Université Claude Bernard, Lyon 1

Rapport de Stage
Master 2 Maths en Action

Etude de la variabilité de la réponse immunitaire T CD8 à l'aide de modèles à effets mixtes.

Guillaume Metzler
Stage effectué à Inria

Encadrants : Dr. Fabien Crauste
Dr. Olivier Gandrillon

9 Mars 2015 — 31 Août 2015

Résumé

Suite à l'infection de l'organisme par un pathogène, les cellules T CD8 de notre système immunitaire sont activées afin de détruire le pathogène. Des mécanismes complexes conduisent à la différenciation des lymphocytes T CD8 de l'état naïf jusqu'à un phénotype mémoire. Une fois activée, ces cellules présentent de grandes capacités à proliférer. En outre, une fois le stade effecteur atteint, les cellules effectrices acquièrent des propriétés cytotoxiques leur permettant de lutter contre l'agent infectieux. Les cellules mémoires jouent un rôle clef en cas de nouvelle infection par le même pathogène, elles sont activées plus rapidement entraînant une réponse plus efficace. Nous confrontons un modèle décrivant les dynamiques des différents phénotypes des cellules T CD8 à des données expérimentales afin voir si le modèle étudié est capable de rendre compte de la variabilité présente dans des données issues de trois lignées de souris. L'approche proposée repose sur l'utilisation des modèles à effets mixtes permettant l'estimation de paramètres d'un modèle. A l'aide de cette approche nous sommes en mesure d'identifier les paramètres à l'origine de la variabilité de la réponse. Une analyse de sensibilité complète l'étude de notre modèle et de la variabilité. Suite à l'estimation des valeurs de paramètres, nous définissons également une distance entre les individus et entre les lignées. Finalement, nous discutons de la capacité prédictive du modèle à partir de données mesurées au pic de la réponse T CD8, là où le nombre de cellules est le plus important.

Remerciements

Je tiens à remercier toute l'équipe du Master Mathématiques pour la Biologie et la Médecine pour m'avoir offert l'occasion d'effectuer un stage dans la recherche ainsi que le directrice de l'Institut Camille Jordan.

Je souhaiterai plus particulièrement remercier les membres de l'équipe Dracula (INRIA) de m'avoir reçu dans leurs locaux à Lyon en tant que stagiaire. Ils m'ont accueilli dans une ambiance plus que chaleureuse et ont fait en sorte que ce stage se déroule dans des conditions optimales. Ils m'ont permis de découvrir le monde de la recherche mais aussi à échanger avec les membres de l'équipe. J'ai également eu l'occasion de m'exercer un bon nombre de fois à l'exercice de la présentation orale que ce soit à travers des réunions, le groupe de travail ou des présentations à des biologistes.

Un grand merci à Fabien Crauste et à Olivier Gandrillon de m'avoir accepté comme stagiaire M2. La tache s'annonçait difficile étant donnée mon aisance à l'oral. Ils ont fait preuve d'une grande patience (et de courage) en m'entrainant à cet exercice un grand nombre de fois (quand on se rappelle de la première répétition pour une réunion avec les biologistes qui a eu lieu en anglais). J'ai également eu la chance de participer à un colloque Maths-Bio au mois de Juillet et d'y faire une présentation de mon travail. Cela a permis de rencontrer de nouvelles personnes et de discuter de certains points (théorique ou pratique) sur le sujet abordé. Fabien et Olivier ce sont toujours montrés d'une grande disponibilité et furent présents pour répondre à chacune de nos questions et de nos doutes au niveau des thèses ou du stage. Ils m'ont poussé à faire certaines choses, comme la participation à un colloque, que je n'aurais jamais pensé faire si l'on ne m'avait pas un petit peu forcé la main.

Je souhaiterai également remercier l'équipe de biologistes avec qui nous avons travaillé pendant ce stage : Jacqueline Marvel, Christophe Arpin et Margan Grau. Ils ont fortement participé à ce stage à travers la génération des données mais aussi à travers les diverses réunions qui ont pu avoir lieu et qui ont soulevées beaucoup de questions. Les premiers échanges furent difficiles à cause de mon ignorance dans le sujet, mais heureusement Christophe et Olivier étaient là pour m'expliquer le vocabulaire employé. Je tiens également à remercier les immunologistes pour m'avoir fait visiter leurs locaux ainsi que de m'avoir présenté le matériel employé pour processus de numération des cellules. Je leur souhaite une bonne continuation dans leurs travaux de recherches et un bon courage et une bonne réussite à Morgan Grau pour sa thèse.

Et finalement un grand merci aux stagiaires : Simon, Aurélien et Matthieu avec qui nous avons pu échanger tout au long du stage nos moments de doute ou d'angoisse mais aussi les bons moments (comme les parties de baseball, basketball,...). Un grand merci aussi aux étudiants en thèse dans l'équipe qui ont su nous conseiller tout au long de ce stage et nous aider quand nous avions besoin d'aide.

Table des matières

1 Introduction

Dans les deux premières sections nous introduisons le contexte biologique dans lequel s'inscrit ce stage. Nous présentons la réponse immunitaire de façon générale puis plus spécifiquement la réponse immunitaire T CD8. Cette introduction se base sur un ouvrage d'immunologie [1], sur la thèse d'Emmanuelle Terry [2] ainsi que sur une référence en ligne [3]. Le lecteur désireux d'en connaître davantage est invité à se référer à ces ouvrages.

1.1 Réponses immunitaires

L'étude du système immunitaire a permis des avancées majeures dans la société notamment à travers le développement de vaccins permettant de lutter contre des virus. Mais qu'est-ce que l'immunité ? Ce terme fait référence aux mécanismes de défenses qui interviennent pour supprimer toute menace, telle une infection, qui représente un risque de survie pour l'organisme. On distingue deux types d'immunité, l'immunité innée et l'immunité acquise.

L'immunité **innée** est non spécifique et peut agir contre nombre de micro-organismes sans tenir compte du type de pathogène rencontré. Cette immunité est présente dès la naissance chez les individus et constitue une première barrière de défense. Elle fait intervenir les muqueuses ainsi que les cellules phagocytaires, comme les macrophages, qui sont chargées d'éliminer les cellules infectées par l'agent infectieux. L'immunité **acquise** est une réponse adaptative qui intervient après que l'organisme ait été exposé à un élément étranger spécifique. Des cellules spécialisées, appelées lymphocytes, sont chargées d'éliminer l'agent pathogène de l'organisme. Il existe plusieurs familles de lymphocytes qui interviennent dans le cadre de la réponse adaptative, les lymphocytes B, les lymphocytes T. Les lymphocytes B luttent contre l'agent pathogène en sécrétant des anticorps spécifiques à l'antigène rencontré. L'antigène est un fragment issu de l'agent pathogène qui permet le déclenchement de la réponse immunitaire. L'antigène peut se présenter sous la forme d'un fragment de protéine, appelé peptide, présent à la surface de cellules compétentes appelées APC (*Antigen Presenting Cell*, voir Figure1).

L'efficacité de la réponse immunitaire, innée ou acquise, tient en partie à la capacité des cellules de l'organismes à distinguer le soi du non-soi. Tandis que certaines cellules, tels que les macrophages, sont actives indépendamment de la présence du pathogène, les lymphocytes ne s'activent qu'en présence de l'agent infectieux. Il arrive cependant que l'organisme active des lymphocytes contre des éléments du soi. Les lymphocytes attaquent alors des cellules de l'organisme, entrainant des maladies auto-immunes. Nous ne considérerons pas ce cas et nous nous intéresserons uniquement au rôle des lymphocytes T dans la lutte contre le non-soi. Les mécanismes de développement des lymphocytes T CD8 sont présentés dans la section suivante.

1.2 Réponse immunitaire T CD8

Les lymphocytes T CD8 jouent un rôle majeur dans la réponse immunitaire. En présence du pathogène, ces lymphocytes sont capables d'enclencher leurs mécanismes de différenciation [1]. Au cours de leur différenciation, les lymphocytes vont acquérir des propriétés cytotoxiques qui vont leur permettre de lutter contre l'agent infectieux. On distingue trois grandes étapes, également appelées phénotypes, dans le développement des lymphocytes T CD8 : l'état naïf, l'état effecteur et l'état mémoire. Le passage de l'un à l'autre des états se fait selon deux hypothèses :
(1) soit par une différenciation "parallèle", c'est-à-dire qu'une cellule naïve se différencie soit en cellule effectrice, soit en cellule mémoire. Cette vision est peu répandue dans la littérature et n'est donc pas celle que nous retiendrons ici.
(2) soit par une différenciation "linéaire" où la cellule passe par l'état effecteur avant d'acquérir le phénotype mémoire [2]. Il s'agit de l'hypothèse que nous conserverons dans la suite.

Comme toutes les cellules immunitaires, les cellules T CD8 sont produites au niveau de la moelle osseuse.

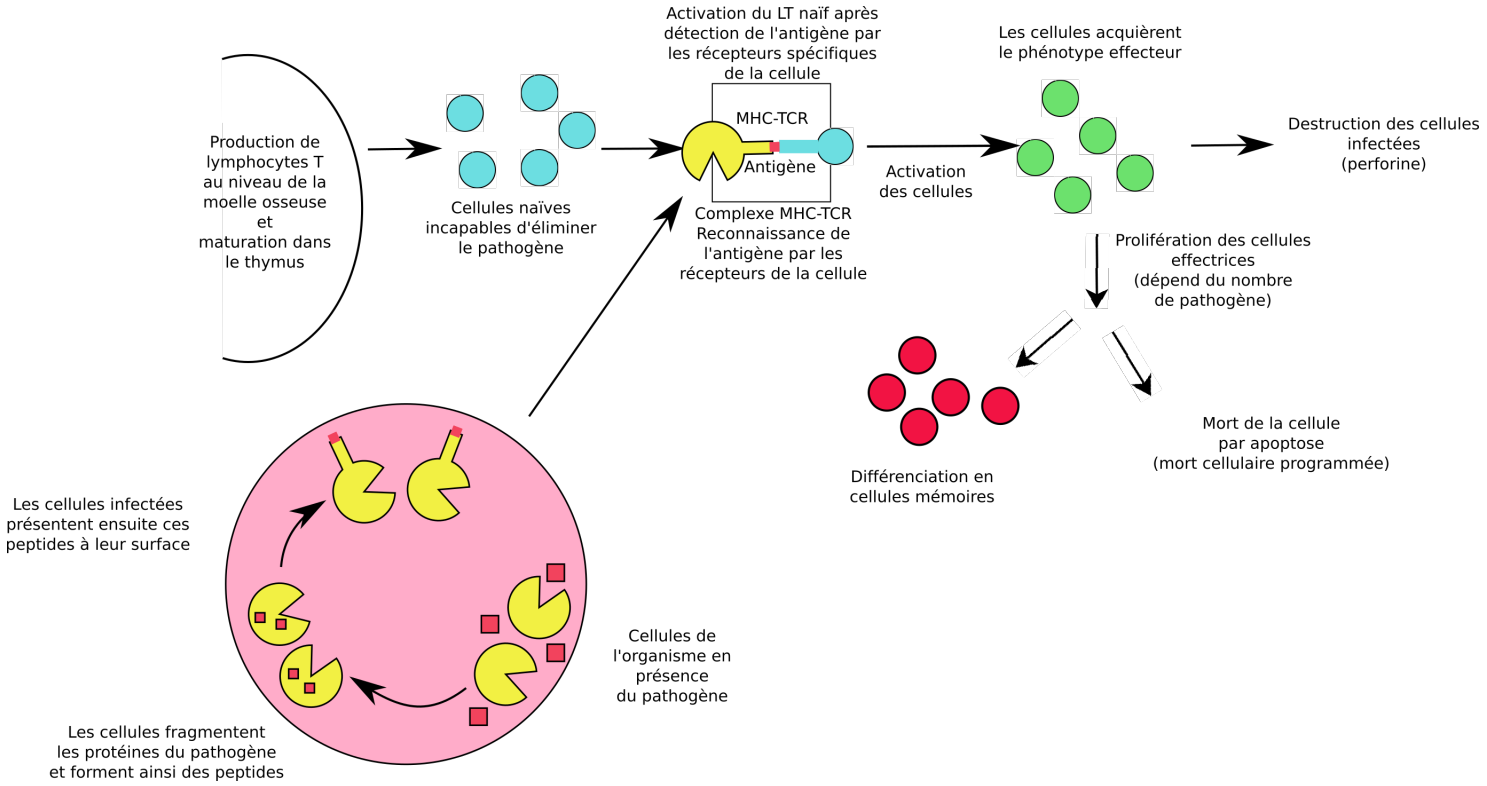


FIGURE 1 – Schéma bilan de la réponse immunitaire des cellules T CD8. Les cellules bleues représentes les cellules naïves. Nous représentons les cellules activées par la couleur verte et les cellules mémoires par la couleur rouge. Le cercle rose en bas à gauche de la figure est un insert présentant présentant la façon dont l'agent infectieux est exprimé à la surface de certaines cellules, comme les macrophages, devenant ainsi des cellules présentatrices d'antigènes (APC).

A leur naissance ces lymphocytes T sont dits immatures. Ils migrent ensuite vers le thymus où va débuter un processus de maturation. C'est lors de ce processus que les cellules vont développer leurs récepteurs, les "récepteurs des cellules T" appelés TCR (*T Cell Receptors*). Ce sont ces récepteurs qui vont permettre aux cellules naïves de reconnaître un antigène. Le caractère aléatoire du développement de ces récepteurs permet à l'organisme de créer une barrière contre un large spectre d'antigènes [1, 3]. Cependant, lors de cette même phase de maturation, les lymphocytes développent également des TCR capables de reconnaître des antigènes du soi, c'est pourquoi cette phase de maturation est suivie d'une phase de sélection appelée sélection thymique [3, 4]. Pendant cette phase, les lymphocytes immatures ayant développé des récepteurs leurs permettant de lutter contre les cellules du soi et représentant un danger pour l'organisme sont éliminés, cela représente environ 90% des lymphocytes [4]. On appelle cela la sélection négative. Au contraire, les cellules capables de reconnaître le non-soi sont libérées dans les organes lymphoïdes secondaires comme les ganglions ou la rate, on parle de sélection positive.

A ce stade là, les lymphocytes libérés dans l'organisme n'ont pas encore rencontré d'antigène et ne possèdent pas la propriété de lutter contre une infection. Pour acquérir cette faculté, les cellules naïves doivent rencontrer une cellule présentatrice de l'antigène spécifique aux récepteurs développés par la cellule naïve. Les cellules présentatrices d'antigène, appelées aussi APC, forment un réseau de sentinelles et peuvent, grâce à un système de reconnaissance, capter divers agents pathogènes par le biais de récepteurs (voir Figure 1). Les cellules naïves rencontrant ainsi une APC forment ce que l'on appelle un complexe MHC-TCR. Le MHC (*Major Histocompatibility Complex*) est un système de reconnaissance du soi présent chez les APC [3]. La cellule présentatrice peut ainsi présenter un peptide du pathogène à sa surface qui pourra être reconnu par un lymphocyte présentant des récepteurs spécifiques à cet antigène. Cette rencontre se fait dans les organes lymphoïdes secondaires comme les ganglions ou la rate. Les lymphocytes naïfs y circulent de façon continue et y arrivent par la circulation sanguine.

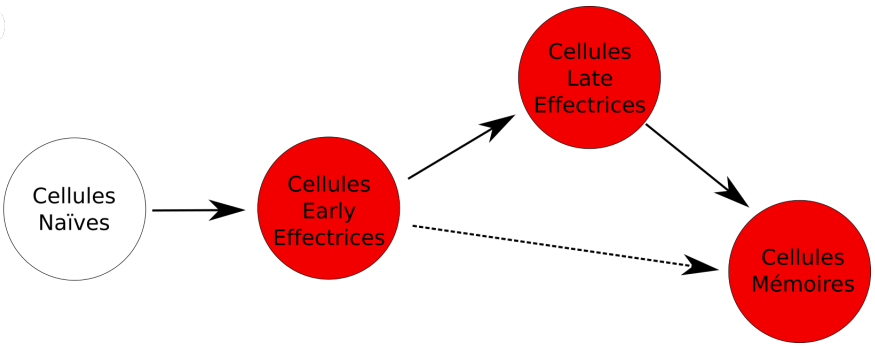


FIGURE 2 – Schéma de différenciation des cellules T CD8. Les stades de différenciation en rouge représentent des cellules dites activées. Les flèches en trait plein représentent le processus de différenciation. Une flèche en trait pointillé indique que le phénomène de différenciation existe mais qu'il est moins fréquent que les autres [5].

Les APC se trouvant dans ces organes sont chargées de peptides antigéniques capturés dans les tissus se trouvant autour de la zone lymphoïde. Ces peptides sont alors présentés aux lymphocytes T et déclenchent leur système de reconnaissance. Si la cellule T est incapable de reconnaître les peptides présentés par l'APC, alors elle se détache de l'APC. Dans le cas contraire elle reste fixée à l'APC et commence sa phase d'activation [4].

Lors de la la phase d'activation, plusieurs mécanismes se mettent en place au niveau moléculaire. Au contact de l'APC, le gène *IL2* de la cellule est stimulé ce qui provoque la production d'une cytokine du même nom, permettant la communication entre les cellules. Parallèlement à cela, la cellule développe également des récepteurs spécifiques à cette cytokine. La cytokine vient se fixer sur les récepteurs pour les activer. Lorsqu'un grand nombre d'entre eux sont activés, la cellule passe dans un état dit activé. Lors de ce même processus d'activation, l'activation des récepteurs à l'IL2 stimule le gène *Tbet* de la cellule. La stimulation de ce gène entraîne la sécrétion d'une protéine du même nom. Lorsque la concentration en Tbet de la cellule est suffisante, la cellule passe alors à l'état effecteur. Les mécanismes moléculaires sont présentés dans Prokopiou et al. [6].

Nous diviserons ici les cellules effectrices en deux sous-familles, les cellules early effectrices et les cellules late effectrices. Le schéma de différenciation linéaire se présente comme illustré sur la Figure 2. La différence entre les cellules early et late effectrices réside au niveau moléculaire et se traduit par leur capacité à proliférer ou non. Contrairement aux cellules early effectrices, les cellules late effectrices se trouvent dans un état non prolifératif [5].

Lorsqu'une cellule effectrice rencontre une cellule infectée, elle commence à proliférer, on assiste à une phase d'expansion clonale [2, 7] où le nombre de lymphocytes T CD8 augmente de façon exponentielle (voir Figure 3). Cette phase d'expansion permet une lutte plus efficace contre le pathogène et multiplie aussi les chances de rencontrer des cellules infectées afin de les éliminer. La stimulation par contact avec l'antigène augmente la vitesse de différenciation des cellules T CD8 [5].

Après l'élimination du pathogène, s'en suit une phase de contraction [2], phase pendant laquelle le nombre de cellules effectrices diminue considérablement. Environ 90% des cellules effectrices sont éliminées pendant la phase de contraction. Les mécanismes permettant d'expliquer l'élimination des T CD8 effecteurs sont explicités dans [6]. Le pourcentage de cellules restant constitue un stock de cellules mémoires. Ces cellules jouent un rôle primordial en cas de nouvelle infection par un pathogène déjà rencontré par l'organisme, elles garantissent une réponse plus rapide et plus efficace, comme l'illustre la Figure 3.

1.3 Etat de l'art et objectifs

La modélisation de la réponse immunitaire permet de mieux comprendre l'importance des mécanismes qui interviennent lors des différentes phases de la réponse. Des modèles déterministes et stochastiques sont alors développés pour étudier différents phénomènes.

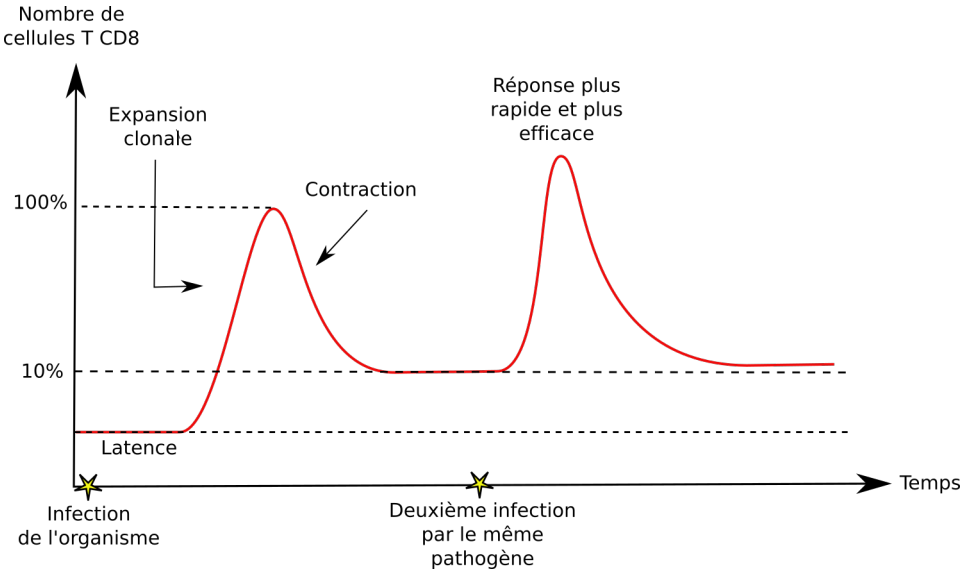


FIGURE 3 – Schéma de l'évolution du nombre de cellules T CD8 au cours du temps dans un organisme. Lors de la première rencontre avec un pathogène, on observe une période de latence pendant laquelle le nombre de cellules T CD8 n'évolue pas significativement, elle correspond à la période d'activation des lymphocytes naïfs par les APC. Une fois les cellules activées, on assiste à la phase d'expansion clonale qui se traduit par une augmentation importante de cellules T CD8, elle correspond à la prolifération des cellules effectrices au contact des cellules infectées. Lorsque les cellules infectées et le pathogène sont éliminés de l'organisme, les cellules effectrices sont ensuites éliminées par l'organisme, c'est le phénomène d'apoptose et cela correspond à la phase de contraction. A la fin de la phase de contraction subisistent majoritairement des cellules mémoires. En cas de nouvelle infection par le même pathogène, les cellules mémoires vont également être activées par les APC permettant une réponse plus rapide et donc plus efficace dans la lutte contre le pathogène.

L'utilisation de modèles déterministes est très courante dans la modélisation de la réponse immunitaire. La plupart de ces modèles sont construits à partir d'équations différentielles qui permettent de décrire les dynamiques des différentes populations de cellules rencontrées. Dans le cadre de l'étude d'une infection, Nowak et Bangham [8]présentent successivement plusieurs modèles pour décrire les dynamiques des populations de cellules infectées et non infectées ainsi que la dynamique du pathogène. Pour modéliser la réponse immunitaire T CD8, certains auteurs se basent sur des modèles du type proie-prédateur pour décrire les dynamiques des différentes populations, une population de cellules effectrices constituant la population des prédateurs et la population de cellules infectées (comprenant le pathogène) représentant les proies [8, 9]. On cherche alors à décrire les interactions entre les différentes populations. Dans [9], Antia et al. considèrent quatre populations de cellules : les cellules naïves, les cellules effectrices et les cellules mémoires ainsi que le pathogène. Ils proposent un modèle structuré en âge pour décrire les différentes étapes de différenciation de la cellule naïve à la cellule mémoire. Les auteurs font également le choix de faire dépendre le taux de prolifération des cellules en fonction du nombre de divisions déjà effectuées. Un autre modèle entièrement défini à l'aide d'équations différentielles est présenté par De Boer et Perelson [10]. Dans ce modèle, De Boer et Perelson font le choix d'introduire un terme source pour décrire la production continue de cellules naïves au niveau du thymus. Ils tiennent également compte de la concentration en pathogène dans l'organisme pour modéliser le phénomène de différenciation des cellules naïves.

Cependant, lorsqu'un individu est infecté, les cellules présentatrices d'antigène ont pour rôle d'activer les cellules naïves par contact. Or ce phénomène est tout à fait aléatoire et ne concerne pas la population entière de cellules. Certains auteurs proposent alors une approche stochastique pour modéliser la réponse immunitaire [11]. Si la base des modèles reste la même pour ce type de modèle, à savoir des équations différentielles qui permettent de décrire les dynamiques des différentes populations, les phénomènes de différenciation des cellules sont rendus stochastiques. Au lieu de considérer un taux de différenciation comme cela est le cas dans les modèles détermi-

nistes, on définit cette fois-ci une probabilité pour chaque cellule de se différencier ou de mourir par exemple. Ces approches stochastiques reposent sur des modèles à base d'agent comme présenté dans [11]. Ils permettent de prendre en compte la variabilité génétique présente chez les individus se traduisant par des réponses différentes à une infection. Cependant, contrairement à une approche déterministe, son implémentation est très coûteuse en temps de calcul lorsque l'on considère un très grand nombre de cellules. Si les modèles stochastiques sont un bon moyen de modéliser la variabilité, ce type de modèle reste minoritaire.

Très souvent, les modèles déterministes sont confrontés à des données expérimentales pour tester leur robustesse[7, 10]. L'une des étapes clefs de la modélisation à l'aide de modèles déterministes est la détermination des valeurs de paramètres permettant de décrire les données. Ces deux équipes s'intéressent également à la prise en compte de la variabilité dans leur modèle pour savoir si ce dernier est capable de reproduire, en fonction des valeurs de paramètres, les différents comportements observés à l'échelle des individus.

Une étude de la variabilité de la réponse immunitaire au niveau des valeurs de paramètres est présentée dans [7]. La variabilité, présente au sein des données étudiées, conduit les auteurs à chercher des ensembles de valeurs de paramètres qui permettent de reproduire ces données à l'aide d'un système d'équations différentielles. Le système employé décrit la dynamique des différents phénotypes des cellules T CD8 : naïf, effecteur et mémoire. Les auteurs déterminent plusieurs ensembles de valeurs de paramètres afin de déterminer la valeur moyenne de chacun d'entre eux et de confronter leurs résultats avec ceux de la littérature. La méthode employée pour l'estimation des paramètres a l'avantage de ne supposer aucune loi quant à la distribution de ces derniers, elle se base uniquement sur les données et repose sur une méthode des moindres carrés. En outre, elle permet une première étude de la variabilité de la réponse immunitaire en terme de valeurs de paramètres du modèle. Cependant, la méthode employée est très coûteuse en terme de temps de calcul et nécessite de tester un très grand nombre de jeux de paramètres. De plus, elle ne permet d'estimer le meilleur jeu de paramètre qui permettrait de reproduire les dynamiques à l'échelle des individus, mais seulement à l'échelle de l'ensemble des individus (c'est-à-dire de l'ensemble des données).

Nous proposons une approche différente permettant d'estimer les paramètres du modèle, elle repose sur l'utilisation des modèles à effets mixtes. Cette méthode permet à la fois d'estimer les paramètres à l'échelle d'une population (en estimant les éléments caractérisant les distributions des différents paramètres) mais également à l'échelle des individus.

Les modèles à effets mixtes sont très utilisés dans des domaines tels que l'agronomie [12], la cancérologie [13]. Dans [13], les modèles à effets mixtes sont utilisés dans le but d'étudier la variabilité du taux de croissance des cellules chez différentes populations d'individus. On souhaite savoir comment la diversité génétique va influencer la croissance et ainsi adapter le traitement selon la population dont est issu notre individu. Cependant, ces modèles ont historiquement été conçu pour un usage pharmaceutique et plus particulièrement en pharmacocinétique pour étudier l'élimination d'un médicament chez les individus. Dans [14], les auteurs utilisent des modèles pharmaco-cinétiques ainsi que des modèles à effets mixtes pour étudier l'évolution de la concentration d'une substance présente dans le plasma sanguin. Ils cherchent entre autre à déterminer le temps d'absorption d'un médicament par l'organisme ainsi que la vitesse d'élimination de ce dernier. Le but étant d'étudier la variabilité de ces valeurs en fonction des différents facteurs phénotypiques des individus donc en créant des groupes d'individus. Ces outils permettent donc d'adapter le traitement selon les différentes caractéristiques d'un individu.

Ils permettent également de décrire plusieurs types de variabilités lors de l'estimation de paramètres dans un modèle et des variabilités présentes à plusieurs échelles : celle de l'individu et celle de la population.

La variabilité inter-individuelle est présente entre les individus d'une même population, elle se traduit par la variation d'un ou plusieurs paramètres du modèle autour d'une valeur typique qui caractérise la population. Chaque individu possède ainsi sa propre valeur de paramètre permettant ainsi de mieux décrire les données qui lui sont relatives. Cette variabilité est naturellement présente au sein des individus d'une même population et se reflète directement à travers les observations.

La variabilité entre populations intervient lorsque l'on souhaite étudier des données issues de différentes populations (par exemple des souris issues de différentes lignées. La variabilité va alors être décrite au niveau de la valeur typique caractérisant la population. Les variabilités inter-individuelles et entre populations sont illustrées par la Figure 4.

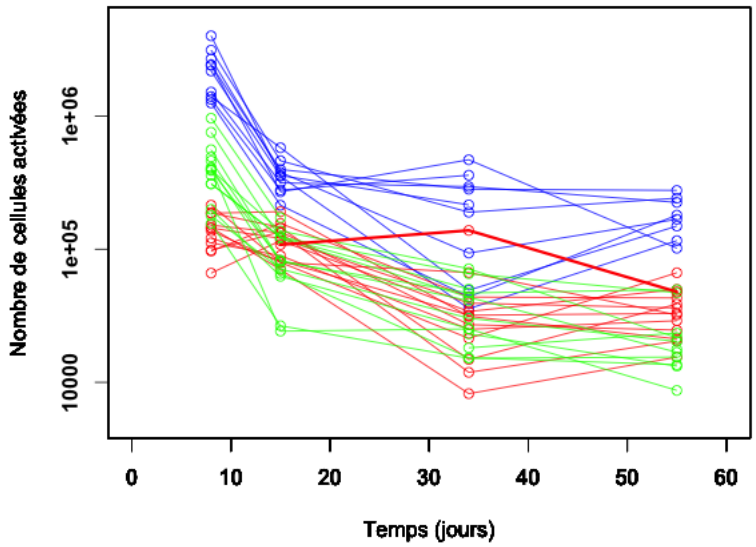


FIGURE 4 – Trajectoires individuelles pour différentes souris de différentes lignées : C57Bl/6, BalbC et OF1. On représente le nombre total de cellules activées mesurées au cours du temps, aux jours 8,15, 34 et 55, en échelle semi-logarithmique. Le nombre total de cellules activées est supposé nul au jour 0 : le pathogène n'a jamais été rencontré par l'organisme auparavant. Cette valeur n'est pas représentée sur le graphique à cause de l'utilisation l'échelle logarithmique sur l'axe des ordonnées. Le trait épais de couleur rouge représente un individu de la lignée BalbC, que l'on appellera individu 20 dans la suite, dont le comportement diffère de celui des autres individus de la lignée.

Le travail présenté se base sur les travaux effectués dans [7]. Nous considérons ici non plus trois différents phénotypes pour les cellules T CD8 mais quatre phénotypes [5], nous séparons la population de cellules effectrices en deux sous-populations : early effectrices et late effectrices. La distinction entre les cellules early et late se fait au niveau de l'expression de certains gènes. Les hypothèses qui permettent de distinguer ces deux sous-populations sont présentées en Section 2.2. Nous utiliserons les modèles à effets mixtes (voir section 2.3) afin d'étudier et de comprendre d'où est issue la variabilité de la réponse immunitaire T CD8 en terme de valeurs de paramètres. Nous tenterons finalement d'exploiter les valeurs estimées pour expliquer les différentes dynamiques observées (voir Figure 4).

2 Matériel, modèles et méthodes

2.1 Données

Les données qui vont nous permettre d'étudier la variabilité de la réponse immunitaire T CD8 sont fournies par l'équipe I2V dirigée par Jacqueline Marvel, Inserm U851, à laquelle appartiennent Christophe Arpin et Morgan Grau lesquels ont participé la génération des données. Elles sont présentées en Figure 4 et proviennent de souris issues des lignées C57Bl/6, OF1 et BalbC. Pour chaque lignée et chaque individu, nous disposons du nombre de cellules activées, c'est-à-dire la somme du nombre de cellules effectrices (early et late) et mémoires à différents temps : 0, 8, 15, 34 et 55 jours post-infection.

Pour les obtenir, un agent pathogène est injecté à la souris, ici le virus de la vaccine au jour 0. On compte le nombre de cellules activées chez la souris aux différents temps. Cette numération du nombre de cellules acti-

vées se fait à l'aide d'un marqueur fluorescent. Si l'expression de ce marqueur, autrement dit si l'intensité de la fluorescence, est suffisamment élevée, alors la cellule est dite activée. Il est important de comprendre qu'il n'est pas possible de distinguer les différents phénotypes cellulaires des cellules activées à l'aide de ce marqueur, c'est pourquoi nos données représentent simplement le nombre de cellules activées.

Nous appellerons *trajectoire* le suivi du nombre de cellules activées au sein d'une souris au cours du temps. Ainsi nous disposons de 10 trajectoires provenant d'individus de la lignée C57Bl/6, 13 trajectoires décrivant la dynamique chez des souris issues de la lignée BalbC et 15 trajectoires de souris issues de la lignée OF1.

Au jour 0, on mesure un nombre de cellules activées chez chaque souris, cette valeur mesurée est en fait un bruit. En effet, les souris sur lesquelles sont effectuées les mesures sont supposées ne jamais avoir rencontré le virus de la vaccine et ne possèdent aucune cellule activée spécifique à ce virus. Pour corriger ces données, nous avons alors retranché la valeur mesurée à jour 0 aux valeurs mesurées aux différents jours.

A travers les données présentées en Figure 4, on peut déjà observer la variabilité de la réponse à deux échelles différentes : à l'échelle des individus issus d'une même lignée mais aussi entre les différentes lignées. En effet, pour un point de mesure donné, la réponse, en terme de nombres de cellules activées, peut varier d'un log d'un individu à un autre, par exemple, pour les individus de la lignée BalbC au jour 34, le nombre de cellules activées varie entre 10^4 et 10^5. Cette différence peut même être de plusieurs log entre individus issus de lignées différentes. Cela est par exemple le cas entre les individus de la lignée C57Bl/6 et les individus de la lignée BalbC. Le nombre maximum de cellules mesuré à jour 8 est de 5.10^6 pour la lignée C57Bl/6 alors qu'il est de l'ordre de 5.10^4 pour les individus de la lignée BalbC (voir Figure 4). Nous pouvons également constater qu'une lignée semble se détacher des deux autres en terme de nombre de cellules activées. En effet, celui-ci est plus important chez les individus issus de la lignée C57Bl/6 que chez les OF1 et BalbC.

2.2 Modèle

Concentrons-nous d'abord sur le modèle qui servira à décrire les données. Ce dernier repose sur le schéma de différenciation présenté en Figure 2 ainsi que sur les travaux d'Emmanuelle Terry [15] et d'Antoine Burg au cours de son stage en 2014 [16]. Nous reprenons le système présenté par Antoine en modifiant la façon dont il est paramétré.

Pour ce faire, nous avons également besoin d'introduire des hypothèses qui régissent les interactions entre les différentes populations de cellules ainsi que les phénomènes de différenciation. Les hypothèses sont les suivantes :

- Les cellules early effectrices agissent sur la mort par apoptose des cellules early effectrices et des cellules late effectrices avec des taux différents [17].
- Les cellules late effectrices agissent sur la mort par apoptose des cellules early effectrices et des cellules late effectrices avec des taux différents [17].
- Seules les cellules early effectrices sont capables de proliférer. Les cellules late effectrices sont supposées se trouver dans un état non prolifératif [5].
- Les cellules early effectrices et les late effectrices agissent toutes les deux sur la mort du pathogène mais à des taux différents [18].
- Les cellules early effectrices et les cellules late effectrices sont capables de se différencier en cellules mémoires mais avec des taux différents [18].

Dans le modèle biologique décrivant les interactions entre les différentes populations de cellules présentées en Figure 5 nous désignons par N la population de cellules naïves, par E la population de cellules early effectrices, par L la population de cellules late effectrices. Les lettres M et P désignent respectivement la population de cellules mémoires et le pathogène.

Le modèle présenté en Figure 5 est décrit par le système différentiel présenté en (1).

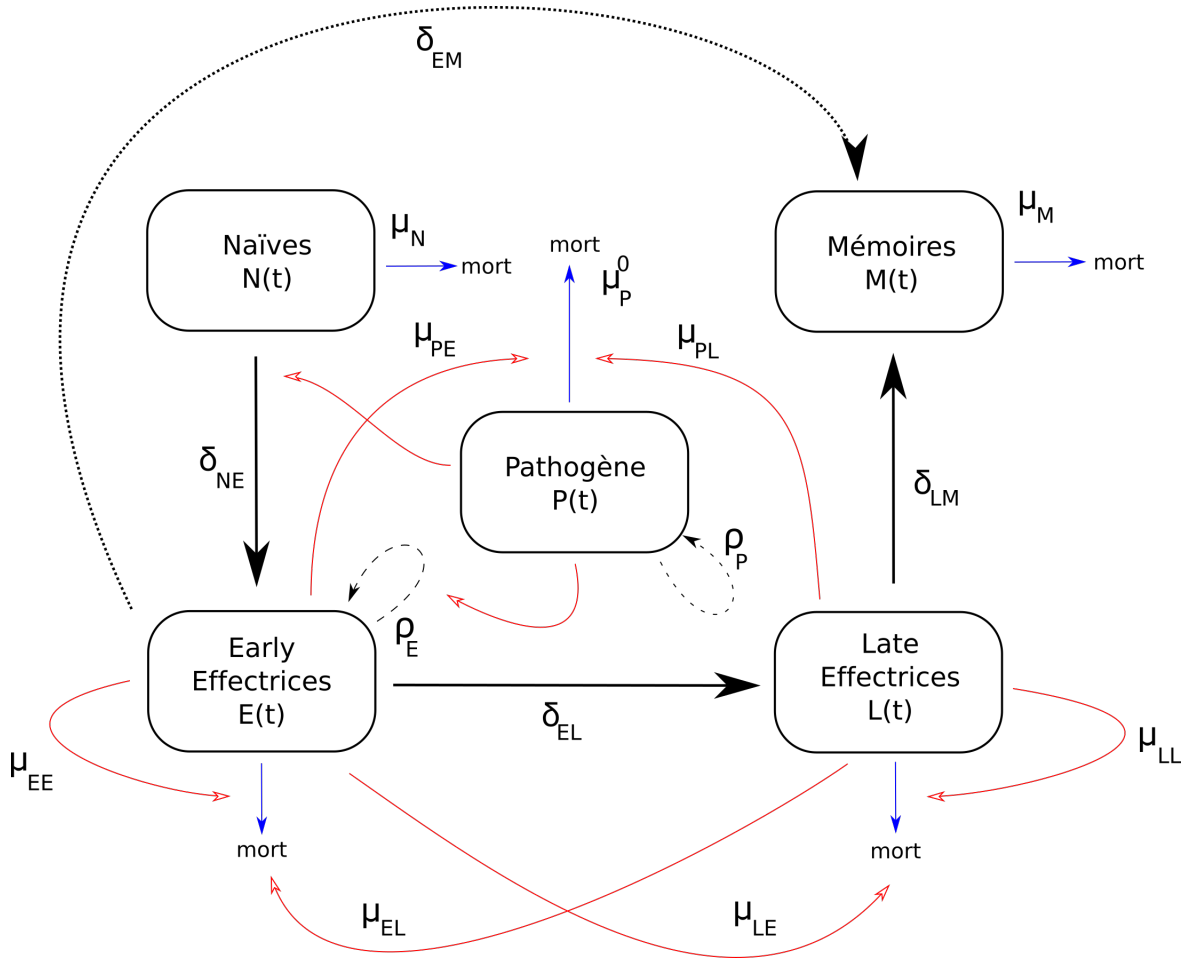


FIGURE 5 – Schéma du modèle décrivant les interactions entre le pathogène et les différentes populations de cellules. La lettre P désigne le pathogène, N les cellules naïves, E les cellules early effectrice, L cellules late effectrices et M les cellules mémoires. Les paramètres μ (respectivement δ et ρ) caractérisent les taux de mort (respectivement les taux de différenciation et de prolifération) des populations de cellules indiquées en indice. Les flèches épaisses de couleurs noires indiquent un processus de différenciation. Les flèches pointillés représentent le processus de prolifération. Une flèche rouge indique l'influence de la population se trouvant en amont de la flèche sur le processus présenté à l'autre extrêmité de la flèche. Une flèche bleu représente le processus de mort.

$$
\begin{aligned}
\frac{dN}{dt} &= -(\mu_N + \delta_{NE}P)N, & N(0) &= N_0,\\
\frac{dE}{dt} &= \delta_{NE}PN + E(\rho_E P - \mu_{EE}E - \mu_{EL}L - \delta_{EM} - \delta_{EL}), & E(0) &= 0,\\
\frac{dL}{dt} &= \delta_{EL}E - L(\mu_{LL}L + \mu_{LE}E + \delta_{LM}), & L(0) &= 0,\\
\frac{dM}{dt} &= \delta_{LM}L - \mu_M M + \delta_{EM}E, & M(0) &= 0,\\
\frac{dP}{dt} &= (\rho_P P - \mu_{PE}E - \mu_{PL}L - \mu_P^0)P, & P(0) &= 1.
\end{aligned}
\tag{1}
$$

Les différents paramètres présents au sein du système servent à décrire les différents mécanismes interve-

nant dans la réponse immunitaire au niveau cellulaire (différenciation, prolifération et mort). Le système comporte 16 paramètres, 15 d'entre eux sont des paramètres présents dans les différentes équations et le dernier paramètre représente le nombre de cellules naïves présent initialement dans l'organisme.

Les quantités $E(0), L(0)$ et $M(0)$ sont supposées nulles car nous étudions le cas où l'organisme n'a jamais rencontré le pathogène par lequel il est infecté. Il n'a ainsi pu développer aucune défense spécifique pour lutter contre ce pathogène.

Dans la suite nous supposerons que le paramètre μ_M, taux de mort des cellules mémoires, est égal à 0. Nous le choisissons ainsi car l'échelle du temps d'étude de la réponse, ainsi que des données, est beaucoup trop court pour qu'il soit nécessaire de prendre en compte ce phénomène, ce choix est également guidé par une remarque effectuée dans [7]. De plus, nous supposerons que le nombre initial de cellules naïves est le même chez tous les individus et indépendamment de la lignée. Le fait d'imposer cette valeur peut être compensé par une valeur du taux de mort des cellules naïves plus ou moins importante.

Nous chercherons à estimer les valeurs de paramètres qui permettront de décrire de la façon la plus exacte possible les données (voir Figure 4) à l'échelle de la population, c'est-à-dire pour une lignée de souris. Nous ferons ensuite de même en cherchant les valeurs de paramètres qui décrivent au mieux les données d'un individu.

Introduisons maintenant les outils statistiques nécessaires à l'estimation des paramètres du modèle.

2.3 Outils statistiques : modèles à effets mixtes

Les modèles à effets mixtes sont une généralisation des modèles de régression, ils permettent, à l'aide de données issues de différents individus sur une échelle de temps plus ou moins semblable de distinguer si des phénomènes (biologiques) peuvent présenter de la variabilité au sein des données utilisées et quelles sont les sources de cette variabilité. Afin de mieux comprendre quelles sont les sources de variabilités observées en Figure 4 nous employons ces modèles à effets mixtes pour étudier la variabilité des paramètres de notre modèle décrit par le système (1).

2.3.1 Modèles à effets mixtes

La présentation suivante se base sur celle du wiki dédié à popix [19]. Nous désignons par n le nombre d'individus et n_i le nombre d'observations dont nous disposons pour l'individu i. Les observations pour l'individu i sont notées $y_i = (y_{i1}, ..., y_{in_i}),\ 1 \leq i \leq n$. Les modèles à effets mixtes se présentent de la façon suivante :

$$y_{ij} \quad = \quad f(t_{ij}, p_i) + g(t_{ij}, p_i)\varepsilon_{ij} \tag{2}$$

où :

- $p_i \in \mathbb{R}^M$ désigne le vecteur des paramètres de l'individu i,
- t_{ij} est le temps de la mesure j chez l'individu i,
- ε_{ij} est une variable aléatoire qui est supposée indépendante des paramètres que l'on cherche à estimer. Ce terme permet de décrire l'erreur commise sur l'observation y_{ij}. On suppose que :

$$\varepsilon_{ij} \sim \mathcal{N}(0,1), \quad 1 \leq i \leq n, 1 \leq j \leq n_i.$$

- f est une fonction que l'on appelle modèle structurel. Il s'agit d'une fonction qui est définie pour décrire au mieux les données individuelles ainsi que la variabilité propre à chaque sujet. Le modèle structurel peut-être une fonction linéaire ou non linéaire en les paramètres p_i. Elle peut, par exemple, être définie à l'aide d'un système d'EDO.

- g est une fonction qui constitue le terme résiduel de notre modèle. Elle permet de prendre en compte l'erreur d'estimation, c'est-à-dire l'erreur commise entre les estimations fournies par le modèle f et les observations.
  Ce terme résiduel peut par exemple dépendre ou non de l'estimation effectuée. Il est possible de choisir une fonction g constante pour tout t_{ij} et tout p_i : $g(t_{ij}, p_i) = a,\ a \in \mathbb{R}$ ou encore proportionnelle à l'estimation effectuée : $g(t_{ij}, p_i) = bf(t_{ij}, p_i)., b \in \mathbb{R}$.

Après avoir défini la structure de notre modèle, nous présentons comment modéliser les paramètres. Le choix de la modélisation des paramètres passe par le choix d'une distribution pour chaque paramètre.
Soit p_i notre vecteur des paramètres de l'individu i, alors :

$$p_i = m(\mu, \beta, c_i) + \eta_i,$$

où

- μ est le vecteur des effets fixes. Il s'agit de la valeur qui caractérise la population. C'est autour de cette valeur que vont être estimés les valeurs de paramètres individuels.
- c_i désigne un ensemble de covariables pour l'individu i. Dans le cadre présent, il s'agit de covariables dites catégorielles qui servent à caractériser l'appartenance d'une souris à une lignée.
- β sert à quantifier l'impact de la covariable sur la valeur caractérisant la population par rapport à une population de référence caractérisée par le paramètre μ. Il sert donc à quantifier la variabilité entre différentes populations de souris dans notre cas.
- η_i est le vecteur des effets aléatoires. C'est ce vecteur aléatoire qui permet de décrire la variabilité inter-individuelles. Cet effet aléatoire est supposé distribué de la façon suivante :

  $$\eta_i \sim \mathcal{N}(0, \Omega^2),\ 1 \leq i \leq n.$$

  $\Omega^2 \in \mathbb{R}^{k \times k}$ désigne la matrice de variance covariance des effets aléatoires. Cette dernière est aussi à déterminer.

Tel que présenté ainsi, notre vecteur de paramètres suit une loi gaussienne de moyenne $m(\mu, \beta, c_i)$ et de variance Ω^2. Cependant, la distribution de nos paramètres n'est pas forcément gaussienne, nous sommes donc amenés à considérer une transformation de nos paramètres pour que ces derniers suivent une loi normale.

Dans notre cas, et pour des contraintes de positivité des paramètres qui traduisent une propriété biologique, nous avons choisi une distribution log-normale pour nos paramètres.

Une fois l'étape de modélisation terminée, nous passons maintenant au coeur du sujet, l'estimation des paramètres.

2.3.2 Estimation des paramètres

Nous devons estimer les paramètres suivants :

- le vecteur des effets fixes, appelé aussi paramètre de population μ,
- la matrice de variance covariance Ω servant à définir les effets aléatoires η,
- les paramètres β qui dépendent des covariables,
- les paramètres du modèle d'erreur, c'est-à-dire les paramètres qui caractérisent la fonction g.

Dans la suite nous désignerons par θ l'ensemble des paramètres de la population, autrement dit $\theta = (\mu, \beta, \Omega, g)$, $\theta \in \Theta$. Il permet de décrire le comportement moyen de nos populations desquelles sont issues nos observations. Une fois ce paramètre θ estimé, nous pourrons alors estimer les paramètres individuels à l'aide d'un algorithme stochastique. C'est ce vecteur de paramètres que l'on va chercher à estimer dans un premier temps à l'aide du logiciel Monolix [20].

Ce logiciel a été développé en 2003 par des équipes de statisticiens issus de l'INSERM en collaboration

avec INRIA. Ce logiciel utilise des modèles statistiques pour étudier la variabilité que l'on peut observer dans les données. L'algorithme à la base de l'estimation des paramètres de population est l'algorithme $SAEM$, il s'agit d'une version stochastique de l'algorithme EM qui est basée sur l'estimation du maximum de vraisemblance des observations.

Dans notre contexte, le vecteur de paramètre p_i est aussi appelé vecteur des données manquantes. En effet il s'agit d'une donnée pour laquelle nous ne disposons d'aucune information et qui permet de définir totalement le modèle (2).
Notons $L(y, p; \theta)$ la vraisemblance des données complètes et $d(y_{ij}, p_i, \theta)$ la valeur de la densité d pour chaque observation. La vraisemblance des données complètes est calculée de la façon suivante :

$$L(y, p; \theta) = \prod_{i=1}^{n} L(y_i, p_i; \theta) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} d(y_{ij}, p_i; \theta).$$

En outre,

$$d(y_i, p_i; \theta) = d(p_i \mid \theta) \cdot d(y_i \mid p_i; \theta).$$

Le premier terme représente le choix de la modélisation des paramètres et le second terme est défini par le choix du modèle concernant la structure de nos données effectué en (2).

En pratique, nous cherchons à maximiser la vraisemblance des données observées $L(y \mid \theta)$ définie en intégrant la vraisemblance des données complètes par rapport aux données manquantes :

$$d(y \mid \theta) \quad = \quad \int d(y, p; \theta)\, dp. \tag{3}$$

On cherche dans ce cas à résoudre le problème :

$$\hat{\theta} \quad = \quad \underset{\theta \in \Theta}{argmax} \quad d(y \mid \theta).$$

Dans le cas des modèles non linéaires à effets mixtes il est quasiment impossible d'avoir une expression analytique du problème (3) à maximiser. Il est parfois plus simple de chercher à maximiser la vraisemblance des données complètes. Plusieurs algorithmes sont utilisées dans le cadre d'estimation des données manquantes, c'est le cas des algorithmes EM (*Expectation Maximization*) et sa version stochastique SAEM *Stochastic Approximation of EM*, que nous allons maintenant présenter.

2.3.3 Algorithmes EM et SAEM

L'algorithme EM est un algorithme reposant sur l'estimation du maximum de vraisemblance de la log-vraisemblance des données complètes. Cependant, il est impossible d'estimer directement la valeur de la vraisemblance des données complètes $L(y, p; \theta)$. En effet, les paramètres p ne sont pas connus.

Une façon de contourner ce problème est de maximiser son espérance conditionnelle connaissant les données observées y et un paramètre de population $\tilde{\theta}$ fixé initialement. Nous regardons alors la quantité :

$$Q(\theta \mid \tilde{\theta}) \quad = \quad \mathbb{E}_p(\log d(y, p; \theta) \mid y; \tilde{\theta}). \tag{4}$$

Les relations (3) et (4) sont reliées comme suit : une croissance de la fonction Q entraîne une croissante de la vraisemblance des données observées.

Il s'agit du principe de fonctionnement de l'algorithme EM. On va chercher à augmenter la valeur de cette fonction Q par itérations successives. Lorsque la valeur de la vraisemblance n'augmente plus, on considère que la fonction a atteint son maximum et que nous avons atteint la solution du problème.

Pour pouvoir implémenter l'algorithme EM il est nécessaire d'avoir une expression explicite de la quantité à maximiser. Ce qui n'est en général pas le cas dans le cadre des modèles non linéaires à effets mixtes. L'algorithme $SAEM$ est une variante de l'algorithme EM permettant de contourner cette difficulté. Il s'agit de remplacer

l'étape *Expectation* de l'algorithme EM par une étape de simulation dans laquelle on va chercher à approcher la distribution des paramètres à estimer $d(\psi_i \mid y; \theta)$.

L'approximation de la distribution des paramètres non observés (p_i) se fait à l'aide de l'algorithme de *Metropolis-Hastings*. Afin d'approcher au mieux la loi conditionnelle de nos paramètres et d'obtenir une réalisation de notre vecteur de paramètres individuels p, nous employons des chaines de Markov permettant de converger vers la distribution recherchée. Pour employer cet algorithme nous devons choisir une distribution initiale, notée $q(. \mid .)$ de notre vecteur p. Notons $((\tilde{p})_m)_{m\in\mathbb{N}}$ et soit M un nombre qui représente le nombre de passage dans l'algorithme de Metropolis Hastings :

- *Initialisation* : posons $\tilde{p}_0 = p_j$, où j est l'indice représentant l'itération dans l'algorithme $SAEM$.
- *Simulation* : pour tout entier $m = 1, ..., M$ un tirage aléatoire de $\tilde{p}_m$ est effectué sous la distribution conditionnelle $q(. \mid .)$. On définit ensuite un seuil de probabilité α par :
$$\alpha = \frac{d(\tilde{p}_m \mid y; \theta_j) q(\tilde{p}_m \mid p_j)}{d(p_j \mid y; \theta_j) q(p_j \mid \tilde{p}_m)}$$
- *Actualisation* :
$p_{j+1} = \tilde{p}_M$ avec probabilité α,
$p_{j+1} = p_j$ avec probabilité $1 - \alpha$.

Nous présentons maintenant la procédure de l'algorithme SAEM [21] :

- *Initilisation* : on pose un vecteur de paramètre de population $\theta^{(0)}$
- *Simulation Step* : soit $k \geq 1$, on tire un vecteur $p_i^{(k)}$ de la distribution conditionnelle $d(p_i \mid y_i; \theta_{k-1})$ à l'aide de l'algorithme de Metropolis-Hasting.
- *Stochastic Approximation Step* : on pose :
$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k(log(d(y, p^{(k)}; \theta) - Q_{k-1}(\theta)).$$
$(\gamma_k)_{k\in\mathbb{N}}$ est une suite de nombres réels qui doit vérifier certaines hypothèses afin que l'algorithme converge.
- *Maximisation Step* : on maximise la quantité Q_k par rapport à θ pour obtenir un nouvel estimateur :
$$\theta^{(k+1)} = \underset{\theta\in\Theta}{argmax} \quad Q_k(\theta).$$

Une fois que les paramètres de population sont estimés on cherche à nouveau à estimer les paramètres (le vecteur de paramètres p_i) qui caractérisent chacun des individus. On cherche alors à approcher la distribution conditionnelle $d(p_i \mid y; \hat{\theta})$. Pour tout individu, la valeur du paramètres p_i est alors égale au mode de la distribution approchée à l'aide l'algorithme de Metropolis-Hastings.

Remarques : Lorsque la valeur du nombre d'individu est trop faible on simule plusieurs vecteurs de paramètres individuels et on effectue une approximation par méthode de Monte-Carlo. L'étape d'*approximation stochastique* devient alors :
$$Q_k(\theta) = Q_{k-1}(\theta) + \frac{1}{L}\sum_{l=1}^{L} \gamma_k(log(d(y, p^{(k,l)}; \theta) - Q_{k-1}(\theta))$$

Afin d'assurer la convergence de l'algorithme, la suite $(\gamma_k)_{k\in\mathbb{N}}$ doit vérifier :
$$\sum \gamma_k = \infty \quad et \quad \sum \gamma_k^2 < \infty.$$

Les hypothèses supplémentaires ainsi que la preuve de la convergence de l'algorithme se trouvent dans [22].

Maintenant que tous les outils nécessaires à l'estimation des paramètres sont présentés, nous pouvons définir des critères nous permettant de choisir notre modèle statistique et plus particulièrement le modèle d'erreur pour les estimations de paramètres ainsi que la façon dont nous allons introduire nos covariables dans les estimations.

2.4 Choix d'un modèle d'erreur et introduction de la covariable

2.4.1 Choix du modèle d'erreur

Pour le choix de notre modèle d'erreur plusieurs possibilités s'offrent à nous. J'ai choisi d'effectuer un choix parmi les modèles suivant qui sont les plus utilisés en pratique :

- un modèle constant : $g(t_{ij}, p_i, c_i) = a$, l'erreur d'estimation est la même pour tous les individus à chaque temps de mesure et indépendante de l'estimation ;
- un modèle proportionnel : $g(t_{ij}, p_i, c_i) = bf(t_{ij}, p_i, c_i)$, l'erreur est proportionnelle à l'estimation effectuée à l'aide de notre modèle ;
- un modèle combiné : $g(t_{ij}, p_i, c_i) = a + bf(t_{ij}, p_i, c_i)$, l'erreur résiduelle a une partie constante et est en partie proportionnelle à l'estimation effectuée.

Notre objectif est de trouver le modèle d'erreur qui, couplé à notre fonction f définie par le système (1), permettra d'expliquer au mieux les données. Pour discriminer l'un ou l'autre des modèles nous utiliserons les données d'une lignée de souris ainsi que les critères suivant :

1) Comparaison des valeurs AIC/BIC.

   Les critères AIC (*Akaike Information Criterion*) et BIC (*Bayesian Information Criterion*) sont des mesures de la qualité d'un modèle. Plus cette valeur est faible, meilleur est le modèle. Le logiciel Monolix emploie les formules suivantes pour le calcul des critères AIC/BIC :

   $$AIC = -2l(\theta; y) + 2k \qquad et \qquad BIC = -2l(\theta; y) + 2\log(n)k. \tag{5}$$

   où k désigne le nombre de paramètres à estimer.
   Ces critères permettent de comparer deux modèles entre eux. On préfèrera choisir un modèle dont la valeur de ces critères statistiques est minimale. Contrairement à l'AIC, la valeur du critère BIC dépend également du nombre d'observations.

2) Calcul de l'écart quadratique entre les données observées et les données estimées.

   Nous estimons les paramètres du modèle à l'aide des données de la lignée considérée. Nous conservons le vecteur du paramètre de population μ qui décrit le comportement moyen de la lignée. Nous pouvons alors estimer le nombre de cellules activées, noté $\tilde{z_j}$, $1 \leq j \leq 5$, à chaque temps où nous avions des observations, c'est-à-dire aux jours 0, 8,15, 34 et 55 en résolvant le système différentiel (1). Avec ces mêmes données, notées y_{ij}, nous calculons la moyenne z_j à chacun des jours mentionnés précédemment et l'écart-type σ_j.
   L'écart quadratique total ξ est alors la somme des écarts quadratiques à chaque temps, autrement dit :

   $$\xi \quad = \quad \sum_{j=1}^{5} \frac{\bar{y_{.j}} - z_j}{\sigma_j}, \tag{6}$$

   où $\bar{y_{.j}}$ représente la moyenne des observations au temps j.

Nous retiendrons le modèle qui minimise ces deux critères et nous pourrons ainsi estimer les valeurs des paramètres pour les différentes lignées de souris.

Il pourrait également être intéressant de considérer la totalité des données pour l'estimation des paramètres en espérant ainsi améliorer la précision des estimations. Cependant, nous avons vu en Figure 4 que nos données varient fortement d'une lignée à une autre c'est pourquoi il est nécessaire de tenir compte de la lignée de nos données, appelée covariable dans notre contexte.

2.4.2 Introduction de la covariable dans le modèle

La prise en compte de la covariable dans l'estimation des paramètres entraîne l'estimation d'un paramètre supplémentaire qui mesure l'effet de la covariable sur le paramètre. On voit donc que l'approche naïve consistant à dire que tous nos paramètres dépendent de la covariable nous fait perdre le bénéfice de considérer un nombre plus important de données.

La procédure ci-dessous et schématisée en Figure 6 présente la façon dont est introduite la covariable dans le modèle.

- Etape initiale :
  Nous étudierons les trois lignées ensemble. Dans un premier temps nous ne ferons pas dépendre les paramètres de la covariable mais nous observerons les valeurs des paramètres individuels et de populations d'une lignée de souris à une autre. Nous retiendrons les paramètres qui ont une variabilité significative d'une lignée de souris à une autre. Cette appréciation est subjective mais elle permet d'effectuer un premier tri au niveau des paramètres et ainsi d'éviter de lancer un grand nombre de simulations lorsque le nombre de paramètres du modèle est important. Cela nous donnera également une valeur de référence de la log-vraisemblance.

- Etape Forward :
  Pour $k \geq 0$, nous désignons par H_0 l'hypothèse sous laquelle, dans notre modèle, k paramètres dépendent de la covariable et H_1 : l'hypothèse sous laquelle, un paramètre en plus dépend de la covariable. Ce test permet de comparer deux modèle présentant un nombre différent de paramètres. Ainsi un modèle est donc préférable à un autre si la différence des valeurs de la log-vraisemblance est suffisamment grande. Ce seuil à partir duquel la différence devient significative dépend de la différence du nombre de paramètre dans les deux modèles.
  On note $\log(\mathcal{L}(\hat{\theta_0}))$ la valeur de la log-vraisemblance sous l'hypothèse H_0. La statistique de test employée pour effectuer ce test est :

  $$T \quad = \quad 2\log(\mathcal{L}(\hat{\theta_0})) - 2\log(\mathcal{L}(\hat{\theta})).$$

  Sous l'hypothèse H_0 la loi de T peut-être approchée par une loi du χ^2 à un degré de liberté ($k+1-k$). On se fixe un seuil α à 5% (qui correspond à un risque de première espèce : rejeter l'hypothèse H_0 à tort) ce qui correspond à une valeur critique de $\chi^2_\alpha(1) = 3.84$. Si la valeur t de la statistique de test T est plus grande que 3.84, on rejette l'hypothèse H_0.
  Cette étape s'arrête dès lors que l'ajout d'un paramètre supplémentaire n'implique plus de changement significatif au niveau de la log-vraisemblance.

- Etape Backward :
  Nous reprenons le procédé précédent mais dans le sens inverse. On évalue la valeur de la vraisemblance de notre modèle en retirant à chaque fois le paramètre le moins significatif, c'est-à-dire celui qui n'entraîne pas de changement significatif au niveau de la valeur de la log-vraisemblance. Nous continuons jusqu'à ce la log-vraisemblance du modèle évolue de façon significative.

L'introduction de la covariable dans le modèle va également nous permettre de savoir quels sont les paramètres qui dépendent significativement de la lignée.

Nous avons maintenant toutes les estimations nécessaires à la définition et à l'étude d'une distance entre les différentes souris et nos trois lignées.

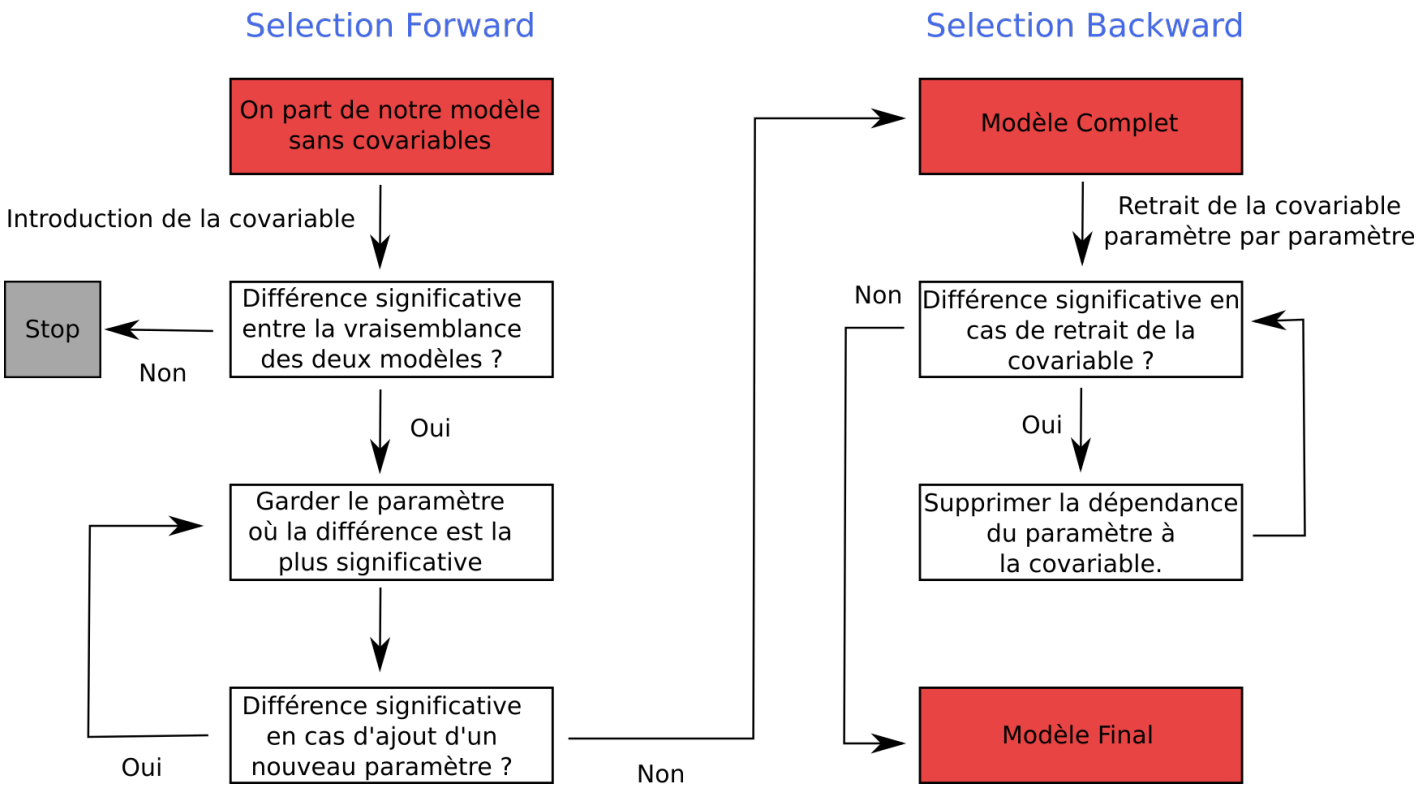


FIGURE 6 – Schéma illustrant la procédure employée pour l'introduction de la covariable dans le modèle.

2.5 Etude de la variabilité et définition d'une distance

Une fois les valeurs des paramètres estimées pour les individus des différentes lignées, nous voudrions savoir quelles informations nous pouvons en tirer. Nous pouvons regarder la variabilité relative des valeurs de chacun des paramètres pour comparer quels sont les plus variables et ainsi définir une distance qui permettrait d'expliquer les différences observées au niveau de la réponse immunitaire T CD8.

2.5.1 A l'échelle des individus

Après estimation des valeurs de paramètres, nous disposons d'un nombre M (nombre de paramètres) d'échantillons de taille n (nombre d'individus). Nous voulons déterminer, parmi ces M échantillons, quels sont ceux qui présentent la plus grande variabilité. Le problème est que nos échantillons n'ont pas tous la même échelle, il ne suffit donc pas de calculer la variance de l'échantillon. Afin de palier à ce problème d'échelle, nous utilisons le coefficient de variation défini par :

$$V(p^{(m)}) = \frac{sd(p^{(m)})}{\frac{1}{n}\sum_{i=1}^{n} p_i^{(m)}},$$

où sd désigne l'écart-type de notre échantillon et $p^{(m)}$ est le vecteur des valeurs du $m - ième$ paramètre. Cette quantité permet de comparer la variance de plusieurs échantillons dont les valeurs n'ont pas le même ordre de grandeur. On s'en sert pour affecter un poids $w^{(m)}$ au paramètre m,

$$w^{(m)} = \frac{V(p^{(m)})}{\sum_{k=1}^{M} V(p^{(m)})}.$$

Nous supposons ensuite que les paramètres qui sont les plus variables sont ceux qui expliquent le plus la distance entre les observations des différents individus, en terme de nombre de cellules activées au cours du temps.

Considérons p l'ensemble des valeurs des paramètres des individus étudiés. Nous définissons la distance D_{ij} entre les individus i et j par :

$$D_{ij} = \sum_{m=1}^{M} w^{(m)} \frac{\mid p_i^{(m)} - p_j^{(m)} \mid}{\frac{1}{n}\sum_{k=1}^{n} p_k^{(m)}}. \quad (7)$$

La relation (7) vérifie les axiomes d'une distance. Nous récupérons ensuite une matrice de distance symétrique de taille $n \times n$. Nous projetons ensuite cette matrice sur un plan deux dimensions à l'aide d'une méthode de *multidimensional scaling* [23]. La plus connue d'entre elle est en fait l'analyse en composante principale, ce que nous aurions pu effectuer ici à l'aide de notre matrice des valeurs des paramètres.

La méthode employée ici est la suivante. Nous cherchons à déterminer les coordonnées C_{xy}^i de nos individus, à l'aide de notre matrice de distance D définie par la formule (7), qui minimise l'erreur suivante :

$$Err = \sum_{i<j} (\left\| C_{xy}^i - C_{xy}^j \right\| - D_{ij})^2.$$

La représentation de la distance entre les individus dans le plan se fait à l'aide de la fonction "cmdscale" du logiciel R.

2.5.2 A l'échelle de la lignée

Nous voudrions maintenant faire de même mais à l'échelle des lignées et définir une distance entre les les différentes lignées.

Pour ce faire nous n'utiliserons plus les valeurs des paramètres estimées pour les individus mais les distributions qui caractérisent les paramètres. Afin d'évaluer la distance entre les lignées, nous employons la distance de Kantorovich [24] qui sert à calculer une distance entre distributions. Cependant, dans ce qui suit, nous utiliserons une version discrète de cette distance entre distributions. Le principe de calcul de la distance entre deux distributions est le suivant : considérons deux distributions de probabilité f_1 et f_2. On considère ensuite une partition de l'intervalle de définition de ces deux distributions que l'on note $[x_1, ..., x_L]$. La distance de Kantorovich entre les deux distributions, notées $K(f_1, f_2)$ est définie par :

$$K(f_1, f_2) = (x_L - x_1)^2 \sum_{l=1}^{L} (\mid f_1(x_l) - f_2(x_l) \mid). \quad (8)$$

Avec cette même définition on pourra également identifier les paramètres qui expliquent le plus la distance entre les différentes lignées. A nouveau, et pour éviter les problèmes dus à une différence d'échelle, nous diviserons les valeurs de la densité par sa valeur moyenne pour pouvoir comparer les distributions des différents paramètres entres elles.

Maintenant que nous avons défini une distance entre nos différents individus à l'aide des paramètres estimés, il serait également intéressant d'étudier l'influence de chacun de ses paramètres sur la variabilité de la dynamique de la réponse.

2.6 Analyse de sensibilité

Dans cette partie, notre objectif est double. Nous souhaitons mettre en évidence le rôle des différents paramètres aux différentes étapes de la réponse immunitaire, à savoir lors de la phase d'expansion, de la phase de contraction et à la fin de la réponse immunitaire. Nous voulons également simplifier notre modèle en fixant les paramètres qui ont peu d'influence sur la dynamique des cellules activées. Pour ce faire, nous nous appuyons sur l'analyse de sensibilité présentée en [25]. Elle s'appuie sur le calcul des indices de Sobol des différents paramètres dont nous présentons la méthode.

Dans cette section, nous désignons par Y la sortie de notre modèle, à savoir le nombre de cellules activées au cours du temps, par P_i, $1 \leq i \leq M$, les différents paramètres du modèle et h une fonction intégrable qui correspond à notre modèle, ainsi :

$$Y = f(P_1, ..., P_M).$$

Les variables P_i sont supposées deux à deux indépendantes et distribuées uniformément. Nous cherchons ensuite à déterminer de combien la variance de Y décroît lorsque que l'on fixe l'un des paramètres P_i, i.e on s'intéresse à la quantité : $V(Y \mid P_i = p_i)$. où V désigne la variance. Afin de contourner le choix de la valeur de p_i, on s'intéresse plutôt à la quantité $\mathbb{E}[V(Y \mid P_i = p_i)]$, ainsi plus notre paramètre a une importance sur la variance de notre sortie Y plus cette quantité sera petite. Cependant, pour le calcul des indices de Sobol, nous serons amenés à calculer la quantité $V(\mathbb{E}[Y \mid P_i])$. Par la formule de la variance totale $V(Y) = V(\mathbb{E}[Y \mid P_i]) + \mathbb{E}[V(Y \mid P_i)]$ nous voyons que plus le paramètre P_i a de l'importance sur la variabilité de Y plus la quantité $V(\mathbb{E}[Y \mid P_i])$ sera grande. On peut ainsi définir l'indice de sensibilité (d'ordre 1) du paramètre i par la relation suivante :

$$S_i = \frac{V(\mathbb{E}[Y \mid P_i])}{V(Y)} \in [0, 1].$$

Comme $V(\mathbb{E}[Y \mid P_i]) = \mathbb{E}[\mathbb{E}[Y \mid P_i]^2] - \mathbb{E}[Y]^2$ par définition de la variance, nous allons estimer la quantité

$$U_i = \mathbb{E}[\mathbb{E}[Y \mid P_i]^2]$$

à l'aide de la méthode de Sobol qui repose sur une estimation par méthode de Monte-Carlo.

Pour ce faire nous générons deux échantillons $p^{(1)}$ et $p^{(2)}$ de taille N de valeurs de paramètres, $p^{(1)}$ et $p^{(2)}$ sont des matrices de taille $N \times M$. On calcule alors U_i à l'aide de la relation suivante :

$$U_i \simeq \frac{1}{N} \sum_{k=1}^{N} f(p_{k1}^{(1)}, ..., p_{k(i-1)}^{(1)}, p_{ki}^{(1)}, p_{k(i+1)}^{(1)}, ..., p_{kM}^{(1)}) f(p_{k1}^{(2)}, ...p_{k(i-1)}^{(2)}, p_{ki}^{(1)}, p_{k(i+1)}^{(2)}, ..., p_{kM}^{(2)}).$$

La moyenne et la variance de Y sont également calculés par une méthode de Monte-Carlo :

$$\mathbb{E}[Y] \simeq \frac{1}{N} \sum_{k=1}^{N} f(p_{k1}, ..., p_{kM}),$$

$$V(Y) \simeq \frac{1}{N} \sum_{k=1}^{N} f(p_{k1}, ..., p_{kM})^2 - \mathbb{E}[Y]^2.$$

Cependant, les indices de sensibilité de premier ordre ne suffisent pas pour analyser la variabilité de la sortie de notre modèle. Pour compléter cette étude, nous calculerons également les indices de sensibilité d'ordre total $U_{\sim i}$. Il est défini comme la somme de tous les indices de sensibilité relatifs au paramètre P_i. Nous avons donc $U_i \leq U_{\sim i}$. Ce dernier est calculé de la façon suivante :

$$U_{\sim i} \simeq \frac{1}{N} \sum_{k=1}^{N} f(p_{k1}^{(1)}, ..., p_{k(i-1)}^{(1)}, p_{ki}^{(1)}, p_{k(i+1)}^{(1)}, ..., p_{kM}^{(1)}) f(p_{k1}^{(1)}, ...p_{k(i-1)}^{(1)}, p_{ki}^{(2)}, p_{k(i+1)}^{(1)}, ..., p_{kM}^{(1)}).$$

Cet indice prend en compte les effets de du paramètre étudié sur la sortie du modèle tout en tenant compte des effets avec tous les autres paramètres. Il se peut que certains paramètres, à eux seuls, n'aient aucune influence sur la sortie du modèle, mais combinés avec d'autres paramètres, il peut expliquer une grande partie de la variabilité du modèle.

	Modèle constant	Modèle proportionnel	Modèle combiné
AIC	1382	393	391
BIC	1391	401	400

Table 1 – Comparaison des critères statistiques AIC/BIC pour les trois modèles d'erreurs.

	Modèle proportionnel	Modèle combiné
ξ	0.144	44.3

Table 2 – Erreur quadratique pour le modèle proportionnel et le modèle combiné (voir Figure 7).

3 Résultats et simulations

3.1 Choix du modèle statistique

Avant d'effectuer les estimations de paramètres pour chaque lignée, nous devons d'abord déterminer le modèle d'erreur le plus approprié à l'aide des critères présentés en Section 2.4.1. Pour le choisir, nous considérons les données relatives à la lignée C57Bl/6 puis nous estimons les paramètres p_i relatifs à ces données. Parallèlement à l'estimation des paramètres, le logiciel Monolix calcule également la valeur de la log-vraisemblance et donc des critères AIC/BIC selon la relation (5).

Nous lançons trois estimations différentes, une pour chacun des modèles d'erreur. Les valeurs des critères AIC et BIC obtenus sont présentées dans la Table 1 et permettent d'écarter le modèle constant. En effet les valeurs des critères statistiques de ce dernier sont beaucoup plus élevées que dans le cas d'un modèle proportionnel ou combiné.

Ce critère là ne permet cependant pas de discriminer le modèle combiné ou le modèle proportionnel pour lesquels les valeurs des critères sont similaires. Nous allons donc calculer l'erreur quadratique ξ pour les deux modèles à l'aide des estimations et de la formule (6).

Nous présentons les fits de la population pour les deux modèles d'erreur en Figure 7. Sur cette figure, on observe que l'on approche beaucoup mieux les données à l'aide d'un modèle d'erreur proportionnelle qu'avec un modèle d'erreur combinée. En effet si le premier modèle estime très bien les données, avec un modèle d'erreur combinée, le nombre de cellules activées est surestimé à la fin de la réponse. Les résultats sont présentés dans la Table 2. Avec une erreur quadratique égale à 0.144, le modèle d'erreur proportionnelle est alors celui que nous retiendrons pour estimer les paramètres des différentes lignées.

A partir de maintenant et jusqu'à la fin du rapport, le modèle statistique employé pour l'estimation des paramètres sera :

$$y_{ij} = f(t_i j, p_i) \cdot (1 + b \cdot \varepsilon_{ij}), \qquad \varepsilon_{ij} \sim \mathcal{N}(0, 1).$$

La constante b de ce modèle est choisie et égale à 0.3 qui est la valeur utilisée et estimée par Monolix pour le choix du modèle d'erreur proportionnelle.

3.2 Estimation pour chacune des lignées et distance entre les individus

Nous estimons maintenant, lignée par lignée, les valeurs des paramètres qui la caractérisent ainsi que les valeurs pour chaque individu. Les valeurs initiales de l'algorithme ainsi que les résultats sont présentés dans la Table 3. Les valeurs initiales de l'algorithme sont celles estimées par Fabien Crauste lors de précédents travaux et sur un jeu de données différent. Nous rappelons que tous nos paramètres suivent une loi-log-normale. Le fit des données obtenues pour chaque lignée à l'aide des estimations des paramètres est représenté en Figure 8. Les valeurs estimées des paramètres pour les différentes lignées sont présentées dans la Table 3.

Les valeurs de paramètres estimées permettent d'approcher de façon plus que satisfaisante le comportement moyen des différentes lignées. En effet, la dynamique estimée, présentée en Figure 8 approche de façon

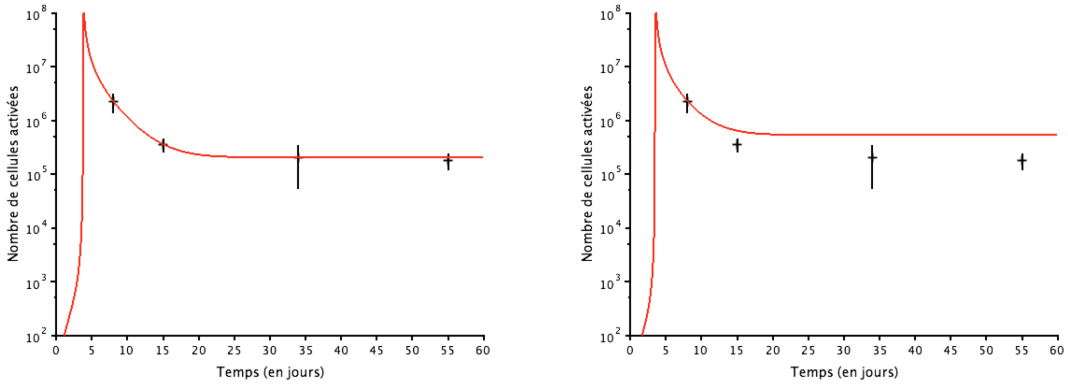


FIGURE 7 – Dynamique des cellules activées au cours du temps. Le nombre de cellules activées est représenté en échelle logarithmique. La courbe en rouge représente le nombre de cellules activées au cours du temps. Cette courbe est obtenue en estimant les valeurs de paramètres à l'aide du modèle d'erreur proportionnelle (à gauche) et d'un modèle d'erreur combinée (à droite). Les croix noires représentent la valeur moyenne du nombre total de cellules activées plus ou moins l'écart-type.

précise les valeurs des données, pour les trois lignées.

A l'aide de la Table 3, on peut voir que les différences observées au niveau du nombre de cellules T CD8 activées aux différentes phases de la réponse se retrouvent au niveau des valeurs des paramètres. Prenons l'exemple du nombre de cellules activées au pic de la réponse : nous constatons en Figure 8.A que ce nombre est beaucoup plus important pour la lignée C57Bl/6 que pour les lignées BalbC et OF1. Cela se traduit par un paramètre de prolifération des cellules early effectrices ρ_E ainsi que par des paramètres de mort μ_{EE} et μ_{PE} plus faibles pour la lignée C57Bl/6 que pour les deux autres lignées.

Paramètre	Initialisation	Valeurs estimées		
		C57Bl/6	BalbC	OF1
μ_N	0.72	0.222	0.527	0.658
δ_{NE}	0.022	0.00729	0.00347	0.00235
ρ_E	0.73	0.524	0.778	0.754
μ_{EE}	$5.3.10^{-8}$	$5.05.10^{-8}$	$1.26.10^{-7}$	$1.99.10^{-7}$
μ_{EL}	$3.7.10^{-7}$	$6.09.10^{-8}$	$3.76.10^{-7}$	$5.57.10^{-7}$
δ_{EL}	0.55	0.315	0.0947	0.215
δ_{EM}	0.0001	0.000358	0.000379	0.000433
μ_{LL}	$6.8.10^{-5}$	$5.37.10^{-5}$	$2.95.10^{-5}$	6.10^{-5}
μ_{LE}	$6.8.10^{-5}$	$6.23.10^{-6}$	$2.92.10^{-6}$	$1.65.10^{-5}$
δ_{LM}	0.02	0.308	0.0332	0.111
ρ_P	0.128	0.275	0.272	0.27
μ_{PE}	$1.6.10^{-5}$	$3.59.10^{-6}$	$5.62.10^{-5}$	$2.03.10^{-5}$
μ_{PL}	$1.6.10^{-5}$	$7.06.10^{-6}$	$1.24.10^{-5}$	$1.75.10^{-5}$
μ_P^0	0.014	0.0187	0.0059	0.0125

TABLE 3 – Estimation des valeurs des paramètres de population qui caractérisent les différentes lignées. Les valeurs présentées dans la colonne initialisation correspondent à la valeur de l'effet fixe, la valeur de l'effet aléatoire, nécessaire à l'initialisation de l'algorithme, est fixée à 0.1 pour l'ensemble des paramètres. Nous ne donnons pas les valeurs estimées des effets aléatoires pour les différents paramètres, cette information n'étant pas directement utilisée dans ce rapport.

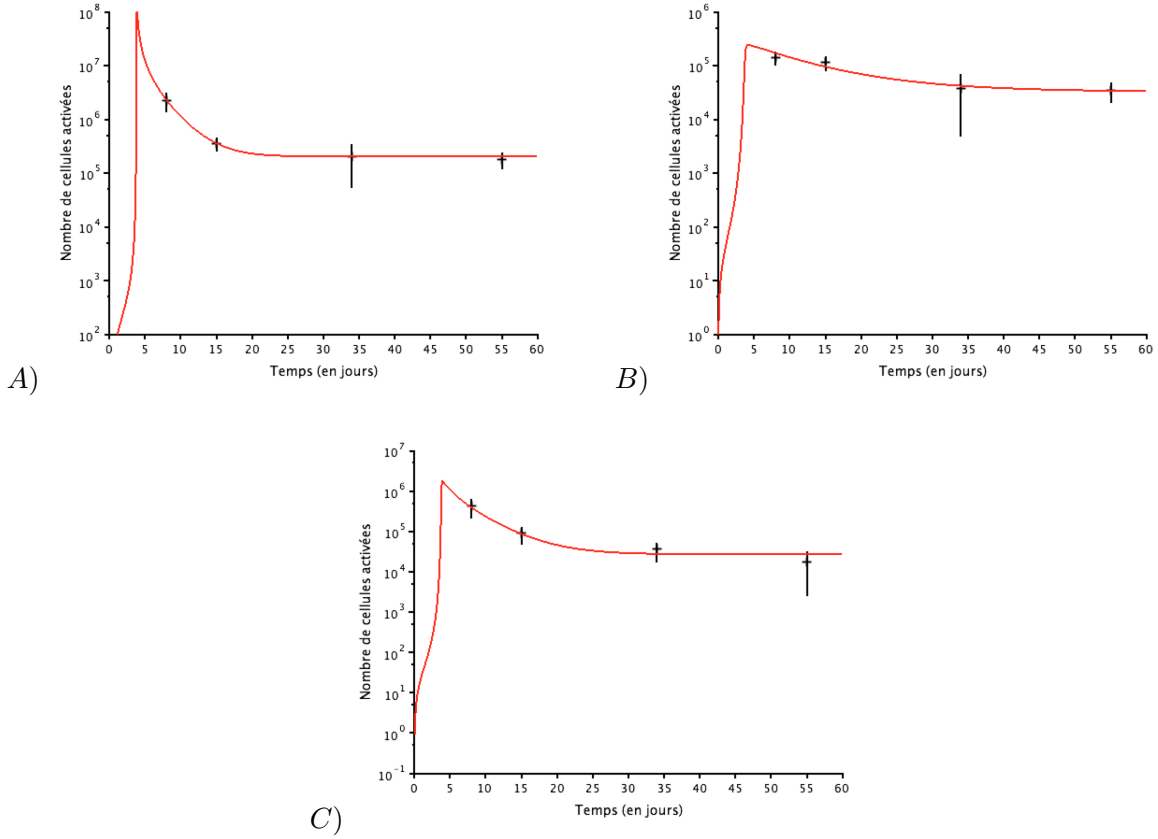


FIGURE 8 – Représentation de la dynamique des cellules activées obtenues à l'aide des valeurs de paramètres de population, présentées dans la Table 3, pour les différentes lignées de souris au cours du temps. La courbe rouge représente la dynamique estimée pour la lignée C57Bl/6 en A), la lignée BalbC en B) et la lignée OF1 en C). Les croix noires représentent la valeur moyenne du nombre total de cellules activées plus ou moins l'écart-type. Les valeurs de paramètres estimées permettent d'approcher les données de façon très satisfaisante pour les différentes lignées. On note également que le nombre de cellules activées estimé est plus important chez la lignée C57Bl/6, ce que nous observions déjà à travers les données. La phase de contraction est également plus marquée chez cette même lignée, comparée aux deux autres.

Les paramètres ρ_E et μ_{PE} étant plus faibles que pour les deux autres lignées, cela permet au pathogène de survivre plus longtemps dans l'organisme au tout début de la réponse et de proliférer davantage, cela se traduit par une quantité de pathogène plus importante au moment du pic de la réponse (voir Figure 15 en Annexe). L'augmentation de la quantité de pathogène au début de la réponse va ainsi entraîner une croissance plus rapide et plus importante du nombre de cellules early effectrices pendant la phase d'expansion cellulaire. Estimer le paramètre μ_{EE} plus faiblement pour cette lignée permet de maintenir un nombre important de cellules early effectrices plus longtemps.

Si les données sont bien reproduites à l'échelle des lignées, cela est également le cas à l'échelle individuelle, comme le montre la Figure 9 où nous avons choisi deux souris issues de la lignée BalbC, pour l'un des individus nous ne possédions que des données pour trois temps différents, Figure 9.A), et quatre temps pour le deuxième, Figure 9.B).

Maintenant que nous avons toutes les informations relatives à chaque individu de chaque lignée, nous allons pouvoir étudier la variabilité de chacun des paramètres et utiliser la distance introduite en Section 2.4.1 pour étudier la distance entre les différents individus mais également entre les lignées. Nous chercherons, dans un premier temps, à déterminer quels sont les paramètres les plus variables et quels sont ceux qui expliquent le

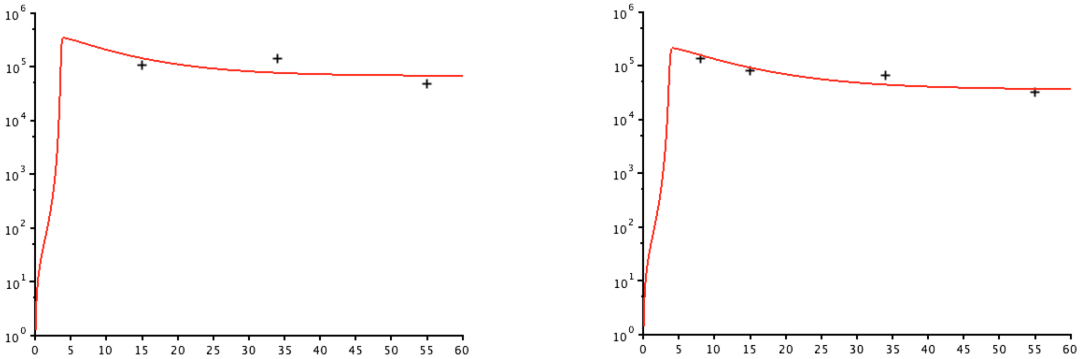


FIGURE 9 – Représentation de la dynamique des cellules activées au cours du temps pour deux individus de la lignée BalbC. Les courbes rouges représentent les dynamiques des cellules activées obtenues à l'aide des valeurs des paramètres individuels. Les données relatives aux individus sont représentées par les croix noires.

plus la distance entre deux lignées.

La Figure 10 est une projection de la matrice de distance entre les individus. Pour l'obtenir nous avons estimé les paramètres de l'ensemble des individus sans prendre en compte l'étiquette des données, c'est-à-dire l'appartenance à une lignée donnée. Nous pouvons voir qu'une lignée semble clairement se démarquer des deux autres, à savoir la lignée C57Bl/6. Par contre, la distance utilisée ne permet pas de faire la distinction entre les lignées BalbC et OF1 où les nuages de points restent mélangés. Nous avons déjà pu constater cela à travers la représentation des données en Figure 4 où la dynamique des lignées BalbC et OF1 sont très semblables en terme de nombre de cellules activées.

Cependant nous remarquons qu'un individu de la lignée BalbC, l'individu 20 (voir Figure 4) se rapproche des individus de la lignée C57Bl/6. Cela peut s'expliquer par le fait que pour cet individu, nous ne possédons pas d'information quant au nombre de cellules activées à jour 8. De plus, le nombre de cellules activées au jour 34 pour cet individu est proche du nombre de cellules activée des souris de la lignée C57Bl/6.

Nous avons également tenté d'autres méthodes de classification afin de savoir s'il était possible de classer les individus dans les différentes lignées en fonction des valeurs des paramètres. Nous avons essayé à l'aide d'un algorithme de type K-means, mais les résultats n'étaient pas satisfaisants. En effet si l'on demande à l'algorithme d'effectuer 3 classes, il est incapable de trier correctement les individus. Si on lui demande de faire deux classes distinctes, il parvient à séparer les souris C57Bl/6 des deux autres. En retirant les individus de la lignée C57Bl/6, l'algorithme n'est tout de même pas capable de distinguer la lignée OF1 des BalbC. L'approche développée ici a permis de trouver les informations contenues dans les données, à savoir la proximité des lignées BalbC et OF1 et le détachement de la lignée C57Bl/6.

Quittons maintenant l'échelle des individus et intéressons-nous aux valeurs caractéristiques des différentes lignées. L'estimation des paramètres des différentes lignées (appelés paramètres de population) nous donne la possibilité de connaître les paramètres caractérisants les distributions des paramètres des différentes lignées. Pour déterminer la distance entre les différentes lignées de souris nous calculons alors les différentes distance entre les distributions de valeurs de paramètres à l'aide de la distance de Kantorovich (voir Section 2.5.2).

Dans un premier temps, nous regardons ce qui se passe paramètre par paramètre. Ainsi, pour un paramètre donné, nous calculons la distance entre les différentes distributions et nous attribuons la valeur 1 pour la distance la plus grande et 0 pour la distance la plus petite. Un point intermédiaire représente simplement la distance relative par rapport aux distances extrémales. L'objectif étant de déterminer quels sont les paramètres qui expliquent le rapprochement ou l'éloignement entre deux lignées.

On pourrait s'attendre à ce que la distance soit minimale entre les lignées OF1 et BalbC pour chacun des paramètres étant donnée la proximité observée au niveau des données. Or cela n'est pas le cas, comme le montre la Figure 11 seulement 8 paramètres sur les 14 que comporte le modèle montrent que les lignées OF1 et BalbC sont

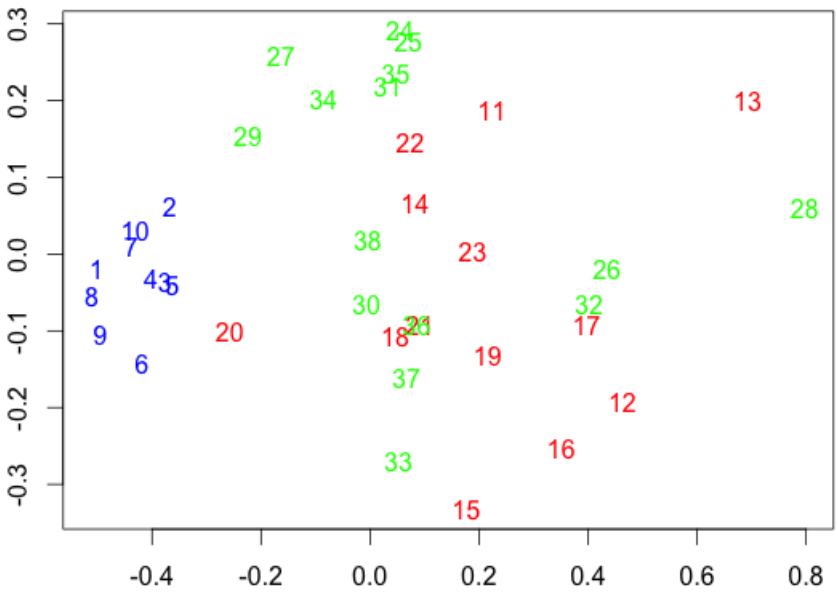


FIGURE 10 – Représentation des individus dans le plan euclidien. Cette représentation est obtenue par *multidimensional scaling*. En abscisse et en ordonnée sont représentées les coordonnées des différents individus dans le plan. Ces coordonnées sont obtenues à l'aide de la méthode présentée dans [23] et de la matrice de distance D définie par la formule (7). Les couleurs sont utilisées pour identifier la lignée à laquelle appartient un individu : C57Bl/6, BalbC, OF1. Les nombres 1 à 38 représentent les différents individus. La lignée C57BL/6 (en bleu) se détache des deux autres lignées et constitue un premier groupe. Cependant les individus de la lignée BalbC et OF1 sont confondus dans un même nuage de points. Seul l'indivdu 20 se détache de ce nuage et s'approche des individus C57Bl/6. En effet nous avons pu voir (Figure 4) que le nombre de cellules activées au jour 34, pour cet individu, est proche de celui observé pour les souris C57Bl/6, en outre nous ne disposons pas d'informations à jour 8.

proches et éloignées de la lignée C57Bl/6. Si l'on compare maintenant ces résultats à ceux présentés sur la Figure 10, on peut voir que certains individus des lignées BalbC et OF1 sont très éloignés les uns des autres, ce qui peut expliquer pourquoi la distance est parfois maximale entre ces deux lignées pour certains paramètres du modèle. Nous présentons la distance entre les différentes lignées dans la Table 4, les valeurs illustrent bien la remarque effectuée précédemment.

Si la plupart des paramètres permettent de justifier la distance entre deux lignées étudiées, il serait intéressant de voir quels sont ceux dont la valeur moyenne dépend spécifiquement des lignées et qui permettent, de façon significative, de justifier un tel écart entre les lignées. Nous allons donc introduire la covariable dans notre modèle. Cette dernière va nous permettre d'effectuer des estimation de paramètres pour l'ensemble de nos données étiquetées. Nous serons ainsi capable de recalculer une distance entre les individus et les lignées.

3.3 Prise en compte de la covariable

Pour introduire la covariable dans le modèle, nous appliquons la procédure présentée dans la Figure 6 en Section 2.4.2. Nous avons tout d'abord lancé une simulation en n'affectant la covariable à aucun des paramètres

	BalbC-OF1	C57Bl/6-OF1	C57Bl/6-BalbC
Distance	0.232	0.410	0.549

TABLE 4 – Distance entre les différentes lignées. Cette distance est obtenue en sommant les différentes distance de Kantorovich entre les deux lignées considérées. Comme pour l'étude de la variabilité entre différents échantillons, nous avons normalisé chaque distributions en divisant par la valeur moyenne.

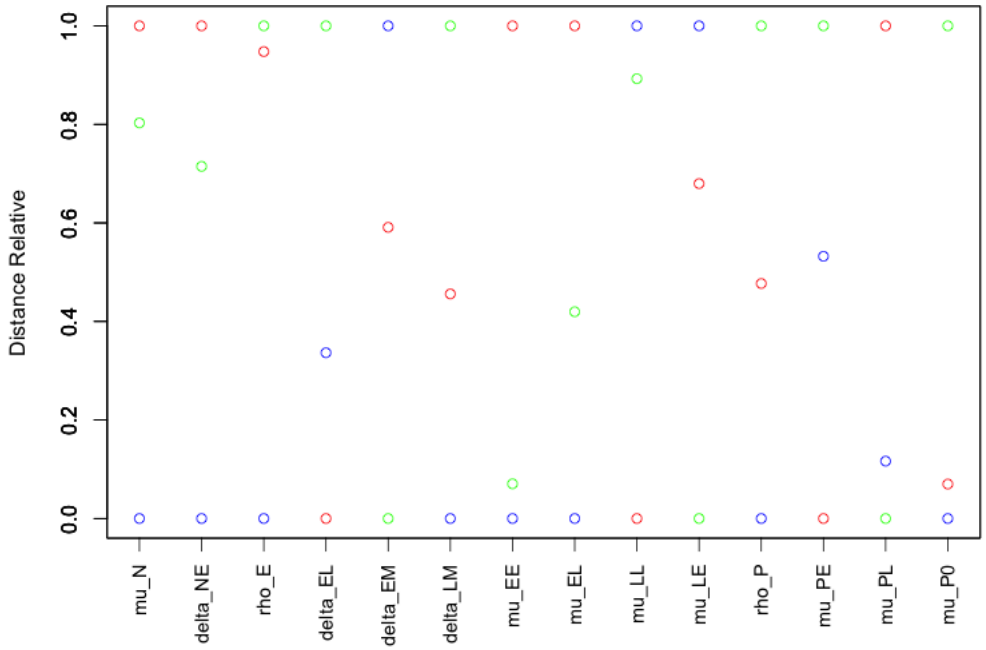


FIGURE 11 – Distance relative entre les lignées pour chaque paramètre. Le calcul de la distance, pour un paramètre donné, s'effectue à l'aide de la formule (8). Un point représente la distance relative entre deux lignées données : C57Bl/6-BalbC en vert, C57Bl/6-OF1 en rouge et BalbC-OF1 en bleu. Un point se situant sur la droite d'équation $y = 0$ indique, pour le paramètre en question, que la distance entre les deux lignées est la plus petite. La distance entre les deux lignées est la plus grande lorsque, pour le paramètre donné, le point se trouve sur la droite $y = 1$. Une position intermédiaire indique une distance relative par rapport aux distances minimale et maximale. On remarque que pour 8 paramètres sur les 14 que comportent le système, la distance entre les lignées OF1 et BalbC est la plus petite. Cette proximité en terme de distance entre distributions de valeurs de paramètres est à mettre en lien avec la proximité observée au niveau des données (cf. Figure 4).

Paramètres dont nous avons testés la dépendance à la covariable	Paramètres retenus comme dépendants de la covariable
$\rho_E, \delta_{EL}, \delta_{EM}, \mu_{EE}, \mu_{EL}$ $\mu_{LE}, \mu_{PE}, \mu_{PL}$	ρ_E, δ_{EL} μ_{EE}, μ_{PL}

TABLE 5 – Résultats de la procédure permettant de choisir quels sont les paramètres dépendant de la covariable.

du modèle puis conservé les paramètres qui variaient le plus d'un individu à l'autre. Les candidats potentiels ainsi que les paramètres retenus par la procédure sont présentés dans la Table 5.

Nous remarquons maintenant que seuls quatre paramètres dépendent de façon significative de la covariable : les paramètres $\rho_E, \delta_{EL}, \mu_{EE}$ et μ_{PL}. Nous discuterons de la signification de ces quatre paramètres dans

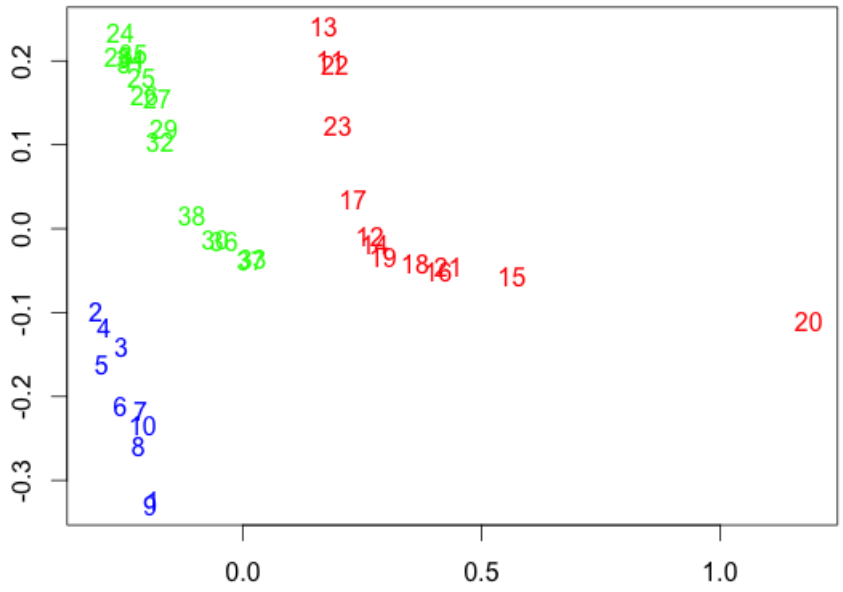


FIGURE 12 – Représentation des individus dans le plan euclidien en tenant compte de la covariable. La légende est la même que celle employée en Figure 10.

la partie discussion ainsi que d'éventuelles hypothèses que l'on pourra formuler par la suite.

Nous pouvons maintenant estimer les nouvelles valeurs des paramètres du modèle en prenant en compte l'effet de la covariable sur les paramètres susmentionnés. Nous n'affichons pas les nouvelles valeurs des paramètres estimées mais nous regardons directement la distance qui sépare les individus des différentes lignées en tenant compte de la covariable. Une représentation dans le plan de cette distance entre les individus est présentée en Figure 12. Si l'on compare cette figure à la Figure 10 on remarque que l'on arrive maintenant à faire la distinction entre les lignées BalbC et OF1. Nous calculons recalculons la distance entre les lignées à partir de ces nouvelles distributions, en tenant compte uniquement des distributions de paramètres qui dépendent de la covariable, cette distance est présentée dans la Table 6. En tenant compte de la covariable, on remarque que la la lignée OF1 est à mi-distance des lignées C57Bl/6 et BalbC. Ce résultat est à mettre en parallèle avec la Figure 12. On peut effectuer la même remarque en regardant le nombre de cellules activées mesurées au sein des souris à jour 8 (voir Figure 4). Nous développerons ce dernier point dans la section suivante.

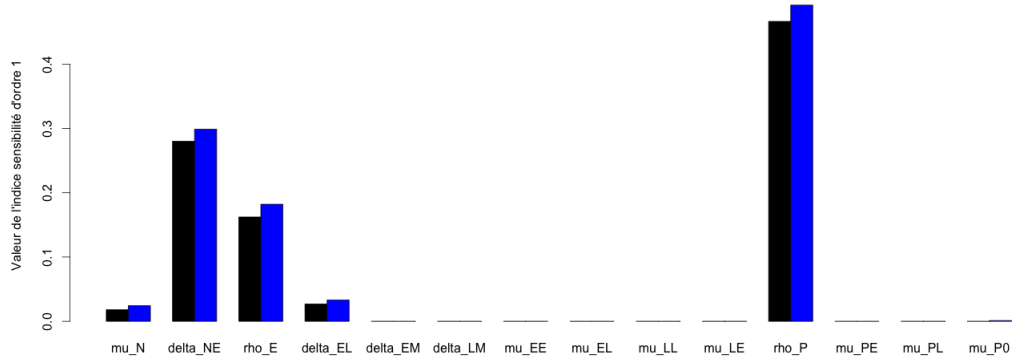
Finalement, nos quatre paramètres permettent de bien séparer les différentes lignées ainsi que les individus. Nous finirons cette présentation des résultats en regardant l'influence des paramètres sur les différentes phases de la dynamique de la réponse.

	BalbC-OF1	C57Bl/6-OF1	C57Bl/6-BalbC
Distance	0.119	0.121	0.237

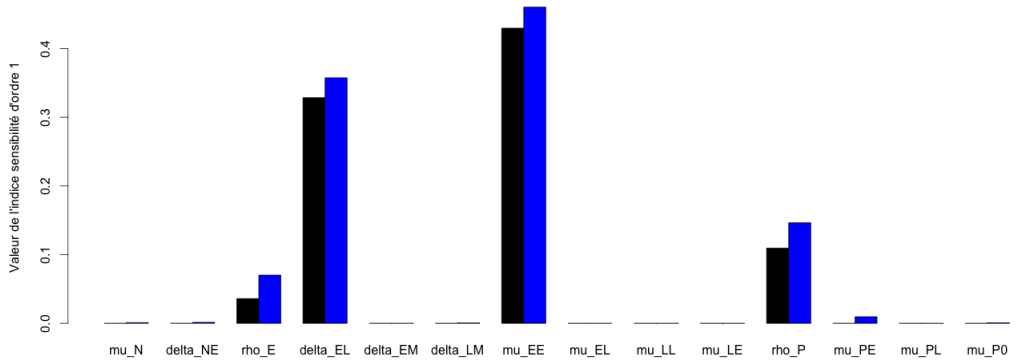
TABLE 6 – Distance de Kantorovich entre les lignées en prenant en compte la covariable.

3.4 Influence des paramètres sur la dynamique de la réponse

A l'aide d'une analyse de sensibilité (voir Section 2.6), nous souhaitons déterminer quels sont les paramètres qui expliquent le plus la variabilité lors des différentes phases de la réponse immunitaire. Cette analyse de sensi-



A)



B)

FIGURE 13 – Analyse de sensibilité de la réponse T CD8. La couleur noire est utilisée pour représenter l'indice de Sobol d'ordre 1 et l'indice d'ordre total est représenté en bleu. Plus la taille de la colonne est importante plus le paramètre a de l'influence sur la variabilité de la réponse. En A) sont présentés les paramètres qui jouent un rôle majeur dans la variabilité de la réponse lors de la phase d'expansion cellulaire (jour 3) . En B), ceux qui expliquent la variabilité de la réponse pendant la phase de contraction (jour 9). Ainsi, pendant la phase d'expansion cellulaire, le nombre de cellules activées est très sensible aux paramètres δ_{NE}, ρ_E et ρ_P . Lors de la phase de contraction ce même nombre est sensible aux paramètres δ_{EL} et μ_{EE}

.

bilité a été effectué au jour 3, pendant la phase d'expansion cellulaire, au jour 9 lors de la phase de contraction et au jour 25 qui correspond à la phase mémoire de la réponse autour des valeurs estimées pour la lignée C57Bl/6. Nous avons généré des valeurs de paramètres se trouvant à plus ou moins 25% autour de la valeur moyenne.

Au jour 3 (voir Figure 13.A), lors de la phase d'expansion cellulaire, ce sont les paramètres δ_{NE}, ρ_E et ρ_P qui expliquent le plus la variabilité de la réponse. Ces paramètres sont ceux qui interviennent dans la dynamique des cellules early effectrices de façon directe (δ_{NE}, ρ_E) ou indirecte (ρ_P) et jouent un rôle dans l'expansion du nombre de ces cellules

Au jour 9 (voir Figure 13.B) lors de la phase de la contraction, ce sont les paramètres δ_{EL} et μ_{EE} qui sont à l'origine de la variabilité. Il s'agit à nouveau de paramètres qui concernent la population de cellules early effec-

trices mais qui jouent sur la décroissance de la population (δ_{EL} et μ_{EE}). La population de cellules early effectrices diminue au profit des cellules late effectrices.

Lors de la phase mémoire au jour 25 (voir Figure 14), correspondant à la fin de la réponse immunitaire, ce sont les paramètres de différenciation en cellules mémoires (δ_{LM}) et de mort des cellules effectrices (μ_{EE} et μ_{LE}) qui expliquent la variabilité du nombre de cellules mémoires généré.

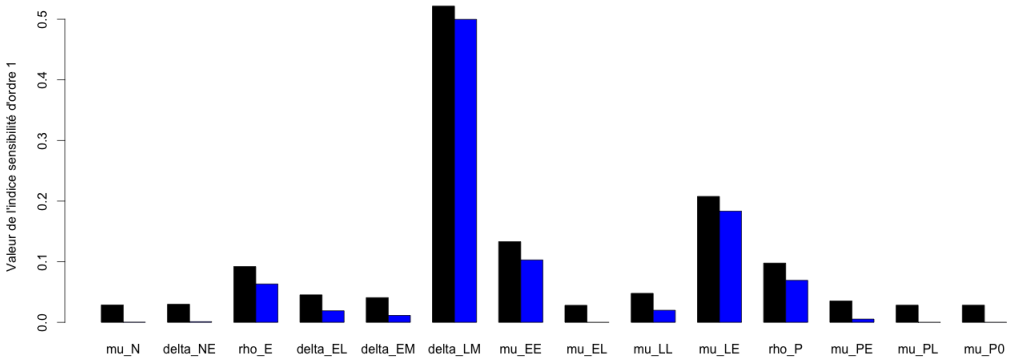


FIGURE 14 – Analyse de sensiblité au jour 25 de la réponse T CD8. La couleur noire est utilisée pour représenter l'indice de Sobol d'ordre 1 et l'indice d'ordre total est représenté en bleu. Plus la taille de la colonne est importante plus le paramètre a de l'influence sur la variabilité de la réponse. Ce jour correspond à la phase mémoire de la réponse (jour 25), les paramètres qui expliquent le plus la variabilité de la réponse sont δ_{LM} et μ_{LE}. Dans cette figure l'indice de sensibilité total est plus petit que l'indice de sensibilité d'ordre 1, alors que ce devrait être le contraire. Cela résultat certainement d'un problème numérique. Il reste cependant possible de tirer des informations de ce graphique.

4 Discussion

L'objectif était de tester la robustesse du modèle présenté en (1) décvrivant les dynamiques des différentes populations de cellules : naïves, early effectrices, late effectrices et mémoires. Ce modèle a ensuite été confronté à des données expérimentales afin de tester son efficacité à rendre compte de la variabilité présente dans les données. Pour ce faire, nous avons estimé ses paramètres à l'aide de modèles à effets mixtes.

A travers l'utilisation d'un tel outil statistique, nous avons pu mettre en évidence les valeurs caractéristiques des paramètres des différentes lignées de souris en considérant chaque lignée séparément. Comme cela est montré à travers les Figures 8 et 9, les résultats fournis par cette approche sont très satisfaisants, aussi bien à l'échelle de la population qu'à l'échelle des individus, en dépit d'un faible nombre d'observations.

Nous avons ensuite défini une distance entre les différents individus basée sur les valeurs estimées des paramètres. En représentant la distance séparant les individus, nous avons pu remarquer un détachement des souris de la lignée C57Bl/6 alors que les souris des lignées BalbC et OF1 sont confondues dans un nuage de point (voir Figure 10) ce qui marque une proximité entre les lignées. En outre, les individus de la lignée C57Bl/6 forment un nuage de point plus dense que les individus des deux autres lignées. Nous avons également pu mettre en évidence (voir Figure 11) les paramètres qui expliquent d'avantage la distance entre les lignées BalbC et OF1 en calculant la distance entre les distributions de paramètres de ces différentes lignées.

L'étape suivante a consisté à considérer l'ensemble des données et de regarder quels sont les paramètres dont la valeur dépend de la lignée étudiée. Cela est rendu possible en introduisant la covariable dans le modèle et lors de l'estimation des paramètres. A l'aide de la procédure présentée en Figure 6, quatre paramètres ce sont

révélés dépendant à la covariable : ρ_E, δ_{EL}, μ_{EE} et μ_{PL}. Trois d'entre eux interviennent directement dans la dynamique des cellules early effectrices. Cela suggère que c'est au niveau de la dynamique de ces cellules que l'on peut observer les différences de comportement des lignées. Or, cette population est prédominante lors de la phase de d'expansion cellulaire et au pic de la réponse. Si l'on se tourne à nouveau vers les données de la Figure 4, et que l'on considère les données à jour 8, nous observons que les différentes lignées se distinguent en terme de nombre de cellules activées. Nous nous sommes alors demandé s'il était possible de tirer une information directement à partir de la connaissance du nombre de cellules au pic de la réponse. Nous avons considéré un modèle de régression logistique multinomiale pour le vérifier. Si l'on considère moins de 50% des données pour la construction et qu'on le teste sur les données restantes, il est impossible de distinguer les lignées OF1 et BalbC. En considérant une part plus importante des données (environ 70 %), le modèle permet de déterminer, sans erreur, la lignée des observations restantes. Ce dernier résultat corrobore notre hypothèse. En considérant une souris et en mesurant son nombre de cellules activée à jour 8 nous serions capable de déterminer de quelle lignée, la dynamique de cette souris se rapproche le plus. Cependant, le modèle employé reste discutable. En testant l'hypothèse de nullité des coefficients du modèle, ils se trouvent que ces derniers ne sont pas significativement différents de zéro.

En plus d'être dépendants de la covariable, les paramètres ρ_E et δ_{EL} se trouvent faire partie de ceux qui expliquent la variabilité du nombre de cellules T CD8 activées pendant la phase d'expansion cellulaire (voir Figure 13.A). Il y a donc probablement une réelle information à tirer au niveau de ces paramètres. L'origine de la variabilité de la réponse et des différents comportements observés peut se trouver dans ces paramètres. Or ces paramètres représentent des mécanismes cellulaires et moléculaires qui sont régis par des gènes. Il serait donc intéressant d'étudier la variabilité à la source, c'est-à-dire identifier les gènes qui interviennent dans les mécanismes de prolifération, mort et différenciation des cellules early effectrices. On pourra ainsi mesurer l'expression de ces gènes d'une lignée afin de mieux comprendre l'origine de la variabilité observée.

Cependant, si l'on souhaite prévoir le profil mémoire de nos souris, la seule connaissance du nombre de cellules au pic de la réponse est insuffisant. Il faudrait être en mesure de prévoir comment va se dérouler la phase de contraction de la réponse. La variabilité observée lors de cette phase est majoritairement expliquée par les paramètres δ_{EL} et μ_{EE} (voir Figure 13.B). Prédire le déroulement de la phase de contraction nécessiterait de savoir comment sont liés les paramètres ρ_P, δ_{NE} et ρ_E (paramètres importants de la phase d'expansion) aux paramètres δ_{EL} et μ_{EE}. Cependant, une étude de la corrélation entre ces différents paramètres ne montre aucune corrélation entre ces derniers (la corrélation est étudiée sur les valeurs estimée, les résultats sont présentées dans la Figure 16 en Annexe).

Lors de la présentation de l'analyse de sensibilité, nous avons mentionné qu'il était également possible d'utiliser les résultats afin de simplifier le modèle. Le modèle est peu sensible à un certain nombre de paramètre (δ_{EM}, μ_{LL}, μ_{LE}, μ_{PL} et μ_P^0). En combinant ces résultats à ceux obtenus par l'introduction de la covariable, la simplification du modèle consisterait à fixer les paramètres indépendant de la covariable et qui influent faiblement sur la dynamique de la réponse. Nous pourrons affecter à ce paramètre les valeurs présentées dans la Table 3. Prenons l'exemple du paramètre de différenciation des cellules early effectrices en cellules mémoires δ_{EM}, ce paramètre n'a pas d'impact sur la variabilité de la réponse et sa valeur n'est que très faiblement modifiée d'une lignée à une autre. Il est donc possible de fixer la valeur de ce paramètre. Cependant, pour les paramètres dont la valeur varie de façon significative d'une lignée à une autre (valeur pouvant doubler) il pourrait être nécessaire de fixer des critères permettant de sélectionner ceux qui peuvent être fixés. Une approche possible est le calcul de l'erreur moyenne commise sur l'ensemble des données.

L'étude de la variabilité pourrait être améliorée en tenant compte de la variabilité intra-individuelle, elle permet de décrire la variabilité des observations au cours du temps pour un individu donné. En effet, les mécanismes de la réponse, comme la prolifération, varient au cours du temps. Cela se traduit par une dépendance en temps des valeurs de paramètres. Pour ce faire, il nous faudrait disposer de données qui s'étendent sur une échelle de temps suffisamment longue. Nous pourrions diviser cette échelle de temps en plusieurs sous intervalles qui correspondraient à différentes étapes de la vie des souris. Nous disposerions ainsi d'une valeur de paramètres par intervalle de temps permettant ainsi de suivre l'évolution des mécanismes de la réponse immunitaire au cours de différentes périodes de vie d'un individu.

Références

[1] C. Janeway, K. Murphy, P. Travers, P. Masson et M. Walport, *Immunobiologie.* De Boeck Supérieur, 2009.

[2] E. Terry, *Mathematical modeling of T CD8 immune response dynamics, at cellular and molecular scales.* Thèse doctorat, Université Claude Bernard - Lyon I, 2012.

[3] « Ouvrage L2 immunologie générale de assim », 2011. Disponible sur `http://www.assim.refer.org/raisil/raisil/L02.html`.

[4] N. LEGRAND, *Central selection, survival and peripheral selection of CD8+ ab T lymphocytes.* Thèse de doctorat, Université Pierre et Marie Curie - Paris VI, sept. 2002.

[5] F. Crauste, J. Mafille, L. Boucinha, S. Djebali, E. Gallica, O. Gandrillon, J. Marvel et C. Arpin, « Identification of emerging memory CD8 T cells during a primary response. »,

[6] S. Prokopiou, L. Barbarroux, S. Bernard, J. Mafille, Y. Leverrier, C. Arpin, J. Marvel, O. Gandrillon et F. Crauste, « Multiscale modeling of the early CD8 T-Cell immune response in lymph nodes : An integrative study », *Computation*, vol. 2, no. 4, p. 159–181, 2014.

[7] F. Crauste, E. Terry, I. L. Mercier, J. Mafille, S. Djebali, T. Andrieu, B. Mercier, G. Kaneko, C. Arpin, J. Marvel et O. Gandrillon, « Predicting pathogen-specific CD8 T cell immune responses from a modeling approach », *Journal of Theoretical Biology*, vol. 374, p. 66 – 82, 2015.

[8] M. A. Nowak et C. R. M. Bangham, « Population dynamics of immune responses to persistent viruses », *Science*, vol. 272, no. 5258, p. 74–79, 1996.

[9] R. Antia, C. T. Bergstrom, S. S. Pilyugin, S. M. Kaech et R. Ahmed, « Models of CD8+ responses : 1. what is the antigen-independant proliferation program ? », *Journal of Theoretical Biology*, vol. 221, p. 585–598, 2003.

[10] R. J. D. Boer et A. S. Perelson, « Quantifying T lymphocyte turnover », *Journal of Theoretical Biology*, vol. 327, p. 45 – 87, 2013.

[11] D. L. Chao, M. P. Davenport, S. Forrest et A. S. Perelson, « A stochastic model of cytotoxic T cell responses », *Journal of Theoretical Biology*, vol. 2, no. 228, p. 227–240, 2004.

[12] H. P. Piepho, A. Büchse et K. Emrich, « A hitchhiker's guide to mixed models for randomized experiments », *Journal of Agronomy and Crop Science*, vol. 189, no. 5, p. 310–322, 2003.

[13] H. K. Im, E. R. Gamazon, A. L. Stark, R. S. Huang, N. J. Cox et M. E. Dolan, « Mixed effects modeling of proliferation rates in cell-based models : consequence for pharmacogenomics and cancer », *PLoS Genetics*, vol. 8, no. 2, p. e1002525, 2012.

[14] T. Koue, M. Kubo, T. Funaki, T. Fukuda, J. Azuma, M. Takaai, Y. Kayano et Y. Hashimoto, « Nonlinear mixed effects model analysis of the pharmacokinetics of aripiprazole in healthy japanese males », *Biological and Pharmaceutical Bulletin*, vol. 30, no. 11, p. 2154–2158, 2007.

[15] E. Terry, J. Marvel, C. Arpin, O. Gandrillon et F. Crauste, « Mathematical model of the primary CD8 T cell immune response : stability analysis of a nonlinear age-structured system », *Journal of Mathematical Biology*, vol. 65, no. 2, p. 263–291, 2012.

[16] A. Burg, « Contribution to a multi-scale model of CD8 T cells immune response », mémoire de master, Ecole Centrale de Lyon, 2014.

[17] T. Yajima, K. Yoshihara, K. Nakazato, S. Kumabe, S. Koyasu, S. Sad, H. Shen, H. Kuwano et Y. Yoshikai, « Il-15 regulates CD8+ T Cell contraction during primary infection », *The Journal of Immunology*, vol. 176, no. 1, p. 507–515, 2006.

[18] N. S. Joshi et S. M. Kaech, « Effector CD8 T cell development : a balancing act between memory cell potential and terminal differentiation », *The Journal of Immunology*, vol. 180, no. 3, p. 1309–1315, 2008.

[19] M. Lavielle, *Mixed Effects Models for the Population Approach : Models, Tasks, Methods and Tools.* Chapman and Hall/CRC, 2014.

[20] « Lixoft », 2013.

[21] M. Duval, *Modelisation and estimation of heterogeneous variances in nonlinear mixed models.* Thèse de doctorat, AgroParisTech, déc. 2008.

[22] B. Delyon, M. Lavielle et E. Moulines, « Convergence of a stochastic approximation version of the EM algorithm », *Ann. Statist.*, vol. 27, no. 1, p. 94–128, 1999.

[23] E. Pękalska et R. Duin, *The Dissimilarity Representation for Pattern Recognition : Foundations and Applications.* Series in machine perception and artificial intelligence, World Scientific, 2005.

[24] Bobkov, Sergey, Ledoux et Michel, *One-dimensional empirical measures, order statistics and Kantorovich transport distances.* 2014. Book in preparation. Disponible sur http ://perso.math.univ-toulouse.fr/ledoux/files/2013/11/Order.statistics.10.pdf.

[25] J. Jacques, *Contributions à l'Analyse de Sensibilité et à l'Analyse Discriminante Généralisée.* Thèse de doctorat, Université Joseph Fourier-Grenoble 1, 2005.

Annexe

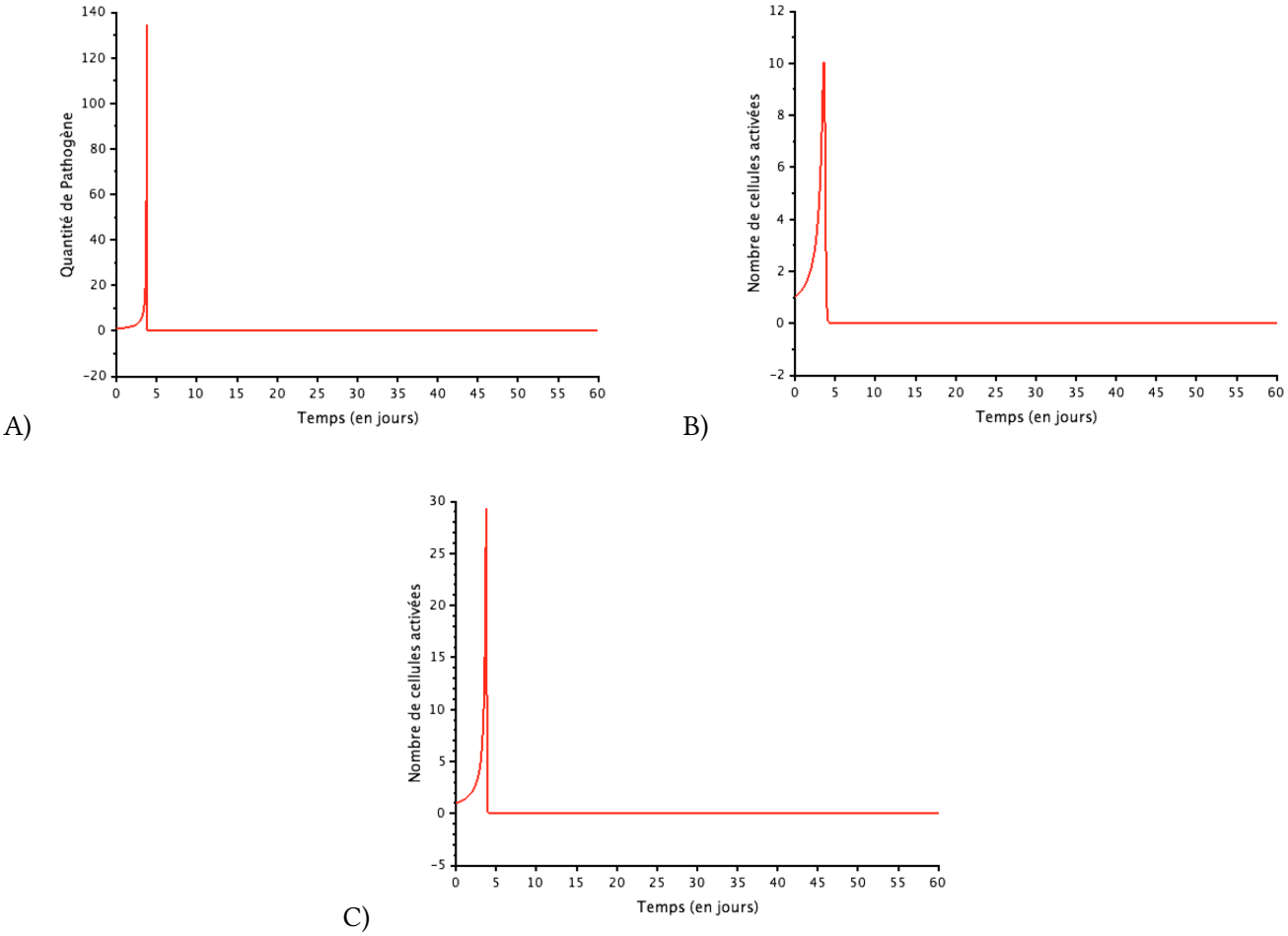


FIGURE 15 – Evolution de la quantité de pathogène au cours du temps chez les différentes lignée de souris : A) C57Bl/6, B) BalbC et C) OF1. On remarque que la quantité de pathogène est beaucoup plus importante au moment du pic de la réponse (à jour 4-5) chez la lignée C57Bl/6 par rapport aux deux autres lignées. Cependant, pour toute les trois lignées, la quantité de pathogène est maximale au moment du pic de la réponse.

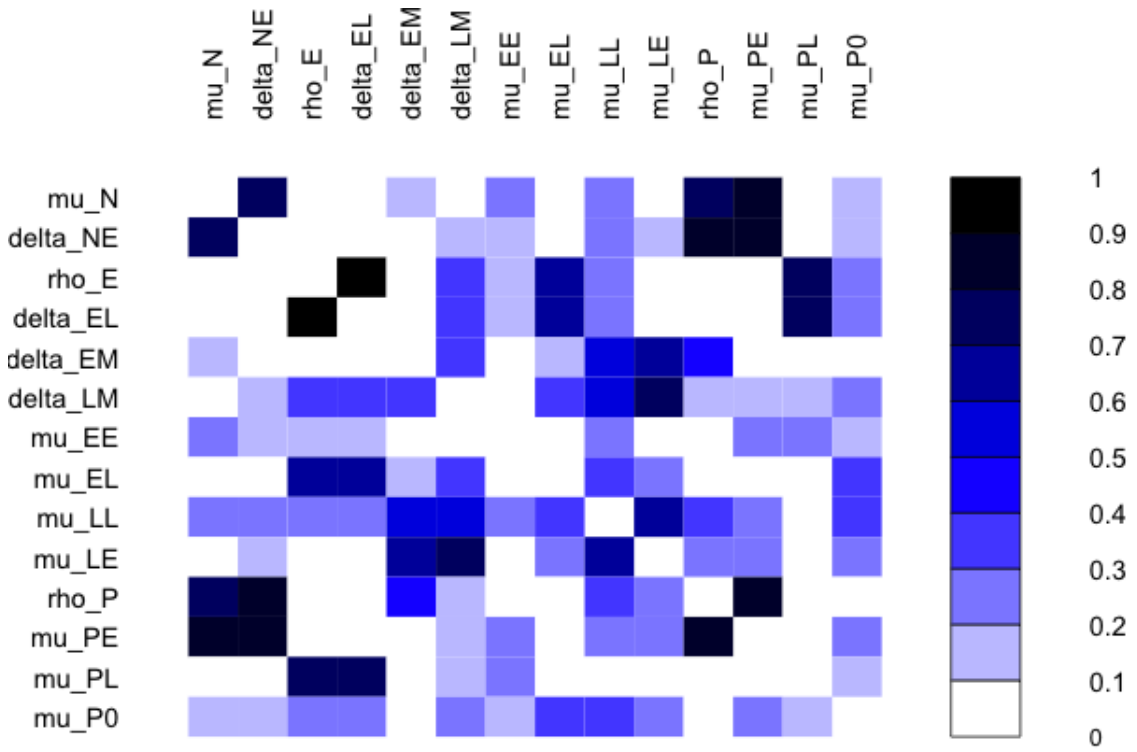


FIGURE 16 – Graphique de la corrélation entre les différents paramètres du modèle. Le coefficient de corrélation est représenté en valeur absolue et est calculé à partir de l'ensemble des valeurs individuelles de paramètres. Plus la case est foncée plus la corrélation est importante. Les élément diagonaux sont représentés en blanc pour plus de lisibilité mais devraient être représentés en noir.