Introduction au Domaine de Recherche - Modélisation de la diversité génétique dans les populations en expansion

Apolline Louvet

5 novembre 2018

Table des matières

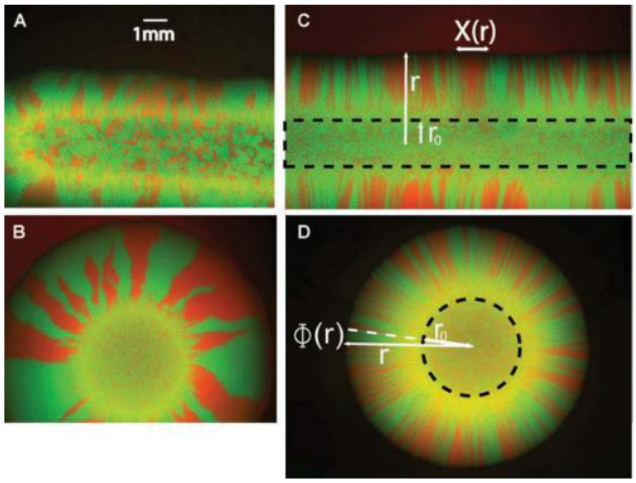

FIGURE 1 – Sectorisation observée dansdes populations de bactéries, dans le cadre de l'expérience présentée dans [17]. Les couleurs rouges et vertes représentent deux souches de bactéries ne différant que par leur fluorescence. Source : [17]

1 Motivations

1.1 Motivations biologiques

En biologie, les expansions de population constituent un phénomène générique qui recouvre un grand nombre de situations, de l'écologie à la cancérologie en passant par l'épidémiologie. L'étude de leur impact sur la diversité génétique à l'intérieur d'une population peut permettre entre autres de reconstruire l'historique d'une expansion[31, 19], ou par exemple d'optimiser le traitement des tumeurs cancéreuses[20, 24, 32, 6, 21, 28]. De façon générale, il est connu que la diversité génétique est décroissante dans la direction d'expansion de la population, du fait d'événements de fondation successifs : les premiers individus partis sont les premiers à pouvoir se reproduire. Par exemple, c'est de cette façon qu'il est possible de retrouver la direction des expansions humaines hors d'Afrique [31].
Quid de la direction transverse à l'expansion ? Dans l'article [17] est présentée une expérience dans laquelle deux souches de bactéries ne différant que par la couleur de leur fluorescence sont introduites ensemble dans le même milieu de croissance. L'expansion de la colonie fait apparaître une sectorisation de la population. Au niveau du front de propagation, certaines zones sont peuplées uniquement de bactéries d'une couleur, d'autres de bactéries de l'autre couleur. Si l'une des souches de bactéries est avantagée sélectivement, seule la forme des secteurs change. Si l'origine stochastique de ce phénomène est claire, la loi du nombre de secteurs, l'évolution de la frontière entre deux secteurs successifs, et le lien avec les paramètres démographiques de la population sont encore mal compris.

1.2 Motivations mathématiques

Les approches déterministes pour les populations en expansion sont utilisées depuis longtemps, et impliquent souvent l'équation de Fisher-KPP [15], qui permet de modéliser l'expansion d'une population en toutes dimensions. Cependant, l'expérience des deux souches de bactéries fluorescentes [17] nous montre que la stochasticité joue un rôle essentiel dans l'apparition de la sectorisation. Ce type de phénomène ne peut donc pas être compris avec des modèles déterministes. L'équation de Fisher-KPP ne possédant un équivalent stochastique qu'en dimension 1, il est nécessaire de construire de nouveaux modèles stochastiques, particulièrement pour traiter le cas des dimensions 2 et plus. En effet, d'un point de vue biologique, les dimensions supérieures à 3 ne présentent que peu d'intérêt. De plus, il n'est pas possible de définir des modèles de reproduction en dimension 2 en partant d'une approche individu-centrée sans rajouter d'interactions locales entre les individus. Cet effet est connu sous le nom de *pain in the torus* [14].
A ces difficultés à modéliser les processus de reproduction se rajoute les problématiques liées à l'étude des expansions. En effet, l'étude de la diversité génétique peut par exemple nécessiter de suivre les

descendants d'un groupe d'individus. Le comportement reproductif dépendant de leur position par rapport au front, il faudra suivre à la fois leur position et celle du front de population. De même, une autre approche de mesure de la diversité génétique d'un échantillon consiste à au contraire remonter dans le temps et à suivre les positions des ancêtres. Ceci nécessite de trouver un moyen d'inverser le sens du temps pour le front de population aussi - ou de développer des approches permettant de contourner ce problème.
En d'autres termes, l'étude des populations en expansion est aussi un problème mathématiquement riche. Beaucoup de questions d'intérêt restent encore à traiter. Ainsi, comment peut-on modéliser la vague d'invasion ? Où sont situés les individus faisant avancer le front ? Quelle est la forme des généalogies au niveau du front, et plus en arrière ?

2 Le modèle de Moran

2.1 Définition

Présentons un modèle sans structuration spatiale, qui sera ensuite adapté pour les populations en expansion. Il s'agit d'un modèle classique en génétique des populations : le *modèle de Moran*, introduit par Patrick Moran en 1958 [27].
Dans celui-ci, nous suivons une population vérifiant les hypothèses suivantes, notées (H) dans la suite :

- la population est constituée de N individus, pouvant être de deux types différentes, notés dans la suite 0 et 1 (**hypothèse de taille finie**).
- la population est *panmictique*, c'est à dire qu'un individu donné a autant de chances de se reproduire avec n'importe lequel des autres membres de la population, ceci indépendamment de leurs types ou de la distance entre eux (**hypothèse de panmixie**)
- les types sont neutres sélectivement, c'est à dire que la loi du nombre de descendants d'un individu donné ne dépend pas de son type (**hypothèse de neutralité**)

Point biologique 1 *- Les types mentionnés dans la définition du modèle de Moran, correspondent, d'un point de vue biologique, aux différents allèles d'un gène. Un gène peut être défini grossièrement comme une unité indivisible et héritable d'information génétique, les allèles étant les différentes version possibles pour ce gène.*
Point Biologique 2 *- Pour être plus précis d'un point de vue biologique, nous supposons que les individus sont haploïdes, c'est à dire qu'ils ne portent qu'une seule copie de chaque gène. Par exemple, les humains sont diploïdes, portent deux copies de chaque gène, et ne transmettent que l'un des deux à leur progéniture. De même, les amateurs d'huitres du bassin d'Arcachon auront entendu parler des controversées huitres triploïdes, qui quant à elles portent trois copies de chaque gène.*
L'intérêt de l'étude des populations haploïdes est double : de cette façon, chaque individu n'a qu'un seul parent, et de plus, les cas multiploïdes peuvent etre ramenés à des cas haploïdes.

Le modèle de Moran se définit alors de la façon suivante.

Définition 1. Soit une population de N individus satisfaisant les hypothèses (H). Nous dirons que la population se reproduit selon le modèle de Moran si son processus de reproduction est le suivant.
Soit $(\Pi_t)_{t\geq 0}$ un processus de Poisson d'intensité $\binom{N}{2}$. A chaque instant de saut du processus, une paire d'individus est tirée dans la population. L'un d'entre eux est choisi pour être le *parent*, l'autre étant alors *l'enfant*. Le parent donne alors son type à l'enfant.

La quantité d'intérêt est la proportion d'individus de type 1 dans la population, notée $(u_t^N)_{t\geq 0}$. L'indice N permet de rappeler la dépendance en la taille de la population. Nous disposons alors d'une limite grande population pour la densité :

Théorème 1. *Soit $(u_t)_{t\geq 0}$ le processus de générateur infinitésimal :*

$$Af(x) = \frac{1}{2}x(1-x)f''(x)$$

Alors, $(u_t^N)_{t\geq 0}$ converge vers $(u_t)_{t\geq 0}$ en distribution lorsque $N \longrightarrow +\infty$.

Une preuve de ce résultat figure dans le livre [13], qui constitue de façon plus générale une introduction aux divers modèles utilisés en génétique des populations, et contient les principaux résultats connus.
Pour rappel, le *générateur infinitésimal* A d'un processus markovien $(X_t)_{t\geq 0}$ à valeurs dans $\mathbb{R}^n$ est défini pour toute fonction $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ dans le domaine de définition de A par :

$$Af(x) = \lim_{t\to 0} \frac{\mathbb{E}_x[f(X_t)] - f(x)}{t}$$

Cet objet permet souvent de caractériser le processus associé et son comportement, sous certaines conditions liées entre autres à son domaine de définition. Le générateur obtenu est très classique en génétique des populations. Il s'agit en effet de celui de la *diffusion de Wright-Fisher*, décrivant l'évolution de la fréquence d'un allèle dans une population infinie respectant les hypothèses du modèle de Moran (panximie, taille finie, neutralité).
Point biologique *- La diffusion de Wright-Fisher modélise le phénomène biologique connu sous le nom de dérive génétique, dans lequel la fréquence d'un allèle neutre sélectivement varie sous l'effet du hasard. L'allèle en question peut en particulier être perdu ou se fixer dans la population.*
Vu comme processus stochastique, $(u_t^N)_{t\geq 0}$ est un exemple de processus de saut. Il s'agit d'une catégorie de processus stochastiques constants par morceaux, évoluant par sauts d'amplitude aléatoire pouvant dépendre de l'état du système, lors d'instants eux aussi aléatoires. Il s'agit de plus d'une martingale : pour $0 \leq s \leq t$, l'espérance de u_t^N sachant $(u_{s'}^N)_{s'\leq s}$ est égale à celle de u_s^N.

Point mathématique *- Comme le processus est défini sur une durée infinie, nous devrions plutôt parler de* ***martingale locale****, c'est à dire de processus auquel peut être associé une série de temps d'arrêts tendant vers l'infini, chacun d'entre eux rendant le processus une martingale. Cependant, l'ensemble des théorèmes et résultats que nous présenterons supposent, dans leur énoncé ou dans leur preuve, que nous nous restreignons à un intervalle de temps borné. Dans la suite, nous parlerons donc tout le temps de martingale par abus de langage.*
Introduisons quelques définitions supplémentaires.

Définition 2. Soit $(X_t)_{t\geq 0}$ un processus stochastique. X est dit à *variation finie* si sa variation est presque sûrement bornée sur tout intervalle de temps fini.
Il est alors possible de calculer la *variation quadratique* du processus. Si $(X_t)_{t\geq 0}$ est un processus de saut, alors sa variation quadratique sur l'intervalle $[0, t]$ est la somme des carrés des amplitudes des saut réalisés sur cet intervalle de temps.

Définition 3. Si $(X_t)_{t\geq 0}$ est une martingale, alors son *crochet* $(< X >_t)_{t\geq 0}$ est l'unique processus à variation finie tel que $X_t^2 - < X >_t$ est une martingale.

2.2 Définition alternative via les processus ponctuels de Poisson

Il est possible de redéfinir le modèle de Moran en adoptant un point de vue différent. En effet, précédemment, nous avions défini le processus en tirant les temps auxquels se produisent des événements, puis la nature de chaque événement. Mais il est aussi possible d'adopter une autre représentation, en associant *à chaque paire ordonnée parent-enfant* un processus de Poisson de paramètre $\frac{1}{2}$. Alors, pour une paire donnée, à chaque temps de saut du processus de Poisson associé, le parent donne son type à l'enfant.
Notons alors $(\xi_t^i)_{1\leq i\leq N}$ les types des N individus au temps t, et $T > 0$ un instant. Il est possible d'écrire le type de l'individu $1 \leq i \leq N$ au temps T à partir des types de chacun des autres individus, en utilisant des *processus ponctuels de Poisson*.
De façon générale, un processus ponctuel permet de modéliser la répartition aléatoire de points dans un espace E. Il s'agit d'une mesure aléatoire P sur cet espace, pouvant s'écrire presque sûrement sous la forme :

$$P = \sum_{i\in X} \delta_{x_i}$$

où $X \subset \mathbb{N}$ et $(x_i)_{i\in X}$ est un ensemble de points de E.
Les processus ponctuels de Poisson sont un cas particulier de processus ponctuels.

Définition 4. Processus ponctuel de Poisson
Soit E un espace métrique localement compact, et μ une mesure sur E finie sur tout compact, sans atomes. Alors, le *processus de Poisson sur E d'intensité μ*, noté P, est un processus ponctuel sur E vérifiant :

— Pour tout compact $B \in E$, $P(B)$ est une variable aléatoire de Poisson de paramètre $\mu(B)$
— Pour tous $n \geq 1$, $(B_1, ..., B_n)$ compacts disjoints de E, les variables aléatoires $P(B_1), ..., P(B_n)$ sont indépendantes.

De plus, lorsque l'intensité du processus est de la forme αdl avec $\alpha > 0$ et dl la mesure de Lebesgue, nous dirons que le processus est de taux α

Une présentation plus complète des processus ponctuels de Poisson peut être trouvée dans [1].
Dans le cas particulier du modèle de Moran, les processus ponctuels de Poisson utilisés seront définis sur l'espace $\mathbb{R}_+$, à partir de la mesure $\frac{1}{2}dt$ où dt représente la mesure de Lebesgue.
Notons donc, pour tous $1 \leq i, j \leq N$, $P^N_{(i,j)}$ un processus ponctuel de Poisson sur $\mathbb{R}_+$ de taux $\frac{1}{2}$. Alors, pour $1 \leq i \leq N$:

$$\xi^I_T = \xi^i_0 + \sum_{j \neq i} \int_0^T (\xi^j_s - \xi^i_s) dP^N_{(i,j),s}$$

2.3 Ajout de la sélection

Grâce au formalisme des processus ponctuels de Poisson, il nous est maintenant possible de modifier le modèle de Moran de façon à rajouter de la sélection naturelle. L'un des deux types se reproduira plus souvent que l'autre - par convention, le type 1. D'un point de vue biologique, ceci peut être interprété de deux façons :

— soit le type 1 a un avantage sélectif sur le type 0
— soit le type 0 correspond à l'absence d'individus, N étant alors le nombre total d'individus que peut accueillir le milieu

Ceci peut être intégré à notre modèle précédent en considérant qu'une paire ordonnée (i, j) est impliquée dans un événement de sélection à taux θs_N, où s_N dépend de la taille totale de la population. Lors d'un tel événement, j ne prend le type de i que si i est de type 1.

Point biologique *A priori, la deuxième interprétation proposée pour les types peut paraître assez absurde. Cependant, il ne faut pas perdre de vue que les événements de reproduction dans lequel un individu* 1 *prennent le type d'un autre correspondent en fait à une mort de l'individu en question, suivie de la colonisation de l'espace qu'il occupait par un nouvel individu. Un événement "parent 0 - enfant 1" modélise en fait la mort de l'individu de type* 1*, non suivie d'une recolonisation. Enfin, les événements "parents 0 - enfant 0", qui ne correspondent à rien d'un point de vue biologique, sont en fait introduits pour permettre de tirer des événements de reproduction* ***potentiels****, indépendamment du type des individus.*

Chaque paire (i, j) se voit donc associer un nouveau processus de Poisson $\tilde{P}^N_{(i,j)}$ sur $\mathbb{R}_+$ de taux θs_n. L'équation donnant le type de l'individu i an temps T devient :

$$\xi^i_T = \xi^i_0 + \sum_{j \neq i} \int_0^T (\xi^j_s - \xi^i_s) dP^N_{(i,j),s} + \sum_{j \neq i} \int_0^T \xi^j_s (\xi^j_s - \xi^i_s) d\tilde{P}^n_{(i,j),s}$$

Dans la suite, les événements générés par les processus de Poisson $\tilde{P}^N_{(\cdot,\cdot)}$ seront appelés événements de sélection, et ceux générés par $P^N_{(\cdot,\cdot)}$ événements de vote. Cette nomenclature est liée à celle utilisée en théorie des systèmes de particules en interaction.

3 Ajout d'une structuration spatiale

3.1 Présentation du modèle

Le modèle de Moran modifié qui vient d'être introduit ne permet pas de rajouter de structure spatiale. Ceci est lié à l'hypothèse de panximie sous-tendant le modèle. A cause de celle-ci, un individu a autant de chances de se reproduire avec un voisin proche qu'avec un individu éloigné. Il va donc être nécessaire de modifier la loi de reproduction, afin qu'elle prennent en compte la distance entre deux individus formant une paire.
L'approche utilisée dans l'article [10], que nous exposons ici, permet de traiter le cas à une dimension dans lequel la population est répartie le long d'une ligne. L'interprétation des types choisie sera que le type 0 correspond à des individus potentiels. Ceci permettra de supposer une population infinie tout en ayant toujours une interprétation biologique. La ligne est divisée en sous-unités de largeur constante $\frac{1}{L_n}$, appelées *dèmes* et contenant M_n individus. Les positions de chacun des dèmes seront notées ω, et prennent leurs valeurs dans $L_n^{-1}\mathbb{Z}$. Un individu sera représenté sous la forme $x := (\omega, i) \in L_n^{-1}\mathbb{Z} \times \{1, ..., M_n\}$.
Chaque individu (ω, i) ne pourra interagir qu'avec des individus dans les dèmes $\omega - 1$ et $\omega + 1$, et interagira avec même probabilité avec l'ensemble de ces individus. Le formalisme des processus ponctuels de Poisson peut alors être utilisé pour décrire le processus de reproduction.
Chaque paire ordonnée (x, y) d'individus séparés par un seul dème peut être affectée par deux types d'événements. Le processus ponctuel de Poisson générant les événements de vote, noté $P_{n,\cdot}^{x,y}$, est de taux r_n, tandis que celui générant les événements de sélection, noté $\tilde{P}_{n,\cdot}^{x,y}$, est de taux $\frac{\theta}{R_n}$.
Point mathématique *- La population est ici infinie, et une infinité de changements potentiels d'état se produit à chaque instant. Il est donc nécessaire de vérifier que le processus de reproduction est bien défini, dans le sens suivant :*

- *les changements de type doivent mettre un certain temps à se propager. En d'autres termes, l'état d'une particule pendant une durée ϵ finie ne doit dépendre que de celui de voisins assez proches.*
- *un ensemble fini fixé de particules ne doit changer d'état qu'un nombre fini de fois pendant une durée finie ϵ.*

Ceci permet de s'assurer qu'à presque tout instant, l'état du système est bien défini. Le livre [25] présente les conditions exactes à vérifier pour que le processus soit bien défini, et introduit les outils permettant d'analyser de façon plus générale les systèmes de particules en interaction.

Notons $\xi_t(x)$ le type de la particule $x := (\omega, i) \in L_n^{-1}\mathbb{Z} \times \{1, ..., M_n\}$ au temps t. Notons de plus, pour $y \in L_n^{-1}\mathbb{Z} \times \{1, ..., M_n\}$, $y \sim x$ si x et y sont voisins. Comme précédemment, pour $T > 0$, le type de la particule en (ω, i) est égal à :

$$\begin{aligned}\xi_T(x) = \xi_0(x) &+ \sum_{y\sim x} \int_0^T (\xi_s(y) - \xi_s(x)))dP_{n,s}^{x,y} \\ &+ \sum_{y\sim x} \int_0^T \xi_s(y)(\xi_s(y) - \xi_s(x))d\tilde{P}_{n,s}^{x,y}\end{aligned}$$

La quantité dont nous cherchons à calculer la limite est cette fois-ci la densité en individus de type 1 dans chaque dème $(u_t^n)_{t\geq 0}$ définie par :

$$\forall t \geq 0, u_t^n : \omega \mapsto \frac{1}{M_n} \sum_{i=1}^{M_n} \xi_t(\omega, i)$$

prolongée à $\mathbb{R}$ par interpolation.
L'existence d'une limite nécessite les conditions suivantes sur la convergence des paramètres :

$$\begin{aligned}\frac{r_n M_n}{L_n^2} &\underset{n\to+\infty}{\longrightarrow} \alpha \in (0, \infty) \\ \frac{M_n}{R_n} &\underset{n\to+\infty}{\longrightarrow} \beta \in [0, \infty)\end{aligned}$$

$$\frac{r_n}{L_n} \underset{n\to+\infty}{\longrightarrow} \gamma \in [0,\infty)$$
$$L_n \underset{n\to+\infty}{\longrightarrow} +\infty$$
$$L_n R_n \underset{n\to+\infty}{\longrightarrow} +\infty$$

Sous ces conditions, nous disposons alors d'un résultat de convergence pour u_t^n. Ceci nous renseigne entre autres sur la vitesse d'expansion, et sur la forme du front. Mais plutôt que de présenter ce résultat, nous en exposerons un plus général, permettant de suivre les descendants d'un sous-groupe de la population. Les applications de cela sont nombreuses : il est par exemple possible de mesurer la diversité génétique à l'intérieur du front, ou encore de d'estimer la localisation par rapport au front des individus l'ayant généré.
Pour ce faire, nous pouvons utiliser la méthode des *traceurs*, introduite entre autres dans [18], qui consiste simplement à introduire un nouveau type, noté 1^*, transmis de la même façon que le type 1. En d'autres termes, lors d'un événement de sélection, le parent ne transmet son type que si il est de type 1 ou 1^*.
Nous modifions alors quelque peu nos notations, u_t^n devenant la densité en individus de type 1 ou 1^*. Introduisons de plus la notation l_t^n pour la densité en individus de type 1^*. Nous disposons alors d'un résultat de convergence pour le couple $(u_t^n, l_t^n)_{t\geq 0}$.

Théorème 2. *Supposons que les hypothèses précédentes de convergence des paramètres sont vérifiées, et que* $(u_0^n, l_0^n) \longrightarrow (f_0, g_0)$ *dans* $C_{[0,1]}(\mathbb{R})$. *Alors,* $(u_t^n, l_t^n)_{t\geq 0}$ *converge en distribution vers* $(u_t, l_t)_{t\geq 0}$ *continu à valeurs dans* $C_{[0,1]}(\mathbb{R}) \times C_{[0,1]}(\mathbb{R})$ *solution faible de l'EDS :*

$$\delta_t u = \alpha\Delta u + 2\theta\beta u(1-u) + |4\gamma l(1-u)|^{\frac{1}{2}}\dot{W}^0 + |4\gamma(u-l)(1-u)|^{\frac{1}{2}}\dot{W}^1 \tag{1}$$
$$\delta_t l = \alpha\Delta l + 2\theta\beta l(1-u) + |4\gamma l(1-u)|^{\frac{1}{2}}\dot{W}^0 + |4\gamma l(u-l)|^{\frac{1}{2}}\dot{W}^2 \tag{2}$$

et de conditions initiales :

$$(u_0, l_0) = (f_0, g_0) \tag{3}$$

W^0, W^1 *et* W^2 *étant des bruits blancs espace-temps indépendants sur* $[0,\infty)\times\mathbb{R}$.
De plus, il y a unicité de la solution si $\gamma \geq 0$.

3.2 Idée générale de la preuve

Nous présentons maintenant les grandes étapes de la preuve, et les différents outils utilisés classiquement pour résoudre ce type de problème.
Point mathématique *- Il s'agit d'une présentation extrêmement succinte de la preuve. Toutefois, la démarche générale de la preuve est classique, et se retrouve dans de nombreux articles sur des thèmes similaires ([8, 7]). La connaissance des étapes de la preuve permet donc de traiter beaucoup de cas différents, dont les preuves ne diffèrent que dans les calculs techniques.*
L'approche utilisée est présentée dans [8]. Le livre [26] explique aussi cette approche, et donne plusieurs exemples d'application.

Etape 1 - Ecriture du problème de martingale approché A partir de la formule donnée pour $\xi_t(x)$, par intégration par parties en utilisant une fonction test adaptée $\phi_t(\omega)$, il est possible d'obtenir l'équation en formulation faible suivante pour $u_t(x)$:

$$\begin{aligned}\int_0^t < u_s^n, \partial_s\phi_s > ds =& < u_t^n, \phi_t > - < u_0^n, \phi_t > \\ &- \frac{1}{M_n L_n}\sum_{x\in\Lambda_n}\sum_{y\sim x}\int_0^t (\xi_{s^-}^n(y) - \xi_{s^-}^n(x))\phi_s(x) dP_{n,s}^{x,y} \\ &- \frac{1}{M_n L_n}\sum_{x\in\Lambda_n}\sum_{y\sim x}\int_0^t (\xi_{s^-}^n(y) - \xi_{s^-}^n(x))\phi_s(x) d\tilde{P}_{n,s}^{x,y}\end{aligned}$$

Le but est de réécrire les différents termes afin de faire apparaître des quantités ressemblant à des quantités connues :

— un terme correspondra à un mouvement brownien
— un terme correspondra à un laplacien, en mettant en évidence son équivalent discret
— un terme correspondra à un terme de sélection/reproduction, de la forme *constante* $\times u(1-u)$.

Point biologique *- La forme du troisième terme peut être interprétée de la façon suivante : l'effet de la sélection ne se fait sentir que lorsque les deux parents potentiels sont de type différents.*

Les *processus de Poisson compensés* seront utilisés pour isoler une composante déterministe et une composante stochastique. La partie déterministe correspond à la moyenne du nombre de points tirés par le processus ponctuel de Poisson, et la partie stochastique aux fluctuations autour de cette moyenne. Il s'agit d'une martingale, dont le crochet caractérisera la limite. En particulier, si le crochet tend vers 0 lorsque $n \longrightarrow +\infty$, alors la composante stochastique disparaît à la limite.
Les processus de Poisson compensés se construisent en retirant au processus de Poisson initial l'intégrale entre 0 et t de son intensité au temps s, ce qui permet d'obtenir une martingale.

Définition 5. Soit $\alpha > 0$. Alors, le processus de Poisson compensé associé au processus de Poisson $(P_t)_{t\geq 0}$ d'intensité αds, noté $(\tilde{P}_t)_{t\geq 0}$ est défini par :

$$\forall t \geq 0, \tilde{P}_t = P_t - \int_0^t \alpha ds = P_t - t\alpha$$

Plus précisément, l'association des termes à des quantités connues se fait en supposant que ceux-ci convergent lorsque $n \longrightarrow +\infty$, et en conjecturant ce qu'est alors cette limite. Le but du reste de la preuve est de montrer qu'il y a bien convergence, et ce vers la limite voulue.
Le cas de $l_t(x)$ se traite exactement de la même façon.

Etape 2 - Choix de la fonction test Il est possible de prendre une fonction test adaptée qui permet de faciliter les calculs dans les étapes suivantes. Ici, le choix adapté est celui des fonctions de Green, définies de la façon suivante.

Définition 6. Soit $(X_t^n)_{t\geq 0}$ la marche aléatoire simple sur $L_n^{-1}\mathbb{Z}$ de taux de saut $2L_n^2$, qui converge vers le mouvement brownien dans $\mathbb{R}$ à taux 2 lorsque $n \to +\infty$. Notons $p_t^n(\omega) := L\mathbb{P}(X_t^n = \omega | X_0^n = 0)$ sa probabilité de transition, $(P_t^n)_{t\geq 0}$ son semi-groupe et $p_t(\omega)$ la densité de transition du mouvement brownien à taux 2. La fonction de Green associée et adaptée à notre problème, notée $\phi_{n,s}^{t,z}(\omega)$ est définie par :

$$\phi_{n,s}^{t,z} := p^n_{\alpha_n(t-s)}(\omega - z)$$

si $s \in [0,t]$et $\phi_{n,s}^{t,z} = 0$ sinon.

L'intéret du choix de cette fonction test est qu'elle permet de "remonter dans le temps" dans la diffusion générée par la marche aléatoire décrite. Ceci permet de simplifier l'équation en formulation faible, en supprimant les termes impliquant des laplaciens et des dérivées en temps, grâce au résultat suivant.

Proposition 3.

$$\partial_s \phi_{n,s}^{t,z} = -\alpha_n \Delta_{L^n} \phi_{n,s}^{t,z}$$

Etape 3 - Tension La troisième étape consiste à s'assurer de la tension de $(u^n, l^n)_{n\geq 1}$, c'est à dire que toute sous-suite possède une sous-suite convergente. Ceci nous assure de l'existence de limites potentielles, et la dernière étape consistera à montrer qu'une seule limite est possible.
Ici, nous aurons besoin d'un critère un peu plus précis que celui de la tension, appelé *critère de C-tension*. Il permet d'obtenir à la fois la tension de la suite et la continuité en espace des limites. Ce critère s'énonce de la façon suivante.

Définition 7. Soient $(f_n)_{n\geq 1} \in D(\mathbb{R}_+ \times C_{[0,1]}(\mathbb{R}))$ et $T > 0$. La séquence $(f_n)_{n\geq 1}$ est C - tendue dans $D([0,T], C_{[0,1]}(\mathbb{R}))$ si $\forall \eta > 0$, $\forall t \geq 0$ rationnel, il existe un compact $\Gamma_{\eta,t} \subset C_{[0,1]}(\mathbb{R})$ tel que :

$$\inf_{n\in\mathbb{N}} \mathbb{P}(f_n(t) \in \Gamma_{\eta,t}) \geq 1 - \eta$$

et si de plus, pour tout $\epsilon > 0$,

$$\lim_{\delta \to 0} \limsup_{n \to +\infty} \mathbb{P}\Big(\sup_{\substack{t_1 - t_2 < \delta \\ 0 \leq t_2 \leq t_1}} ||f^n(t_1) - f^n(t_2)|| > \epsilon \Big) = 0$$

où $||.||$ est la norme définie par $\forall f \in C_{[0,1]}(\mathbb{R})$,

$$||f|| = \sum_{k=1}^{+\infty} 2^{-k} \sup_{|x| \leq k} |f(x)|$$

Dans cette définition, $D(\mathbb{R}_+ \times C_{[0,1]}(\mathbb{R}))$ représente l'ensemble des fonctions càdlàg (continues à gauche, limitées à droite) de $\mathbb{R}_+$ à valeurs dans $C_{[0,1]}(\mathbb{R})$, et est aussi appelé *espace de Skorokhod*. Le livre [4] constitue une bonne référence sur ce type d'espace et les problématiques associées.
La première condition, appelée critère de *weak compact containment*, provient de ce que les densités sont bornées par 1, et du choix de la norme choisie. La deuxième se démontre en majorant les moments d'ordre p pour $p \geq 1$, et ce pour chacun des termes intervenant dans l'équation en formulation faible. Le choix effectué pour la fonction test des probabilités de transition d'une marche aléatoire simple permet d'utiliser tous les résultats connus sur ces probabilités de transition, rappelés dans [10]. Pour les termes de martingale, un résultat permet de passer d'inégalités sur le supremum à des inégalités sur l'espérance du crochet de la martingale, qui peut être calculé ou majoré facilement : *l'inégalité de Burkholder-Davis-Gundy.*

Théorème 4. *Soit $(X_t)_{t \geq 0}$ une martingale locale partant de 0. Notons $< X >_t$ sa variation quadratique lorsqu'elle est définie, et X_t^* son supremum sur $[0, t]$. Soit p entier naturel, et supposons que X^* admet un moment d'ordre p. Il existe alors des constantes c_p, C_p telles que, pour tout temps d'arrêt τ :*

$$c_p \mathbb{E}[< X >_\tau^{\frac{p}{2}}] \leq \mathbb{E}[(X_\tau^*)^p] \leq C_p \mathbb{E}[< X >_\tau^{\frac{p}{2}}] \tag{4}$$

Etape 4 - Unicité de la limite Une fois l'existence de limites montrée, il reste à prouver qu'une seule limite est possible. Ceci se fait grâce à un outil présenté dans le paragraphe suivant.

3.3 Dual et unicité de la solution

L'outil nécessaire pour montrer l'unicité de la solution est un outil classique en génétique des populations, appelé *dual*. L'idée générale est de suivre les ancêtres d'un échantillon d'individus, afin de reconstruire son arbre généalogique. De cette façon, l'arbre généalogique des individus et l'origine géographique des ancêtres sont retrouvés. Ceci peut par exemple permettre d'estimer la diversité génétique dans l'échantillon : les mutations portées par une partie seulement des individus ne peuvent être apparues qu'après le dernier ancêtre commun à l'ensemble de l'échantillon. De même, si la répartition initiale des types des individus n'est pas homogène en espace, connaître la localisation potentelle de l'ancêtre d'un individu permet d'estimer la probabilité qu'il soit d'un type fixé.
Dans le modèle le plus classique, respectant les mêmes hypothèses que le modèle de Moran (population de taille finie fixée, panximie, pas de sélection), l'objet obtenu est un arbre composé initialement de N branches, dans lequel chaque paire de branches fusionne indépendamment à taux $\binom{N}{2}$. Il s'agit du *coalescent de Kingman*
Point biologique *- Le terme coalescent est utilisé car sous ces hypothèses, un échantillon d'individus finit toujours par se retrouver un ancêtre commun. Cet objet apparaît comme limite des généalogies dans beaucoup de modèles de reproduction, y compris certains dans lesquelles les hypothèses du coalescent de Kingman sont a priori violées. Souvent, les généalogies se comportent comme un coalescent de Kingman à condition de considérer comme taille de la population une "taille de population effective" encodant différents effets.*
Lorsqu'une structuration spatiale est ajoutée, il n'y a plus forcément coalescence de l'ensemble des lignées à la limite. Le cas le plus simple étant celui de deux îles isolées l'une de l'autre : deux lignées situées sur deux îles différentes ne se trouveront alors jamais d'ancêtre commun. Dans des situations

similaires à notre modèle, dans lequel les individus sont identifiables par leur position et ne se reproduisent qu'avec leurs plus proches voisins, il est possible de repérer les ancêtres par leur position géographique dans l'espace. Les lignées ancestrales se comportent alors à la limite comme des mouvements browniens, et deux individus se trouvent un ancêtre commun lorsque les mouvements browniens ont passé un certain temps en la même position, au sens des *temps locaux*, définis de la façon suivante.

Définition 8. Soit X une semimartingale continue, et soit $a \in \mathbb{R}$. Le *temps local de X en a*, noté $(L_t^a(X))_{t\geq 0}$, est le processus croissant à valeurs dans $\mathbb{R}_+$ vérifiant les trois identités suivantes :

$$|X_t - a| = |X_0 - a| + \int_0^t \text{sgn}(X_s - a)dX_s + L_t^a(X)$$
$$(X_t - a)^+ = (X_0 - a)^+ + \int_0^t \mathbb{I}_{X_s > a} dX_s + \frac{1}{2}L_t^a(X)$$
$$(X_t - a)^- = (X_0 - a)^- - \int_0^t \mathbb{I}_{X_s \leq a} dX_s + \frac{1}{2}L_t^a(X)$$

Plusieurs propriétés utiles des temps locaux figurent dans [23].
Il est nécessaire d'utiliser les temps locaux car lorsque le mouvement brownien atteint un point, il y reste un temps nul. Il n'est donc pas possible de sommer directement les temps de séjour consécutifs. Les temps locaux permettent d'obtenir un équivalent de la durée de séjour.
Enfin, dans le cas où de la sélection est présente, les événements de reproduction associés sont caractérisés par la présence de deux parents potentiels. Seule la connaissance de leurs types permet de conclure. Or, l'un des objectifs de la construction d'un dual est justement d'estimer le type d'un individu. Il est donc nécessaire de suivre *les deux parents potentiels*, ce qui se traduit par un branchement du mouvement brownien associé à la position de l'ancêtre. Cette idée est une adaptation de l'Ancestral Selection Graph [22, 29], définit sous des hypothèses de reproduction similaires à celles du modèle de Moran.
Lorsque la population est localement finie, donc hors de la limite $n \longrightarrow +\infty$, le dual est défini en utilisant la théorie des représentations graphiques, développée dans [16]. Chaque individu se voit associer un tuyau, les événements de reproduction correspondant à des connexions établies entre eux. Dans le cas d'un événement de sélection, le tuyau du parent est connecté à celui de l'enfant, de façon à ce que tout fluide circulant dans le tuyau du parent passe dans celui de l'enfant. Dans le cas d'un événement de vote, le tuyau de l'enfant est de plus bouché juste avant la connexion. A partir de cette construction, le dual est obtenu en injectant du fluide au temps présent dans les tuyaux correspondant aux individus dont nous voulons connaître les généalogies. L'unicité du dual, que nous noterons dans la suite ζ_n, est claire. La figure 2 présente un exemple de construction du dual à partir de la représentation graphique.
La limite de ce dual, dans le cadre du modèle avec structuration spatiale que nous avons introduit, est la suivante.

Théorème 5. *Sous les conditions énoncées précédemment pour les limites des paramètres quand $n \to +\infty$ et si $\gamma > 0$, ζ^n converge en distribution vers une limite dans laquelle les particules bougent selon des mouvements browniens sur $\mathbb{R}$, branchants à taux $\theta\beta$. Chaque paire de particules coalesce lorsque le temps local en 0 de la différence entre leurs positions devient supérieure à $\frac{\alpha\tau}{\gamma}$, où $\tau \sim Exp(1)$ est indépendante du mouvement des particules.*
Dans le cas particulier où $\gamma = 0$, les particules ne coalescent plus, mais le reste de la limite reste inchangé.

Une preuve de ce résultat peut se trouver dans l'article [11].

4 Perspectives

4.1 Une autre approche du problème : le processus Λ-Fleming Viot spatial

L'approche que nous avons présentée précédemment n'est pas la seule permettant de modéliser les populations en expansion. Nous présentons maintenant rapidement un autre modèle, qui peut servir

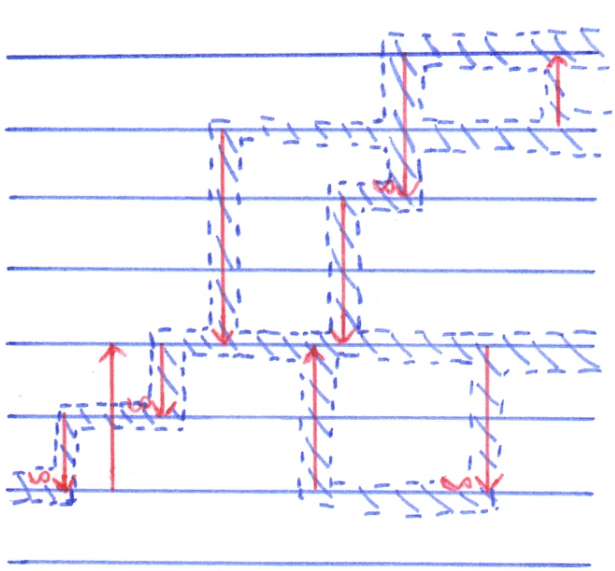

FIGURE 2 – Construction du dual à partir de la représentation graphique. Le sens d'écoulement du temps est de gauche à droite si nous considérons le dual, et de droite à gauche pour le processus de reproduction. La zone hachurée représente les localisations des ancêtres potentiels pour l'individu considéré, situé initialement sur la deuxième ligne en partant du bas.

de base à l'étude des populations en expansion : le processus Λ-Fleming Viot spatial avec sélection, introduit dans [12]. Il s'agit d'une variante du processus Λ-Fleming Viot spatial (sans sélection) défini dans [2].
Nous travaillons cette fois-ci dans un espace de dimension d quelconque. Les individus sont toujours de deux types 0 et 1, où 1 est avantagé sélectivement et où 0 peut représenter des individus potentiels. Cette fois-ci, la population sera supposée de taille infinie en chaque point, afin de pouvoir travailler avec des densités. Notons, pour $t \geq 0$ et $x \in \mathbb{R}^d$, $\omega_t(x)$ pour la densité d'individus de type 0 en x au temps t - il est ici plus simple de suivre la densité en individus de type 0 que de type 1, pour des raisons qui apparaitront plus tard.
La stratégie utilisée pour définir le processus de reproduction consiste à travailler de façon *événement-centrée*. Un événement de reproduction va affecter une boule de centre x et de rayon r. Il est caractérisé par un paramètre d'impact u, qui représente la proportion des individus dans la boule qui va être affectée par l'événement, une position z du parent qui va remplacer cette portion de la boule par ses descendants, et une deuxième position parentale z' qui servira pour les événements sélectifs. Chaque événement est de plus neutre avec une probabilité $1-s$, et sélectif avec une probabilité s. Dans ce deuxième cas, le parent en z ne se reproduit que si il est de type 1. Sinon, c'est le parent en z' qui se reproduit.
Si nous notons τ et τ' les types de deux parents, les densités évoluent alors de la façon suivante :

— lors d'un événement de reproduction neutre : $\forall y \in B_x(r), \omega_t(y) = (1-u)\omega_{t^-}(y) + u\mathbb{I}_{\tau=0}$
— lors d'un événement de reproduction sélectif : $\forall y \in B_x(r), \omega_t(y) = (1-u)\omega_{t^-}(y) + u\mathbb{I}_{\tau=\tau'=0}$

Les événements de reproduction sont générés à partir d'un processus ponctuel de Poisson, défini sur $\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}_+\times(0,1]$ et d'intensité $dt\otimes dx\otimes\mu(dr)\nu_r(du)$ où ν_r dépend du r tiré par μ. Les positions des parents potentiels et le type de l'événement sont tirés au cas par cas après la génération de l'événement.
Sous l'échelle $s_n = \frac{\log(n)}{n}$, l'objet le plus facile à étudier pour mesurer la diversité génétique dans un échantillon est le dual, qui se comporte à la limite comme un système de mouvements browniens branchants, ne coalesçant jamais. Afin de présenter ce résultat, notons $P^n(p)$ l'évolution des positions des ancêtres de l'individu situé en p, et $B(p,V)$ un mouvement brownien branchant à taux V partant de p. Pour $\epsilon > 0$ et $T > 0$, définissons les événements suivants :

$$D_n(\epsilon,T) = \{\forall l \in P^n(p), \exists l' \in BBM(p,V) : \sup_{t\in[0,T]} |l(t) - l'(t)| \leq \epsilon\}$$

$$D'_n(\epsilon, T) = \{\forall l \in BBM(p, V), \exists l' \in P^n(p) : \sup_{t\in[0,T]} |l(t) - l'(t)| \leq \epsilon\}$$

Le résultat suivant est alors démontré dans l'article [12].

Théorème 6. *Soient $T \leq \infty, p \in \mathbb{R}^2$. Alors, $\forall \epsilon > 0, \exists N_\epsilon \in \mathbb{N}, \forall n \geq N_\epsilon$, il existe un couplage entre $BBM(p, \epsilon)$ et $P^n(p)$ tel que :*

$$\mathbb{P}(D_n(\epsilon, T) \cap D'_n(\epsilon, T)) \geq 1 - \epsilon$$

4.2 Questions ouvertes

L'étude de la diversité génétique dans les populations en expansion est, comme nous l'avons vu en introduction, une problématique d'importance. Il y a grand besoin de modèles explicatifs pour les *patterns* observés, mais aussi d'outils statistiques liés.
Dans les deux cas, il reste encore beaucoup à faire. Ainsi, les modèles construits ne correspondent pas toujours à ce qui est observé expérimentalement. Par exemple, la tentative de Durrett de construire un modèle permettant d'expliquer les *patterns* de diversité intratumorale a échoué, prédisant des secteurs trop petits pour être observés par biopsie [9]. De même, la construction d'outils statistiques est en plein développement, permettant pour l'instant de traiter le cas sans expansion de population. Citons par exemple les articles [3, 30] pour l'inférence des paramètres démographiques d'une population dans le cadre du processus de Λ-Fleming Viot spatial. Il reste de plus beaucoup de situations plus complexes que celles déjà étudiées à aborder. Le cas de variations de l'avantage sélectif des individus, causées entre autres par un environnement hétérogène, a commencé à être exploré dans [5].
Enfin, une autre question d'intérêt est l'influence des paramètres démographiques sur l'apparition et la taille des secteurs pouvant être observés dans des conditions similaires à l'expérience sur les bactéries fluorescentes.

Références

[1] Adrian Baddeley, Imre Bárány, and Rolf Schneider. Spatial point processes and their applications. *Stochastic Geometry : Lectures given at the CIME Summer School held in Martina Franca, Italy, September 13–18, 2004*, pages 1–75, 2007.

[2] N. Barton, A. Etheridge, A. Véber, et al. A new model for evolution in a spatial continuum. *Electronic Journal of Probability*, 15 :162–216, 2010.

[3] Nick H Barton, Alison M Etheridge, Jerome Kelleher, and Amandine Véber. Inference in two dimensions : allele frequencies versus lengths of shared sequence blocks. *Theoretical population biology*, 87 :105–119, 2013.

[4] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[5] Niloy Biswas, Alison Etheridge, and Aleksander Klimek. The spatial lambda-fleming-viot process with fluctuating selection. *arXiv preprint arXiv :1802.08188*, 2018.

[6] I. Bozic, J. Reiter, B. Allen, T. Antal, K. Chatterjee, P. Shah, Y. S. Moon, A. Yaqubie, N. Kelly, D. T Le, et al. Evolutionary dynamics of cancer in response to targeted combination therapy. *elife*, 2 :e00747, 2013.

[7] Nicolas Champagnat and Sylvie Méléard. Invasion and adaptive evolution for individual-based spatially structured populations. *Journal of Mathematical Biology*, 55(2) :147, 2007.

[8] C.Mueller and R.Tribe. Stochastic PDE's arising from the long range contact and long range voter processes. *Probability theory and related fields*, 102(4) :519–545, 1995.

[9] R. Durrett. Genealogies in Growing Solid Tumors. *bioRkiv*, page 244160, 2018.

[10] R. Durrett and W. T. Fan. Genealogies in expanding populations. *Annals of Applied Probability*, 26(6) :3456–3490, 2016.

[11] Richard Durrett, Mateo Restrepo, et al. One-dimensional stepping stone models, sardine genetics and Brownian local time. *The Annals of Applied Probability*, 18(1) :334–358, 2008.

[12] A. Etheridge, N. Freeman, S. Penington, and D. Straulino. Branching Brownian motion and selection in the spatial Λ-Fleming-Viot process. *The Annals of Applied Pprobability*, 27(5) :2605–2645, 2017.

[13] Alison Etheridge. *Some Mathematical Models from Population Genetics : École D'Été de Probabilités de Saint-Flour XXXIX-2009*, volume 2012. Springer Science & Business Media, 2011.

[14] Joseph Felsenstein. A pain in the torus : some difficulties with models of isolation by distance. *The American Naturalist*, 109(967) :359–368, 1975.

[15] Ronald Aylmer Fisher. The wave of advance of advantageous genes. *Annals of eugenics*, 7(4) :355–369, 1937.

[16] David Griffeath. *Additive and cancellative interacting particle systems*, volume 724. Springer, 2006.

[17] O. Hallatschek and D. R. Nelson. Life at the front of an expanding population. *Evolution : International Journal of Organic Evolution*, 64(1) :193–206, 2010.

[18] Oskar Hallatschek and David R Nelson. Gene surfing in expanding populations. *Theoretical population biology*, 73(1) :158–170, 2008.

[19] G. Hewitt. The genetic legacy of Quaternary ice ages. *Nature*, 405(6789) :907, 2000.

[20] Y. Iwasa, M. A. Nowak, and F. Michor. Evolution of resistance during clonal expansion. *Genetics*, 172(4) :2557–2566, 2006.

[21] K. Kaveh, Y. Takahashi, M. A Farrar, G. Storme, M. Guido, J. Piepenburg, J. Penning, J. Foo, K. Z Leder, and S. K Hui. Combination therapeutics of nilotinib and radiation in acute lymphoblastic leukemia as an effective method against drug-resistance. *PLoS computational biology*, 13(7) :e1005482, 2017.

[22] S.M Krone and C. Neuhauser. Ancestral processes with selection. *Theoretical population biology*, 51(3) :210–237, 1997.

[23] Jean-François Le Gall et al. *Brownian motion, martingales, and stochastic calculus*, volume 274. Springer, 2016.

[24] K. Leder, J. Foo, B. Skaggs, M. Gorre, C. L Sawyers, and F. Michor. Fitness conferred by bcr-abl kinase domain mutations determines the risk of pre-existing resistance in chronic myeloid leukemia. *PloS one*, 6(11) :e27682, 2011.

[25] Thomas Milton Liggett. *Interacting particle systems*, volume 276. Springer Science & Business Media, 2012.

[26] Sylvie Meleard and Vincent Bansaye. *Stochastic Models for Structured Populations : Scaling Limits and Long Time Behavior*, volume 1. Springer, 2015.

[27] Patrick Alfred Pierce Moran. Random processes in genetics. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 54, pages 60–71. Cambridge University Press, 1958.

[28] S. M Mumenthaler, J. Foo, K. Leder, N. C Choi, D. B Agus, W. Pao, P. Mallick, and F. Michor. Evolutionary modeling of combination treatment strategies to overcome resistance to tyrosine kinase inhibitors in non-small cell lung cancer. *Molecular pharmaceutics*, 8(6) :2069–2079, 2011.

[29] C. Neuhauser and S.M Krone. The genealogy of samples in models with selection. *Genetics*, 145(2) :519–534, 1997.

[30] Raazesh Sainudiin and Amandine Veber. Full likelihood inference from the site frequency spectrum based on the optimal tree resolution. *bioRxiv*, page 181412, 2018.

[31] A. Templeton. Out of Africa again and again. *Nature*, 416(6875) :45, 2002.

[32] C. Tomasetti and D. Levy. Role of symmetric and asymmetric division of stem cells in developing drug resistance. *Proceedings of the National Academy of Sciences*, 107(39) :16766–16771, 2010.