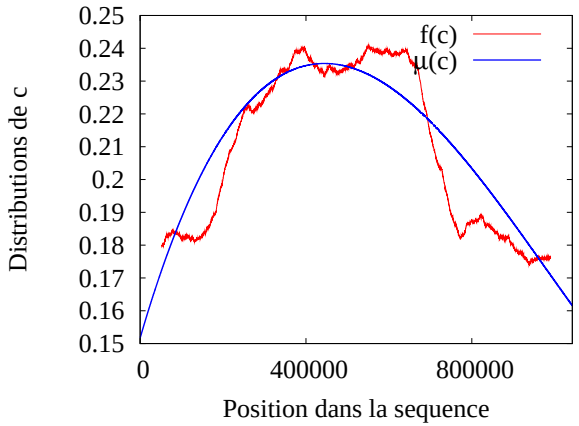


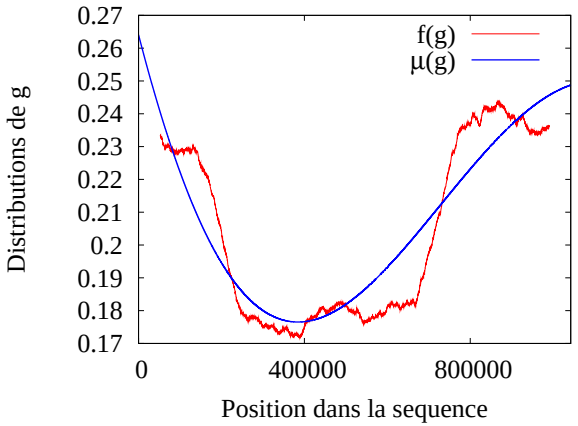


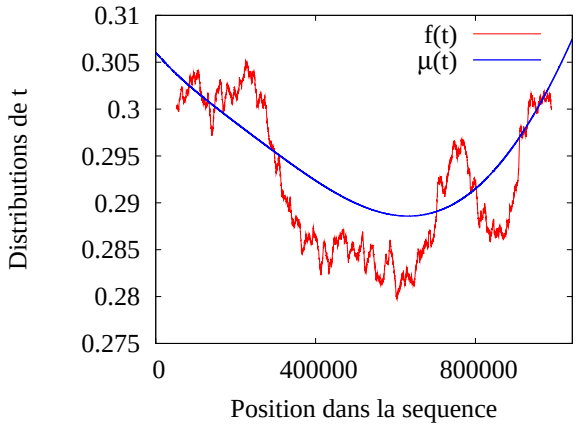


Fréquences / Lois stationnaires (degré 3)

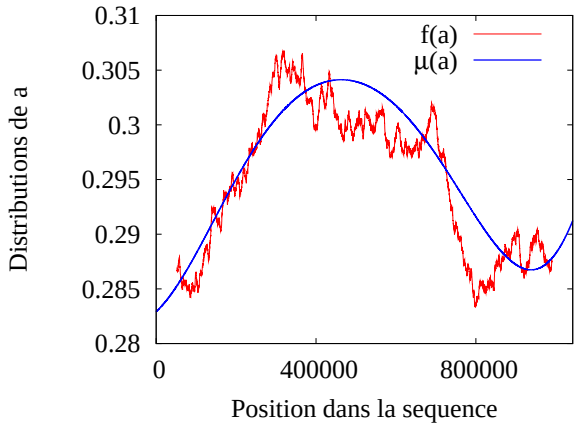


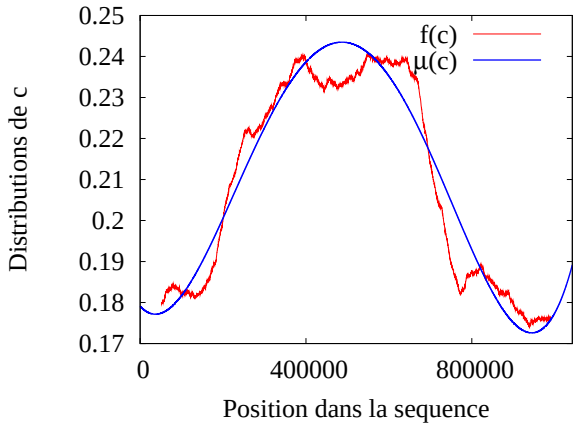


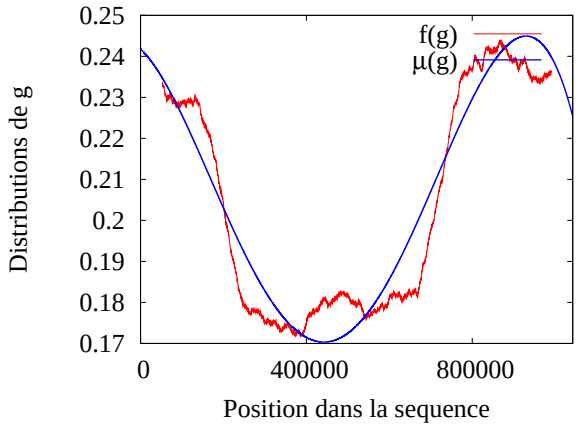


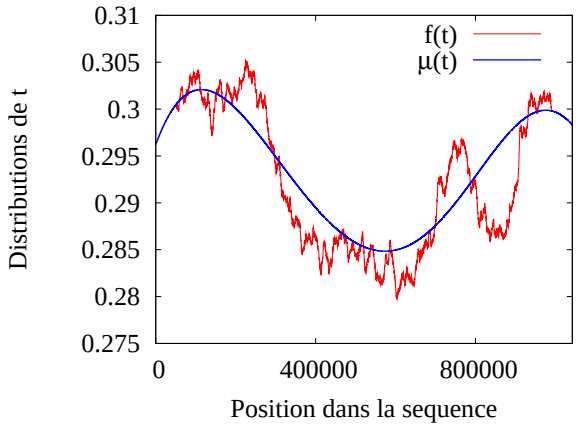


Fréquences / Lois stationnaires (degré 4)

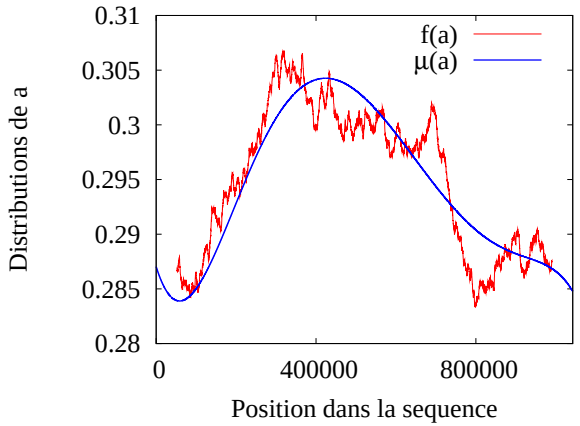


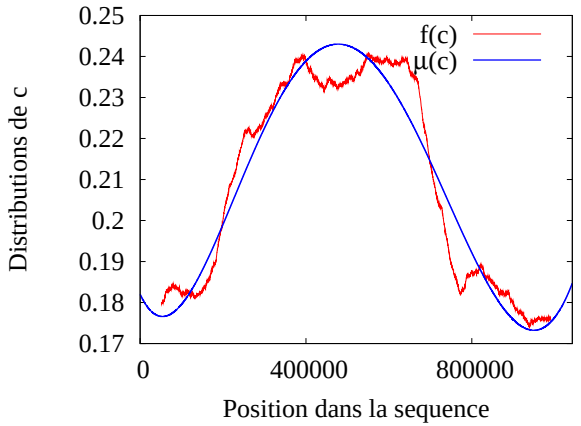


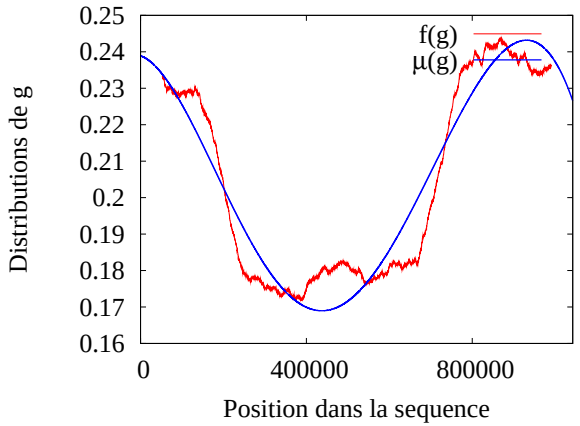


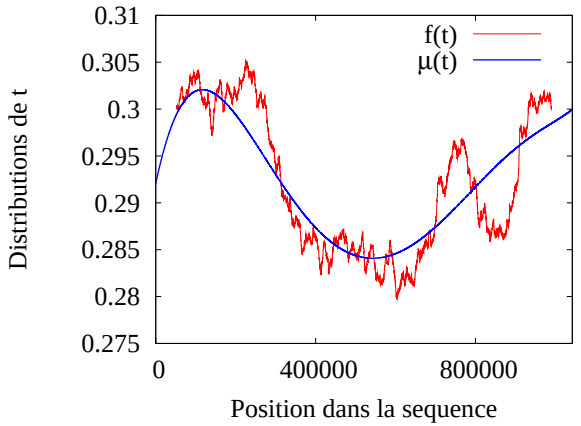


Fréquences / Lois stationnaires (degré 5)

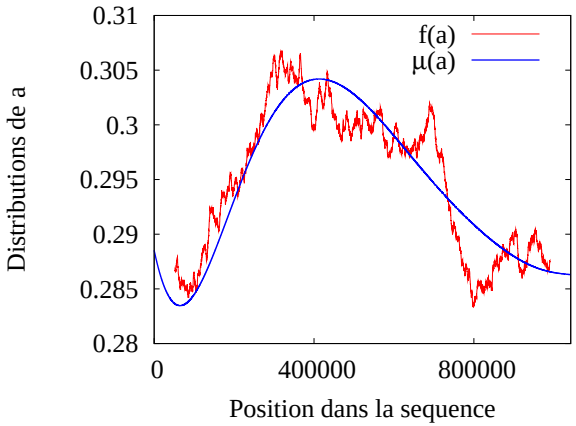


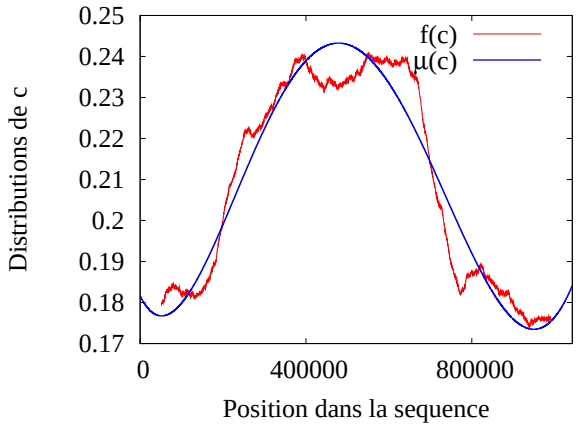


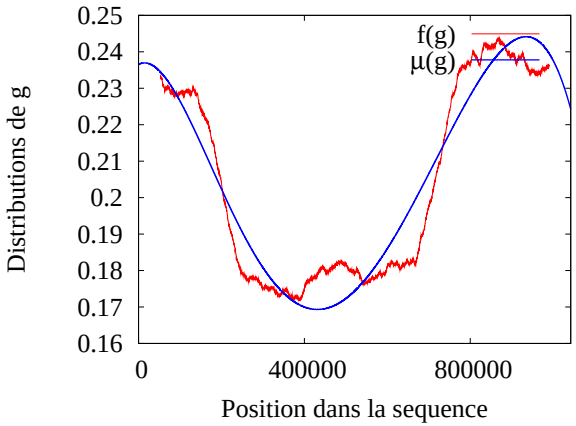


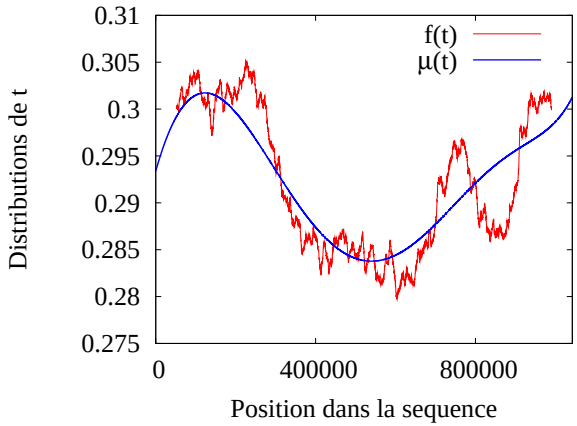


Fréquences / Lois stationnaires (degré 6)

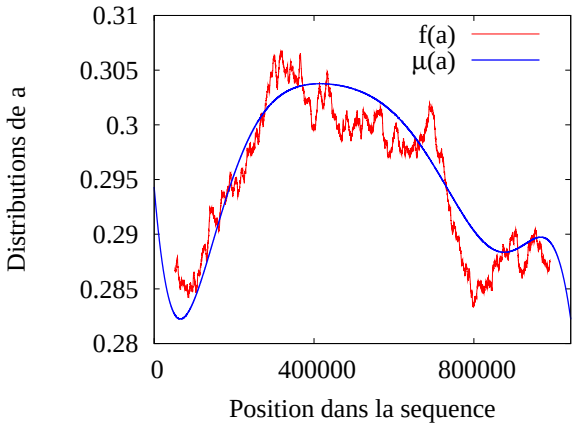


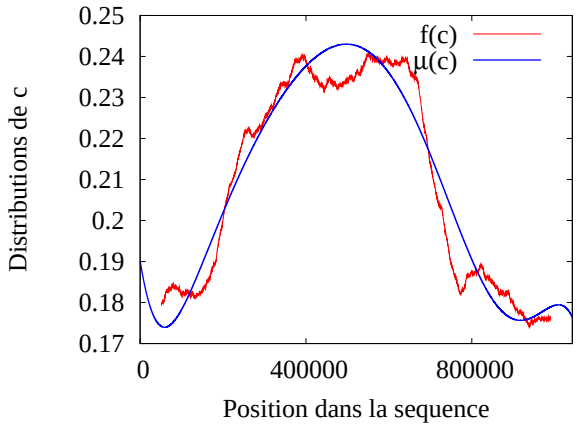


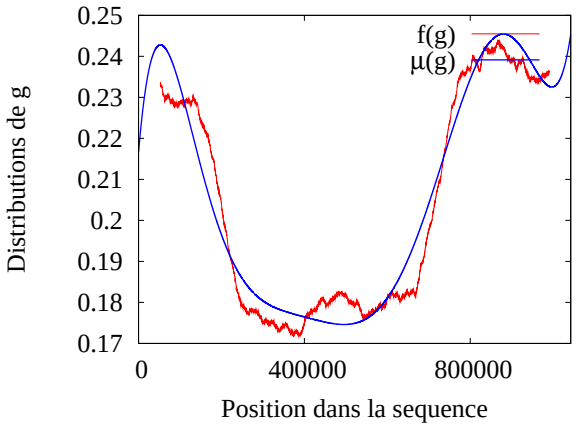


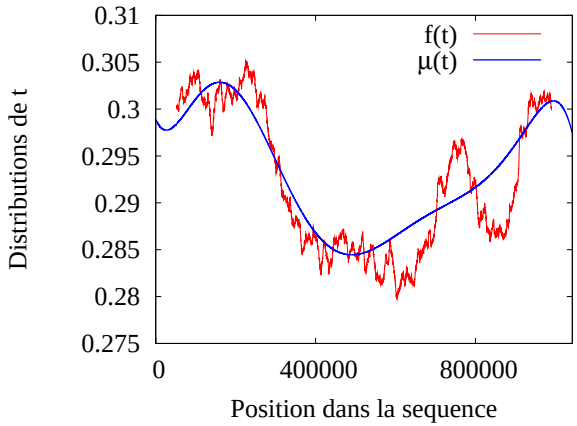


Fréquences / Lois stationnaires (degré 7)

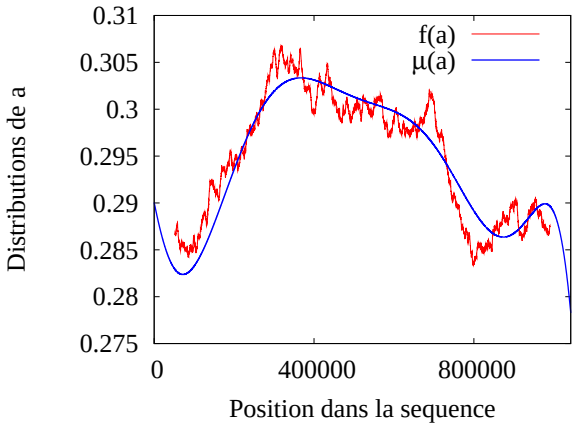


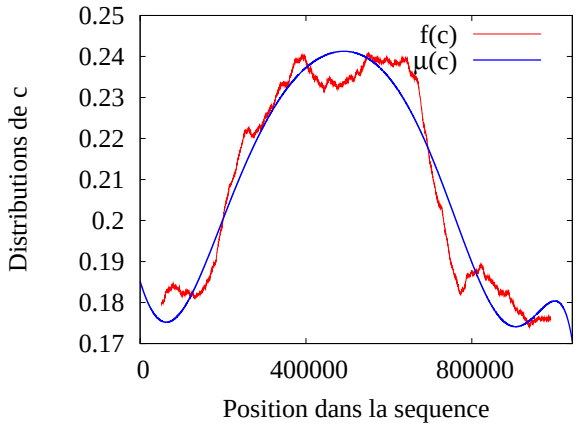


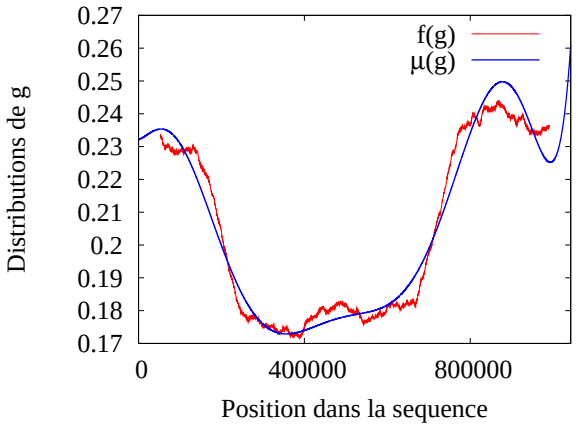


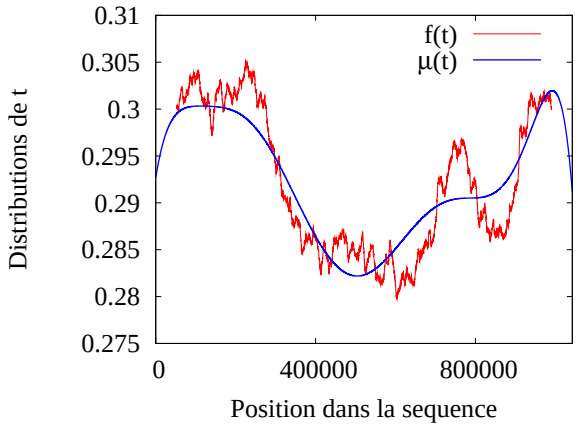


Fréquences / Lois stationnaires (degré 8)

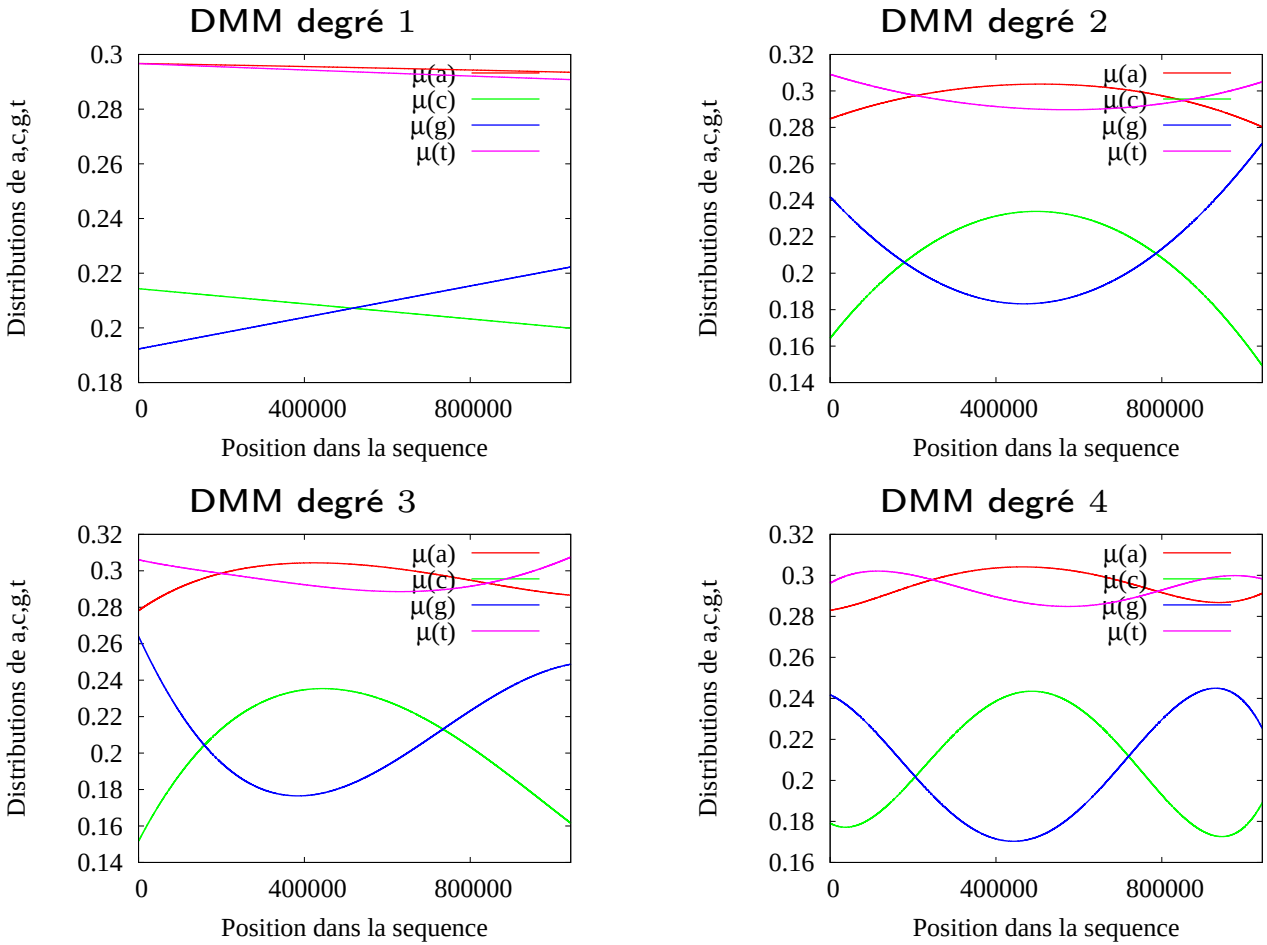




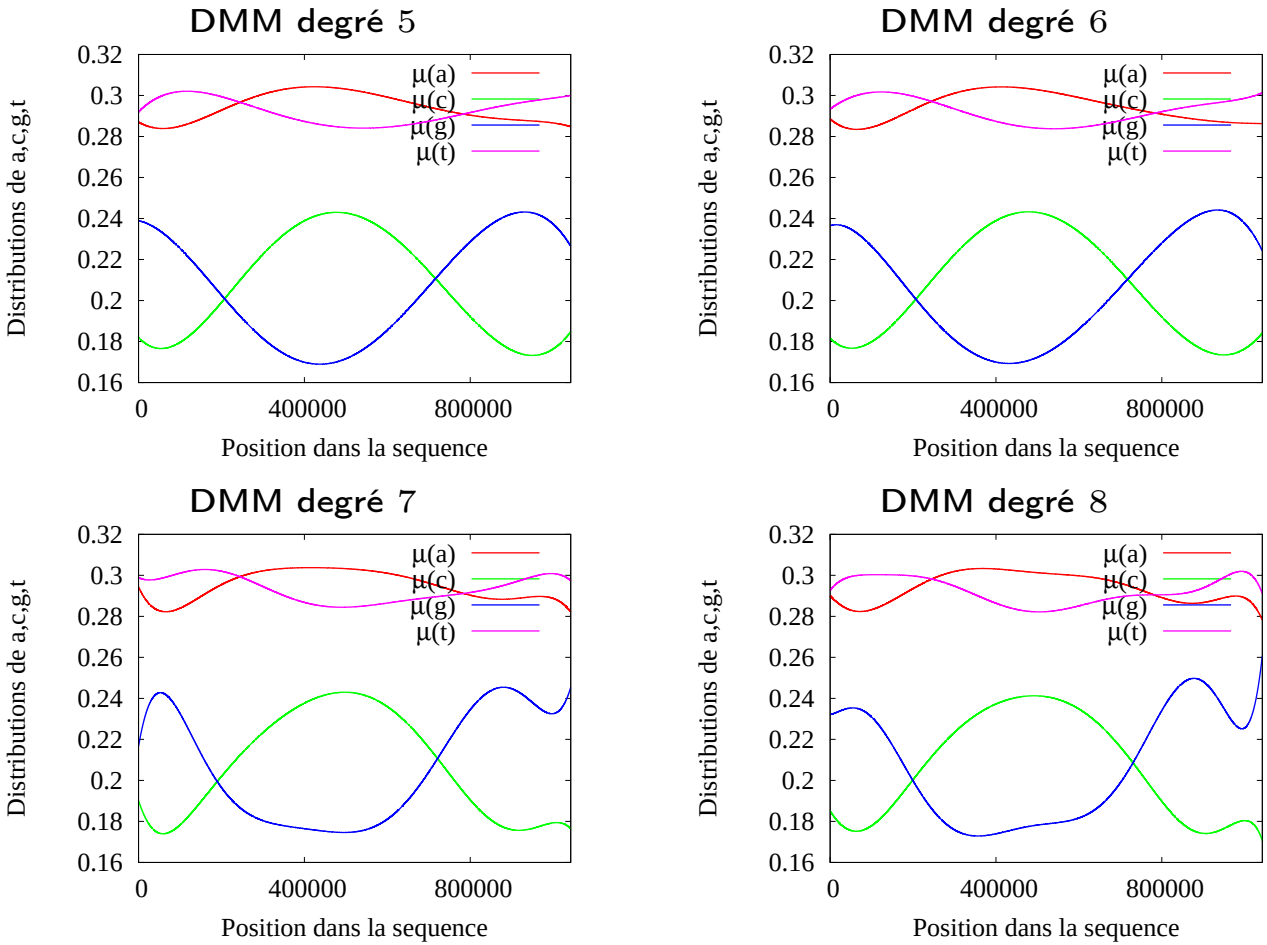




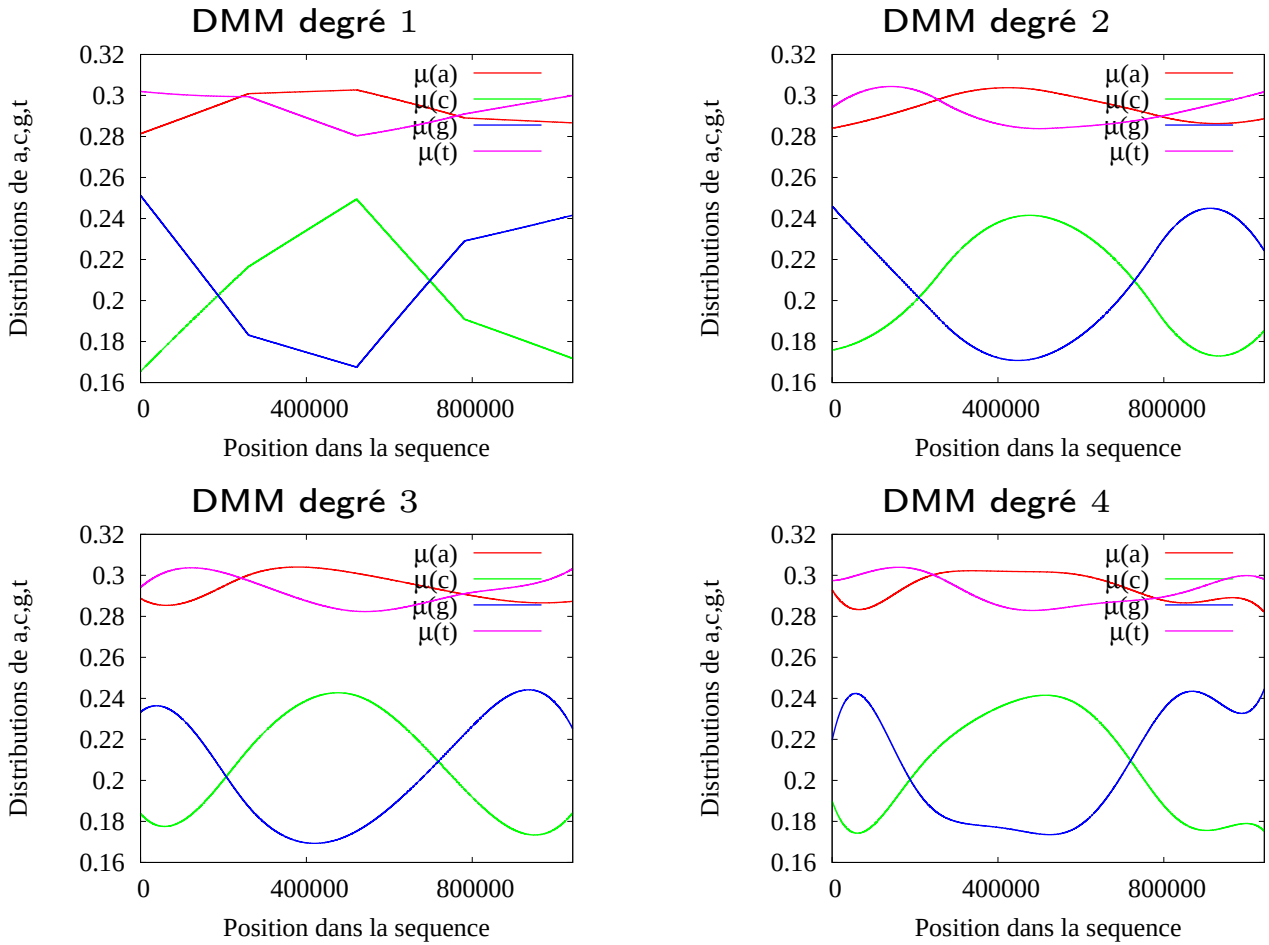
Dérive polynomiale sur *Chlamydia trachomatis*



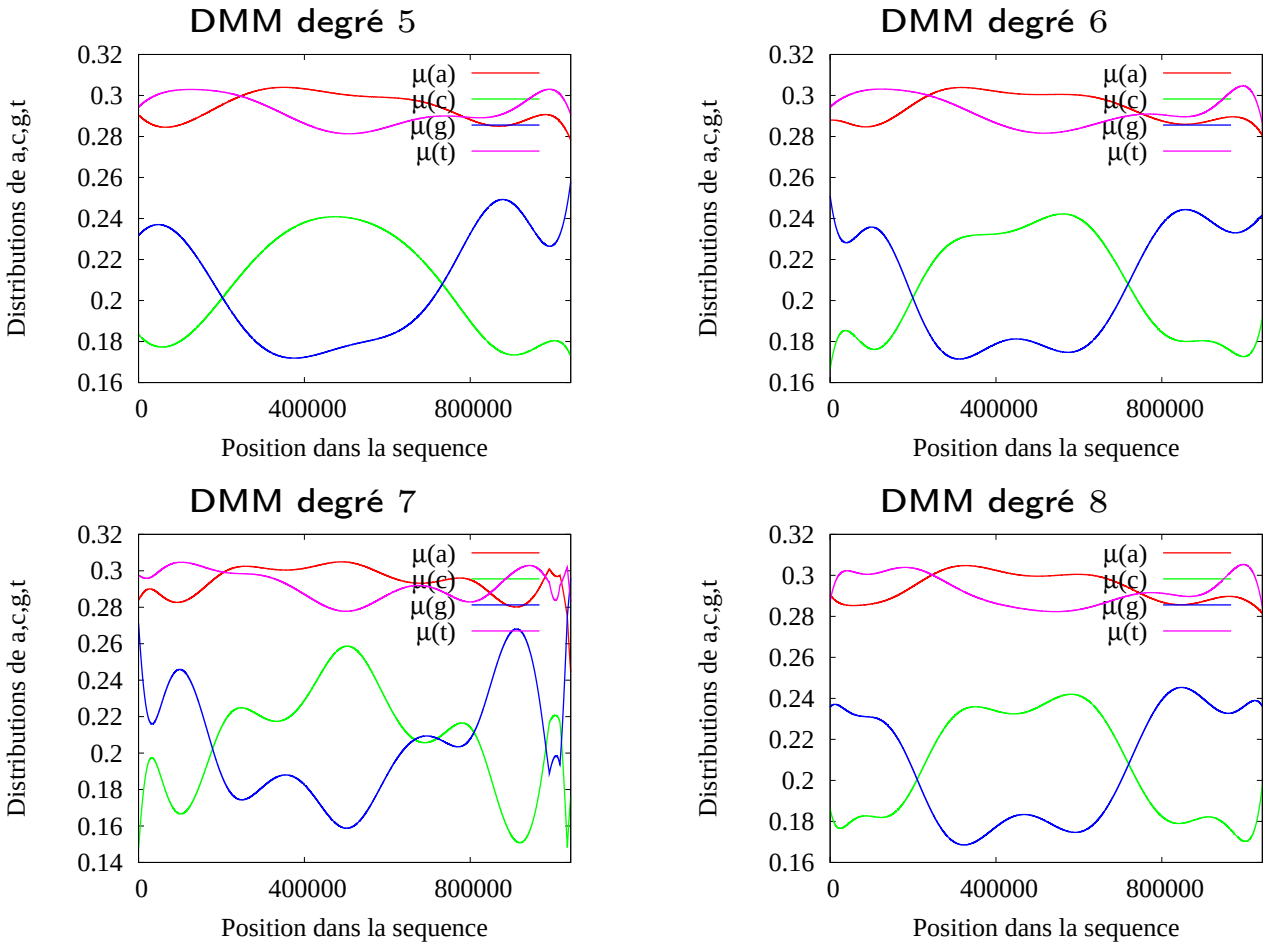
Dérive polynomiale sur *Chlamydia trachomatis*



Dérive par splines polynomiales sur *Chlamydia trachomatis*



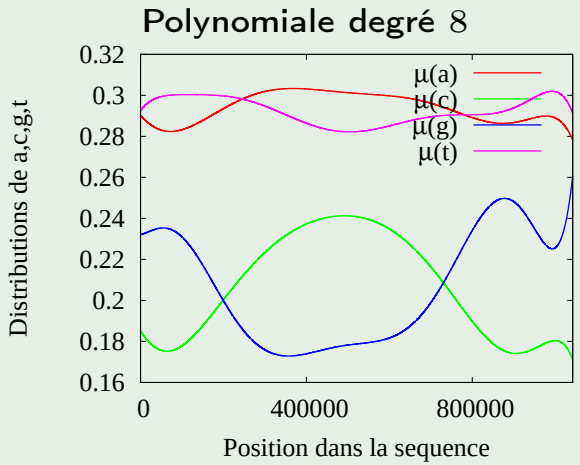
Dérive par splines polynomiales sur *Chlamydia trachomatis*

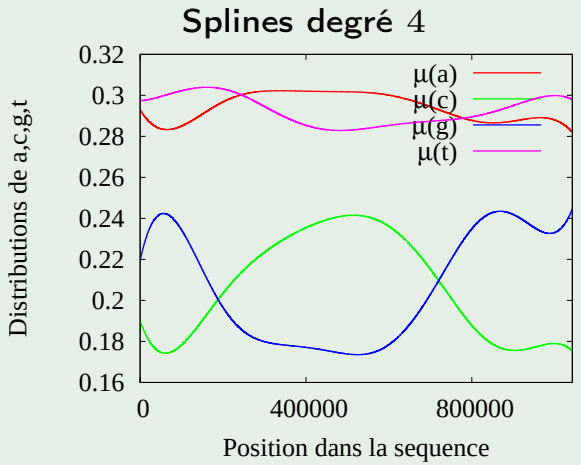


Dérive polynomiale / Dérive par splines polynomiales

Remarque

Une dérive par splines polynomiales de degré 4 avec 4 segments, fournit des courbes semblables à une dérive de degré 8.

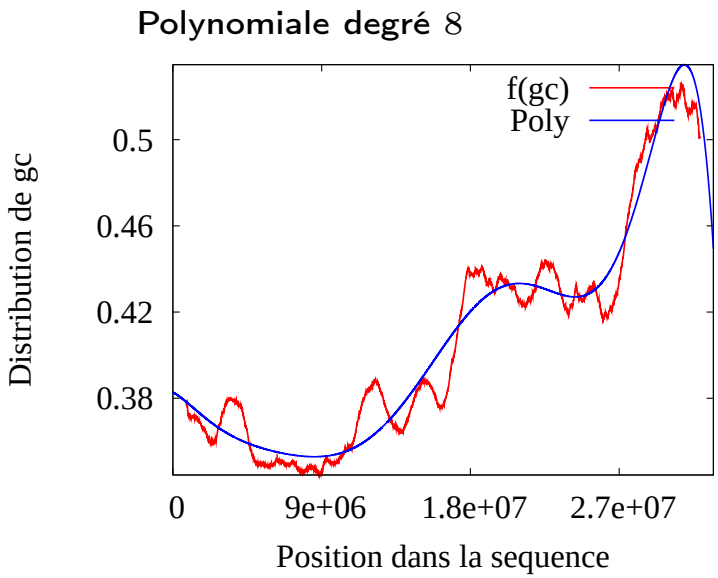


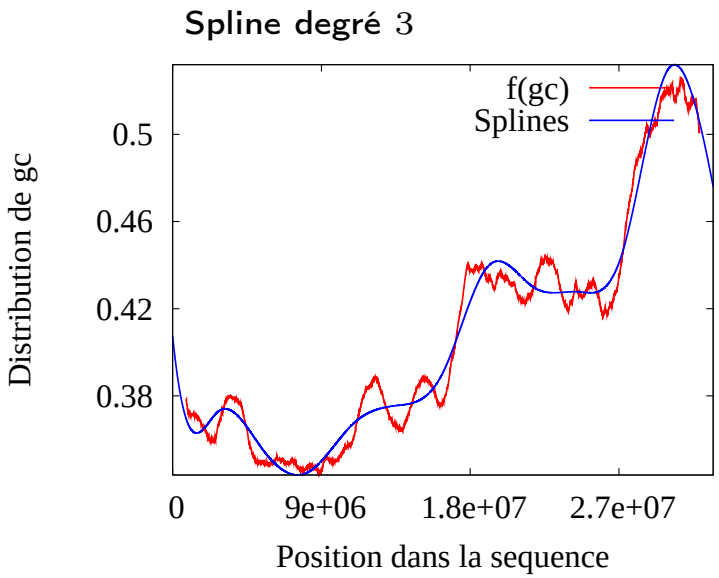


Dérive polynomiale / Dérive par splines polynomiales

Méthodes retenues

- Dérive polynomiale : estimation point par point
- Dérive par splines polynomiales : estimation globale

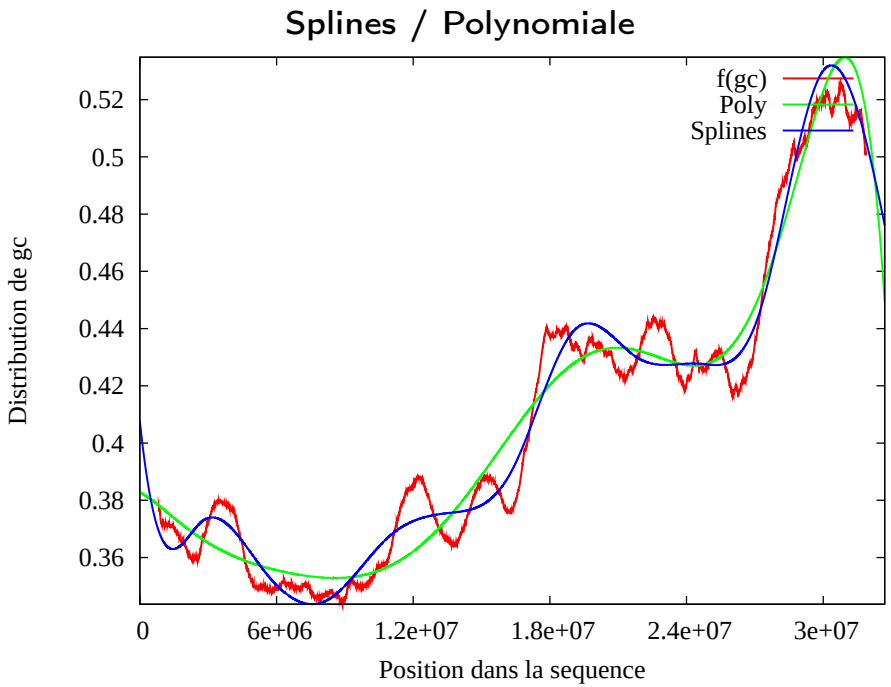




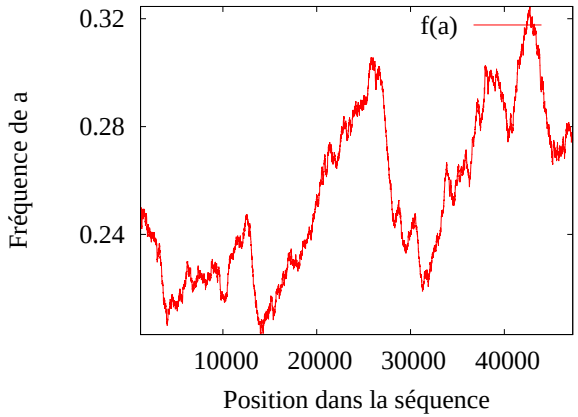
Dérive polynomiale / Dérive par splines polynomiales

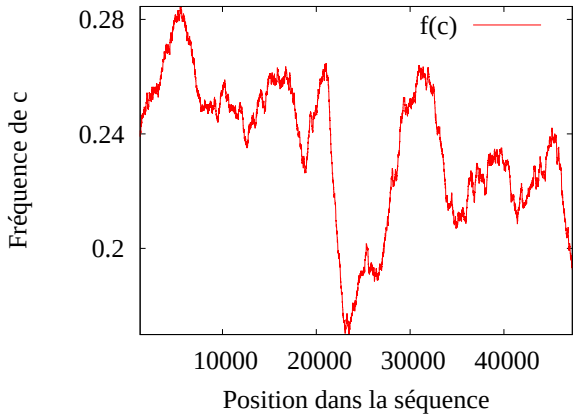
Avantage des splines

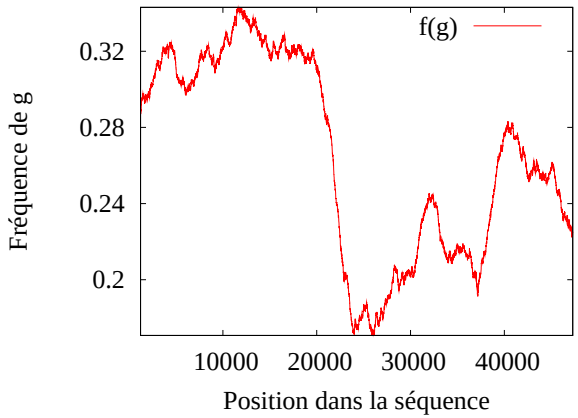
La dérive par splines polynomiales permet des oscillations cohérentes avec les fréquences alors que la dérive polynomiale offre un tracé plus lisse.

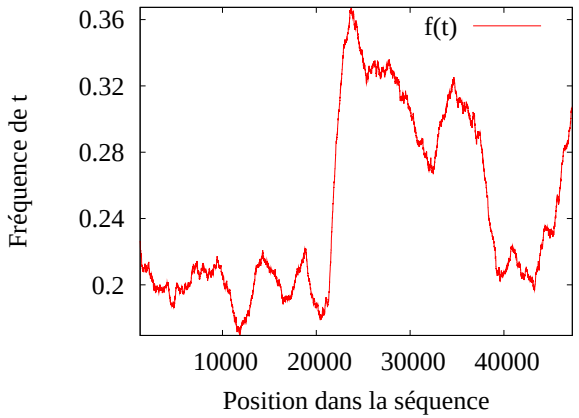


Fréquences des nucléotides

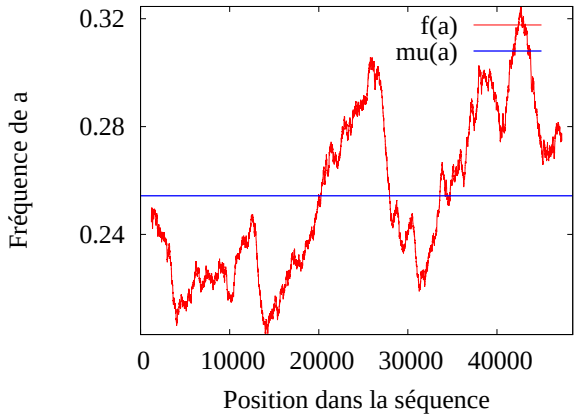


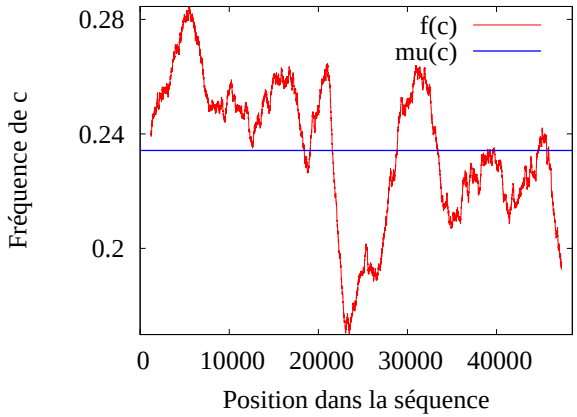


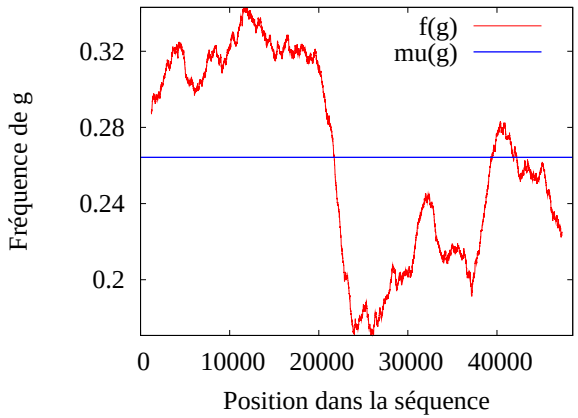


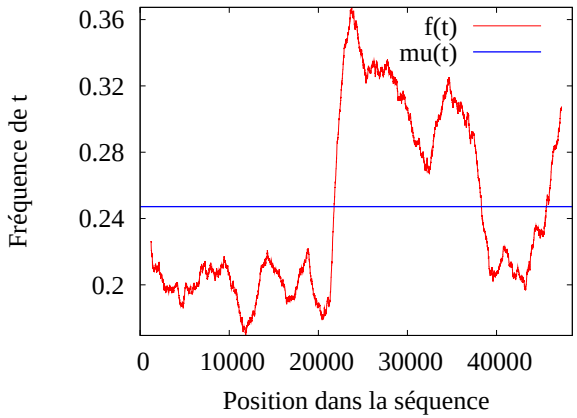


Fréquences des nucléotides et Markov classique

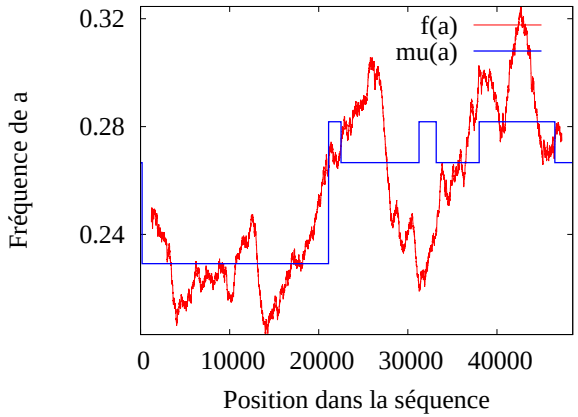


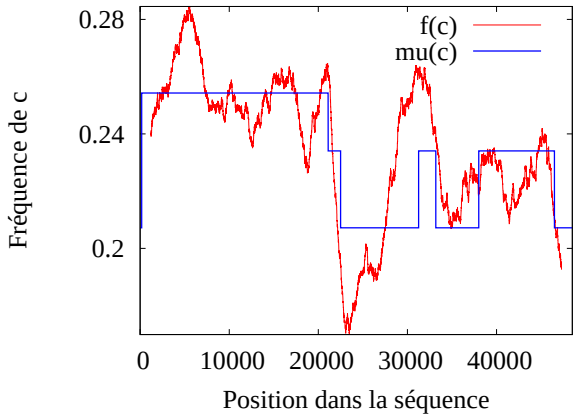


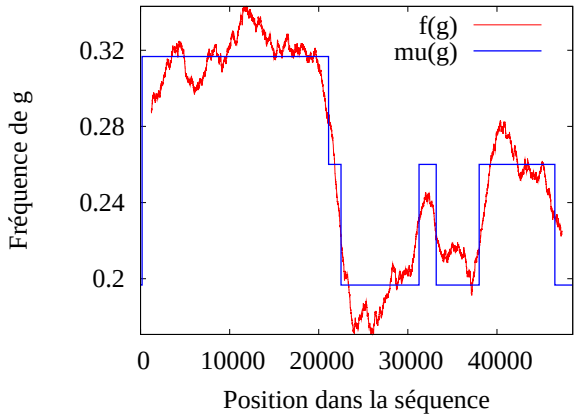


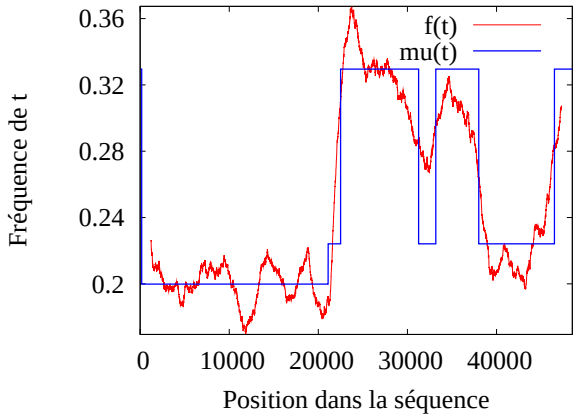


Fréquences des nucléotides et Markov caché

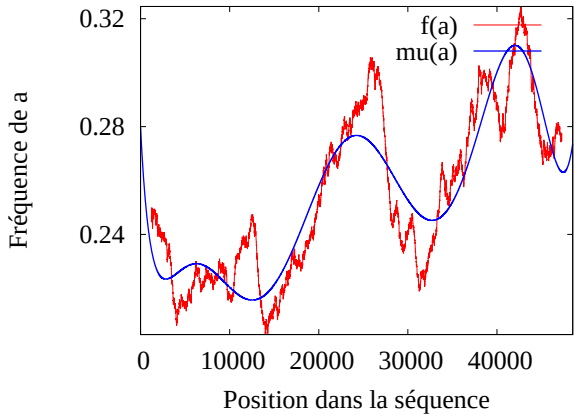


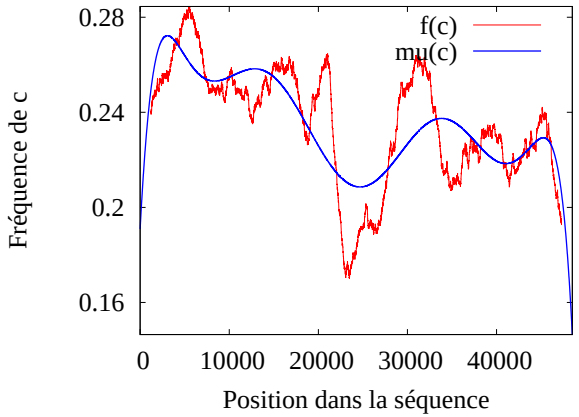


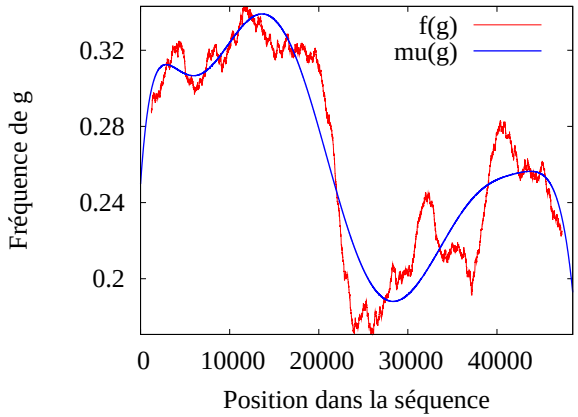


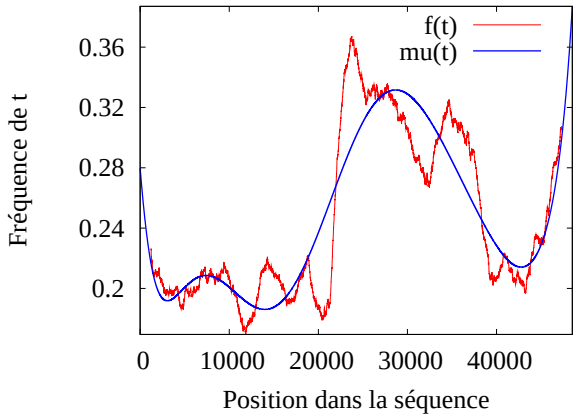


Fréquences des nucléotides et Markov régulé









Distances distributions / fréquences

Distance

$$d_{df} = \sum_{v \in \mathcal{A}} \sum_{t \in \mathcal{P}} (f_t(v) - \mu_t(v))^2.$$

Exemple sur le phage T4

- HMM à 3 états cachés : $d_{df} = 5.873$
- DMM de degré 3 : $d_{df} = 5.865$

Distances distributions / fréquences

Distance

$$d_{df} = \sum_{v \in \mathcal{A}} \sum_{t \in \mathcal{P}} (f_t(v) - \mu_t(v))^2.$$

Exemple sur le phage T4

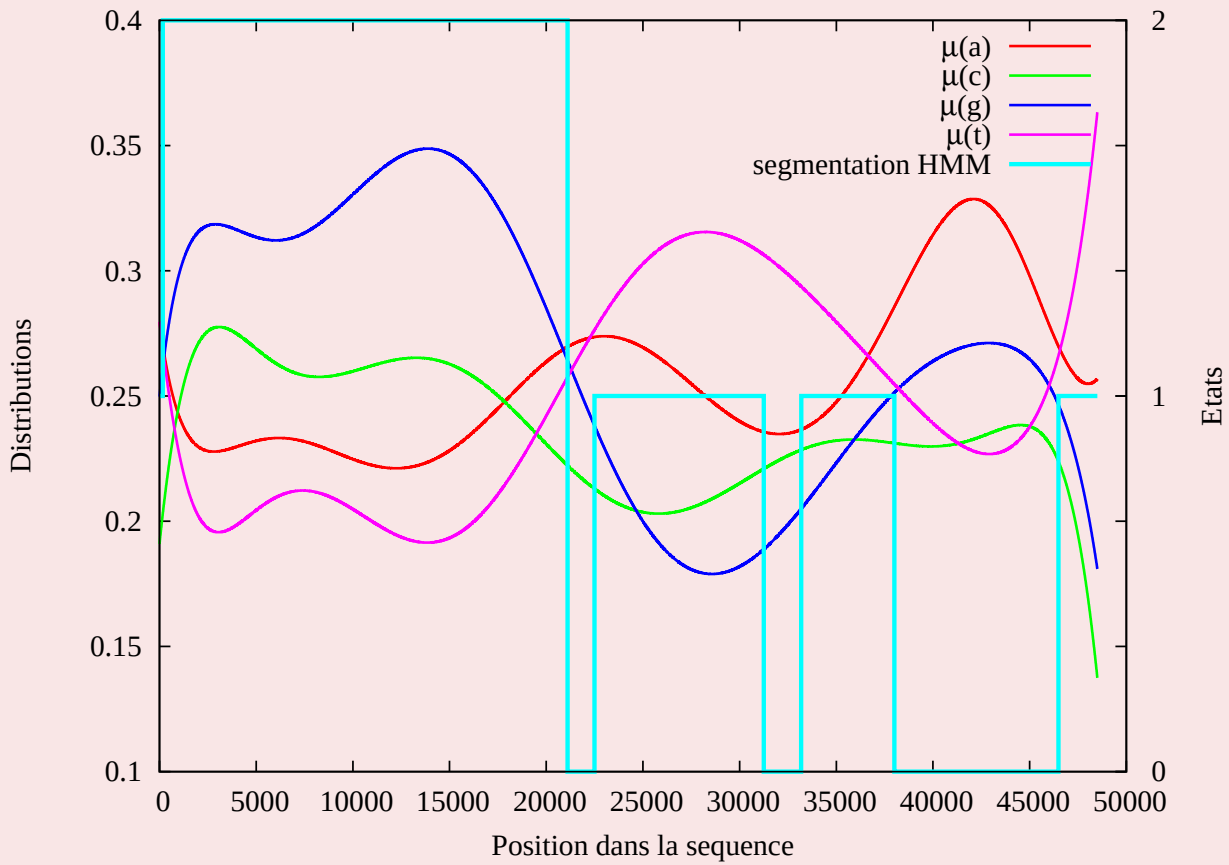
- HMM à 3 états cachés : $d_{df} = 5.873$
- DMM de degré 3 : $d_{df} = 5.865$
- DMM de degré 8 : $d_{df} = 3.391$

Conclusion

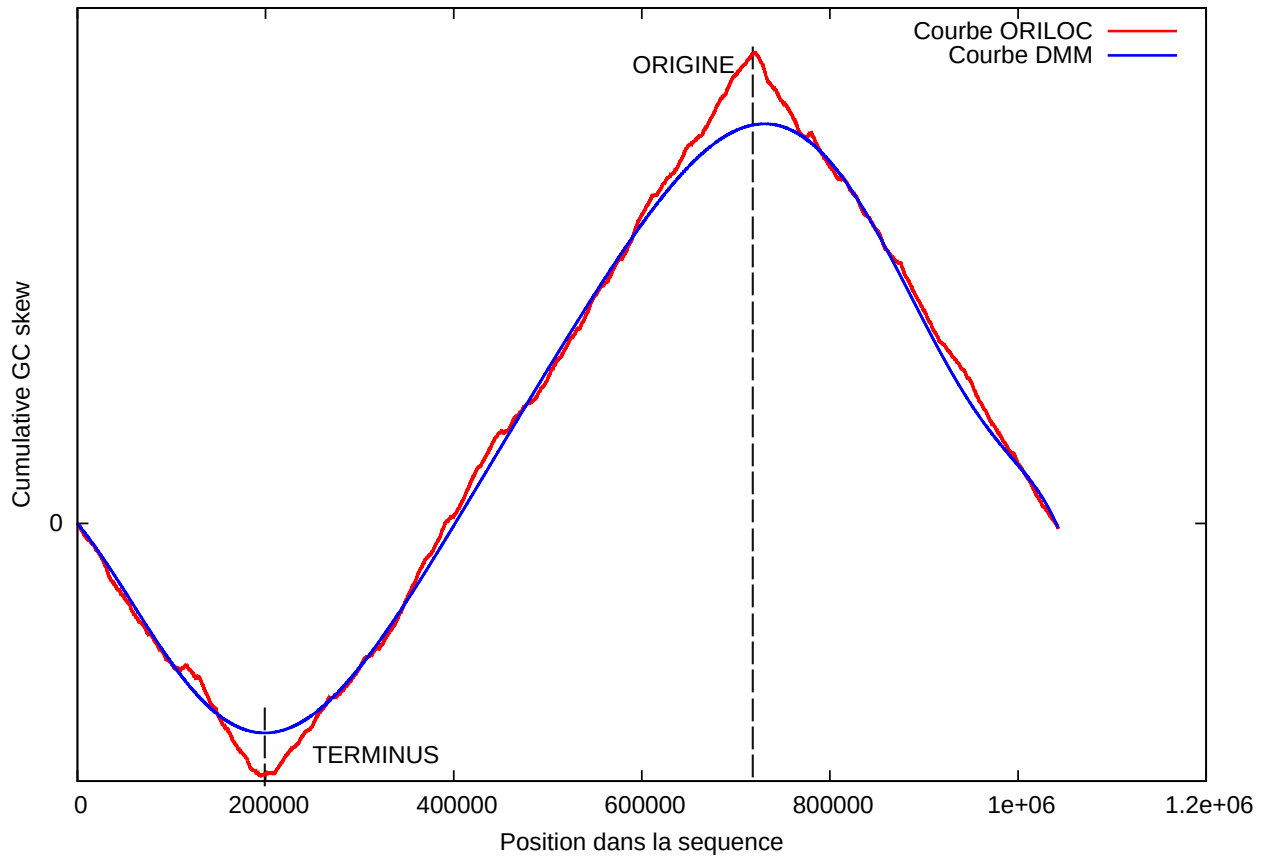
Les modèles régulés offre une modélisation plus flexible, plus proche de la réalité.

Lois stationnaires / HMM

Limite des HMM



Chlamydia trachomatis : Origine de réplication

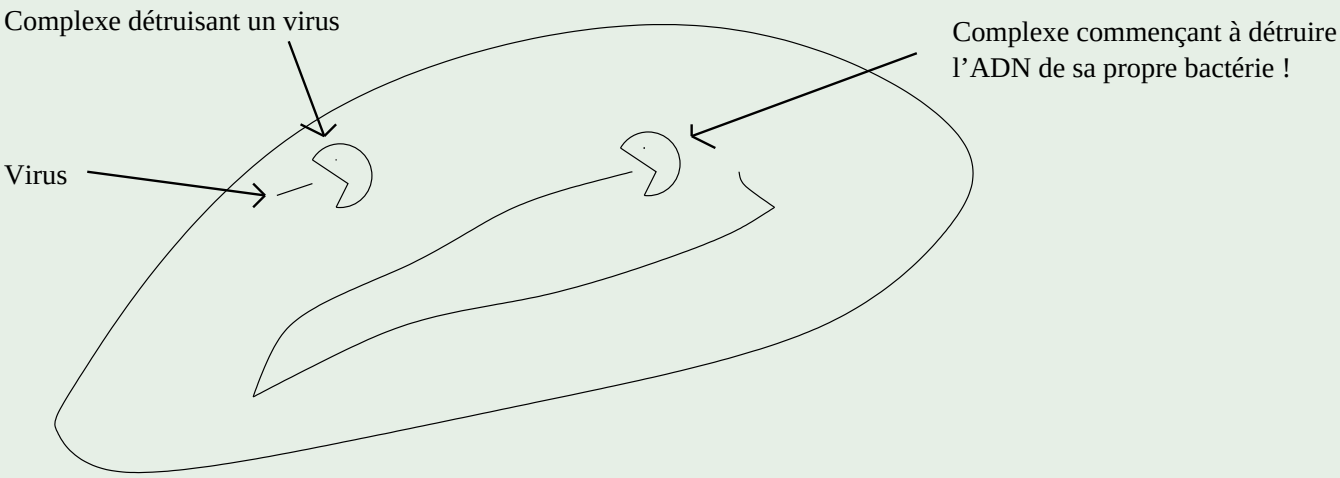


Mots exceptionnels

Idées basiques

- Mot sur-représenté : important pour la survie de l'organisme.
- Mot sous-représenté : nuit à l'organisme.

Le Chi



p-valeurs

Définition

$$S = \begin{cases} -\log_{10} \mathbb{P}(N > N_{obs}) & \text{si } N_{obs} > \mathbb{E}(N) \\ +\log_{10} \mathbb{P}(N < N_{obs}) & \text{si } N_{obs} < \mathbb{E}(N) \end{cases} .$$

Le Chi de *E. coli* : gctggtgg

Ordre	Degré	Espérance	S
1	0	70.10	240.814
1	1	70.26	240.398
1	2	71.88	238.766
1	3	71.87	238.774
1	8	71.94	238.605
2	0	173.84	88.902
2	1	174.03	88.747
2	2	175.16	87.837
2	3	175.10	87.881
2	8	175.31	87.717

Choix d'un modèle

Classification des mots de taille 5 dans le génome complet du *phage Lambda*

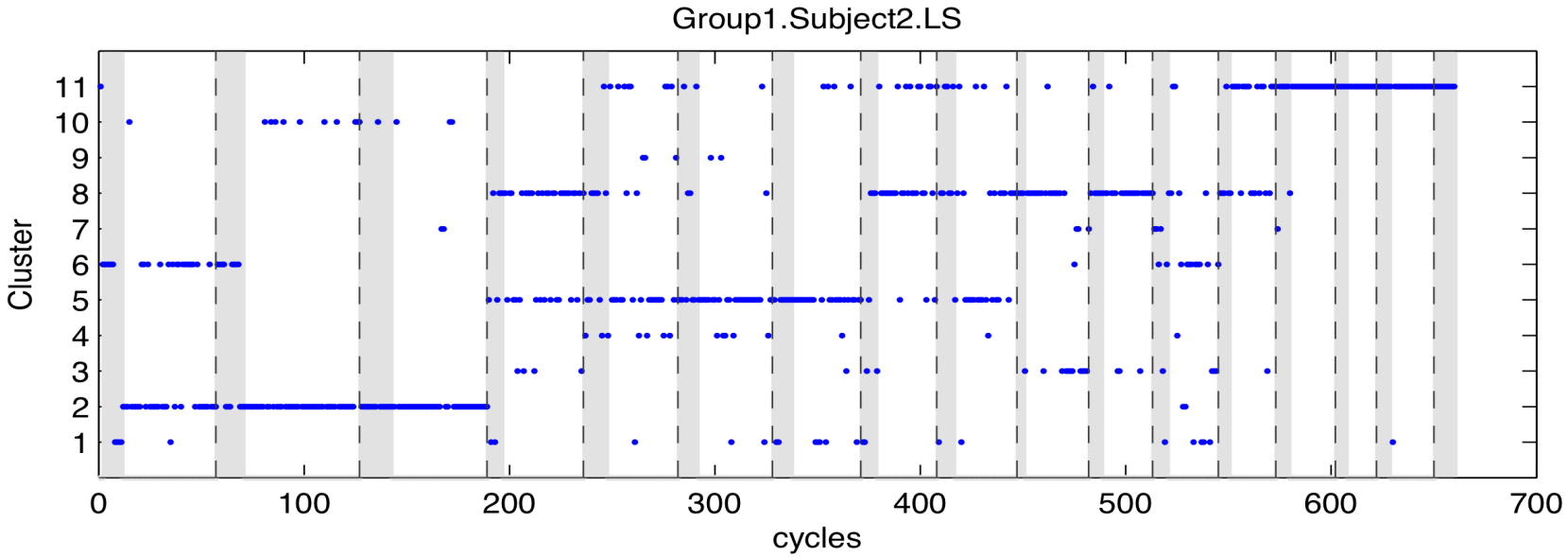
MM				HMM 3 états				DMM degré 1			
Mots	N_{obs}	$\mathbb{E}(N)$	S	Mots	N_{obs}	$\mathbb{E}(N)$	S	Mots	N_{obs}	$\mathbb{E}(N)$	S
aattg	32	88.22	−11.41	aattg	32	83.38	−10.07	aattg	32	86.53	−10.94
ttggg	20	65.12	−10.33	acttg	13	47.59	−8.57	ttgga	21	64.94	−9.76
ttgga	21	66.70	−10.29	tctag	2	24.60	−8.19	ttggg	20	62.94	−9.66
acttg	13	50.74	−9.59	ttgga	21	59.47	−8.15	acttg	13	50.27	−9.44
taggg	3	29.60	−9.21	tcgag	9	39.01	−8.11	tcgag	9	40.69	−8.68
gccgg	114	53.97	12.13	gctgg	127	65.44	14.23	gctgg	127	64.80	11.77
ctgaa	124	61.02	12.16	ctgaa	124	61.34	14.90	ctgaa	124	60.85	12.21
tccgg	100	39.98	15.08	ccgga	112	44.00	20.58	tccgg	100	38.81	16.18
ccgga	112	43.11	17.93	tccgg	100	36.50	20.65	ccgga	112	43.57	18.10
gcaga	141	57.51	20.20	gcaga	141	58.35	22.66	gcaga	141	57.59	20.31

Que choisir ?

Un modèle reflétant au mieux la réalité.

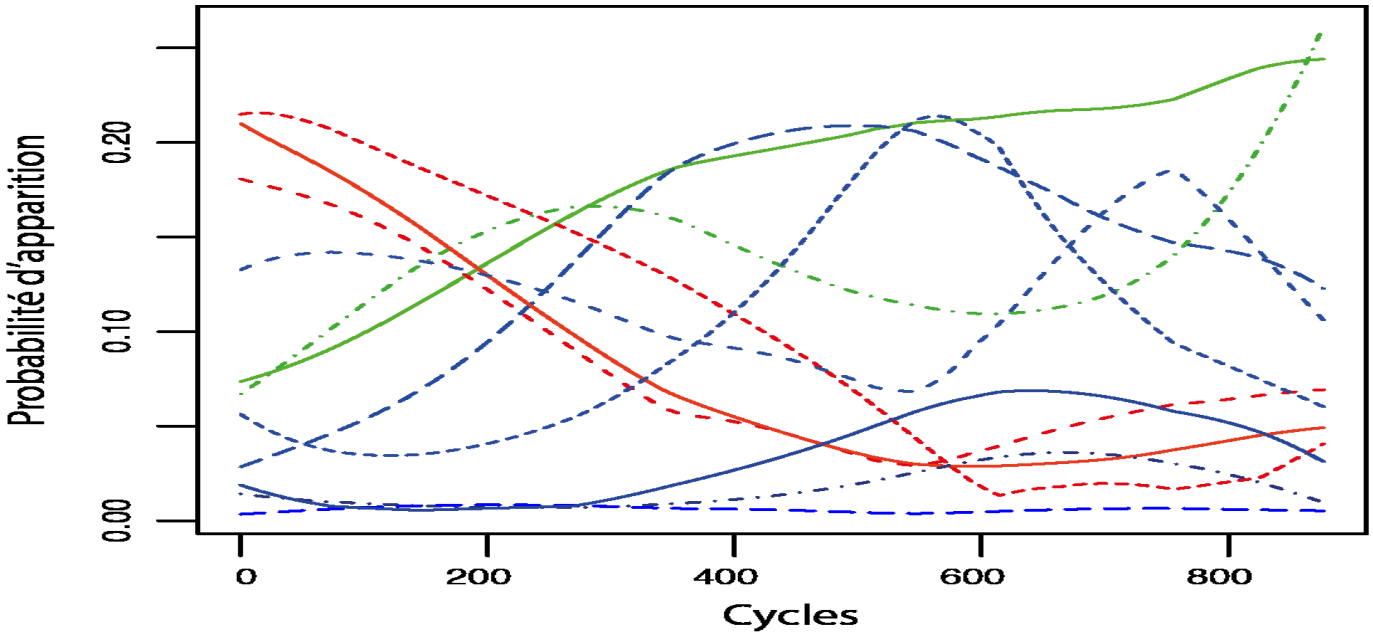
Plan

Natation : apprentissage de la nage (J. Komar)



- 30 nageurs, 20 séances de 250m (2200 cycles) : coordination bras-jambe
- 11 états : 11 comportements moteurs caractéristiques
- Ici, la coordination 11 s'est installée progressivement chez ce nageur

Natation : 3 phénomènes constitutifs de l'apprentissage



- Disparition (rouge) des comportements débutants
- Apparition (vert) de comportement experts
- Exploration motrice (bleu) : apparition et disparition de comportements (nécessité ?)
- Perspective : étude de ces comportements en fonction des conditions d'apprentissage

Fiabilité et survie : définitions (V. Barbu)

Modèle, temps d'entrée et fiabilité

- On note $U = \{1, \ldots, s_1\}$ les états de marche et $D = \{s_1 + 1, \ldots, s\}$ les états de panne.
- Par exemple on partitionne ainsi :

$$\Pi_0 = \begin{array}{c} \\ U \\ D \end{array} \begin{array}{c} \begin{array}{cc} U & D \end{array} \\ \begin{pmatrix} \Pi_0^{UU} & \Pi_0^{UD} \\ \Pi_0^{DU} & \Pi_0^{DD} \end{pmatrix} \end{array}, \qquad \alpha = \begin{array}{c} \begin{array}{cc} U & D \end{array} \\ \begin{pmatrix} \alpha^U & \alpha^D \end{pmatrix} \end{array}.$$

- Temps d'entrée :

$$T_D := \inf\{l \in \mathbb{N}; \quad X_l \in D\}, \text{ with } \inf \emptyset := \infty,$$

- Fiabilité (probabilité d'un fonctionnement sans panne entre 0 et k) :

$$R(k) := \mathbb{P}(T_D > k) = \mathbb{P}(X_l \in U, l = 0, \ldots, k).$$

Fiabilité et survie : Markov régulées

Fiabilité, disponibilité et maintenance au temps k

$$R(k) = \alpha^U \prod_{l=1}^{k} \left(\left(1 - \frac{l}{n} \right) \Pi_0^{UU} + \frac{l}{n} \Pi_1^{UU} \right) \mathbb{1}^U,$$

where $\mathbb{1}^U = (\underbrace{1, \cdots, 1}_{s_1})^\top$.

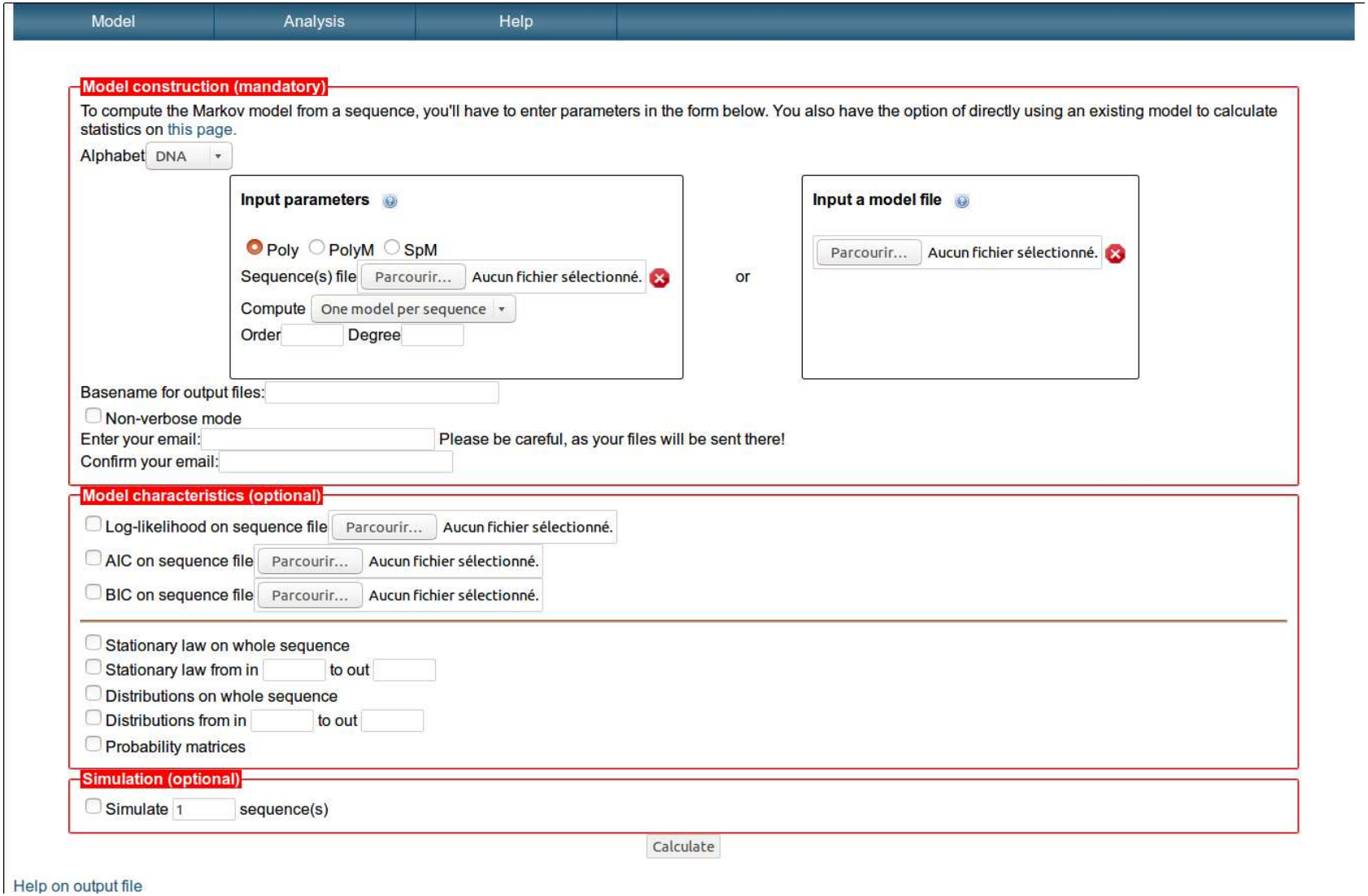
$$A(k) = \alpha \prod_{l=1}^{k} \left(\left(1 - \frac{l}{n} \right) \Pi_0 + \frac{l}{n} \Pi_1 \right) \mathbb{1}^{E,U},$$

where $\mathbb{1}^{E,U} = (\underbrace{1, \cdots, 1}_{s_1}, \underbrace{0, \cdots, 0}_{s - s_1})^\top$.

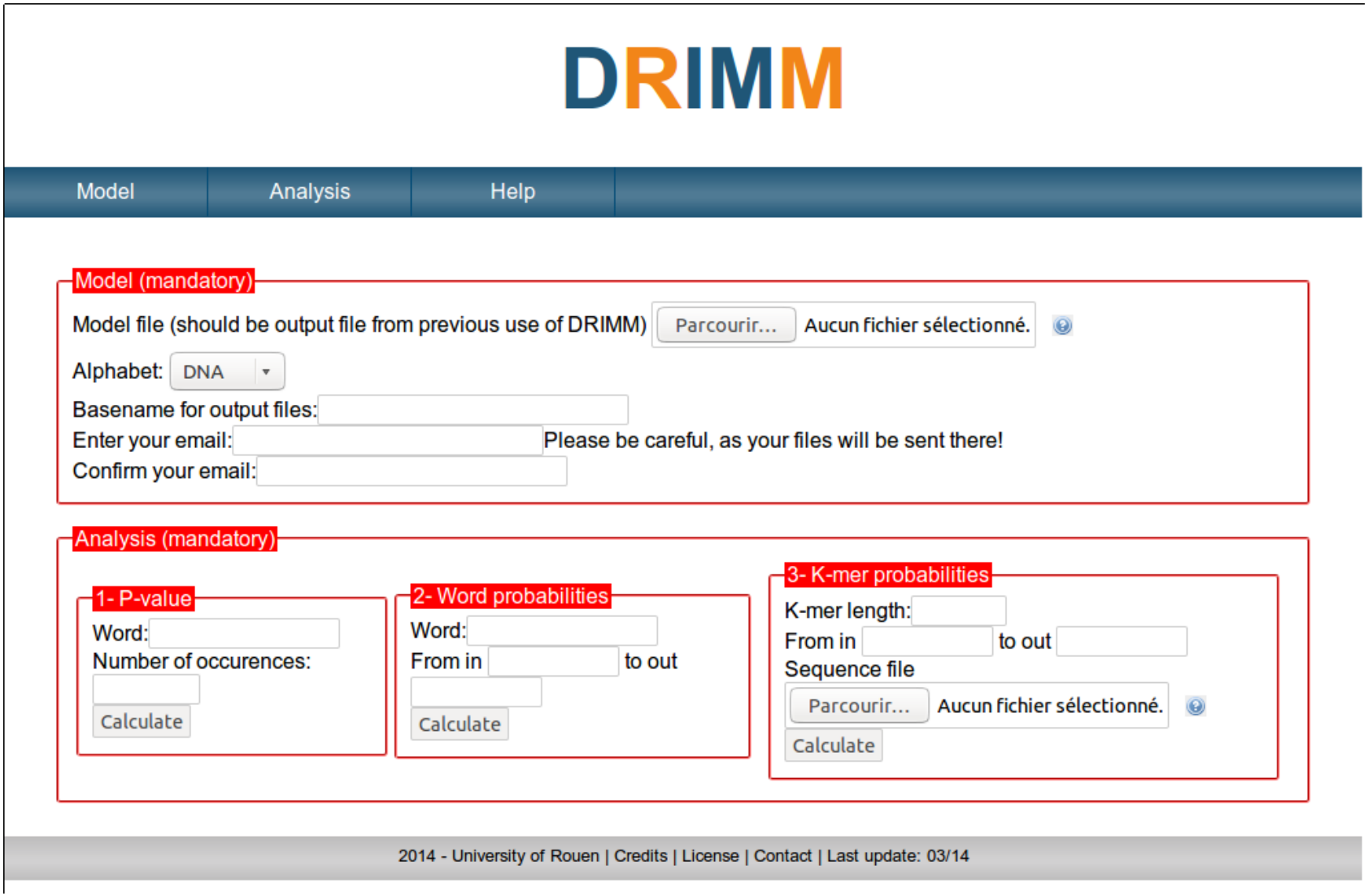
$$M(k) = 1 - \alpha^D \prod_{l=1}^{k} \left(\left(1 - \frac{l}{n} \right) \Pi_0^{DD} + \frac{l}{n} \Pi_1^{DD} \right) \mathbb{1}^D,$$

where $\mathbb{1}^D = (\underbrace{1, \cdots, 1}_{s - s_1})^\top$.

Interface : modélisation (A. Lefèbvre)



Interface : analyses



Conclusion et perspectives

Pour conclure : de nouveaux modèles plus proches de la réalité

- Recherche de mots exceptionnels
- Chaînes de Markov régulées + chaînes de Markov cachées (V. Barbu)
- Chaînes de Markov régulées + semi-chaînes de Markov (V. Barbu)
- Chaînes de Markov régulées + chaînes de Markov cachées + semi-chaînes de Markov (V. Barbu)

Référence + Logiciel

- N. Vergne (2008). Drifting Markov Models with Polynomial Drift and Applications to DNA Sequences. *Statistical Applications in Genetics and Molecular Biology*, **7**(1), `http://www.bepress.com/sagmb/vol7/iss1/art6/`
- DRIMM (Drifting Markov Models) : `http://stat.genopole.cnrs.fr/software/drimm`