Modélisation statistique des lignages cellulaires

Mathieu Pradel et Amandine Véber

sous la direction de F.Taddei et G.Paul

Table des matières

Introduction

Le vieillisement chez la bactérie E. Coli

Le problème est posé par l'unité de recherche INSERM U571. Une colonie bactérienne provenant d'une unique cellule initiale peut se représenter comme un arbre binaire : une bactérie se divise en deux bactéries filles, et on peut alors considérer que chaque élément de la colonie (sauf l'ancêtre commun) possède exactement une mère et une soeur. La figure 1 représente le schéma d'organisation obtenu.

La division étant semble-t-il symétrique, deux soeurs, qui possèdent le même bagage génétique, devraient avoir des caractéristiques similaires (croissance, capacité de reproduction, mortalité). Or, lorsqu'on regarde les taux de croissance représentés sous forme de lignages, on observe une hétérogénéité dans le taux de croissance individuel qui n'est pas aléatoire : la cellule ayant hérité du vieux pôle croît moins vite que sa soeur (voir figures 2 et 3). Entre cellules génétiquement identiques, il y a donc une variabilité qui trouve peut-être son origine dans des mécanismes de sous- ou sur-expression de protéines (pour la plupart présentes en petit nombre dans une bactérie) acquis par hérédité ou requis pour faire face à l'évolution de l'environnement. De même, des expériences ont montré que la mort de bactéries était liée à la fois à des phénomènes aléatoires et à des caractéristiques du lignage : à côté de morts isolées, on trouve la mort simultanée des 8 petites-petites filles d'une même bactérie ! Ceci est d'autant plus étonnant que les notions de vieillissement et de mort programmée ne sont pour le moment pas associées à des organismes se divisant de manière morphologiquement symétrique et sans phase juvénile (période de maturation précédant une phase adulte). Pour en savoir plus, voir [6].

La modélisation des lignages cellulaires a pour but d'étudier les corrélations entre les individus qui composent l'arbre de la colonie lorsqu'on suit l'évolution de la valeur prise par une grandeur X (cf figure 3 pour une illustration). On regardera par exemple la corrélation entre mère et fille, entre soeurs, entre cousines plus ou moins éloignées ou au sein d'une génération, afin d'en tirer des indications quant à l'influence de facteurs héréditaires, géographiques et/ou aléatoires sur une grandeur (typiquement, la taille de la cellule lors de sa division). Pour ce faire, on utilise des modèles dits *statistiques* qui permettent d'envisager plusieurs types de variabilités des données.

Différents types de variations

Lorsqu'on s'intéresse à un phénomène réel, il est rare que les observations disponibles soient continues dans le temps. Souvent indirectes, elles sont répétées à intervalles de temps variables et on obtient alors une série discontinue de valeurs prises à différentes dates. La suite obtenue est appelée *série temporelle*, et chaque terme est indicé par la date de l'observation ou par un entier si les dates sont équidistantes. L'objectif est de construire, à partir des données récoltées, un modèle permettant de

– mieux comprendre la nature du système qui a généré la série temporelle ainsi que les interactions entre plusieurs systèmes grâce à des tests d'hypothèses ;
– prévoir au mieux le comportement d'un système identique ;
– évaluer les effets d'une politique de contrôle qui le manipulerait pour en modifier l'évolution.

Les modèles les plus simples sont ceux obtenus par régression linéaire : il s'agit de trouver une relation explicite de la forme $Y_i = a.X_i + b$, qui relie l'évolution de la grandeur Y (dont la série temporelle des valeurs mesurées est notée $(Y_i)_{i\in I}$) à celle d'une autre grandeur X. Pour généraliser au cas où plusieurs variables sont supposées explicatives, on peut écrire chaque X_i comme un vecteur de taille fixe, et a est un vecteur de paramètres qu'il faudra estimer au mieux. Il se pose alors plusieurs problèmes :

– il est quasiment impossible d'obtenir exactement les mêmes résultats après deux expériences, même si les conditions initiales sont identiques. Le système évolue aléatoirement (ou du moins on le voit ainsi) autour d'une moyenne. Pour modéliser cette déviation de la réponse par rapport à la réponse prédite, on introduit une variable aléatoire dite *terme d'erreur*. Pour la construction du modèle, on choisit le type de loi que suit ce terme, puis il peut être intéressant d'en estimer certaines caractéristiques comme la moyenne (qui est généralement choisie nulle) ou la variance. A défaut, on traite ces grandeurs comme des paramètres de nuisance. Par exemple, on pourrait tenter de voir l'influence de la quantité de substrat disponible X sur la vitesse de croissance Y des bactéries d'une colonie. Le modèle le plus simple s'écrit

$$Y_i = a.X_i + e_i \tag{1}$$

où $(e_i)_{i\in I}$ est une suite de variables aléatoires indépendantes et identiquement distribuées de loi normale $\mathcal{N}(0, \sigma^2)$. L'erreur ainsi introduite fait partie des effets fixes sur les mesures : elle est commune à toutes les observations que l'on peut faire, même en changeant de colonie bactérienne ou de conditions d'expériementation, et permet par exemple de prendre en compte les imprécisions dues à la mesure.

– les observations qui composent une série temporelle ne sont généralement pas indépendantes. Par exemple, on peut étudier plusieurs colonies bactériennes et obtenir des vecteurs explicatifs (notés a précédemment) différents pour chaque série temporelle obtenue. Ceci s'explique par les différences innées entre les colonies comme, par exemple, une meilleure capacité de l'une à affronter une pénurie de substrat, alors qu'une autre proliférera plus rapidement dans un milieu très favorable. Les observations de chaque série sont donc corrélées, et il n'est pas toujours possible de mettre au point une expérience contournant ce problème. On tient donc compte de cet effet aléatoire entre les groupes (ici, un groupe est une colonie, mais la répartition pourrait se faire selon le paramètre modifié et regrouper plusieurs colonies) au moyen d'une deuxième variable aléatoire. Pour re-

prendre l'exemple précédent, on écrirait alors

$$Y_{ij} = a.X_{ij} + b_j.Z_j + e_i \tag{2}$$

pour le $i$ème individu de la colonie j. e_i a la même signification et suit la même loi que précédemment, a est toujours un vecteur de paramètres correspondant aux effets fixes, b_j est un vecteur de coefficients dépendant du groupe considéré et qui suit une loi normale $\mathcal{N}(0, B)$, et Z_j est le vecteur des paramètres de la colonie j censés expliquer les effets aléatoires.
On peut aussi modifier la structure du terme d'erreur pour faire apparaître une corrélation temporelle des erreurs (typiquement, l'erreur sur une bactérie corrélée à celle sur sa mère), et la généralité de ces modèles dits *mixtes* permet de conjuguer différents termes d'erreur et d'expliquer la variance des données. En effet, celle-ci dépend à la fois de la variance des effets fixes et de celle des effets aléatoires, ce qui permet de tester la significativité de ces effets sur l'évolution du système.

– spécialement dans le cadre des lignages cellulaires, le nombre d'observations peut être limité (on ne dépasse pas 7 générations par arbre, et il est coûteux de multiplier les expériences). Il est donc nécessaire d'avoir une idée du nombre de données qui suffisent pour obtenir la précision voulue. Selon la méthode d'estimation des paramètres du modèle que l'on veut construire et les lois données aux différents termes d'erreur, il existe des théorèmes généraux quant à la vitesse de convergence des estimateurs et leurs propriétés asymptotiques. Celles-ci servent, outre à estimer les paramètres du modèle, à construire des tests d'hypothèse à taille d'échantillon donnée, tests qui permettent d'obtenir d'autres informations sur le système avec la précision recherchée (jamais infinie, évidemment ...).

Le premier type de modèles présenté dans cet exposé ne tient compte que des effets fixes, mais la structure des termes d'erreur met en avant la corrélation des valeurs présentes et passées. Les modèles BARMA sont une généralisation des modèles ARMA aux données structurées en arbres telles que les lignages cellulaires. Puis on présentera des extensions de ces modèles qui permettent de modifier la structure trop rigide des erreurs et d'ajouter des effets aléatoires aux effets fixes. Enfin, on se penchera sur un type de modèles plus général, les modèles à état, qui conservent le caractère autorégressif du modèle tout en permettant d'introduire une différence entre la grandeur mesurée et celle étudiée.

1 Modèles autorégressifs $BARMA(p, q)$

1.1 Modèles $ARMA(p, q)$

1.1.1 Préliminaires

Considérons une famille de variables aléatoires réelles $(X_t)_{t\in T}$ définies sur un espace de probabilité $(\Omega, \mathcal{F}, P)$, appelé processus stochastique. Dans la suite,

nous nous intéresserons plus particulièrement aux familles indexées par $\mathbf{Z}$, appelées séries temporelles car elles peuvent être interprétées comme une variable temporelle discrète ($t < 0$ correspond au passé, $t > 0$ au futur). Souvent dans cette partie, le paramètre t désignera la $t$ième génération obtenue par division cellulaire à partir d'une cellule mère.

Définition 1 *La fonction de covariance γ_X (ou γ) : $T^2 \to \mathbf{R}$ est définie par*

$$\gamma_X(r,s) = Cov(X_s, X_t) = E[(X_r - EX_r)(X_s - EX_s)] \tag{3}$$

Définition 2 *Un processus stochastique $(X_t)_{t\in T}$ est dit stationnaire (ou stationnaire du deuxième ordre) s'il vérifie les trois conditions suivantes :*

1. $E|X_t|^2 < \infty \qquad \forall t \in \mathbf{Z}$
2. $EX_t = m \qquad \forall t \in \mathbf{Z}$
3. $\gamma(r,s) = \gamma(r+t, s+t) \qquad \forall t,s,r \in \mathbf{Z}$

Remarquons alors que $\gamma(r,s) = \gamma(r-s,0),\ \forall r,s \in \mathbf{Z}$. La fonction d'autocovariance peut donc être vue comme une fonction d'une seule variable $\gamma(h) = \gamma(h,0)$, qui mesure la covariance entre des valeurs distantes de h unités de temps. Cela permet d'introduire la fonction d'autocorrélation

$$\rho(h) = \gamma(h)/\gamma(0) = Corr(X_{t+h}, X_t) \tag{4}$$

1.1.2 Définition du processus $ARMA(p,q)$

Les processus ARMA jouent un rôle important dans la modélisation de séries temporelles. Une raison à cela est leur densité, au sens suivant : pour toute fonction d'autocovariance $\gamma : \mathbf{Z} \to \mathbf{R}$ et pour tout entier n, il existe un processus ARMA $(X_t)_{t\in\mathbf{Z}}$ tel que $\gamma(k) = \gamma_X(k) \quad \forall k \in \{0,\cdots,n\}$.

Définition 3 *Un processus $(\epsilon_t)_{t\in\mathbf{Z}}$ est appelé bruit blanc (white noise), de moyenne nulle et de variance σ^2 ssi pour tout t, $E\epsilon_t = 0$ et si $\gamma_\epsilon(h) = \sigma^2$ pour $h = 0$ et 0 sinon. On note $(\epsilon_t) \sim WN(0,\sigma^2)$.*

Définition 4 *Le processus $(X_t)_{t\in\mathbf{Z}}$ est dit $ARMA(p,q)$ si $(X_t)_{t\in\mathbf{R}}$ est stationnaire et si $\forall t \in \mathbf{Z}$,*

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = \epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_q\epsilon_{t-q} \tag{5}$$

avec $(\epsilon_t) \sim WN(0,\sigma^2)$.

Par extension, $(X_t)_{t\in\mathbf{Z}}$ est dit $ARMA(p,q)$ de moyenne m si $(X_t - m)_{t\in\mathbf{Z}}$ est $ARMA(p,q)$. En notant B^j l'opérateur défini par $B^jX_t = X_{t-j}$, la condition s'écrit

$$\forall t \in \mathbf{Z}, \qquad \phi(B)X_t = \theta(B)\epsilon_t \tag{6}$$

avec $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ et $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$.

A priori, le processus $(X_t)_{t\in\mathbf{Z}}$ dépend des valeurs passées, présentes et futures prises par le bruit blanc. On va voir que sous certaines conditions sur les polynômes ϕ et θ, X_t va pouvoir s'exprimer comme fonction des seules valeurs passées du bruit blanc. Pour cela, on va définir la notion de causalité.

1.2 Unicité de la solution stationnaire

1.2.1 Causalité

Définition 5 *En gardant les mêmes notations, un processus $ARMA(p,q)$ est dit causal s'il existe une suite réelle ou complexe $(\psi_i)_{i\in\mathbf{N}}$ sommable telle que*

$$X_t = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i} \qquad \forall t \in \mathbf{Z} \tag{7}$$

La notion de causalité n'est pas intrinsèque au processus, il faut bien la comprendre comme une relation entre deux processus, $(X_t)_{t\in\mathbf{Z}}$ et $(\epsilon_t)_{t\in\mathbf{Z}}$; ainsi, un processus causal peut être déterminé à partir de la connaissance des valeurs prises jusqu'à l'instant t par une autre variable. Souvent dans cette partie, cette dernière représente une caractéristique de l'environnement dans lequel baignent les cellules.

On va maintenant s'attacher à caractériser la causalité des processus ARMA. Pour cela, on n'a besoin que d'une proposition que l'on donne sans preuve.

Proposition 1 *Si $(X_t)_{t\in\mathbf{Z}}$ est un processus stochastique tel que $\sup_t E|X_t| < \infty$, et si*

$$\sum_{i=-\infty}^{+\infty} |\psi_i| < \infty$$

alors la série

$$\sum_{i=-\infty}^{+\infty} \psi_i X_{t-i} = \psi(B) X_t$$

converge absolument, presque sûrement.

Théorème 1 *Soit $(X_t)_{t\in\mathbf{Z}}$ un processus $ARMA(p,q)$ tel que les zéros des polynômes ϕ et θ sont distincts. Alors le processus est causal ssi ϕ ne s'annule pas sur la boule unité de $\mathbf{C}$. En outre, la suite $(\psi_i)_{i\in\mathbf{N}}$ est donnée par le développement en série entière en 0 de la fonction rationnelle θ/ϕ (ie*

$$\sum_{i=0}^{\infty} \psi_i z^i = \theta(z)/\phi(z)$$

pour tout z tel que $|z| \leq 1$).

Démonstration : Dans un premier temps, supposons que $\phi(z) \neq 0$ si $|z| \leq 1$. Alors la fonction $1/\phi(z)$ est développable en série entière en 0 de rayon de convergence strictement supérieur à 1. En d'autres termes, il existe $\delta > 0$ tel que

$$1/\phi(z) = \sum_{i=0}^{\infty} \zeta_i z^i \qquad \forall |z| < 1+\delta. \tag{8}$$

Mais alors, il existe une constante K réelle telle que $|\zeta_i| < K(1+\delta/2)^{-i}$ pour tout entier i. Il vient donc

$$\sum_{i=0}^{\infty} |\zeta_i| < \infty$$

et $\zeta(z)\phi(z) = 1$ pour tout z dans la boule unité de $\mathbf{C}$. On peut alors appliquer l'opérateur $\zeta(B)$ à l'égalité $\phi(B)X_t = \theta(B)\epsilon_t$ pour finalement obtenir : $X_t = \zeta(B)\theta(B)\epsilon_t$.

On a bien obtenu la représentation cherchée :

$$X_t = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}$$

.

Réciproquement, supposons que le processus $(X_t)_{t\in\mathbf{Z}}$ est causal. On a donc

$$X_t = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}$$

pour une certaine suite $(\psi_i)_{i\in\mathbf{N}}$ dont la somme est absolument convergente. Il vient alors, en remplaçant X_t dans l'égalité $\phi(B)X_t = \theta(B)\epsilon_t$:

$$\theta(B)\epsilon_t = \phi(B)\psi(B)\epsilon_t. \tag{9}$$

En notant

$$\eta(z) = \phi(z)\psi(z) = \sum_{i=0}^{\infty} \eta_i z^i,$$

l'égalité devient

$$\sum_{i=0}^{\infty} \theta_i \epsilon_{t-i} = \sum_{i=0}^{\infty} \eta_i \epsilon_{t-i}.$$

Soit k un entier. Multiplions les membres de l'égalité par ϵ_{t-k} : en prenant la covariance de chaque terme et en utilisant le fait que $Cov(\epsilon_i, \epsilon_j) = 0$ pour $i \neq j$ car $(\epsilon_t)_{t\in\mathbf{Z}}$ est un bruit blanc, on trouve finalement que $\eta_k = \theta_k$ si $k \leq q$ et 0 sinon. Il vient donc pour tout nombre complexe z de module inférieur ou égal à 1

$$\theta(z) = \eta(z) = \phi(z)\psi(z). \tag{10}$$

Puisque θ et ϕ ne s'annulent pas simultanément et que la quantité $|\psi(z)|$ est finie pour tout z de module inférieur à 1, ϕ ne peut s'annuler sur la boule unité de $\mathbf{C}$.

Remarque 1 *Les cas où θ et ϕ ont des zéros communs se traite facilement : on se ramène au cas précédent lorsqu'il n'y a pas de zéro sur le cercle unité. Dans le cas contraire, l'équation ARMA peut avoir plus d'une solution stationnaire,*

ce qui ne nous intéressera pas. En outre, le théorème montre que si $(X_t)_{t\in\mathbf{Z}}$ *est une solution stationnaire de l'équation ARMA avec* $\phi(z) \neq 0\ \forall |z| \leq 1$*, alors*

$$X_t = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}$$

avec les notations précédentes. Inversement, si on a

$$X_t = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i},$$

alors $(X_t)_{t\in\mathbf{Z}}$ *vérifie l'équation ARMA. Cette somme est donc l'unique solution stationnaire de l'équation ARMA.*

Remarque 2 *En pratique, puisque* (ψ_i) *est sommable, les valeurs de la suite sont rapidement négligeables et la somme utilisée est tronquée plus ou moins arbitrairement pour alléger les calculs.*

On va maintenant introduire une notion proche de celle de causalité : l'inversibilité.

1.2.2 Inversibilité

Définition 6 *Un processus* $ARMA(p,q)$ *défini par l'équation* $\phi(B)X_t = \theta(B)\epsilon_t$ *est dit inversible s'il existe une suite* $(\alpha_i)_{i\in\mathbf{N}}$ *sommable telle que*

$$\epsilon_t = \sum_{i=0}^{\infty} \alpha_i X_{t-i} \qquad \forall t \in \mathbf{Z}. \tag{11}$$

ϵ_t *s'interprète ici comme l'innovation du processus.*

De la même manière, cette définition caractérise une propriété reliant les processus $(X_t)_{t\in\mathbf{Z}}$ et $(\epsilon_t)_{t\in\mathbf{Z}}$. Cette fois, les valeurs passées de X expliquent l'innovation présente ϵ_t. Nous allons de même caractériser l'inversibilité pour un processus ARMA.

Théorème 2 *Soit* $(X_t)_{t\in\mathbf{Z}}$ *un processus* $ARMA(p,q)$ *pour lequel* ϕ *et* θ *n'ont pas de zéro commun. Alors le processus* $(X_t)_{t\in\mathbf{Z}}$ *est inversible ssi* $\theta(z) \neq 0$ *pour tout* z *dans la boule unité de* $\mathbf{C}$*.*

Démonstration : Elle est quasiment identique à précédente.

Théorème 3 *Si* $\phi(z) \neq 0$ *pour tout complexe* z *de module 1, alors l'équation ARMA* $\phi(B)X_t = \theta(B)\epsilon_t$ *possède une et une seule solution stationnaire :*

$$X_t = \sum_{i=-\infty}^{+\infty} \psi_i \epsilon_{t-i} \tag{12}$$

où les coefficients ψ_i *sont déterminés par le développement en série entière en 0 de* θ/ϕ*.*

Démonstration : Le processus $(X_t)_{t\in\mathbf{Z}}$ est bien défini par la proposition précédente, et est stationnaire. Appliquons l'opérateur $\phi(B)$. On obtient bien :

$$\phi(B)X_t = \theta(B)\epsilon_t \tag{13}$$

Inversement, soit $(X_t)_{t\in\mathbf{Z}}$ une solution stationnaire. Puisque $\phi(z) \neq 0$ pour tout complexe z de module 1, il existe $\delta > 0$ tel que la série

$$\sum \zeta_i z^i = \theta(z)/\phi(z)$$

converge absolument pour tout z tel que $1/\delta < |z| < \delta$. En appliquant l'opérateur $\zeta(B)$ à l'égalité précédente, on retrouve $X_t = \psi(B)\epsilon_t$.

1.3 Modèle $BARMA(p, q)$

Nous allons maintenant nous intéresser plus particulièrement à un modèle dérivé du modèle $ARMA(p, q)$: le modèle $BARMA(p, q)$, pour *bifurcating autoregressive model*, plus approprié pour modéliser les lignages cellulaires.

Partant d'une cellule mère, on arrive à observer sur une échelle de temps assez large les divisions successives, tout en collectant quelques informations. Celles-ci permettent ensuite d'estimer les liens entre mères et filles ou entre soeurs. Ces liens peuvent avoir différentes origines : héritage légué par la mère aux deux filles (qui se retrouve dans la corrélation mère-fille), environnement dans lequel les deux cellules filles se développent (indiqué par la corrélation soeur-soeur conditionnellement à la valeur correspondant à la mère), ...

Ainsi, Cowen et Staudte ont introduit le modèle $BAR(1)$ défini ainsi : considérant une cellule t, ses cellules filles sont notées $2t$ et $2t + 1$. Partant d'une caractéristique initiale X_1 associée à l'individu 1, on définit alors

$$X_t = \theta X_{[t/2]} + \epsilon_t \tag{14}$$

Les couples de variables aléatoires $(\epsilon_{2t}, \epsilon_{2t+1})$ sont indépendants, identiquement distribués et suivent une loi normale. L'hypothèse d'indépendance est motivée par le fait que des cellules soeurs se développent dans un environnement similaire, alors que des membres plus éloignés de la lignée, comme des cellules cousines, vivent dans des environnements différents.

Nous allons maintenant voir une définition plus générale de ce modèle.

1.3.1 Définition

Définition 7 *On définit un nouvel opérateur de bifurcation, b^r pour r entier par :*

$$b^r X_t = X_{[t/2^r]*} \quad si\ t > 0, \qquad X_{t-r}\ sinon \tag{15}$$

avec $[t/2^r]* = [t/2^r]$ si $t/2^r \geq 1$ et $[t/2^r]* = [ln(t/2^r)] + 1$ sinon.

Ainsi, si X_t représente la valeur d'une caractéristique de la cellule t, $b^r X_t$ représente celle du $r$ième ancêtre de t.

Définition 8 *De la même manière que pour le modèle ARMA, un processus* $(X_t)_{t\in\mathbf{Z}}$ *suit le modèle* $BARMA(p,q)$ *s'il vérifie l'équation*

$$\phi(b)X_t = \theta(b)\epsilon_t. \tag{16}$$

Supposons désormais que le processus $Y_i = X_{[t/2^i]*}$ est causal et inversible. On peut alors écrire

$$X_t = \sum_{i=0}^{\infty} \psi_i \epsilon_{[t/2^i]*}$$

avec

$$\sum_{i=0}^{\infty} \psi_i z^i = \theta(z)/\phi(z)$$

.

Notons par ailleurs $c(s,t)$ le dernier ancêtre commun à s et t (ie le plus récent dans le temps), et $g_s(s,t) = g_s(t,s)$ le nombre de générations entre $c(s,t)$ et s. Alors, pour le modèle de Cowan-Staudte, on obtient par un calcul direct :

$$Cov(X_s, X_t) = \sigma^2 \sum_{i=0}^{\infty} \psi_{i+g_t}\psi_{i+g_s} + \phi\sigma^2\psi_{g_t-1}\psi_{g_s-1};$$

le second terme est nul si s est un descendant direct de t, ce qui donne un moyen assez simple de calculer la covariance, connaissant les coefficients ψ.

1.3.2 Modèle à distance

Le modèle précédent souffre toutefois de certaines insuffisances : par exemple, il sous-estime les liens entre cousins. Une des raisons est qu'il ne prend pas en compte le fait que ces cellules sont issues d'une même génération. Pour y remédier, un moyen est d'introduire un modèle qui autorise une covariance due à l'environnement décroissant avec l'éloignement des individus d'une génération.

Pour cela, on introduit une distance entre les individus, standardisée de sorte que $d(s,s) = 0$ et $d(2s, 2s+1) = 1$, par exemple $d(s,t) = \frac{g_t(s,t)+g_s(s,t)}{2}$.

On peut alors reprendre l'exemple du modèle de Cowan-Staudte, ainsi modifié :

- si $d(s,t) = 0$, $Cov(\epsilon_s, \epsilon_t) = \sigma^2$
- si $d(s,t) = 1$, $Cov(\epsilon_s, \epsilon_t) = \phi\sigma^2$
- si $d(s,t) > 1$, $Cov(\epsilon_s, \epsilon_t) = 0$

Les valeurs prises par la fonction d'autocorrélation sont donc également corrigées. Nous allons maintenant introduire un estimateur pour les processus ARMA généraux, qui nous permettra de justifier la validité des modèles.

1.3.3 Estimateur du maximum de vraisemblance

On s'intéresse à un processus qui admet une représentation autorégressive

$$X_t = \sum_{i=0}^{\infty} h_i(\theta)\epsilon_{t-i}$$

où les coefficients $h_i(\theta)$ ne sont pas connus et dépendent d'un paramètre θ, et où $(\epsilon_t)_{t\in\mathbf{Z}}$ est un bruit blanc de variance σ^2 également inconnue. On cherche alors à estimer les différents paramètres. Pour cela, on introduit la log-vraisemblance, connaissant les observations successives $X_1, \cdots, X_T$ du processus :

$$L_T(x,\theta,\sigma^2) = -\frac{T}{2}log2\pi - \frac{1}{2}log\ det\Gamma_T(\theta,\sigma^2) - \frac{1}{2}X_T^t\Gamma_T(\theta,\sigma^2)x_T \qquad (17)$$

où $x_T = (X_1, \cdots, X_T)$ et $\Gamma_T(\theta,\sigma^2)$ est la matrice de covariance de $(X_1, \cdots, X_T)$. La vraisemblance $L_T(x,\theta_1,\sigma_1^2)$ vue comme fonction des deux derniers paramètres exprime la probabilité d'obtenir les valeurs x si l'on prend le couple (θ_1,σ_1^2) comme valeur pour (θ,σ^2). On cherche donc à maximiser cette grandeur.

Toutefois, la valeur exacte est difficile à obtenir : on se contente souvent d'approximations par des lois conditionnelles ou en passant par le domaine des fréquences.

On définit un estimateur du maximum de vraisemblance par $(\hat{\theta},\hat{\sigma}^2)$ tel que $L_T(x,\hat{\theta},\hat{\sigma}^2) = \max_{\theta,\sigma^2} L_T(x,\theta,\sigma^2)$ pris sur les différentes valeurs de θ et de σ^2. Sa valeur permet de calculer les corrélations que l'on cherche, puisque celles-ci sont des fonctions simples de θ (on rappelle qu'il s'agit d'un vecteur de paramètres comprenant les coefficients des polynômes considérés dans le modèle) et de σ^2. On étudie ensuite le comportement asymptotique de cet estimateur qui, dans certains cas, converge vers la "vraie" valeur (θ_0,σ_0^2) du paramètre lorsque le nombre de données disponibles (ici le nombre d'individus dans la colonie) tend vers l'infini. Par exemple, dans le cas des données recueillies par Powell sur les bactéries E.Coli, comparons les valeurs données par l'estimateur du maximum de vraisemblance appliqué au modèle de Cowan-Staudte avec les estimations de Powell (cf.tableau ci-dessous). On s'aperçoit que les liens entre cousines sont fortement sous-estimés. En revanche, en apportant aux modèles $BARMA(2,0)$ la correction du modèle à distance, on trouve des valeurs plus proches de celles de Powell.

	$\rho(H_1)$	$\rho(S)$	$\rho(H_2)$	$\rho(C_1)$	$\rho(C_2)$
Powell	$-0,198$	$0,402$	$0,050$	$0,317$	$0,143$
MLE C-S	$-0,28$	$0,42$	$0,078$	$0,033$	$0,003$
MLE $BARMA(2,0)$ à distance	$-0,26$	$0,45$	$-0,005$	$0,28$	$0,20$

où $\rho(H_1)$ est la corrélation entre mère et fille, $\rho(S)$ entre soeurs, $\rho(H_2)$ entre grand-mère et petite-fille, $\rho(C_1)$ entre cousines proches et $\rho(C_2)$ entre cousines du second degré.

1.4 Exploitation des résultats

Les estimations des différentes corrélations permettent d'avoir des indices quant au lignage étudié et à la transmission de la caractéristique mesurée. En effet, puisqu'une cellule naît de la division de sa mère, la corrélation mère-fille est un indicateur de transmission héréditaire, tandis que la corrélation entre soeurs se compose de l'influence commune de leur mère et de celle de l'environnement commun dans lequel se développent les deux bactéries, environnement dont le rôle est mis en évidence par la corrélation de X_{2t} et X_{2t+1} sachant X_t.

L'avantage des modèles BARMA est que les corrélations sont des fonctions simples des polynômes θ et ϕ. Dans le cas d'un modèle $BARMA(1,0)$ sans distance, ie $X_{2t} = \theta X_t + \epsilon_{2t}$, on a :

- si t est un ancêtre de s et que k générations les séparent, leur corrélation est θ^k ;
- la corrélation entre soeurs est donnée par $\rho = \theta^2 + \eta(1-\theta^2)$, où η est le coefficient de corrélation de ϵ_{2t} et ϵ_{2t+1} ;
- celle entre deux individus qui ne sont pas sur la même lignée vaut $\rho\theta^{g_s+g_t-2}$.

Le tableau précédent suggère que les corrélations entre cousines s'expliquent davantage par des causes environnementales que par un héritage génétique commun, puisque $\frac{\rho(H_2)}{\rho(C_1)} = 0,16$. On peut aussi penser que la corrélation entre soeurs est plus influencée par des facteurs liés à l'environnement. En effet, $\frac{\rho(H_1)}{\rho(H_2)} = 3,96$ alors que $\frac{\rho(S)}{\rho(C_1)} = 1,27$: la corrélation héréditaire directe décroît plus vite avec l'éloignement que la corrélation au sein d'une génération.

2 Une première extension du modèle : erreurs non gaussienne et modèles conditionnels

Les modèles présentés dans le paragraphe précédent ont l'inconvénient d'être assez rigides : d'une part, supposer que le terme d'erreur est gaussien n'est pas toujours satisfaisant. Par exemple, lorsque la variable est une durée de vie (donc une variable positive), on opte plutôt pour une loi gamma ou exponentielle. Lorsque la variable est discrète, par exemple la taille d'une population, une loi discrète semble plus adaptée.

Rappel Soit a, p des réels strictement positifs. La densité d'une loi gamma $G(a,p)$ est donnée par

$$f(x) = \frac{a^p}{\Gamma(p)} x^{p-1} e^{-ax} \delta_{x>0} \qquad avec\, \Gamma(p) = \int_0^\infty x^{p-1} e^{-x} dx. \tag{18}$$

D'autre part, les paramètres intervenant dans l'équation de définition sont supposés indépendants de t, alors que l'influence du passé peut elle aussi varier avec le temps.

Dans cette section, on présente une stratégie de modélisation des lignages cellulaires plus générale qui permet de traiter ces deux critiques du modèle gaussien.

2.1 Modèle général

Faisons les suppositions suivantes :

- $E(X_{2t}|X_t) = E(X_{2t+1}|X_t) = m_t(X_t, \theta)$
- $Var(X_{2t}|X_t) = Var(X_{2t+1}|X_t) = v_t(X_t, \theta, \alpha)$
- $Cov((X_{2t}, X_{2t+1})|X_t) = \gamma_t(X_t, \theta, \alpha)$

où m_t, v_t et γ_t sont des fonctions connues et (θ, α) des (vecteurs de) paramètres inconnus. Les deux paramètres ont une interprétation différente : θ est un vecteur de paramètres influençant le système, comme la température ambiante, alors que α est un paramètre de nuisance, qui influence la variance des données : par exemple, c'est l'erreur de lecture lors d'une mesure. Dans le cas particulier d'un processus BAR(1), on a $\theta = (\phi)$, $\alpha = \sigma, \rho$, $m_t(X_t, \theta) = \phi X_t$, $v_t(X_t), \theta = \sigma^2$ et $\gamma_t(X_t, \theta) = \sigma^2\rho$. La seule hypothèse requise est celle d'un moment d'ordre 2.

2.2 Exemples

2.2.1 Loi Gamma à deux variables

La densité de cette loi est donnée, pour $u, v > 0$ et $\alpha_1, \alpha_2, \alpha_3 > 0$, par :

$$g(u,v) = \frac{e^{-(u+v)}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} \int_0^{min(u,v)} w^{\alpha_3-1}(u-w)^{\alpha_1-1}(v-w)^{\alpha_2-1}e^w dw \quad (19)$$

On montre alors que les lois marginales de U et V sont des lois Gamma de paramètres $(\alpha_1 + \alpha_3, 1)$ et $(\alpha_2 + \alpha_3, 1)$.

On suppose que, conditionnellement à X_t, le couple (X_{2t}, X_{2t+1}) suit cette loi avec $\alpha_1 = \alpha_2 = (1-\rho)m_t(X_t)$ et $\alpha_3 = \rho m_t(X_t)$ où $\rho \in [0, 1[$. Alors en choisissant $m_t = \phi X_t + \lambda$, avec $\lambda > 0$, on obtient $\theta = (\phi, \lambda)$, $\alpha = \rho$, $v_t = \phi X_t + \lambda$ et $\gamma_t = (\phi X_t + \lambda)\rho$.

2.2.2 Coefficients aléatoires

On remplace ici le coefficient ϕ d'un modèle BAR(1) par $\phi + Y_t$ où Y_t est une suite de variables aléatoires iid centrées, de variance τ^2 avec $\phi^2 + \tau^2 < 1$. Les erreurs $(\epsilon_{2t}, \epsilon_{2t+1})$ sont simplement supposées iid, de variance σ^2, de covariance $\rho\sigma^2$ et indépendantes de Y_t à t fixé. On a donc l'équation (celle pour X_{2t+1} est identique) :

$$X_{2t} = \lambda + (\phi + Y_t)X_t + \epsilon_{2t} \quad (20)$$

Ici, $m_t = \lambda + \phi X_t$, $v_t = X_t^2\tau^2 + \sigma^2$ et $\gamma_t = X_t^2\tau^2 + \sigma^2\rho$.

2.3 Comportement asymptotique

Les modifications apportées au modèle compliquent la fonction de vraisemblance. On préfère alors travailler avec la quasi-vraisemblance. Posons, avec les notations précédentes,

$$Z_t(\theta) = \begin{pmatrix} X_{2t} - m_t(X_t, \theta) \\ X_{2t+1} - m_t(X_t, \theta) \end{pmatrix} \quad (21)$$

$$V_t(\theta, \alpha) = \begin{pmatrix} v_t(X_t, \theta, \alpha) & \gamma_t(X_t, \theta, \alpha) \\ \gamma_t(X_t, \theta, \alpha) & v_t(X_t, \theta, \alpha) \end{pmatrix} \quad (22)$$

Si on connaît α, l'équation permettant d'obtenir θ est donnée par

$$\sum_0^n (\frac{dZ_t(\theta)}{d\theta})' V_t^{-1}(\theta, \alpha) Z_t(\theta) = 0 \quad (23)$$

Si α est un paramètre de nuisance inconnu, on le remplace dans cette équation par un estimateur consistent.

Dans le cas où $m_t = \phi X_t$, on peut montrer que la solution de l'équation de quasi-vraisemblance est donnée par

$$\hat{\phi}_{QL} = \frac{\sum_{t=1}^{n}(v_t+\gamma_t)^{-1}X_tU_t}{\sum_{t=1}^{n}(v_t+\gamma_t)^{-1}X_t^2}, \tag{24}$$

où $U_t = \frac{X_{2t}+X_{2t+1}}{2}$.

Proposition 2

$$\sqrt{n}(\hat{\phi}_{QL}-\phi) \rightharpoonup \mathcal{N}(0, \frac{1}{2}(E(\frac{X_t^2}{v_t+\gamma_t}))^{-1}). \tag{25}$$

Par exemple, pour le modèle BAR(1), on obtient

$$\sqrt{n}(\hat{\phi}_{QL}-\phi) \rightharpoonup \mathcal{N}(0, \frac{1}{2}(1+\rho)(1-\phi^2)) \tag{26}$$

et pour l'erreur de loi Gamma, la loi asymptotique est $\mathcal{N}(0, \frac{1}{2}\phi(1+\rho)(EX_t)^{-1})$, un estimateur consistent de EX_t étant

$$\frac{1}{n}\sum_{t=1}^{n} X_t.$$

Remarque 3 *En pratique, la faible quantité de données (arbres de 7 ou 8 générations) limite les indices permettant de trouver la structure de corrélation des erreurs.*

3 Modèles à états

En se servant de ce qui a été fait précédemment, on cherche à construire un modèle permettant d'étudier un système en fonction des données accessibles.

Définition 9 *Un modèle linéaire à état est un modèle du type*

$$\begin{cases} Y_t = A_tX_t + B_tU_t \\ X_{t+1} = C_tX_t + D_tU_t \end{cases}$$

où pour tout t, X_t, Y_t, Y_t sont des vecteurs et A_t, B_t, C_t, D_t des matrices, tous dépendant du temps.

3.1 Idée intuitive

– Ce qu'on mesure ne correspond pas toujours à la grandeur que l'on veut observer : on commence donc par chercher comment ce que l'on observe (Y_t ici) dépend de ce qui est caché (X_t). C'est ce que rend l'*équation de mesure*

$$Y_t = A_tX_t + B_tU_t. \tag{27}$$

– Puis on cherche comment la grandeur Y_t dépend de ses états passés :

$$X_t = C_t X_{t-1} + D_t U_t \tag{28}$$

est l'*équation du système*.

– Il reste à déterminer les données initiales.

Ces trois étapes définissent donc le schéma du modèle linéaire à états, dans lequel U_t représente l'*input*, Y_t l'*output* et X_t le *vecteur d'état*.

3.2 Exemple : modèle à états linéaire et gaussien

Les résultats sont toujours donnés conditionnellement à la filtration liée au processus (Y_t). Ici, Y_t est supposé être un scalaire, et X_t un vecteur de dimension p. On définit le modèle par

– équation d'observation : $Y_t = a_t' X_t + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, N_t)$

– équation du système : $X_t = C_t X_{t-1} + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}_p(0, E_t)$

– information initiale : $X_0 \sim \mathcal{N}_p(b_0, E_0)$.

On suppose connus le paramètre N_t, les vecteurs a_t (qui se composent par exemple d'observations précédentes et de paramètres connus) et b_0, ainsi que les matrices E_0, C_t et E_t. On suppose aussi que η_t et ϵ_t sont chacun constitués de vecteurs indépendants et que X_0, η_t et ϵ_t sont mutuellement indépendants.

On peut retrouver ainsi un processus $BARMA(p,q)$: si l'évolution de X est donnée par l'équation $\phi(b)X_t = \epsilon_t$, où b est l'opérateur de retard $X_t \mapsto X_{\lfloor t/2 \rfloor}$, la représentation modèle d'états correspondante est

$$\begin{cases} Y_t = \theta(b) X_t \\ \theta(b) X_t = \phi(b) \epsilon_t \end{cases} \tag{29}$$

Par exemple, pour $p = 3$ et $q = 1$, le système s'écrit matriciellement

$$\begin{cases} Y_t = (0, \theta_1, 1) \begin{pmatrix} X_{t-2} \\ X_{t-1} \\ X_t \end{pmatrix} \\ \begin{pmatrix} X_{t-2} \\ X_{t-1} \\ X_t \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \phi_3 & \phi_2 & \phi_1 \end{pmatrix} \begin{pmatrix} X_{t-3} \\ X_{t-2} \\ X_{t-1} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \epsilon_t \end{pmatrix} \end{cases}$$

Ainsi, $(Y_t)_{t \in \mathbf{Z}}$ suit un processus $BARMA(p,q)$.

3.3 Estimation des paramètres du modèle linéaire gaussien

On suppose que les paramètres b_0, E_0, C_t, N_t, E_t dépendent d'un vecteur de paramètres θ indépendant du temps.

Notations Soient s et t deux entiers. On note

$$Y_{t|s} = E[Y_t|\mathcal{F}_s] \tag{30}$$

la moyenne de Y_t conditionnellement à $\mathcal{F}_s$ et

$$P_{t|s} = E[(Y_t - Y_{t|s})(Y_t - Y_{t|s})'] \tag{31}$$

sa matrice de précision. On peut obtenir ces grandeurs par récurrence.

On se base alors sur la vraisemblance de θ, qui est obtenue à partir de la distribution des innovations

$$\alpha_t = Y_t - E[Y_t|\mathcal{F}_{t-1}] = Y_t - a_t' C_t Y_{t-1|t-1} \tag{32}$$

qui sont des variables aléatoires iid de loi $\mathcal{N}(0, \sigma_t^2(\theta))$. A une constante près, la log-vraisemblance de θ pour la série $(Y_1, \cdots, Y_N)$ est alors :

$$\log L_y(\theta) = -\frac{1}{2}\sum_{t=1}^{N} \log \sigma_t^2(\theta) - \frac{1}{2}\sum_{t=1}^{N} \frac{\alpha_t^2}{\sigma_t^2}(\theta). \tag{33}$$

Maximiser cette vraisemblance est une des méthodes pour estimer θ.

Pour des compléments sur ce sujet, voir [5].

4 Conclusion

Nous nous sommes donc intéressés à différentes manières de modéliser des lignages cellulaires, tout en en mesurant les limites. Aussi naturels que puissent paraître certains modèles, leur validité est souvent difficilement appréciable en raison du grand nombre de paramètres à prendre en compte et de la quantité de données nécessaires pour valider certains tests.

Assez fréquemment lorsqu'on étudie un phénomène, on doit se contenter d'informations partielles : par exemple, on ne peut n'avoir accès à certaines caractéristiques d'un système qu'à des temps discrets. Parfois même, en particulier lorsqu'on suit les divisions successives d'une cellule mère, le nombre de mesures possibles est limité : les cellules finissent par se développer en dehors du champ visuel du microscope quand elles ne s'entassent pas les unes sur les autres, rendant toute mesure impossible. Dans la majorité des cas, l'expérimentateur est confronté à des problèmes pratiques qui l'empêchent d'avoir un accès direct à toutes les données souhaitées.

Quand, par exemple, on essaie d'apprécier la valeur des paramètres avec la méthode du maximum de vraisemblance, une étude asymptotique est nécessaire pour obtenir avec une bonne précision des résultats non biaisés ; quelque soit la taille n du système (ie la longueur de la série temporelle disponible), ces estimateurs peuvent être biaisés. Ainsi, des théorèmes limites permettent de construire des tests asymptotiques pour les estimateurs, mais là encore, l'échantillon à analyser doit être de dimension "infinie". A taille fixée, on doit se contenter de simulations, paramétriques (méthode de Monte Carlo) ou non.

Références

[1] I.V. Basawa, R.M. Huggins : *Extensions of the bifurcating autoregressive model for cell lineage studies*

[2] I.V. Basawa, J. Zhou : *Non-gaussian bifurcating models and quasilikelihood estimation*, février 2003

[3] P.J. Brockwell, R.A. Davis : *Time series : theory and methods*, Springer, 1990

[4] E. Demidenko : *Mixed models : Theory and applications*, Wiley, 2004

[5] B. Kedem, K. Fokianos : *Regression models for time series analysis*, Wiley-, 2002

[6] E.J. Stewart, R. Madden, G. Paul, F. Taddei(février 2005) : *Aging and death in an organism that reproduces by morphologically symmetric division*, PLOS biology, **9**, 0295-0300

5 Annexe : figures

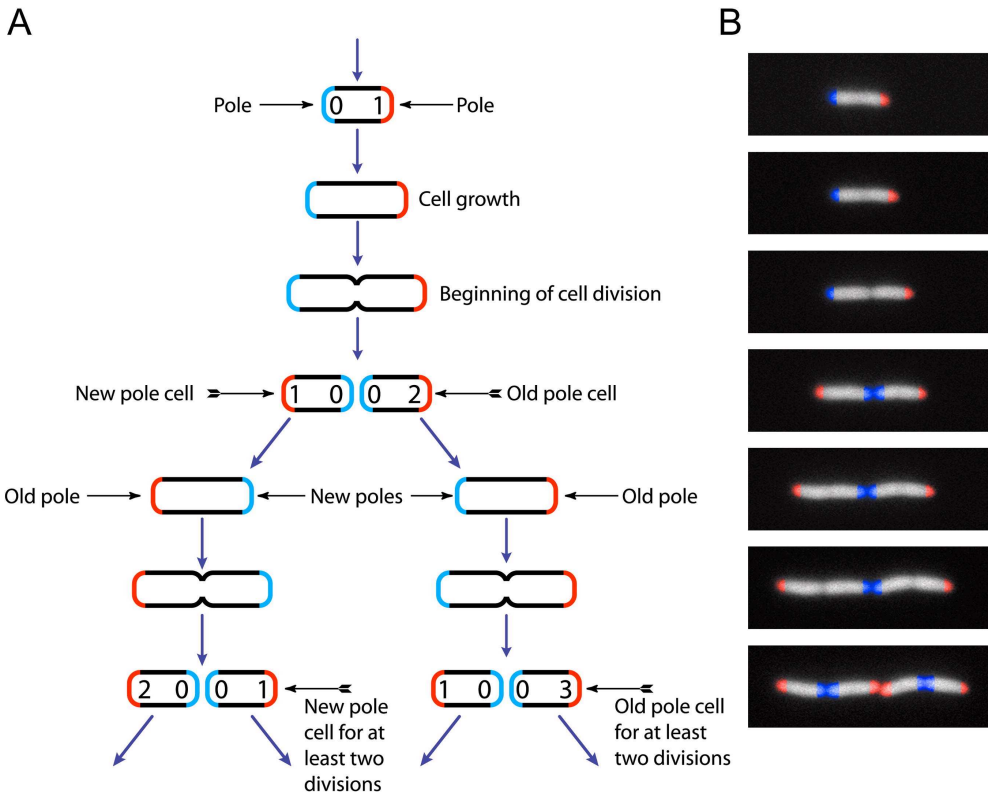

FIG. 1 – Schéma de division bactérienne

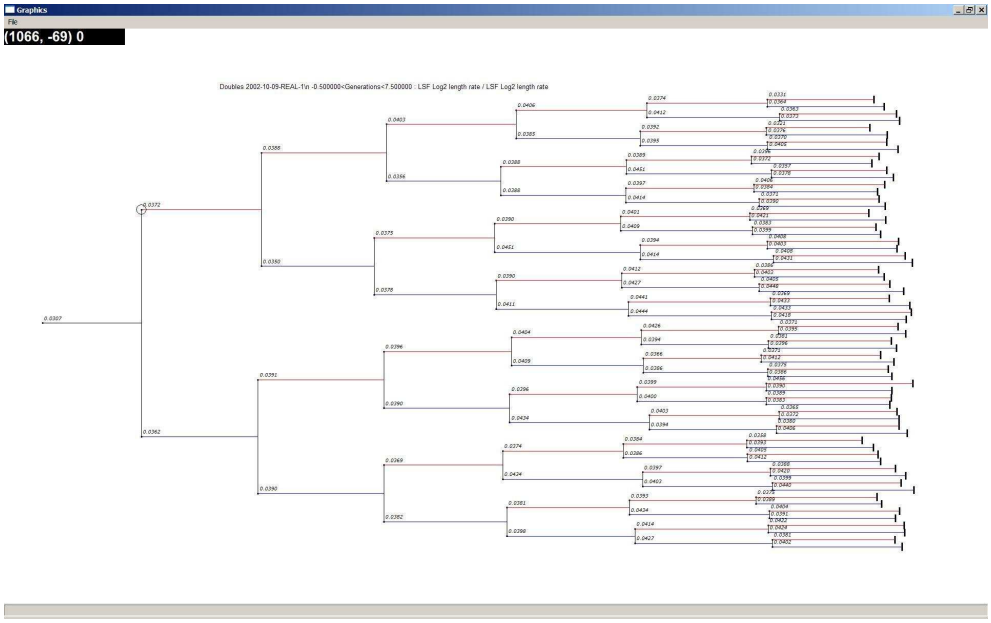

FIG. 2 – Représentation des taux de croissance sous forme de lignages. La longueur de la ligne séparant une bactérie de ses filles est proportionnelle à son taux de croissance. Cette représentation permet de voir l'accumulation des différences entre les bactéries héritant du vieux pôle (en bleu) et celles recevant le jeune pôle (en rouge).

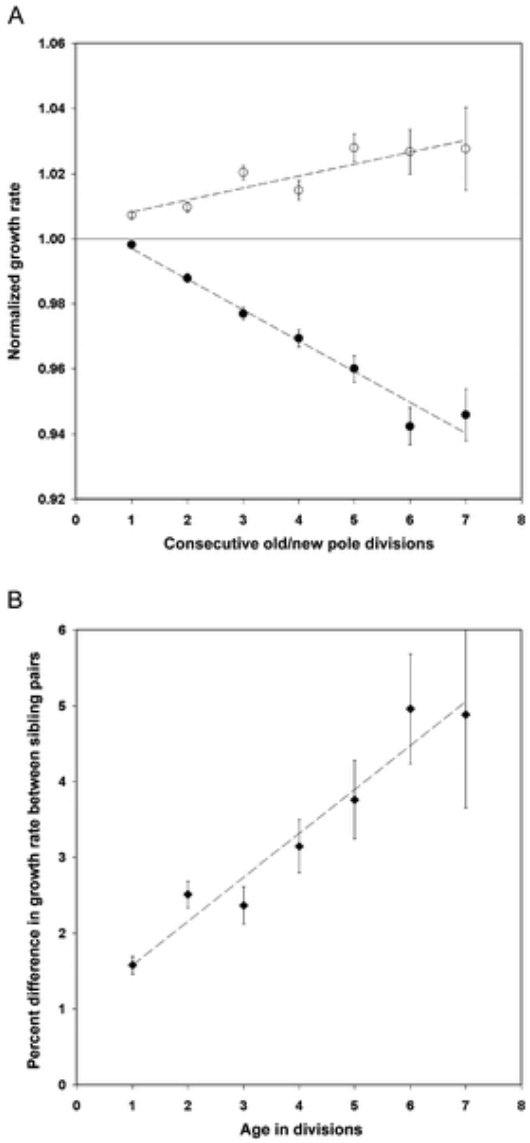

FIG. 3 – Différences de taux de croissance entre soeurs