Conception et identification à partir de données biologiques d'un modèle dédié aux interactions entre cellules souches hématopoïétiques et stromales

Julie Favre

Rapport de stage

Master en mathématiques appliquées

Conception et identification à partir de données biologiques d'un modèle dédié aux interactions entre cellules souches hématopoïétiques et stromales

Auteur :
Julie Favre
N° Sciper 226481

Superviseur :
Jean Clairambault

Octobre 2016 - Février 2017

TABLE DES MATIÈRES

1 | ENVIRONNEMENT DU STAGE ET OBJECTIFS

Ce rapport a pour but de donner un aperçu du travail réalisé lors du stage requis pour le Master en Ingénierie mathématique à l'EPFL. Mon stage, qui a duré 4 mois (octobre 2016 - février 2017), s'est déroulé à Paris au sein du Laboratoire Jacques-Louis Lions (*LJLL*) et de l'équipe de recherche MAMBA - *Modelling and Analysis for Medical and Biological Applications* de l'INRIA.

L'équipe MAMBA, en collaboration avec plusieurs équipes de biologistes de l'hématopoïèse et cliniciens de la leucémie aiguë myéloïde ont soumis en juillet 2016 une réponse à un appel à projets national sur le thème de l'Hétérogénéité des Tumeurs dans leur Ecosystème (HTE), appel auquel ils ont été lauréats. Ce stage s'inscrit donc dans le cadre de ce projet HTE qui durera quatre ans et qui a pour but ultime de mieux comprendre l'hétérogénéité entre les cellules dans la population tumorale, avec ou sans traitement médicamenteux, et comment les échanges avec le micro-environnement tumoral déterminent la croissance et l'hétérogénéité des populations de cellules, afin de définir des schémas thérapeutiques prenant en compte ces échanges et évitant l'émergence de populations cellulaires tumorales résistantes.

Initialement, mon rôle pour ce stage consistait à utiliser le matériel et les données provenant du laboratoire *Migration et Différenciation des Cellules Souches Hématopoïétiques* pour faire le lien avec un modèle EDP ayant pour but de représenter les échanges entre les cellules souches hématopoïétiques, saines ou leucémiques, et les cellules stromales de soutien. Plus précisément, le cahier des charges comprenait : l'assimilation théorique du sujet étudié dans le cadre du projet HTE et des différentes méthodologies pour le traitement des données ; la définition du passage de la représentation statistique de ces données à une structuration des populations cellulaires en traits phénotypiques pertinents, afin de créer une interface entre les données et les modèles EDP structurés en phénotypes ; la modélisation numérique des modèles EDP afin d'étudier le comportement asymptotique des populations cellulaires, en fonction de variations phénotypiques contrôlées et, si possible, de messages contrôlés, entre les deux types de populations cellulaires. Nous verrons au cours de ce rapport que certains objectifs n'ont pas été réalisables dans le cadre de ce projet, notamment dû au fait que le projet HTE n'était qu'à son commencement.

Concernant l'environnement de ce stage, dont la langue de travail était le français, celui-ci était supervisé par le Dr. Jean Clairambault du LJLL, membre de l'équipe MAMBA et directeur de recherche à l'INRIA, et co-encadré par Delphine Salort du Laboratoire de Biologie Quantitative et Computationnelle - LBQC. Les interactions durant ce stage étaient donc principalement avec les deux personnes mentionnées, mais également avec les six autres membres principaux du projet HTE : interactions faites par le biais de diverses réunions HTE, mais également pour certains par des explications et observations de leur travail. Finalement, trois dernières personnes sont également à citer et à remercier de leur aide dans le cadre de ce stage : Hicham Janati, stagiaire de l'équipe MAMBA, qui m'a permis, notamment par son rapport de stage [12], d'assimiler différentes méthodes pour le traitement statistique des données biologiques (voir chapitre 4), Camille Pouchol, qui a apporté son aide concernant l'analyse du modèle mathématique et sa modélisation numérique (voir chapitre 5), notamment par son rapport de stage également [18], et enfin le Dr. Hugues Richard, du LBQC, dans mes recherches sur le sujet des *Single Cell* (voir chapitre 4.3).

2 | LE PROJET HTE

Il a déjà été mentionné ci-dessus lors de la présentation de l'environnement de ce stage, et plus particulièrement du projet HTE, certains éléments qu'il est important de définir et d'approfondir. Ce chapitre présente donc de manière théorique les différents concepts associés au projet HTE, notamment les notions de *cellules souches hématopoïétiques*, de *cellules stromales (de soutien)* et de *leucémie aiguë myéloïde (LAM)*, ainsi qu'une introduction plus précise des objectifs et hypothèses du projet HTE.

2.1 LES CELLULES SOUCHES HÉMATOPOÏÉTIQUES ET L'HÉMATOPOÏÈSE

Une Cellule Souche Hématopoïétique (CSH), est un type de cellule à l'origine de tous les types de cellules sanguines : *érythrocytes* ou globules rouges, *leucocytes* ou globules blancs et *plaquettes*. Chez l'Homme, elles sont situées, dans la vie intra-utérine, dans le tissu conjonctif jusqu'au $2^{ème}$ mois, dans le foie fœtal du $2^{ème}$ au $6^{ème}$, dans la moelle osseuse à partir du $4^{ème}$ mois. Après la naissance, elles se situent exclusivement dans la moelle osseuse. Elles possèdent des capacités de prolifération, d'autorenouvellement et de différenciation en lignées multiples.

L'hématopoïèse désigne le processus physiologique de formation, de maturation et de remplacement continu et régulé des cellules sanguines, à partir de cellules souches hématopoïétiques. Elle décrit les étapes de prolifération et de différentiation des CSH qui engendrent des générations de précurseurs dont la différentiation terminale fournit les lignées de cellules sanguines matures mentionnées ci-dessus (érythrocytes, leucocytes et plaquettes). La prolifération consiste en quatre phases : G_1, S, G_2, M et à la fin de la phase M a lieu la division cellulaire où deux cellules filles son produites, soit des progéniteurs, dont la production est appelée *différentiation*, soit avec les mêmes propriétés biologiques que la cellule parente - *autorenouvellement*.

2.2 LA "NICHE" HÉMATOPOÏÉTIQUE ET LES CELLULES STROMALES

Ce sont les propriétés de CSH (multipotence, quiescence, autorenouvellement, différenciation) qui assurent la durabilité de l'hématopoïèse. Or, celles-ci sont non seulement finement régulés par des mécanismes intrinsèques mais également extrinsèques qui font intervenir des interactions avec les constituants du micro-environnement médullaire des CSH appelé *niche hématopoïétique*.

La niche hématopoïétique est une structure complexe associant notamment la matrice extracellulaire, des cytokines, chimiokines, et de différents types cellulaires, morphologiquement distincts, tels que des cellules endothéliales, des adipocytes, des cellules de lignée ostéoblastique et les cellules stromales mésenchymateuses. Les cellules stromales mésenchymateuses, qui sont les progénitures des cellules souches mésenchymateuses, sont le composant le plus important de la niche hématopoïétique. Elles fournissent un support structural et fonctionnel aux CSH.

L'interaction entre les CSH et la niche hématopoïétique, en particulier les cellules stromales, est pensée de façon à équilibrer leur capacité à survivre et s'autorenouveller avec la différenciation multi-lignages. Or, cette balance est primordiale pour la pérennité à long terme des CSH in vivo.

La façon dont les signaux extrinsèques de cellules de la niche affectent les voies de signalisation des cellules souches dans la régulation de leur survie, différentiation et autorenouvellement est l'une des questions essentielles dans l'étude du micro-environnement. Plus généralement, l'étude du dialogue entre les CSH et les cellules stromales est au cœur de nombreuses recherches, notamment concernant la leucémie aiguë myéloïde, comme il sera expliqué dans la suite de ce rapport.

2.3 LA LEUCÉMIE AIGUË MYÉLOÏDE OU "LAM"

Un défaut dans l'hématopoïèse peut provoquer des maladies telles qu'une aplasie des globules rouges ou une hémopathie, dont la forme maligne comprend notamment les lymphomes et les leucémies.

La leucémie aiguë myéloïde ou *LAM* est un cancer des cellules sanguines (de lignée myéloïde) qui se développe dans la moelle osseuse, à partir de CSH ayant acquis de multiples aberrations génomiques ou chromosomiques. La LAM allie le blocage de la maturation et différentiation, laissant apparaitre une accumulation de cellules myéloides immatures, ainsi qu'un avantage à la prolifération conduisant à l'inondation de la moelle osseuse par des cellules immatures et proliférantes.

On peut distinguer de manière fonctionnelle et génétique trois catégories de cellules dans la LAM : les cellules souches pré-leucémiques, les cellules souches leucémiques et les *blastes*. Pour aboutir à une leucémie complète, il faut une accumulation d'événements génétiques dans des cellules souches hématopoïétiques, qui transforment ces CSH en cellules pré-leucémiques voire leucémiques. Ainsi, le développement de LAM (qui est par définition rapide, par opposition à une leucémie chronique qui se développe au fil des mois voire des années) est combinaison de deux processus en parallèle : une invasion progressive (et incontrôlée) de la moelle osseuse ainsi qu'un processus d'acquisition de lésions génétiques, qui façonnent l'architecture du clone malin.

2.4 HYPOTHÈSES ET OBJECTIFS DU PROJET HTE

Comme il a été déjà mentionné plus tôt, la biologie des cellules souches hématopoïétiques est régulée par d'autres cellules de la niche, notamment les cellules stromales mésenchymateuses. L'influence de cellules stromales sur les CSH est incontestable et incontestée. Cependant, la bi-directionnalité de ce dialogue n'est de loin pas évidente. Or, la compréhension de ce dialogue (qu'il soit uni- ou bidirectionnel) est nécessaire car la dérégulation de ces mécanismes peut être à l'origine de pathologies du système sanguin telles que la LAM. En effet, lors du développement de la LAM, le dialogue entre les CSH, saines ou mutées, et le micro-environnement peut subir de nombreux changements.

La question principale dans les recherches sur la LAM est de mieux comprendre les processus de son développement et plus particulièrement d'expliquer la raison pour laquelle après un premier événement pré-leucémique, on a une expansion de ces cellules souches pré-leucémiques par rapport aux cellules souches normales.

Ce projet HTE a comme hypothèse la bidirectionnalité du dialogue entre les CSH et les cellules stromales, c'est-à-dire qu'il suppose que les cellules stromales peuvent être influencées par les cellules souches hématopoïétiques. La légitimité de cette assomption provient de plusieurs études où les données obtenues ont montré que non seulement les cellules stromales commu-

niquaient avec les CSH, mais que cela allait également dans l'autre sens, résultat que l'on peut retrouver dans des publications telles que [6] et [11].

En effet, dans le premier se posait notamment la question de savoir si le transcriptome des cellules stromales pouvait être influencé par les CSH, avec comme stratégie de comparer les transcriptomes de cellules stromales en présence de CSH ou non. Le résultat obtenu par cette étude est que la fonction stromale existe avant le contact avec les cellules souches hématopoïétiques dans le cellules stromales qui ont une capacité de soutien, mais elle est enrichie par le contact. Concernant le second article, précisons tout d'abord que dans des conditions d'état stable, les CSH existent dans deux états : un état de quiescence et un état actif. Les CSH quiescents sont activés sous des conditions de stress hématopoïétique telles que les infections, les blessures, les agents cytotoxiques, etc. C'est dans ce contexte que se plaçait cette étude : elle cherchait à mieux comprendre l'interaction des signaux entre les cellules stromales et les CSH, pendant un stress hématopoïétique. Les résultats ont montré que les cellules de la niche hématopoïétique répondent au contact des CSH normales et malignes avec une réponse inflammatoire. Une des composantes de cette réponse est l'augmentation dans les cellules stromales du facteur de croissance de tissu conjonctif (*CTGF*). Or, bien que les effets du CTGF soient incertains, leurs données suggèrent qu'il empêche l'arrêt du cycle cellulaire et qu'il peut promouvoir l'autorenouvellement des CSH. En résumé, le résultat principal de cette recherche est d'avoir trouvé que les cellules hématopoïétiques précoces induisaient dans les cellules stromales un profil d'expression inflammatoire associé aux réponses de stress.

Revenons maintenant à la question principale du projet HTE qui est de connaitre les raisons pour lesquelles après un premier événement pré-leucémique, on constate une invasion des cellules pré-leucémiques et un avantage sélectif des cellules mutées par rapport aux cellules saines. L'hypothèse de travail de ce projet, qu'il cherche à valider, est que les CSH mutées sont capables de reprogrammer le stroma. Plus précisément, on suppose ici que les CSH avec les premières lésions se produisant dans le clone LAM, c'est-à-dire les mutations pré-leucémiques, et avec les mutations ultérieures, altèrent leur environnement et en particulier les cellules stromales mésenchymateuses, et qu'à leur tour, les cellules stromales modifiées favorisent l'expansion des cellules de la leucémie. Ainsi, le développement de la maladie ne dépend pas seulement de l'évolution cellulaire du clone leucémique mais également des mécanismes paracrines dus aux interactions cellules leucémiques - stroma.

Les objectifs de ce projet sont donc d'essayer de comprendre au mieux le dialogue entre les cellules souches hématopoïétiques saines, pré-leucémiques et leucémiques avec le stroma médullaire, et de modéliser ce dialogue, dans le but ultime d'aider à identifier de nouvelles cibles pour éradiques le clone malin de la LAM, c'est-à-dire identifier des nouveaux moyens thérapeutiques contre la LAM. De manière plus concrète, il s'agit tout d'abord de faire la preuve du concept qu'il est possible de mettre en évidence les modifications du dialogue entre les CSH et les cellules stromales mésenchymateuses. L'étape suivante est d'être capable d'identifier chez les patients la signature de ces modifications. Enfin, il s'agit de modéliser de manière mathématique à la fois le dialogue et le développement de la LAM dans le contexte d'une hématopoïèse.

3 MANIPULATIONS BIOLOGIQUES

Concernant la partie biologique de ce projet HTE, il est tout d'abord nécessaire de soulever un aspect qui rend les différentes études sur le sujet assez complexes : les cellules souches hématopoïétiques sont des cellules très rares, il est donc difficile de s'en procurer et les expériences ne peuvent pas être extrêmement nombreuses. Les premières données obtenues dans le cadre de ce projet ont par ailleurs été obtenues à partir de CSH murines, et que ce n'est qu'à partir du mois de novembre que des cellules humaines en provenance d'un service hospitalier ont pu être utilisées.

Sachant que la première partie du projet HTE consiste à faire la preuve du concept et donc d'étudier les lésions pré-leucémiques précoces, la première étape consiste donc à faire des co-cultures, toujours avec des lignées stromales murines, car celles-ci sont disponibles, mais d'utiliser des CSH humaines pré-leucémiques. Cependant, bien que différents tests ont montré que la méthodologie utilisée dans des modèles de co-culture de CSH murines et de lignées stromales murines était adaptée pour des cellules humaines également, les premières manipulations ont eu pour but de vérifier la faisabilité de co-culture entre lignées stromales murines et cellules humaines CD34+ [1]. Le principe général de ces manipulations est de suivre un modèle de co-culture court de quatre jours et à l'issu de ces quatre jours, de faire une analyse par cytométrie en flux [2] et, éventuellement, de procéder à un tri cellulaire (à l'aide de CD45 [3] essentiellement) permettant de séparer les cellules stromales et CSH afin de les envoyer en analyse transcriptomique dans le but d'observer s'il existe une différence avant et après contact, de voir les profils d'expression, et éventuellement de déterminer quels sont les réseaux géniques activés par le contact entre les deux populations (voir Chapitre 4). Pour ces premières manipulations, la cytométrie en flux a permir de voir sur les graphiques qu'il y avait une apparition de petites populations, colonies qui se développaient, c'est-à-dire un foyer d'hématopoïèse, montrant qu'il y a bien un soutien de l'hématopoïèse.

J'ai pour ma part, eu la chance d'assister à l'une des manipulations tests faites par le Dr. Pierre Hirsch à partir d'échantillons de patients, plus précisément la mise en co-culture de cellules stromales murines et de cellules CD34 à partir de cellules congelées d'un patient en rémission complète (après avoir subi de la chimiothérapie) ainsi que l'arrêt de celle-ci, qui m'a permis de comprendre de manière globale (sans rentrer dans des détails trop techniques) le processus.

Pour l'établissement de co-culture dans le cadre de cette manipulation du 15 décembre 2016, le jour avant la mise en co-culture, il a fallu préparer les cellules stromales à partir des boîtes de culture et les compter, puis isoler le nombre de cellules désiré (à noter que dans le cas des cellules stromales, un trop grand nombre est mauvais, car celles-ci se multiplient et se divisent rapidement et lorsqu'elles sont trop nombreuses, elles finissent par mourir), et les remettre dans

1. L'antigène de cellules progénitrices hématopoïétiques CD34 est une protéine codée chez les humaines par le gène CD34. Les cellules CD34+, c'est-à-dire les cellules qui expriment du CD34 sont normalement trouvées dans la moelle osseuse ou le cordon ombilical sous la forme de cellule hématopoïétique.

2. La cytométrie en flux est une technique faisant défiler des cellules en suspension à grande vitesse dans le faisceau d'un laser permettant de les compter et de mesurer les caractéristiques individuelles pouvant être détectées par un composé fluorescent. Elle permet donc d'identifier des sous-populations de cellules spécifiques grâce à leur caractéristiques fluorescentes et propriétés physiques, de les isoler et de les trier.

3. Le CD45 est un antigène (antigène leucocytaire commun), qui est en particulier un marqueur présent au niveau de toutes les cellules souches hématopoïétiques.

un volume de milieu pour le stroma, à 33°C. Le jour de la mise en co-culture, il a d'abord fallu décongeler les cellules du patient (congelées à -80°C dans l'azote liquide), de calculer la proportion de cellules vivantes/cellules mortes (grâce au bleu de trypan, qui est une méthode de coloration des cellules mortes), puis de procéder à un tri cellulaire magnétique CD34[4] pour récupérer les cellules CD34+. Ces cellules ensuite été comptées et mises dans le milieu utilisé pour la co-culture. Après avoir vérifié les cellules stromales au microscope et aspiré le milieu des puits des cellules stromales, il s'agissait finalement de mettre du milieu de culture et les cellules hématopoïétiques dans chaque puits.

Le 19 décembre 2016, après quatre jours de co-culture, il fallait procéder à l'arrêt de la co-culture. Pour cela, la première étape consistait à récupérer les surnageants de chaque puits afin de récupérer les cellules non adhérentes. Puis, de la trypsine[5] était mise dans chaque puits afin de décoller les cellules qui étaient ensuite récupérées dans le même tube, et toutes les cellules ont été comptées. Finalement, sans rentrer dans les détails, il a fallu procéder à une préparation pour la cytométrie en flux, et faire la cytométrie en flux elle-même. Les résultats eux-même de cette manipulation sont pas été très concluants, et dans le détail celle-ci était un peu différente des manipulations faites ensuite dans le cadre du projet HTE, mais ces deux jours au laboratoire m'ont permis de voir le *grand principe* de ces manipulations.

Les manipulations prévues pour la suite sont tout d'abord pour la preuve de concept et donc la mise en co-culture de plusieurs lignées stromales murines avec des CSH pré-leucémiques. Concernant l'obtention de cellules humaines pré-leucémiques, qui sont donc celles qui ne portent que la première anomalie génétique, elle peut se faire de deux façons : la première consiste à tout d'abord identifier des patients sources, ceux qui ont les cellules d'intérêt, c'est-à-dire des patients plusieurs années avant qu'ils ne développent la leucémie, ce qui n'est bien sûr pas utilisable en pratique, ou alors des patients qui ont eu une LAM et sont en rémission (qui sont remis au stade pré-leucémique avec une seule anomalie) - c'est ce matériel de patient qui sera utilisé - et ensuite de récupérer de la moelle de ces patients. La deuxième consiste à mimer des cellules pré-leucémiques : pour cela, on récolte des cellules de cordon, qui est préparé de façon à isoler les cellules CD34+, puis ces cellules sont transduites[6] de façon à apporter le code génétique voulu, c'est à dire les mutations des gènes TET2 et DNMT3A car les mutations pré-leucémiques correspondent aux premières lésions qui se produisent dans le clone de LAM, ciblant les régulateurs épigénétiques TET2/DNMT3A. En résumé, on introduit dans des cellules de cordon une anomalie équivalente à ce qui se passe chez les patients ; cela permet un modèle reproductible à volonté.

Une fois la preuve de concept faite, l'objectif est d'utiliser du matériel humain : non seulement des CSH humaines normales, pré-leucémiques et leucémiques mais également avec du stroma humain normal, pré-leucémique, leucémique. Pour les cellules stromales normales (resp. leucémiques), l'idée est d'essayer de les dériver de patients sains (resp. atteints de LAM). Pour les cellules stromales pré-leucémiques, c'est moins évident : la première hypothèse est d'utiliser des cellules de patients en post-traitement quand il ne reste plus que le clone pré-leucémique (le défaut de ce procédé est que le stroma a subi de la chimiothérapie) et la deuxième est d'utiliser du stroma normal, de l'exposer au CD34 pré-leucémique et de le re-séparer.

4. L'idée est que les anticorps utilisés sont liés à une particule aimantée. Lorsque l'on fait passer les cellules à l'intérieur de la colonne aimanté, les cellules marquées par les particules magnétiques sont retenues sur la colonne, alors que les cellules qui n'ont pas l'anticorps sont éluées dans un premier tube. Après avoir *détaché* les cellules retenues dans la colonne, que l'on met dans un deuxième tube, on a donc la séparation de cellules marquées ou non.

5. La trypsine est une enzyme qui est notamment utilisée en culture cellulaire pour détacher des cellules adhérant sur les flasques de culture, car elle clive les protéines membranaires d'adhésion, les cellules se retrouvant alors en suspension.

6. La transduction correspond à l'introduction de matériel génétique viral dans une cellule.

4 MÉTHODES POUR LE TRAITEMENT DES DONNÉES BIOLOGIQUES

Ce chapitre est consacré aux différentes méthodes pour le traitement statistique des données biologiques qu'il m'a été donné d'apprendre dans ce stage, de manière théorique essentiellement, grâce au rapport de stage de Hicham Janati [12] et aux différents articles conseillés par Hugues Richard [3], [19], [22], [21], [5], [20].

4.1 TECHNIQUES DE VISUALISATION DE DONNÉES

4.1.1 Analyse en composantes principales

L'analyse en composante ou *ACP* est une méthode de la famille de l'analyse de données multidimensionnelles, qui consiste à extraire d'une masse de résultats bruts tel qu'un tableau (n, p) (n individus décrits par p variables) une information pertinente et moins redondante, en décrivant les n individus avec le moins de variable possible. Elle a pour but de déterminer les principales relations linéaires dans un ensemble complexe et de grande dimension, afin de le réduire et de manière à ce que les structures sous-jacentes soient mieux comprises. Autrement dit, on cherche à définir k nouvelles variables, appelées *composantes principales*, combinaisons linéaires des variables initiales, faisant perdre le moins d'information possible. Les axes que les nouvelles variables déterminent sont appelés *axes principaux* et les formes linéaires associées sont les *facteurs principaux*. L'analyse en composantes principales est une approche géométrique, les variables étant représentées dans un nouvel espace, et statistique, car la recherche porte sur des axes indépendants qui expliquent au mieux la variabilité des données.

On part donc d'un tableau X de taille (n, p), où l'on note x_i^j la valeur de la variable X^j, $j = 1, ..., p$ observée sur l'individu i, $i = 1, ..., n$, et donc

$$X = \begin{pmatrix} x_1^1 & ... & x_1^p \\ ... & \ddots & ... \\ x_n^1 & ... & x_n^p \end{pmatrix}.$$

Afin d'observer l'information contenue dans le tableau X, on peut essayer de visualiser le nuage de points qui représentent le n individus dans l'espace $\mathbb{R}^p$. Or, deux problèmes surviennent : p peut être relativement grand, ainsi la forme du nuage des individus n'est pas évidente à visualiser. D'autre part, les relations entre les variables ne peuvent pas être appréhendées. La question se pose donc de savoir comment les synthétiser. La méthode d'ACP permet de trouver une solution à cette problématique double.

Le principe de l'analyse en composantes principales se passe de la manière suivante : on commence par centrer les variables (rappelons qu'une variable centrée est une variable dont la moyenne est nulle) en retranchant à chaque x_i^j la moyenne de sa colonne : $\frac{1}{n}\sum_{i=1}^{n} x_i^j$. Ainsi, le centre de gravité du nuage est à l'origine. Souvent on a même recourt à une représentation centrée réduite, qui détermine une analyse en composantes principales *normée*.

Ensuite, afin de réduire la dimension de l'espace qui porte le nuage d'individus (de façon à pouvoir visualiser celui-ci), l'ACP utilise la projection orthogonale sur des sous-espaces affines. Pour cela, on commence par construire la matrice de variable-covariance (qui est, si les variables sont réduites, la matrice des corrélations) Γ, matrice symétrique de $\mathbb{R}^p$, donnée par $\frac{1}{n}X'X$. Cette matrice est également appelée *matrice d'inertie* du nuage de points car l'inertie du nuage est définie comme $I = \frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2$, c'est-à-dire comme la trace de la matrice d'inertie. On définit également l'inertie du nuage autour d'un sous-espace linéaire H comme $J_H = \frac{1}{n}\sum_{i=1}^{n}\|X_i - P_H X_i\|^2$ ($P_H X_i$ étant la projection orthogonale de X_i sur H), qui mesure la déformation du nuage lorsqu'il est projeté orthogonalement sur H. Or, le but est de minimiser la déformation du nuage (pour que la représentation des données par projection sur H garde le maximum d'information) et donc de minimiser J_H. Notons que par le théorème de Pythagore, minimiser J_H revient à maximiser l'inertie du nuage projeté, c'est-à-dire à maximiser $I_H = \frac{1}{n}\sum_{i=1}^{n}\|P_H X_i\|^2$. Finalement, le problème se résume à trouver un sous-espace linéaire de dimension k, H_k tel que $H_k = \underset{H_k:\dim(H_k)=k}{\operatorname{argmin}} J_H = \underset{H_k:\dim(H_k)=k}{\operatorname{argmax}} I_H$.

Finalement, commençons par noter que la matrice Γ est symétrique, semi-définie positive; diagonalisable, toutes ses valeurs propres sont réelles et il existe une base orthonormale de vecteurs propres de $\mathbb{R}^p$. On note ses valeurs propres, s.p.d.g., par ordre décroissant $\lambda_1 \geqslant ... \geqslant \lambda_p$ et $u_1, ..., u_p$ les vecteurs propres associés. On a alors (voir [8] pour la démonstration) que la meilleure droite de projection du nuage est celle du vecteur directeur u_1 associé à la plus grande valeur propre de Γ. Pour un sous-espace de dimension 2, on cherche ensuite un vecteur directeur orthogonal à u_1 portant l'inertie maximale, et on trouve que celui-ci est donné par u_2, associé à λ_2 et ainsi de suite. Ainsi, $H_k = Vect(u_1, ..., u_k)$ et on a que l'inertie sur H_k est donnée par la somme des inerties sur les k axes propres principaux : $I_{H_k} = \sum_{l=1}^{k} I_{u_l} = \sum_{l=1}^{k}\lambda_l$. On définit également la part d'inertie sur le l-ème axe propre comme λ_l/I et donc l'inertie portée par un sous-espace de dimension k est au mieux $\sum_{l=1}^{k}\lambda_l/I$ pour cent de l'inertie totale. Finalement, on peut définir les p composantes principales comme $C^\gamma = Xu_\gamma \in \mathbb{R}^n$.

4.1.2 Classification hiéarchique

Pour la visualisation de nombreuses données multidimensionnelles, une technique particulièrement appropriée est l'utilisation de cartes de répartition de la chaleur ou *Heat Maps*. Une *Heat Map* est une représentation graphique de données statistiques où les valeurs individuelles contenues dans une matrice sont représentées sous forme de couleurs. Il est en général utile de combiner les *Heat Maps* avec une classification hiérarchique, qui permet de mettre en évidence ou de marquer plus facilement la *Heat Map*.

La classification hiérarchique (*hierarchical clustering analysis* en anglais) est un méthode de classification itérative, utilisée en statistique comme méthode d'analyse typologique (*cluster analysis)*. En particulier, elle constitue une façon de trier les éléments selon une hiérarchie basée sur la distance ou le degré de similarité entre les éléments, et peut être vue comme la création des partitions emboîtées (arbre hiérarchique). On distingue la classification *ascendante* hiérarchique - au départ, chaque élément est dans un groupe distinct et à chaque étape deux groupes sont rassemblées en un seul selon un critère d'agrégation - de la classification *descendante* hiérarchique - au départ, tous les éléments sont dans le même groupe et à chaque étape, un groupe est séparé en deux selon un critère de séparation.

Pour déterminer le critère d'agrégation (resp. de séparation), une mesure de dissimilarité doit être choisie, l'avantage étant que l'on peut choisir un type de dissimilarité adapté au sujet étudié et à la nature des données. En général, cela est fait par le choix d'une métrique appropriée telle que la distance euclidienne par exemple. Lorsque ce choix est fait, la matrice des distan-

ces/dissimilarités est calculée, et le principe de la classification ascendante (resp. descendante) hiérarchique est alors très simple : elle va rassembler (resp. diviser), selon cette matrice, de manière itérative les individus (ou éléments) afin de produire une hiérarchie à structure arborescente, appelée *dendogramme*. Celui-ci permet donc de visualiser le regroupement progressif des données, et on peut alors se faire une idée notamment du nombre adéquat de "classes" dans lesquelles les données peuvent être regroupées.

4.2 RNA-SEQ

Le transcriptome est l'ensemble des ARNm issus de la transcription des gènes, présents dans une population de cellules dans des conditions données. La quantification et caractérisation du transcriptome permettent d'identifier les gènes actifs, de déterminer les mécanismes de régulation d'expression des gènes ainsi que de définir les réseaux d'expression des gènes. Ainsi, comprendre le transcriptome est essentiel dans la compréhension de l'expression différentielle dans les processus normaux et de maladie.

La transcription peut être mesurée et détectée de diverses manières, l'une d'elles étant le séquençage de l'ARN ou *RNA-Seq*, une approche de grande précision récemment développée pour le profilage transcriptomique, permettant d'observer des changements se produisant dans des états pathologiques, en réponse à la thérapeutique, sous différentes conditions expérimentales, etc. Il utilise des techniques de séquençage haut débit (*NGS - Next-Generation Sequencing* en anglais), pour séquencer des transcriptomes entiers, permettant de mesurer la quantité relative d'ARN, issu de la transcription d'un génome à un moment donné.

Le principe de RNA-Seq est le suivant : lorsqu'une cellule a besoin de l'information contenue dans un gène, elle fait la copie de la partie désirée à partir de l'ARN. Ainsi, la quantité d'une molécule ARN spécifique montre combien un gène est exprimé dans un échantillon de cellules. Le RNA-Seq consiste alors en la fragmentation de l'ARN trouvé dans la population. Les fragments obtenus sont ensuite liés au génome et on peut définir une matrice numérique d'entiers non négatifs, regroupant les données du séquençage d'ARN, c'est-à-dire qu'on obtient une matrice donc l'élément (i, j) représente le nombre de fragments liés au gène i trouvés dans l'échantillon j. Les données sont ensuite analysées en faisant des comparaisons par paires d'échantillons pour chaque gène.

4.2.1 Modèle Deseq2

Le modèle Deseq2 est un modèle pour les données de RNA-Seq, utilisant l'hypothèse que la distribution la plus consistante pour le nombre de fragments pour chaque gène est la distribution binomiale négative. Il admet comme supposition l'indépendance des gènes. Le modèle

Deseq2 model

Negative binomial:

$$Y_{ij} \sim \mathrm{NB}(\mu_{ij}, \sigma_{ij}^2)$$

Assumptions on mean and variance:

$$\mu_{ij} = q_{i\rho(j)} s_j$$
$$\sigma_{ij}^2 = \mu_{ij} + \alpha_{i\rho(j)} \mu_{ij}^2$$

Where α_i models within-group variability of gene i.

FIGURE 1 – Présentation du modèle Deseq2, provenant de [12].

Deseq2, présenté dans la Figure 1, où Y_{ij} représente le nombre de fragments dans l'échantillon j correspondant au gène i, $q_{i\rho(j)}$ l'expression de gène i dans la condition biologique de l'échantillon j, s_j sont des facteurs normalisants et où les α_i modélisent la variabilité à l'intérieur du groupe du gène i. L'idée ensuite est d'utiliser ce modèle afin de tester la différence de l'expression des gènes entre différentes conditions biologiques/expérimentales, par exemple, avant et après l'introduction d'un agent chimiothérapeutique dans une population de cellules tumorales, ou avant et après contact entre deux populations de cellules. Ceci revient à tester

$$H_0 : q_{i\rho} = q_{i\rho'} \quad \text{vs} \quad H_1 : q_{i\rho} \neq q_{i\rho'}.$$

On applique ensuite un modèle log-linéaire généralisé $\log_2(q_{ij}) = \sum_r x_{jr}\beta_{ir}$ où les x_j indiquent dans quelle condition l'échantillon j est pris et les β_i sont les coefficients du modèle, qui sont utilisés pour calculer les estimations des ratios logarithmiques de q_{ik}. A partir de ce modèle on peut, en utilisant par exemple l'estimation par maximum de vraisemblance, calculer les estimations des paramètres inconnus, c'est-à-dire la dispersion α_i et les ratios logarithmiques. Ensuite, notons que tester la différence d'expression de gène entre les conditions revient à tester la nullité des ratios logarithmiques, qui eux-même peuvent être vus comme une combinaison linéaire des coefficients du modèle log-linéaire généralisé. Ainsi, pour faire une comparaison du niveau d'expression du gène i entre les échantillons a et b, nous avons les équivalences suivantes :

$$H_0 : q_{ia} = q_{ib} \Leftrightarrow \frac{q_{ia}}{q_{ib}} = 1 \Leftrightarrow \log_2\left(\frac{q_{ia}}{q_{ib}}\right) = 0 \Leftrightarrow \beta_{ib} - \beta_{ia} = 0 \Leftrightarrow V'_{a,b}\beta_i = 0,$$

où $V_{a,b}$ est un vecteur qui vaut 1 à la position b, -1 à la position a et 0 ailleurs. La statistique de Wald est finalement utilisée pour tester la différence du niveau d'expression de gène (sous la forme de la dernière équivalence).

Les tests de Wald entre les différentes conditions permettent de mettre en avant un grand nombre de gènes exprimés de manière différentielle (*DGE - differentially expressed genes* en anglais). On utilise alors l'algorithme DBSCAN - *Density-based spatial clustering of applications with noise* - qui est un algorithme de clustering, permettant de regrouper les gènes selon leur cinétique afin d'observer chaque type d'évolution de gènes. Celui-ci (expliqué plus en détails dans [12]) a pour principe d'utiliser deux paramètres : une distance ϵ et un nombre minimum m de points devant se trouver dans un rayon ϵ pour que ces points soient considérés comme un cluster. L'idée de base est alors de considérer un point donné et de vérifier si son voisinage (à distance ϵ) contient bien m points ou plus. Si c'est le cas, il est considéré comme faisant partie du cluster, et sinon, il est considéré comme *noisy*. Cet algorithme est particulièrement approprié aux données de RNA-Seq, car certains gènes ont des moyennes et variances grandes sans raisons biologiques apparentes, et ceux-ci peuvent être considérés comme *noisy*.

Finalement, après avoir formé des clusters, le rôle de chacun doit être interprété. Pour cela, on utilise ce qu'on appelle *l'analyse d'enrichissement*. Le *Gene Set Enrichment Analysis*, développé ces 20 dernières années par des chercheurs, est méthode computationnelle utilisant un test de surreprésentation des catégories de processus biologiques, c'est-à-dire qu'il s'agit de comparer par mesure statistique un ensemble de gènes groupés avec une liste de gènes connus à priori pour être impliqués dans les mêmes processus biologiques.

4.2.2 Modèle WGCNA

Le modèle Deseq2 utilisait l'indépendance des gènes, ce qui est faux biologiquement parlant car les gènes sont impliqués dans des mécanismes complexes de régulation et co-expression. Le modèle *WGCNA - Weighted Genes Co-expression Network Analysis* a au contraire pour but de développer un réseau pour analyses les gènes en modules et étudier la co-expression de gènes en

analysant leur corrélations.

Le principe général du modèle WGCNA est relativement simple (pour les détails, voir [12]) : il se base sur une mesure de similarité de co-expression de gène, la mesure *TOM - Topological Overlap Measure*, qui, sans rentrer dans les détails de sa définition, est liée à la corrélation entre deux gènes. Les modules de gènes sont ensuite formés par classification hiérarchique (voir la sous-section 4.1.2) à partir de la matrice de similarité basée sur la mesure *TOM*. Une analyse d'enrichissement est également faite dans ce modèle, afin de découvrir les processus biologiques qui distinguent les modules les uns des autres. D'autre part, pour chaque module, un gène fictif représentatif est calculé en prenant la première composante principale (voir sous-section 4.1.1), qui est alors appelé *gène propre du module* (*module eigengene* en anglais). La corrélations de ces gènes propres sont finalement "tracés" sur une Heat Map (voir 4.1.2) afin d'analyser les modules co-régulés.

4.3 SINGLE CELL RNA-SEQ

Le séquençage de l'ARN a déjà apporté de grandes avancées à l'analyse de transcriptome et est devenue rapidement une méthode de choix pour aborder des aspects qualitatifs et quantitatifs à l'expression de gènes. La plupart des études qui sont appelées *bulk RNA-Seq* en anglais (que l'on peut traduire comme RNA-Seq de masse, faute de meilleur terme) ou simplement *RNA-Seq* (voir 4.2) ont été conduites au niveau de population de cellules, considérant donc des transcriptomes moyennés sur une population entière de cellules. Cette approche a cependant certaines limites : des mesures sur un grand nombre de cellules, comme c'est le cas pour ce type de méthode, peuvent masquer des signaux d'intérêt ; les approches *de masse* peuvent ne pas parvenir à détecter des différences subtiles mais potentiellement importantes du point de vue biologique entre des cellules qui semblent identiques. Ainsi, les dernières années ont montrées que des aspects très importants pouvaient être acquis par des approches que l'on appelle *Single-Cell* (que l'on peut traduire littéralement comme unicellulaire), où le principe général est d'isoler les cellules dans des puits et de procéder au profilage d'expression de gènes dans chaque cellule. Étudier les cellules au niveau unicellulaire offre une occasion unique d'analyser les interactions entre les processus cellulaires intrinsèques et les stimuli extérieurs (l'environnement local, les cellules voisines).

En particulier, le *Single-Cell RNA-Seq* (ou séquençage d'ARN unicellulaire) fournit le profil d'expression des cellules individuelles. Il peut permettre, à travers l'analyse de clusters de gènes, d'identifier des types rares de cellules, et ainsi de rendre possible la caractérisation de la structure des sous-populations d'une population cellulaire hétérogène. C'est d'ailleurs l'un des objectifs principaux du Single-Cell RNA-Seq que l'on peut lister (liste non-exhaustive !) de la façon suivante :

- Identification de sous-populations de cellules dans une condition biologique donnée,
- Identification et caractérisation des gènes exprimés différemment entre différentes conditions biologiques,
- Ordre pseudo-temporel (*Pseudo-temporal ordering* en anglais),
- Identification de modules co-régulés de gènes, explication de structure et fonction de réseaux de régulation de gènes.

Concernant le premier point, il consiste à utiliser dans le cadre de Single-Cell RNA-Seq des méthodes qui étaient déjà utilisées dans le RNA-Seq "de masse" telles que l'analyse en composantes principales (voir 4.1.1) ou la classification hiérarchique (voir 4.1.2) sur les gènes exprimés différemment, hautement variables ou hautement exprimés, afin d'identifier des sous-populations de cellules. Par ailleurs, d'autres méthodes ont également été développées pour atteindre cet ob-

jectif, méthodes qui permettent d'éviter certains défauts apparaissant par exemple dans l'ACP (voir [3] pour les détails sur ces méthodes). Lorsque les sous-populations sont détectées, un aspect essentiel est de déterminer si elles correspondent à un type de cellule connu, et pour cela, une analyse d'enrichissement de type cellulaire (de la même manière que le *Gene Set Enrichment Analysis*) a été développée.

Le séquençage en ARN unicellulaire a obtenu des résultats déjà intéressants dans la recherche, au niveau de l'dentification de sous-populations de cellules dans une condition biologique donnée. En effet, comme on peut le voir dans l'article [22], traitant de la reconstruction de hiérarchies de lignées de l'épithélium pulmonaire distal, l'analyse en composantes principales et la classification hiérarchique sur des données de Single-Cell RNA-Seq ont permis l'identification et la caractérisation moléculaire de cinq différentes populations de cellules et d'intermédiaires du développement. D'autre part, en plus de la classification des différentes populations de cellules dans le poumon distal, cela a permis l'identification d'un ensemble de gènes spécifique à chaque population fournissant ainsi un grand nombre de marqueurs qui n'étaient pas encore connus, permettant de distinguer les cellules d'une lignée alvéolaire de celles d'une lignée bronchiolaire.

Concernant le deuxième objectif mentionné plus tôt, on utilise pour cela le même type de méthode que dans le cadre de *bulk RNA-Seq*, c'est-à-dire qu'on cherche à tester les changements dans les distributions unimodales à travers les différentes conditions biologiques. A noter cependant que ces techniques ne sont pas tout à fait optimales, à cause d'un problème de variabilité technique et biologique, ainsi, une méthode (que l'on ne détaille pas ici, voir pour cela [3]) plus spécifique a été développée.

Abordons maintenant le sujet de l'ordre pseudo-temporal. Des processus dynamiques, tels que l'autorenouvellement et la différentiation des cellules souches, sont essentiels, et pourtant leur compréhension est limitée, car il existe une grande variabilité d'expression génique entre les cellules, qui pose problème pour l'analyse de ces processus. Le séquençage d'ARN de masse ayant le désavantage de faire une moyenne sur des millions de cellules, perdant ainsi des signaux d'intérêt, il est donc crucial de procéder au profilage de l'expression du génome dans les cellules individuelles. Le Single-Cell RNA-Seq permet évidemment ce profilage, mais le monitoring continu d'expression du génome dans les cellules individuelles à travers le temps n'est pas possible. Cependant, grâce à différents algorithmes, on a la possibilité de reconstruire des chemines de différentiation d'une population de cellules désynchronisées, en partant de l'idée suivante : à un temps donné, la population de cellules possède des cellules à différentes étapes de différentiation, et donc leurs dynamiques d'expression peuvent être résolues en re-ordonnant les cellules selon leur position dans le chemin de différentiation.

Différentes méthodes pour cet *agencement* peuvent être trouvées dans [3], ainsi que dans [21] qui décrit le fonctionnement de l'algorithme `MONOCLE`, un algorithme augmentant la résolution temporelle des dynamiques transcriptomiques en utilisant des données de Single-Cell RNA-Seq collectées à différents points de temps. Il est en principe utilisé pour récupérer la cinétique d'expression de gène d'une cellule individuelle à partir d'un large éventail de processus cellulaires tels que la différentiation ou la prolifération. Le principe général de `MONOCLE` est le suivant : il commence par représenter le profil d'expression de chaque cellule comme un point d'un espace euclidien de grande taille (une dimension pour chaque gène), puis il réduit la dimension de cet espace en utilisant l'analyse en composantes indépendantes (qu'on ne décrira pas ici). Ensuite, un *MST - Minimum Spanning Tree* est construit sur les cellules (notons que le MST, concept venant de la théorie des graphes, est de manière générale la façon de construire un réseau tel qu'il minimise le poids des arêtes). L'algorithme trouve le chemin le plus long à travers le MST, correspondant à la plus longue séquence de cellules similaires au niveau du transcriptome. `MONOCLE`

utilise alors cette séquence afin de créer une trajectoire de la progression des cellules à travers la différentiation puis il examine les cellules n'étant pas le long de cette trajectoire afin de trouver des trajectoires alternatives dans le MST. Ces sous-trajectoires sont finalement ordonnées et connectées à la trajectoire principale et chaque cellule est associée à une trajectoire et une valeur de pseudo-temps. Notons que l'article [21] montre notamment que `MONOCLE` a permis l'analyse de la trajectoire de différenciation de myoblastes, et a fourni des résultats tout à fait consistants avec ce qui était déjà connu.

Finalement, pour revenir au quatrième objectif mentionné plus haut, soulignons tout d'abord le fait qu'expliquer la structure ainsi que la fonction de réseaux de régulation est un but central de beaucoup d'études et le Single-Cell RNA-Seq possède un grand potentiel dans ce sens-là. On peut en principe faire des analyses très similaires à celles faites dans le cadre de séquençage d'ARN de masse pour l'identification de gènes co-régulés, en remplaçant les échantillons par les cellules individuelles. L'intérêt premier est d'identifier les groupes de nœuds (*nodes*), représentant les gènes, et d'estimer les arêtes (*edges*), représentant l'interaction ou la dépendance entre les gènes, et de déterminer la manière dont le réseau change après une perturbation. Souvent, le modèle WGCNA (sous-section 4.2.2), est également utilisé dans le cadre des Single-Cell (les arêtes représentant alors la co-expression des gènes). Cependant, cela n'apporte pas d'information sur les relations de régulations entre les nœuds, c'est-à-dire entre les gènes : on a besoin typiquement pour cela d'expériences temporelles ou avec une perturbation. Dans cette idée, des méthodes récentes ont utilisé l'information produite par des approches d'ordre pseudo-temporel combinée aux méthodes traditionnelles de reconstruction de réseau, de façon à pouvoir inférer sur les relations de régulation parmi les gènes.

Finalement, notons que les études Single-Cell ont déjà apporté des résultats très intéressants dans différentes branches de la recherche, telles que la différenciation de cellules souches, l'embryogénèse, l'analyse de tissu entier et même sur les études d'organismes entiers.

5 MODÈLE MATHÉMATIQUE

Si les méthodes statistiques permettent d'apporter des réponses à des questions variées directement à partir des données, elles sont inappropriées pour la prédiction du comportement de populations d'intérêt, où une représentation appropriée de la dynamique - évolution normale ou pathologique au cours du temps - de ces populations est nécessaire. Cette représentation *appropriée* constitue un ensemble d'équations dont les solutions sont à même de reproduire des propriétés qualitatives remarquables des populations observées. D'autre part, l'analyse mathématique de ces modèles permet de déterminer certains conditions sur les paramètres dans lesquelles une population doit se trouver, dans le but d'observer un phénomène précis et d'obtenir des prédictions sur le comportement de la population. Finalement, les simulations numériques, quant à elles, permettent d'illustrer les résultats analytiques et d'obtenir une intuition du comportement quand l'analyse est impossible.

Les modèles que nous considérons sont déterministes et ont pour but de prédire l'évolution de populations. Les variables sont des densités de ces populations dont on cherche à prédire le comportement à tout instant à partir d'une donnée initiale, et la limite (comportement asymptotique) de ce comportement en temps grand. Nous verrons ici différents modèles [16], [17] : des équations différentielles ordinaires, permettant d'expliciter l'évolution temporelle de populations de cellules et des équations aux dérivées partielles, plus complexes mais plus complètes car tenant compte de structures inhérentes aux populations. D'autre part, on considérera l'évolution d'une population seule ou de deux populations en interaction, où l'on obtiendra alors un système d'équations couplées.

5.1 MODÈLES EDO

5.1.1 Modèles à une seule population

On considère que l'effectif d'une population au cours du temps est représenté par une fonction réelle $N : \mathbb{R}_+ \to \mathbb{R}_+$. On peut considérer comme loi générale des modèles à une seule espèce que la variation de l'effectif de la population au cours du temps est égal au taux instantané des naissances, auxquels on soustrait les décès et on ajoute les migrations. Le modèle le plus simple est le modèle de Malthus, qui pose comme hypothèse que les taux de naissances et de décès sont proportionnels à la population et qu'il n'y pas de migration. On obtient alors

$$\frac{dN(t)}{dt} = (b - d)N(t), \quad b, d > 0,$$

qui a comme solution $N(t) = N_0 e^{(b-d)t}$, où $N_0 = N(0)$ est la population initiale. Ainsi, si $b > d$, la population croît exponentiellement, et si $b < d$, elle décroît exponentiellement jusqu'à l'extinction.

Le modèle de Verhulst corrige le défaut du manque de réalisme de la croissance exponentielle du modèle de Malthus en considérant une fonction de la capacité du milieu en nourriture :

$$\frac{dN(t)}{dt} = (b - d)N(t)\left(1 - \frac{N(t)}{K}\right), \quad b, d > 0, K > 0,$$

avec K la charge utile de l'environnement, et on constate que lorsque celle-ci tend vers l'infini, nous retrouvons le modèle de Malthus. La solution du modèle de Verhulst est donnée par $N(t) = N_0 \frac{Ke^{(b-d)t}}{K+N_0(e^{(b-d)t}-1)}$. A noter que ce modèle peut être réécrit sous la forme

$$\frac{dN(t)}{dt} = [r - dN(t)]N(t),$$

où r est le taux de prolifération et $dN(t)$ le taux de décès, qui est proportionnel à $N(t)$ dû à la compétition intra-spécifique.

5.1.2 Modèles pour 2 populations en interaction

Il existe trois types de modèles pour deux populations en interaction : les modèles de types *proie-prédateur* - lorsque l'effectif d'une population diminue et que l'autre augmente ; les modèles de populations en *compétition* - les deux populations diminuent ; et les modèles de *mutualisme / symbiose* - les deux populations augmentent. Pour ces populations en interaction, on obtient des systèmes couplés, où les échanges bidirectionnels entre populations sont décrits en rajoutant des termes croisés dans les équations.

Concernant les modèles de types proie-prédateur, le modèle le plus connu et le plus simple est le modèle *Lotka-Volterra* qui a les hypothèses suivantes :

— Sans prédateurs, la population des proies suit un modèle de croissance malthusienne $\frac{dN(t)}{dt} = aN(t)$;
— Sans proies, la population des prédateurs suit un modèle de décroissance malthusienne $\frac{dP(t)}{dt} = -bP(t)$;
— La prédation implique la réduction du taux de croissance des proies, de la forme $-bN(t)P(t)$;
— Les proies contribue au taux de croissance des prédateurs avec un terme de la forme $cP(t)N(t)$,

hypothèses que l'on synthétise sous la forme du modèle Lotka-Volterra

$$\begin{cases} \frac{dN(t)}{dt} = aN(t) - bN(t)P(t) & = N(t)(a - bP(t)) \\ \frac{dP(t)}{dt} = cP(t)N(t) - dP(t) & = P(t)(cN(t) - d). \end{cases} \tag{1}$$

En procédant au changement de variables : $\tau = at$, $\alpha = \frac{d}{a}$, $u(\tau) = \frac{cN(t)}{d}$, $v(\tau) = \frac{bP(t)}{a}$, le système d'équation (1) peut se réécrire sous la forme

$$\begin{cases} \frac{du(\tau)}{d\tau} = u(\tau)(1 - v(\tau)) \\ \frac{dv(\tau)}{d\tau} = \alpha v(\tau)(u(\tau) - 1), \end{cases} \quad \Rightarrow \quad \frac{dv(\tau)}{du(\tau)} = \frac{\alpha v(\tau)(u(\tau) - 1)}{u(\tau)(1 - v(\tau))}, \tag{2}$$

qui a comme valeurs singulières $u = v = 1$ (point d'équilibre stable) et $u = v = 0$ (point d'équilibre instable). On a des trajectoires fermées et ainsi la solution est périodique en τ pour u et v (voir [16] pour une analyse plus détaillée). A noter que dans ce modèle les solutions ne sont pas structurellement stables, chaque petite perturbation va déplacer la solution sur une autre trajectoire qui n'est même pas forcément proche de l'origine, il n'est donc que peu utile pour l'interaction de populations réelles, mais il offre une base théorique importante. D'autre part, bien que peu réaliste (notamment par l'hypothèse de croissance malthusienne), il suggère que des interactions simples de proies-prédateurs peut résulter en un comportement périodique de la population.

Les interactions entre cellules souches hématopoïétiques et cellules stromales sont des interactions mutualistes. On s'intéresse donc maintenant aux modèles de populations en symbiose, la forme la plus simple étant donnée par

$$\begin{cases} \dfrac{dN_1(t)}{dt} = r_1 N_1(t) + a_1 N_1(t) N_2(t) \\ \dfrac{dN_2(t)}{dt} = r_2 N_2(t) + a_2 N_2(t) N_1(t), \end{cases} \tag{3}$$

avec $a_1, a_2, r_1, r_2 > 0$, où l'on voit donc que $\frac{dN_1(t)}{dt}, \frac{dN_2(t)}{dt} > 0$ et donc les deux populations N_1, N_2 ont une croissance non bornée. Un modèle plus raisonnable est obtenu en incluant la capacité de charge limite de chaque population :

$$\begin{cases} \dfrac{dN_1(t)}{dt} = r_1 N_1(t) \left(1 - \dfrac{N_1(t)}{K_1} + b_{12} \dfrac{N_2(t)}{K_1}\right) \\ \dfrac{dN_2(t)}{dt} = r_2 N_2(t) \left(1 - \dfrac{N_2(t)}{K_2} + b_{21} \dfrac{N_1(t)}{K_2}\right). \end{cases} \tag{4}$$

Par un changement de variable $\tau = r_1 t, \rho = \frac{r_2}{r_1}, u_i = \frac{N_i}{K_i}, a_{ij} = b_{ij} \frac{K_j}{K_i}$, $(i, j = 1, 2)$, (4) s'écrit sous la forme

$$\begin{cases} \dfrac{du_1(\tau)}{d\tau} = u_1(\tau)(1 - u_1(\tau) + a_{12} u_2(\tau)) \\ \dfrac{du_2(\tau)}{d\tau} = \rho u_2(\tau)(1 - u_2(\tau) + a_{21} u_1(\tau)), \end{cases} \tag{5}$$

qui a comme $(0,0), (1,0), (0,1)$ comme points d'équilibres instables. D'autre part, si $a_{12} a_{21} < 1$ (i.e. $b_{12} b_{21} < 1$) il existe un quatrième point d'équilibre stable $\left(\frac{1+a_{12}}{1-a_{12}a_{21}}, \frac{1+a_{21}}{1-a_{12}a_{21}}\right)$. Ainsi, toutes les trajectoires dans le quadrant positif tendent vers un point d'équilibre (u_1^*, u_2^*) avec $u_i^* > 1, i = 1, 2$, signifiant donc que les populations ont augmenté leur valeur d'équilibre par rapport à celle en isolation. A noter que si la symbiose est trop forte, c'est-à-dire que $b_{12} b_{21} > 1$, alors les populations suivent une croissance non bornée.

5.2 MODÈLES STRUCTURÉS EN PHÉNOTYPE

Dans les 30 dernières années, les modèles de dynamiques de populations (de cellules en particulier) *structurées de manière physiologique*, c'est-à-dire structurées par un paramètre décrivant un caractère biologique, physiologique ou écologiques des individus, ont beaucoup été étudiés. On appelle *trait* ou *phénotype* un tel caractère lorsque celui-ci est inhérent aux individus. Les phénotypes peuvent être tout à fait différents selon l'angle sous lequel on souhaite prendre notre étude : on peut par exemple avoir une population structurée par *l'âge* dans le cadre d'une étude de l'influence de facteurs de croissance sur les dynamiques du cycle cellulaire [4] (l'intérêt ici étant de faire la distinction entre le temps physiologique ou *âge* et le temps externe). Notons également que la description en phénotype permet notamment des représentations de l'hétérogénéité tumorale, cela ne limite pas à une prise en compte de divers types cellulaires supposés homogènes et dont la diversité décrirait de manière exhaustive l'hétérogénéité, ce qui serait trop simpliste : ainsi l'hétérogénéité peut être mesurée par un ou des traits continus. On appelle *dynamique adaptative* la théorie qui se concentre sur l'évolution phénotypique, avec comme ingrédients principaux le principe de sélection (favorisant la population avec le trait le plus adapté) et les mutations permettant que les descendants possèdent des traits un peu différents.

5.2.1 Modèle à une seule population

Dans de tels modèles, les densités de population (en particulier densité ou nombre de cellules), notés n, ne dépendent plus seulement du temps t mais également de la caractérisation de chacune par une variable x *de structure*, éventuellement multidimensionnelle mais de préférence avec un petit nombre de dimensions, permettant ainsi de décrire, de manière réductrice, la variabilité biologique d'intérêt. Un modèle simple, qui ne considère pas de mutation, consiste en une variante du modèle de Verhulst qu'on structure avec un trait $x \in X \subseteq \mathbb{R}^d$, avec les hypothèses que le taux de prolifération dépend du trait et que le taux de mort est proportionnel à la population totale donnée par $\rho(t) = \int_{\mathbb{R}} n(t,x)dx$. On obtient alors

$$\frac{\partial}{\partial t}n(t,x) = [r(x) - d(x)\rho(t)]n(t,x) := R(x,\rho(t))n(t,x), \tag{6}$$

avec $n(t=0) = n^0(x) \geqslant 0$. On voit que n a comme formule implicite $n(t,x) = n^0(x)e^{\int_0^t R(x,\rho(s))ds}$ et donc que $n(t,x) \geqslant 0 \quad \forall t,x$. On prouve que cette solution existe et est unique, de régularité $C(\mathbb{R}_+, L^1(X))$, en suivant [18], en posant tout d'abord les hypothèses suivantes

— $R \in C^1(\mathbb{R}^d \times \mathbb{R})$
— $n^0 \in L^1(X)$
— Il existe une capacité maximale et minimale, i.e. $\exists \rho^m, \rho^M, d \geqslant 0$ tels que

$$\begin{cases} \forall x \in X, R(x,\rho^m) \geqslant 0 \text{ et } \exists x^m \in X : R(x^m,\rho^m) = 0, \\ \forall x \in X, R(x,\rho^M) \leqslant 0 \text{ et } \exists x^M \in X : R(x^M,\rho^M) = 0, \\ \dfrac{\partial R}{\partial \rho} \leqslant -d. \end{cases}$$

On a alors finalement les résultats suivants (sans preuves ici, voir [18]) :

Théorème 1. En supposant que n^0 est tel que $\rho^m \leqslant \rho^0 := \int_X n^0(x)dx \leqslant \rho^M$, alors il existe une solution globale non-négative unique n de (6), qui satisfait de plus $\rho^m \leqslant \rho(t) \leqslant \rho^M$, $\forall t > 0$.

Proposition 1. La fonction $\rho(t)$ converge et sa limite est donnée par ρ^M.

Théorème 2. Supposons qu'il existe un unique x_0 satisfaisant $R(x_0,\rho^M) = 0$. Alors $n(t,x)$ converge, au sens faible, vers $\rho^M\delta_{x_0}$, ce qui signifie que la population se concentre en x_0 avec la masse ρ^M.

Un modèle un peu plus complet est considéré en ajoutant à (6) un terme de diffusion, ayant pour but de modéliser les épimutations aléatoires :

$$\frac{\partial}{\partial t}n(t,x) = [r(x) - d(x)\rho(t)]n(t,x) + \mu\frac{\partial^2 n(t,x)}{\partial x^2}, \tag{7}$$

où μ quantifie l'instabilité non génétique de la population cellulaire par rapport au phénotype x permettant de rendre la population "plus hétérogène".

5.2.2 Le modèle HTE

Il s'agit maintenant de présenter le modèle mathématique pour les interactions entre les cellules souches hématopoïétiques et les cellules stromales, c'est-à-dire le modèle utilisé dans le cadre du projet HTE. Ce modèle d'interactions mutualistes a pour variables la densité de cellules souches hématopoïétiques $n_h(t,x)$ de phénotype x et la densité de cellules stromales $n_s(t,y)$ de phénotype y. Le phénotype x est dans le cas présent choisi pour être un phénotype de *plasticité cellulaire*. La plasticité cellulaire fait référence, dans un cadre général, à la propriété des cellules à changer d'identité, qui est un phénomène rencontré notamment dans certaines pathologies - la

plasticité dans les cellules cancéreuses représente par exemple l'inversion partielle d'un état de type souche dans les cellules et l'adaptabilité résultant des populations de cellules cancéreuses. Les cellules souches pouvant se différencier en plusieurs types de cellules différents, elles sont considérées comme très *plastiques*. On peut finalement voir la plasticité comme étant par exemple la capacité des cellules à passer d'un état pré-leucémique à leucémique, ce qui justifie le choix de ce phénotype pour les CSH dans ce modèle. Les cellules stromales, quant à elles, fournissent un support structural et fonctionnel aux CSH, et c'est cette capacité de soutien qui est représenté par le phénotype y. Pour des raisons de simplicité, les variables x, y peuvent être considérées comme appartenant à $X = [0,1]$ et on note également que $t \in \mathbb{R}$. On a finalement pour l'évolution de $n_h, n_s : \mathbb{R} \times X \to \mathbb{R}$ le modèle d'interactions mutualistes suivant

$$\begin{cases} \dfrac{\partial n_h(t,x)}{\partial t} = [R_h(x,\tilde{\rho}(t)) + R_{hs}(x,\Sigma(t))]n_h(t,x) + D\dfrac{\partial^2 n_h(t,x)}{\partial^2 x} \\ \dfrac{\partial n_s(t,y)}{\partial t} = [R_s(y,\tilde{\rho}(t)) + R_{sh}(y,P(t))]n_s(t,y) + E\dfrac{\partial^2 n_s(t,y)}{\partial^2 y}, \\ \rho(t) = \displaystyle\int_0^1 n_h(t,x)dx, \quad \sigma(t) = \int_0^1 n_s(t,y)dy, \quad \tilde{\rho}(t) = \rho(t) + \sigma(t), \end{cases} \tag{8}$$

où les conditions initiales dépendent de l'expérience, et où les paramètres D et E, comme μ dans (7) quantifient l'instabilité non génétique des populations cellulaires par rapport aux phénotypes x et y. L'interaction des population est modélisée par la dépendance des termes $R_{hs}(x,\Sigma(t)), R_{sh}(y,P(t))$ aux fonctions $\Sigma(t)$ et $P(t)$ qui ont pour but de représenter le signal chimique des cellules stromales au CSH et inversement, ce qui peut être défini comme

$$\Sigma(t) = \int_0^1 \psi(y)n_s(t,y)dy, \quad P(t) = \int_0^1 \varphi(x)n_h(t,x)dx,$$

où ψ, φ sont des fonctions poids affines de phénotypes x, resp. y. De plus, les interactions étant mutualistes, on suppose que $R_{hs}(x,\Sigma(t)), R_{sh}(y,P(t)) \geqslant 0$, avec égalité si et seulement si $\Sigma(t) = 0$, resp. $P(t) = 0$. On définit alors $R_{hs}(x,\Sigma(t)), R_{sh}(y,P(t))$ comme suit

$$R_{hs}(x,\Sigma(t)) = \gamma(x)\Sigma(t), \quad R_{sh}(y,P(t)) = \delta(y)P(t),$$

où $\gamma(x), \delta(y)$ quantifient la force des interactions : $\gamma(x)$ représente la sensibilité des CSH aux messages des cellules stromales et $\delta(y)$ celles des cellules stromales aux messages de CSH. On utilise ensuite, pour les termes $R_h(x,\tilde{\rho}(t))$ et $R_s(y,\tilde{\rho}(t))$ les mêmes hypothèses que dans le modèle (6), c'est-à-dire que les taux de prolifération des CSH, resp. cellules stromales, dépendent des phénotypes et que les taux de morts sont proportionnels à la population totale (des deux populations) donnée par $\rho(t) + \sigma(t)$:

$$R_h(x,\tilde{\rho}(t)) = \alpha(x) - \mu(x)[\rho(t) + \sigma(t)], \quad R_s(y,\tilde{\rho}(t)) = \beta(y) - \nu(y)[\rho(t) + \sigma(t)],$$

où $\mu(x), \nu(y)$ sont donc des fonctions dépendantes du phénotype affectant les termes de logistique liés aux limitations de croissance dans la population des CSH et le stroma respectivement. On obtient donc finalement le modèle sous sa forme détaillée

$$\begin{cases} \dfrac{\partial n_h(t,x)}{\partial t} = \{\alpha(x) - \mu(x)[\rho(t) + \sigma(t)] + \gamma(x)\Sigma(t)\}n_h(t,x) + D\dfrac{\partial^2 n_h(t,x)}{\partial^2 x} \\ \dfrac{\partial n_s(t,y)}{\partial t} = \{\beta(y) - \nu(y)[\rho(t) + \sigma(t)] + \delta(y)P(t)\}n_s(t,y) + E\dfrac{\partial^2 n_s(t,y)}{\partial^2 y}, \\ \rho(t) = \displaystyle\int_0^1 n_h(t,x)dx, \\ \sigma(t) = \displaystyle\int_0^1 n_s(t,y)dy. \end{cases} \tag{9}$$

Pour l'analyse de ce type de modèle, que l'on peut suivre dans [18] et qui n'est pas détaillée il s'agit de considérer un modèle légèrement plus simple :

$$\begin{cases} \dfrac{\partial n_1(t,x)}{\partial t} = [R_1(x,\rho_1(t)) + R_{12}(x,\varphi_2(t))]n_1(t,x) \\ \dfrac{\partial n_2(t,y)}{\partial t} = [R_2(y,\rho_2(t)) + R_{21}(y,\varphi_2(t))]n_2(t,y) \\ \rho_1(t) = \displaystyle\int_0^1 n_1(t,x)dx, \quad \rho_2(t) = \int_0^1 n_2(t,y)dy, \\ \varphi_1(t) = \displaystyle\int_0^1 \psi_1(x)n_1(t,x)dx, \quad \varphi_2(t) = \int_0^1 \psi_2(y)n_2(t,y)dy, \end{cases} \tag{10}$$

avec $R_1(x,\rho_1(t)) = r_1(x) - d_1\rho_1(t)$, $R_2(y,\rho_2(t)) = r_2(y) - d_2\rho_1(t)$ et $R_{12}(x,\varphi_2(t)) = s_1(x)\varphi_2$, $R_{21}(y,\varphi_2(t)) = s_2(y)\varphi 1$ et en supposant que tous les coefficients sont constants en phénotype, l'analyse de (10) revient alors à l'analyse de la dynamique de (ρ_1,ρ_2) qui suit le modèle simple suivant

$$\begin{cases} \dfrac{d\rho_1}{dt} = [r_1 - d_1\rho_1 + C_{12}\rho_2]\rho_1, \\ \dfrac{d\rho_2}{dt} = [r_2 - d_2\rho_2 + C_{21}\rho_1]\rho_2. \end{cases} \tag{11}$$

5.3 SIMULATIONS

Il a été choisi ici de montrer les résultats des simulations numériques produites pour le modèle d'une population seule, structurée en phénotype, avec un terme de diffusion, c'est-à-dire le modèle (7) et pour le modèle HTE (9). Les simulations numériques sont faites avec `Matlab` [9], en utilisant un schéma aux différences finies implicite-explicite.

Il convient de souligner quelque chose d'important : les simulations présentées ci-dessous ont pour but de donner une prédiction *qualitative* et non *quantitative* du comportement des populations. En effet, elles permettent de voir différents types de comportements qui est possible d'observer en modifiant des paramètres, tels que la croissance exponentielle des populations ou au contraire la convergence de celles-ci, mais elles ne permettent pas une analyse quantitative, dans le sens où les paramètres utilisés ne représentent pas des valeurs réelles au niveau biologique. En effet, le projet HTE n'étant qu'à son commencement, il était trop tôt pour obtenir des valeurs pour les taux de prolifération et d'apoptose par exemple. D'autre part, comme il sera expliqué dans la conclusion, pour certains paramètres, tels que les valeurs d'interactions entre les populations, le lien entre les données biologiques et la valeurs de ceux-ci n'est de loin pas évident. Ainsi, pour les simulations présentées ci-dessous, il suit un principe de parcimonie, c'est-à-dire qu'on a décidé de choisir les paramètres de façon à ne pas complexifier le modèle plus que de raison, n'ayant aucune indication dans ce sens-là. Ainsi, les paramètres structurés en phénotypes sont essentiellement sous forme linéaire, et les paramètres restant ont des valeurs constantes. Les codes `Matlab` peuvent être trouvés en annexe.

Pour ce qui est de la population seule, on peut voir tout d'abord sur la Figure 2, un comportement possible de la population en temps grand, et sans diffusion : la population se stabilise autour d'une valeur, c'est-à-dire qu'on a convergence de la population totale, et celle-ci se concentre autour du phénotype de valeur 1. Sur la Figure 3, on peut voir l'effet de l'ajout d'une diffusion sur le comportement de la population en fonction du phénotype : la diffusion comme effet que la population ne se concentre plus autour d'une valeur unique de phénotype, mais en l'occurrence autour de deux valeurs : la diffusion ajoute une diversité phénotypique, permettant ainsi de modéliser l'hétérogénéité phénotypique de la population. Finalement, concernant

le comportement d'une seule population structurée en phénotype, on a choisi de montrer trois autres comportements : sur la Figure 4, on a choisi un taux de croissance très grand, mais on observe que la population se stabilise quand même très rapidement ; sur la Figure 5, on voit l'effet d'un taux de mort très grand, la population est proche de l'extinction ; et sur la Figure 6, on voit l'effet d'une diffusion trop grande : lorsque celle-ci dépasse un certain seuil, on perd la stabilité du modèle.

Concernant le modèle HTE, on commence par voir sur les Figure 7, 8 l'évolution des populations lorsque les interactions mutualistes ne sont pas très fortes et que l'on considère des termes de diffusion nuls (à noter que les valeurs des paramètres de prolifération et d'apoptose sont pris au hasard et ne reflètent pas une situation biologique réelle !). Ces images suggèrent que le comportement asymptotique des populations dans ce cas là est la convergence des deux populations vers une valeur stable, et la concentration de chacune d'elles autour des phénotypes de valeur 1. La Figure 9 montre l'effet de l'ajout de diffusion, où l'on voit bien que les populations ne concentrent plus autour d'une unique valeur de phénotype. Ensuite, on peut voir sur la Figure 10 deux cas où le mutualisme est trop fort, provoquant une croissance non bornée de deux populations. Enfin, les Figures 11,12 montrent le comportement des deux populations en interaction en temps grand, avec des paramètres de diffusion et d'interaction raisonnables : on a ainsi la convergence des populations, et la concentration autour des phénotypes de valeur 1.

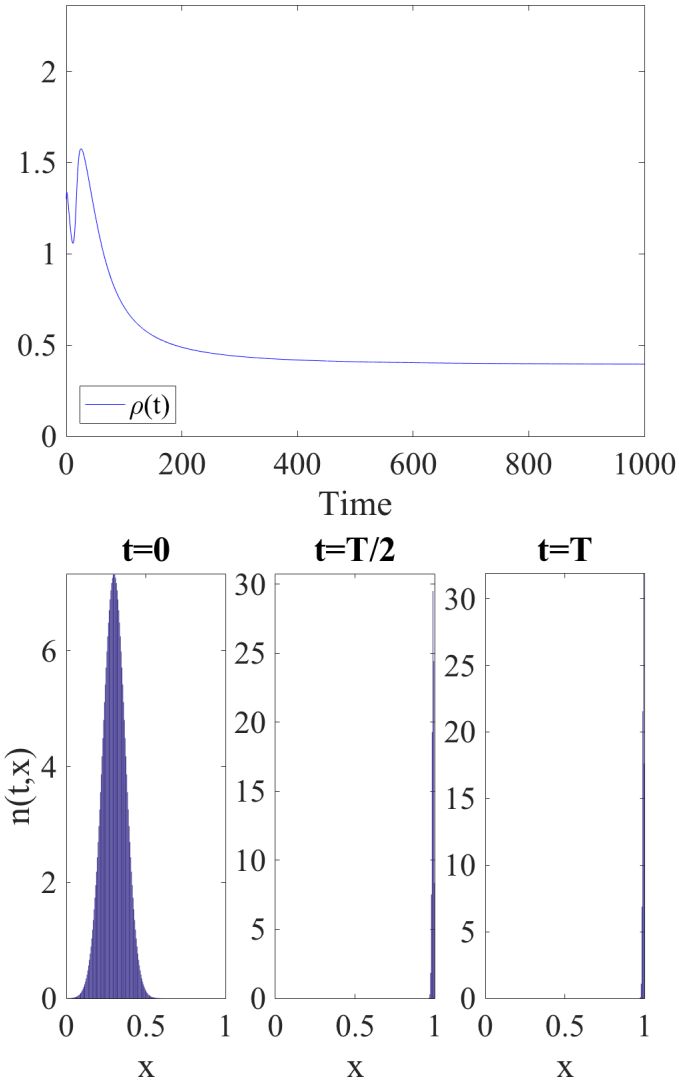

FIGURE 2 – Comportement d'une population, sans diffusion en temps grand (T = 1000). En haut : la population totale en fonction du temps. En bas : la population en fonction de la valeur du phénotype.

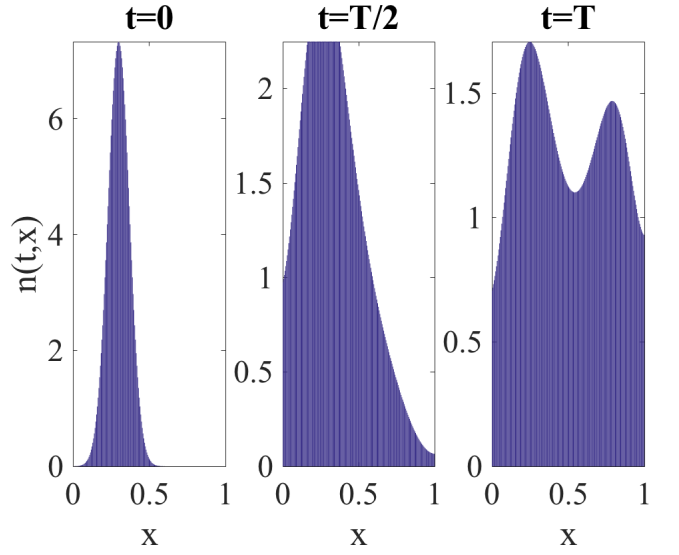

FIGURE 3 – Comportement de la population en fonction de la valeur du phénotype, lorsqu'on ajoute de la diffusion.

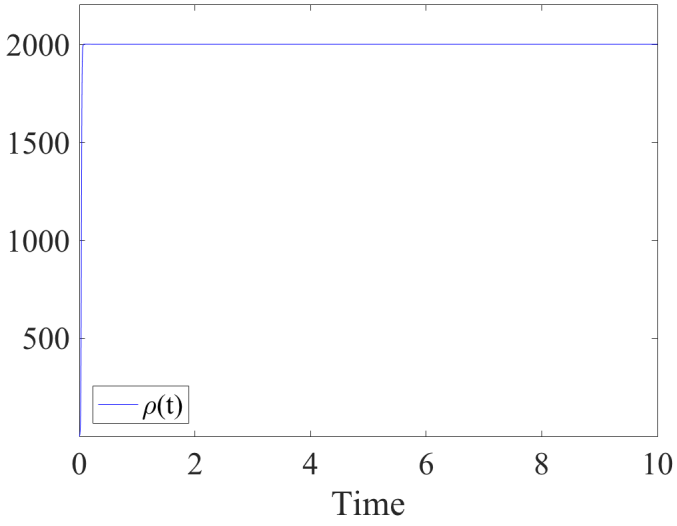

FIGURE 4 – Comportement d'une population avec un taux de croissance élevé, en fonction du temps.

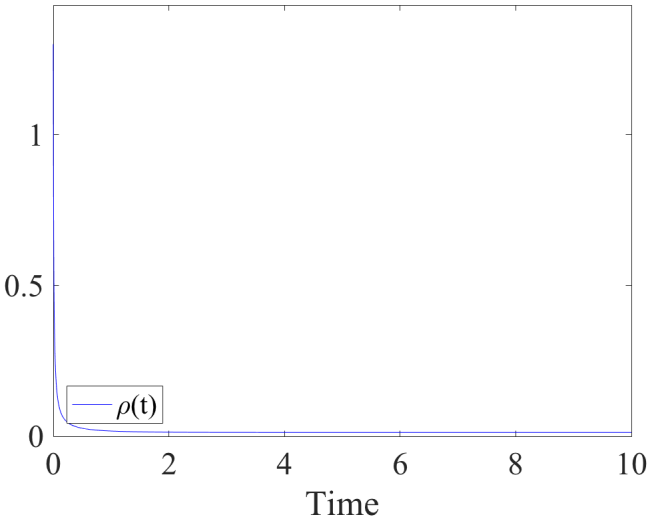

FIGURE 5 – Comportement d'une population avec un taux de mort élevé, en fonction du temps.

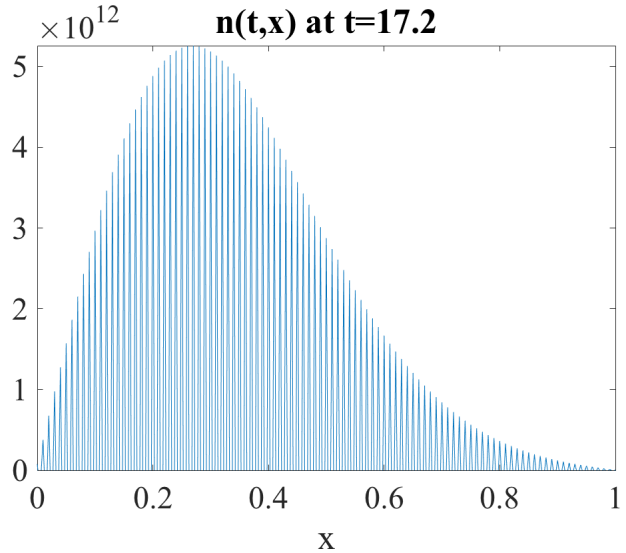

FIGURE 6 – Comportement de la population en fonction de la valeur du phénotype, lorsque la diffusion est trop grande : perte de stabilité.

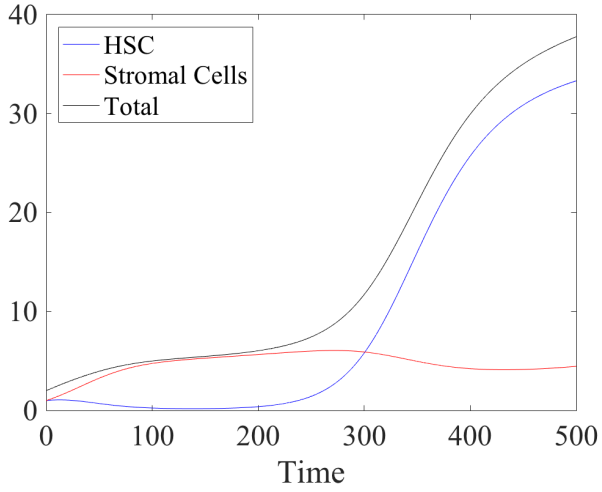

FIGURE 7 – Comportement des populations en fonction du temps. Sans diffusion, paramètres d'interaction faibles.

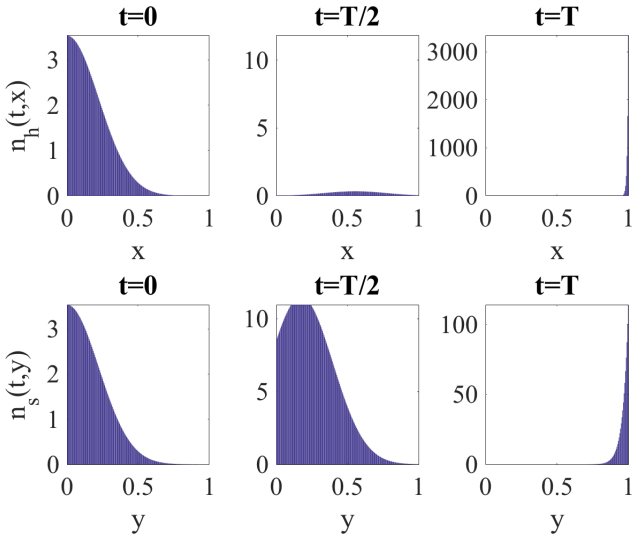

FIGURE 8 – Comportement des populations en fonction des phénotypes. Sans diffusion, paramètres d'interaction faibles.

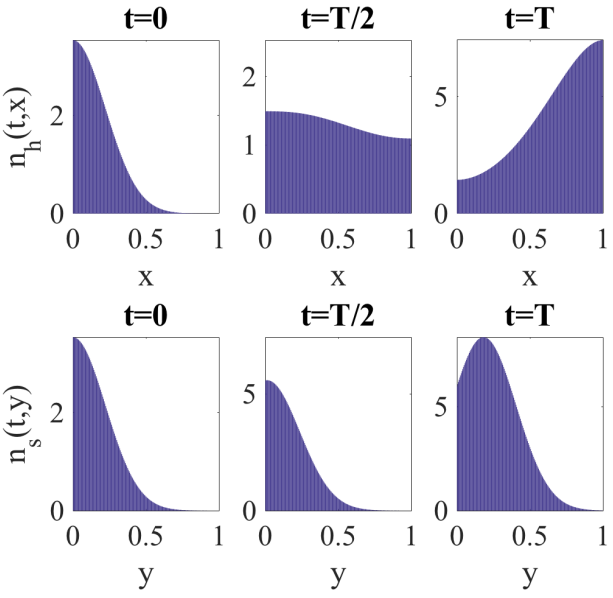

FIGURE 9 – Comportement de la population en fonction de la valeur du phénotype, lorsqu'on ajoute de la diffusion.

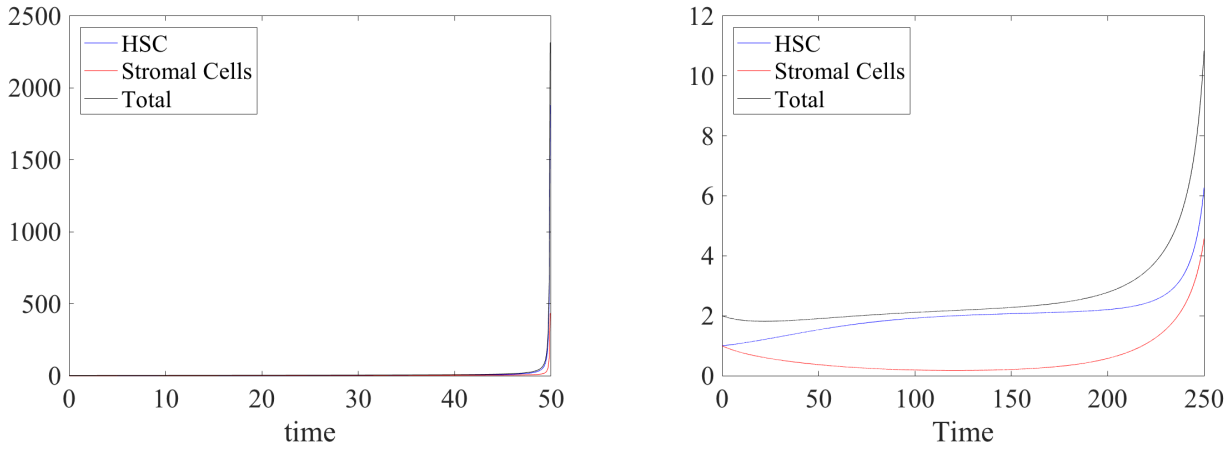

FIGURE 10 – Comportement des populations en fonction du temps, quand le mutualisme est trop fort.

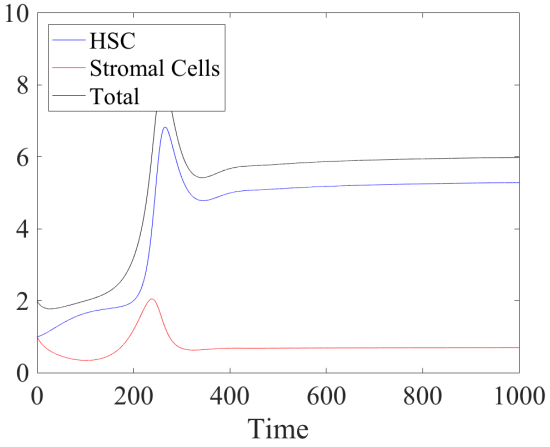

FIGURE 11 – Comportement des populations en temps grand (T = 1000).

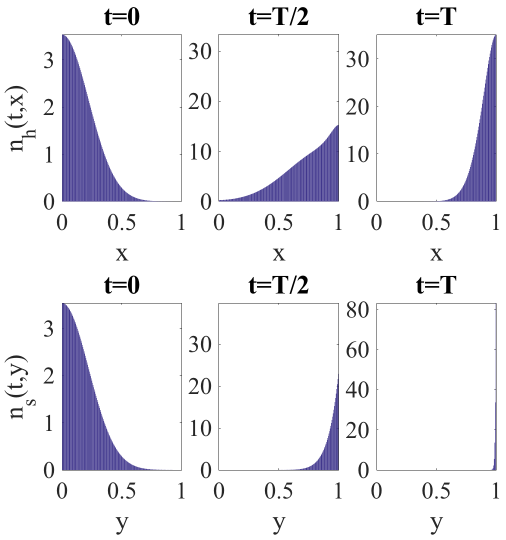

FIGURE 12 – Comportement de la population en fonction de la valeur des phénotypes.

6 CONCLUSION

Ce stage m'a permis d'acquérir de nombreuses connaissances, que ce soit dans le cadre de la biologie et des méthodologies utilisées dans les manipulations biologiques ou notamment les différentes méthodes pour le traitement statistique des données biologiques, grâce à la lecture de nombreux articles, grâce à l'aide de Pierre Hirsch pour la partie biologique, celle de Jean Clairambault, Delphine Salort et Camille Pouchol pour la partie mathématique, grâce au rapport de Hicham Janati et aux articles de Hugues Richard, et grâce aux réunions HTE notamment. Il m'a également apporté l'expérience d'évoluer dans le cadre d'un projet de recherche, avec tous les défis et difficultés que cela engendre : ainsi, les difficultés rencontrées ici ont impliquées que ce stage a été en grande partie théorique et n'a pas pu apporter de résultats concrets dans le cadre de la modélisation numérique. Les raisons sont tout d'abord la durée du stage : en effet, s'agissant d'un stage dans un domaine précis et complexe qu'est la recherche en applications biologiques et médicales, et plus précisément ici la recherche pour la leucémie aiguë myéloïde, un grand nombre de concepts théoriques était à assimiler avant d'entamer concrètement le travail. Ensuite, le projet HTE n'étant qu'à ses débuts, il était trop tôt pour obtenir de la part des biologistes des données pour pouvoir les introduire dans notre modèle HTE, et d'autre part, du côté mathématique, ce début de projet a soulevé beaucoup de questionnements concernant l'interface possible entre les données biologiques et paramètres du modèle, questionnements qui n'ont pas pu être entièrement résolus durant ces quatre mois de stage.

Concernant ce lien entre les données biologiques et paramètres des modèles, il s'agit d'être un peu plus précis : d'un côté, il sera bientôt (lorsque les biologistes auront des résultats concluants) possible de tirer des données biologiques certains paramètres du modèle HTE, tels que les taux de prolifération et d'apoptose, ou même des informations concernant le phénotype de support stromale, qui pourra normalement être relativement facilement identifiable à partir des réseaux de gènes ; de l'autre, un problème plus complexe se pose concernant la quantification des interactions entre les deux populations, et également concernant l'interprétation fonctionnelle des réseaux de gène pour le phénotype de plasticité. En effet, pour ce qui est de la difficulté de quantifier les interactions entre cellules stromales et CSH, cela provient du fait qu'il est possible de cultiver des cellules stromales seules et de faire des co-cultures CSH-cellules stromales, mais qu'il est en revanche impossible de cultiver des CSH seules, car celles-ci meurent. Une possibilité pour résoudre ce problème, est de cultiver des CSH avec des cellules stromales avec des capacités de soutien plus ou moins fortes, afin d'évaluer l'importance des interactions.

Pour ce qui est de l'interprétation fonctionnelle des résultats venant des manipulations biologiques, tels que les réseaux de gènes, notamment concernant le phénotype de plasticité des CSH, il s'agit tout d'abord d'expliquer les raisons qui expliquent un tel manque de lien entre les deux domaines, c'est-à-dire entre les données biologiques et le modèle mathématique. La première raison est qu'il faut comprendre que le but des biologistes et celui des mathématiciens n'est pas tout à fait le même et n'a pas le même angle de vue. Du côté de la biologie, on cherche notamment à trouver des cibles pour la thérapie, de comprendre quels réseaux de gènes sont activés par le contact, et donc on cherche à regarder plutôt ce qu'il se passe dans chaque état (i.e. sain, pré-leucémique et leucémique) avant et après mise en contact des deux populations, alors que du point de vue mathématique, le but est de comprendre ce qu'il se passe **entre** les différents états : quels sont les événements, liés à la plasticité cellulaire, qui font passer les cellules

d'un état pré-leucémique à leucémique, par exemple ? Qu'est ce qui favorise la prochaine mutation ? La deuxième raison du manque de lien, est du au fait que le modèle mathématique n'est pas forcément créé dans un but de prédictions *quantitatives* mais *qualitatives*, c'est-à-dire qu'il a pour objectif de données une idée des comportements possibles des populations (extinction, convergence, croissance non-bornée,...), et ces comportements peuvent ensuite être confirmés ou infirmés par les données biologiques, de manière à améliorer le modèle mathématique.

Certaines pistes sont cependant à explorer dans le cadre de l'interface entre données biologiques et modèle mathématique, notamment l'interprétation des réseaux de gènes, les approches Single-Cell et le concept de *Cold Genes*. Une possibilité pour l'interprétation fonctionnelle de réseaux de gènes est basée sur la mesure d'entropie des clusters ou modules de gènes, c'est-à-dire la distribution des pourcentages de gènes d'un type donné par rapport à l'ensemble des gènes qui peut être calculé pour chaque cluster. Ce genre d'approche pourrait être utile pour obtenir des informations concernant le phénotype de plasticité. Concernant les méthodes Single-Cell, on a déjà montré dans la section 4.3 qu'elles avaient déjà permis des avancées dans des domaines très variés et que, dans cette idée-là et par le fait que c'est un domaine en évolution, elles représentaient des approches intéressantes. De manière plus précise concernant le projet HTE, on peut voir une idée poussée à l'extrême en imaginant l'expression d'un gène comme codant totalement un phénotype : l'analyse Single-Cell permettrait alors d'obtenir la distribution d'expression du gène parmi les différentes cellules et ainsi, par extension, la densité de probabilité du phénotype. Finalement, par la capacité des approches Single-Cell à identifier l'hétérogénéité des populations de cellules stromales et CSH, elles peuvent être d'une grande importance, par exemple dans l'analyse de cette hétérogénéité dans le cadre de la résistance aux médicaments, i.e. dans un but thérapeutique. Enfin, concernant les *Cold Genes*, concept qui apparait dans l'article [23] qui se situe dans le cadre d'un cancer hématologique, *multiple myeloma*, où apparait une résistance en thérapie, ils sont définis comme étant des gènes jamais substitués (par opposition aux *hot genes*, gènes avec des densités de substitution élevées). Dans cette étude, le séquençage ARN des cellules résistantes est utilisé pour examiner l'émergence de *hot* et *cold genes*, révélant une grande fréquence aux substitutions et non-substitutions dans les cellules résistantes émergentes. Cette étude suggère que les *cold genes* avec de grands changements d'expression sont probablement des gènes importants, qui ne peuvent être substitués facilement car jouant un rôle clé dans la survie, la "santé" des cellules, et qu'ils pourraient être des gènes très anciens, représentant ainsi une fonctionnalité essentielle que le cancer utilise pour survivre dans des conditions de stress élevées, comme, vraisemblablement, des formes de vies précoces ont du expérimenter. Ainsi, l'intérêt essentiel de l'étude des *cold genes* est qu'ils peuvent être responsables de l'adaptation de cellules leucémiques à l'environnement : ils sont là pour mettre en jeu des mécanismes de survie et peuvent, par cet aspect, être une aide à la résistance de petites populations cancéreuses.

Pour conclure, le projet HTE qui n'est qu'à son commencement a de nombreux défis à relever et de nombreuses pistes possibles à suivre, telles que celles exposées ci-dessus. Les enjeux à long terme d'un tel projet sont importants : il peut apporter des avancées dans un cadre thérapeutique en identifiant des nouvelles cibles pour éradiquer le clone malin de la LAM, et peut éventuellement apporter des moyens *préventifs* (par le compréhension du dialogue CSH-stroma et ainsi du développement de la LAM entre les états pré-leucémiques et leucémiques par exemple, il serait possible de cibler les sources de ce développement). Ainsi, ce stage, bien que s'inscrivant au début de ce projet, m'a donné la chance de participer et de mieux comprendre les fonctionnements de tels projets de recherche.

ANNEXE

MODÈLE D'UNE POPULATION STRUCTURÉE EN PHÉNOTYPE, AVEC DIFFUSION

```
clc
clear all
close all

set(0,'DefaultAxesFontName', 'Times New Roman')
set(0,'DefaultAxesFontSize', 24)


%% Setup
%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%% Space
N = 200;
x=linspace(0,1,N+1);
dx=(x(2)-x(1));

%% Time
dt = 0.0002;
T=10; %150
time = 0:dt:T;

%% Functions/Parameters
sig=0.01;

a=0.05;
r=1-60*(x-a).*(x-(1/2)).^2 .*(x-(1-a));
d=100;
mu=0.04;
%% Initial conditions
n(:,1) = exp(-(x-0.3).^2/sig);
n(:,1)= 1.3* n(:,1)/(sum(n(:,1))*dx);


%% Evolution
for p=1:T/dt

    rho(p)=sum(n(:,p))*dx;


%     Diffusion

for k=2:N
    n(k,p+1) = n(k,p) + mu*(dt/(dx)^2)*(n(k+1,p)-2*n(k,p)+n(k-1,p));
    % artefacts en 1+dx et -dx pour faire Neumann
    int1=n(1,p);
    int2=n(N+1,p);
    n(1,p+1)= n(1,p) + mu*(dt/(dx)^2)*(n(2,p)-2*n(1,p)+int1);
    n(N+1,p+1)= n(N+1,p)+ mu*(dt/(dx)^2)*(int2-2*n(N+1,p)+n(N,p));
```

```
end



%% Growth (p+1/2-->p+1)
    %% on prend les rho actualisés au temps p+1/2
    roint=sum(n(:,p+1))*dx;

    R = r - d*(roint);
    Rp=max(0,R);
    Rm=-min(0,R);

    n(:,p+1)= 1./(1+dt*Rm').*(n(:,p+1) + dt*n(:,p+1).*Rp');


   if mod(p,1000)==0
    clf
    plot(x,n(:,p))
    axis([0 1 0 max(n(:,p))])
    xlabel('x')
    title(['n(t,x) at t=',num2str(time(p+1))])
    drawnow
    end
end


figure
subplot(1,3,1)
bar(x,n(:,1))
axis([0 1 0 max(n(:,1))])
title('t=0')
xlabel('x')
ylabel('n(t,x)')
subplot(1,3,2)
bar(x,n(:,T/dt*0.25))
axis([0 1 0 max(n(:,T/dt*0.5))])
title('t=T/2')
xlabel('x')
subplot(1,3,3)
bar(x,n(:,T/dt))
axis([0 1 0 max(n(:,T/dt))])
title('t=T')
xlabel('x')


figure
plot(time(1:end-1),rho,'b')
axis([0 T 0 1.1*max(rho)])
xlabel('Time')
legend('\rho(t)','location','southwest')
```

MODÈLE HTE

```
clc
clear all
close all
```

```
set(0,'DefaultAxesFontName', 'Times New Roman')
set(0,'DefaultAxesFontSize', 24)


%% Setup
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%% Space
N = 200;
x=linspace(0,1,N+1);
y=linspace(0,1,N+1);
dx=(x(2)-x(1));
dy=(y(2)-y(1));


%% Time
dt = 0.05;
T=1000; %150
time = 0:dt:T;

%% Functions/Parameters
% HSC parameters

muh = 10^-4;

rh_0=0.05;
rh_1=0.01;
dh_0=0.0248;
dh_1=0.0093;
sh_0=0.01;
sh_1=0.08;
rh=rh_1*x+rh_0*(1-x);
%dh=dh_1*x;
dh=dh_1*x+dh_0*(1-x);
%sh=sh_1*x;
sh=sh_1*x+sh_0*(1-x);

wh=x; %weighting function for phi_h

% Stromal parameters

mus = 0;

rs_0=0.06;
rs_1=0.058;
ds_0=0.0533;
ds_1=0.0016;
ss_0=0.007;
ss_1=-0.01;
rs=rs_1*y+rs_0*(1-y);
ds=ds_1*y+ds_0*(1-y);
%ss=ss_1*x;
ss=ss_1*y+ss_0*(1-y);

ws=y; %weighting function for phi_s

%% Initial conditions
```

```
sig=0.1;
nh(:,1) = exp(-(x-0.0).^2/sig);
rhoh0=sum(nh(:,1))*dx;
nh(:,1)=nh(:,1)/rhoh0;
ns(:,1) = exp(-(y-0).^2/sig);
rhos0=sum(ns(:,1))*dy;
ns(:,1)=ns(:,1)/rhos0;

%% Evolution
for p=1:T/dt

    rhoh(p)=sum(nh(:,p))*dx;
    %phih(p)=sum(nh(:,p)*wh)*dx;
    %phih(p)=(sum(nh(:,p).*x')*dx) /roc(p) ;
    rhos(p)=sum(ns(:,p))*dy;
    %phis(p)=sum(ns(:,p)*ws)*dx;
    rhotot(p) = rhoh(p)+ rhos(p);
    %phia(p)=(sum(na(:,p).*x')*dx) / roa(p);

    %% Diffusion (p--> p+1/2)


for k=2:N
    nh(k,p+1) = nh(k,p) + muh*(dt/(dx)^2)*(nh(k+1,p)-2*nh(k,p)+nh(k-1,p));
    % artefacts en 1+dx et -dx pour faire Neumann
    int1=nh(1,p);
    int2=nh(N+1,p);
    nh(1,p+1)= nh(1,p) + muh*(dt/(dx)^2)*(nh(2,p)-2*nh(1,p)+int1);
    nh(N+1,p+1)= nh(N+1,p)+ muh*(dt/(dx)^2)*(int2-2*nh(N+1,p)+nh(N,p));

    ns(k,p+1) = ns(k,p) + mus*(dt/(dy)^2)*(ns(k+1,p)-2*ns(k,p)+ns(k-1,p));
    % artefacts en 1+dx et -dx pour faire Neumann
    int1=ns(1,p);
    int2=ns(N+1,p);
    ns(1,p+1)= ns(1,p) + mus*(dt/(dy)^2)*(ns(2,p)-2*ns(1,p)+int1);
    ns(N+1,p+1)= ns(N+1,p)+ mus*(dt/(dy)^2)*(int2-2*ns(N+1,p)+ns(N,p));

end


    %% Growth (p+1/2-->p+1)
    %% on prend les rho actualisés au temps p+1/2
    rhohint = sum(nh(:,p+1))*dx;
    phihint = (sum(nh(:,p+1).*wh')*dx);
    rhosint = sum(ns(:,p+1))*dy;
    phisint = (sum(ns(:,p+1).*ws')*dy);
    rhototint = rhohint+rhosint;


    Rh = rh - dh*(rhototint) +  phisint*sh;
    Rs = rs - ds*(rhototint) +  phihint*ss;

    Rhp=max(0,Rh);
    Rhm=-min(0,Rh);

    Rsp=max(0,Rs);
    Rsm=-min(0,Rs);
```

```
    nh(:,p+1)= 1./(1+dt*Rhm').*(nh(:,p+1) + dt*nh(:,p+1).*Rhp');

    ns(:,p+1)= 1./(1+dt*Rsm').*(ns(:,p+1) + dt*ns(:,p+1).*Rsp');


end

figure
subplot(2,3,1)
bar(x,nh(:,1))
axis([0 1 0 max(nh(:,1))])
title('t=0')
xlabel('x')
ylabel('n_{h}(t,x)')
subplot(2,3,2)
bar(x,nh(:,T/dt*0.25))
axis([0 1 0 max(nh(:,T/dt*0.5))])
title('t=T/2')
xlabel('x')
subplot(2,3,3)
bar(x,nh(:,T/dt))
axis([0 1 0 max(nh(:,T/dt))])
title('t=T')
xlabel('x')

subplot(2,3,4)
bar(y,ns(:,1))
axis([0 1 0 max(ns(:,1))])
title('t=0')
xlabel('y')
ylabel('n_{s}(t,y)')
subplot(2,3,5)
bar(y,ns(:,T/dt*0.25))
axis([0 1 0 max(ns(:,T/dt*0.5))])
title('t=T/2')
xlabel('y')
subplot(2,3,6)
bar(y,ns(:,T/dt))
axis([0 1 0 max(ns(:,T/dt))])
title('t=T')
xlabel('y')


figure
plot(time(1:end-1),rhoh,'b')
hold on
plot(time(1:end-1),rhos,'r')
hold on
plot(time(1:end-1),rhotot,'k')
xlabel('Time')
legend('HSC','Stromal Cells','Total','location','northwest')
```

BIBLIOGRAPHIE

[1] Mostafa Adimy, Samuel Bernard, Jean Clairambault, Fabien Crauste, Stéphane Génieys, and Laurent Pujo-Menjouet. Modélisation de la dynamique de l'hématopoïèse normale et pathologique. *Hématologie*, 14(5) :339–350, 2008.

[2] José Louis Avila, Catherine Bonnet, Jean Clairambault, Hitay Özbay, Silviu-Iulian Niculescu, Faten Merhi, Annabelle Ballesta, Ruoping Tang, and Jean-Pierre Marie. Analysis of a new model of cell population dynamics in acute myeloid leukemia. In *Delay Systems*, pages 315–328. Springer, 2014.

[3] Rhonda Bacher and Christina Kendziorski. Design and computational analysis of single-cell rna-sequencing experiments. *Genome biology*, 17(1) :63, 2016.

[4] Frédérique Billy, Jean Clairambault, Franck Delaunay, Céline Feillet, and Natalia Robert. Age-structured cell population model to study the influence of growth factors on cell cycle dynamics. *Mathematical Biosciences and Engineering*, page xx, 2012.

[5] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, 33(2) :155–160, 2015.

[6] Pierre Charbord, Claire Pouget, Hans Binder, Florent Dumont, Grégoire Stik, Pacifique Levy, Fabrice Allain, Céline Marchal, Jenna Richter, Benjamin Uzan, et al. A systems biology approach for defining the molecular framework of the hematopoietic stem cell niche. *Cell Stem Cell*, 15(3) :376–391, 2014.

[7] Rebecca H Chisholm, Tommaso Lorenzi, and Jean Clairambault. Cell population heterogeneity and evolution towards drug resistance in cancer : biological and mathematical assessment, theoretical treatment optimisation. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1860(11) :2627–2645, 2016.

[8] Arnak Dalalyan. Introduction à la statistique et à l'econométrie. Technical report, ENSAE, ParisTech, January 2015.

[9] MATLAB User's Guide. The mathworks. *Inc., Natick, MA*, `http://mathworks.com`, 1998.

[10] Pierre Hirsch, Yanyan Zhang, Ruoping Tang, Virginie Joulin, Hélène Boutroux, Elodie Pronier, Hannah Moatti, Pascale Flandrin, Christophe Marzac, Dominique Bories, et al. Genetic hierarchy and temporal variegation in the clonal history of acute myeloid leukaemia. *Nature communications*, 7, 2016.

[11] Rouzanna Istvánffy, Baiba Vilne, Christina Schreck, Franziska Ruf, Charlotta Pagel, Sandra Grziwok, Lynette Henkel, Olivia Prazeres da Costa, Johannes Berndt, Volker Stümpflen, et al. Stroma-derived connective tissue growth factor maintains cell cycle progression and repopulation activity of hematopoietic stem cells in vitro. *Stem cell reports*, 5(5) :702–715, 2015.

[12] Hicham Janati. *Investigating cancer resistance in a Glioblastoma cell line with gene expression data*. PhD thesis, INRIA, 2016.

[13] Tommaso Lorenzi, Rebecca H Chisholm, and Jean Clairambault. Tracking the evolution of cancer cell populations through the mathematical lens of phenotype-structured equations. *Biology Direct*, 11(1) :43, 2016.

[14] Alexander Lorz, Tommaso Lorenzi, Jean Clairambault, Alexandre Escargueil, and Benoît Perthame. Modeling the effects of space structure and combination therapies on phenotypic heterogeneity and drug resistance in solid tumors. *Bulletin of mathematical biology*, 77(1) :1–22, 2015.

[15] Alexander Lorz, Tommaso Lorenzi, Michael E Hochberg, Jean Clairambault, and Benoît Perthame. Populational adaptive evolution, chemotherapeutic resistance and multiple anti-cancer therapies. *ESAIM : Mathematical Modelling and Numerical Analysis*, 47(2) :377–399, 2013.

[16] J. D. Murray. *Mathematical biology : An introduction*, volume 1. Springer, 2002.

[17] Benoît Perthame. *Transport equations in biology*. Springer Science & Business Media, 2006.

[18] Camille Pouchol. *Modelling interactions between tumour cells and supporting adipocytes in breast cancer*. PhD thesis, UPMC, 2015.

[19] Antoine-Emmanuel Saliba, Alexander J Westermann, Stanislaw A Gorski, and Jörg Vogel. Single-cell rna-seq : advances and future challenges. *Nucleic acids research*, 42(14) :8845–8860, 2014.

[20] Oliver Stegle, Sarah A Teichmann, and John C Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3) :133–145, 2015.

[21] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4) :381–386, 2014.

[22] Barbara Treutlein, Doug G Brownfield, Angela R Wu, Norma F Neff, Gary L Mantalas, F Hernan Espinoza, Tushar J Desai, Mark A Krasnow, and Stephen R Quake. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell rna-seq. *Nature*, 509(7500) :371–375, 2014.

[23] Amy Wu, Qiucen Zhang, Guillaume Lambert, Zayar Khin, Robert A Gatenby, Hyunsung John Kim, Nader Pourmand, Kimberly Bussey, Paul CW Davies, James C Sturm, et al. Ancient hot and cold genes and chemotherapy resistance emergence. *Proceedings of the National Academy of Sciences*, 112(33) :10467–10472, 2015.