Biophysical modeling of bacterial population dynamics and the immune response in the gut

Florence Bansept

Thèse de doctorat
de Sorbonne Université

Spécialité : Physique

École doctorale n°564: Physique en Île-de-France

présentée par :

Florence Bansept

pour obtenir le grade de :

Docteure de Sorbonne Université

Sujet de la thèse :

Biophysical modeling of bacterial population dynamics and the immune response in the gut

Soutenue le 5 Décembre 2018

Devant le jury composé de :

M. Raphaël Voituriez	Directeur de thèse
M. Bahram Houchmandzadeh	Rapporteur
Mme. Agnese Seminara	Rapportrice
Mme. Aleksandra Walczak	Examinatrice
Mme. Silvia De Monte	Examinatrice
M. Andrea Parmeggiani	Examinateur
Mme. Emma Wetter-Slack	Membre invitée
Mme. Claude Loverdo	Membre invitée

Laboratoire Jean Perrin – UMR 8237

À la couleur des blés...

Abstract

The first part of this thesis focuses on the colonization dynamics of a bacterial population in early infection of the gut. The aim is to infer biologically relevant parameters from indirect data. We discuss the optimal observable to characterize the variability in genetic tags distributions. In a first one-population model, biological arguments and inconsistencies between several experimental observables lead to the study of a second model with two-subpopulations replicating at different rates. As expected, this model allows for broader possibilities in observables combination, even though no clear conclusion can be drawn as to a data set on Salmonella in mice. The second part concerns the mechanisms that make the immune response effective. The main effector of the immune system in the gut, IgA (an antibody), enchains daughter bacteria in clonal clusters upon replication. Our model predicting the ensuing reduction of diversity in the bacterial population contributes to evidence this phenomenon, called "enchained growth". Inside the host, the interplay of cluster growth and fragmentation results in preferentially trapping faster-growing and potentially noxious bacteria away from the epithelium, which could be a way for the immune system to regulate the microbiota composition. At the scale of the hosts population, in the context of evolution of antibiotic resistance, if bacteria are transmitted via clonal clusters, the probability to transmit a resistant bacteria is reduced in immune populations. Thus we use statistical physics tools to identify some generic mechanisms in biology.

Résumé

La première partie de cette thèse porte sur les dynamiques de colonisation d'une population bactérienne au début d'une infection intestinale. Le but est de déduire des paramètres biologiquement pertinents de données indirectes. Un modèle simple est étudié, et l'on discute de l'observable optimale pour caractériser la variabilité d'une distribution d'étiquettes génétiques. Des arguments biologiques et des incohérences entre des observables expérimentales avec le premier modèle motivent l'étude d'un second, où deux sous-populations se répliquent à des taux différents, mais on ne peut pas conclure clairement sur le jeu de données utilisé. La seconde partie porte sur les mécanismes de la réponse immunitaire. Le principal effecteur du système immunitaire adaptatif dans l'intestin, l'IgA (un anticorps), enchaîne les bactéries-filles en agrégats clonaux lors de la réplication. Nous avons contribué à prouver ce phénomène par un modèle qui prédit la réduction de la diversité bactérienne qui en découle. Au sein de l'hôte, l'interaction entre la croissance et la fragmentation des agrégats a pour conséquence le piégeage préférentiel des bactéries à croissance rapide, ce qui pourrait permettre au système immunitaire de réguler la composition du microbiote. A l'échelle de la population-hôte, et dans le contexte de l'évolution d'une résistance aux antibiotiques, si les bactéries sont transmises sous forme d'amas clonaux, alors la probabilité de transmettre une bactérie résistante est réduite dans une population immunisée. Ainsi, des outils de physique statistique nous permettent d'identifier des mécanismes génériques en biologie.

Remerciements

Avant de tourner la page, voici venu le temps... des remerciements.

Un immense, immense merci à Claude Loverdo, ma directrice de thèse officieuse, qui a su m'accompagner durant ces trois années (et quelques) avec beaucoup de disponibilité, de bienveillance, de patience et de gentillesse, qui a su canaliser (et contrebalancer) ma nature stressée dans les moments clés, et qui m'a somme toute beaucoup, beaucoup appris. Merci à Anne-Florence Bitbol pour ses conseils précieux, à Loïc pour les bavardages et les encouragements, et pour avoir bien voulu relire des bouts de ce manuscrit. Merci aussi à tous les deux pour une intéressante et fructueuse collaboration sur le dernier chapitre de ce manuscrit. Merci enfin à Raphaël Voituriez, mon directeur de thèse officiel, d'avoir été présent dans les moments clés (notamment lorsqu'il a s'agit de prolonger mon contrat en catastrophe), toujours positif et encourageant, et d'avoir relu ce manuscrit et permis de l'améliorer. Ce fut un plaisir et un privilège de faire partie de cette belle équipe de théorie du Laboratoire Jean Perrin !

Merci aux collègues du bureau B414 (à défaut de B612) : ceux qui sont déjà partis voguer vers d'autres aventures : Adrian, Aurélien, et Johnny, pour les longs et passionnants bavardages. *Penny salutes you !* Et à ceux qui sont encore là pour un temps : Manon, Jean-Baptiste, Anis, Amaury... et une mention toute particulière pour Alexandre, qui a su me supporter face à lui pendant trois ans mieux que quiconque. Partager ce bureau avec vous, mais aussi beaucoup de thé, de repas, de cafés à rallonge, de chocolat parfois, un combat pour un canapé vert, un certain nombre de débats politiques et/ou sociétaux et beaucoup de conversations beaucoup plus futiles, fut un réel plaisir, et une motivation.

Merci à l'ensemble du Laboratoire Jean Perrin, pour une ambiance de travail amicale et privilégiée, et en particulier à Didier Chatenay, garant de cette ambiance propagée au rythme de ses chasses aux bonbons, et à Marie-Nicole Colonnette, dont l'aide précieuse m'a permis entre autres de partir toujours sereinement en conférences. Un merci aussi à tous les doctorants et post-doctorants.

Merci à ceux qui m'ont encouragée et conseillée de plus loin, notamment Jean-François Allemand, parrain de cette thèse, et Jeferson Arenzon. Un remerciement tout particulier pour nos collaborateurs suisses, en particulier Emma-Wetter Slack, pour nous avoir permis de travailler sur un système aussi cool et passionnant, pour son accueil lors d'une visite avant même le début de ma thèse, et pour sa grande gentillesse et sa disponibilité.

Pendant ma thèse, j'ai aussi eu la chance et l'honneur d'enseigner en première et deuxième années d'université. Merci donc aux équipes d'enseignement des UE 1P001-1P002 "Concepts et Méthodes de la Physique", 1P003 "Energies et transformations de la matière" et 2P021 "Electromagnétisme et électrocinétique" pour

leur accompagnement. Et merci surtout à tous mes élèves, qui, s'ils n'ont pas toujours brillé par leur niveau, l'ont toujours fait par leur capacité à me faire sourire.

Enfin merci aux membres du jury qui ont apporté un regard critique et constructif à cette thèse, et merci à ma future équipe de postdoc, pour m'avoir recrutée et avoir manifesté de l'intérêt pour mon travail !

Il me reste maintenant à remercier tous ceux qui n'ont pas directement pris part à mes activités de doctorante, mais dont le rôle fut pourtant essentiel durant ces trois années...

Merci à ma famille, et surtout à mes parents, pour leur amour et leur soutien inconditionnels de parents, et pour m'avoir recueillie et soignée comme un oiseau tombé du nid après l'accident qui a retardé la présentation publique de cette thèse. Merci à Émilie, mon anti-stress personnelle depuis le début de mes études supérieures. Merci à Marilyn, ma "sister" depuis encore plus longtemps. Merci à mes "trois mousquetaires" (dixit ma maman), amis d'école et compagnons de fourires, Frédéric, Louis, et Romain(s), qui après une brève carrière dans la hotline de comptine pour enfant s'est reconverti pendant ma thèse en fournisseur officiel de photos de petits chats mignons, si efficace qu'on se demanderait presque s'il n'est pas légion... Merci à Ashley, Églantine, Sarah et Bruna.

Et puis... merci à ceux qui ont fait partie de mon "autre vie", la musicale. Un merci tout particulier à Matthieu, qui m'a aidée à reprendre le piano où je l'avais laissé, et à tenir ainsi une promesse que j'avais à coeur de tenir, et à Johan, Julien et Fériel. Merci à eux quatre d'avoir partagé et transmis leur art et leurs sensibilités avec tant de générosité, et de m'avoir permis d'entretenir et de nourrir une passion. Merci enfin aux "bisounours" du choeur de mon coeur : Mathilde, Anne, Clara, Simon, Pierre, Marie (qui a bien voulu en prime relire des bouts de ce manuscrit), Arthur, Luc, Gabriel, Anya, Philomène, Guillaume, mes mamans de c(h)oeur Brigitte et Eliza, et tous les autres. Ce serait trop long de tenter d'exprimer toute ma gratitude ici, et je serais presque tentée d'un simple *merci d'exister !* à la JF. Alors simplement merci pour les déjeuners-wallabys si précieux des dernières semaines de rédaction. Et surtout, merci pour toute la musique partagée, dans une multitude de salles et d'arrêts de métro parisiens, à Cambdrige, à Forcalquier, Saint Etienne les Orgues, Jérusalem, Tel Aviv, Hadassah, Chartres, Sofia, Plovdiv, Pazardjik, Saint Brieuc, Binic - Étables-sur-mer, Villé,... du Bach, du Brahms, Fauré, Saint Saens, Charpentier, Beethoven, Beffa, Bernstein, Bizet, Bonis, Borodine, Dvořák, Gershwin, Legrand, Moondog, Anthiome, Ravel, Offenbach, Monk, Haydn, McCartney, Strauss et tant d'autres... Merci pour votre amitié. Merci pour la couleur des blés.

Faire une liste exhaustive serait impossible, alors merci enfin à ceux que je n'ai pas cités ici. A tous, merci d'avoir fait partie de ma vie durant mes trois années de doctorat, et d'en avoir fait trois années si heureuses et épanouies.

Résumé en français

Introduction

Le corps humain abrite un microbiote important. Chaque compartiment anatomique en contact direct avec l'environnement, comme la peau, la plaque dentaire, la salive, les poumons, recèle de très nombreux microorganismes. Il a été estimé [1] qu'un corps humain contient environ 10^{13} bactéries, soit approximativement autant que le nombre de cellules humaines qui composent ce corps. La plupart de ces bactéries (environ 99%) sont situées dans l'appareil digestif. La présence de ces écosystèmes est essentielle à la bonne santé des individus, car ils remplissent plusieurs fonctions importantes [2] : grâce aux effets de compétition, ils protègent l'hôte de l'intrusion d'autres agents potentiellement pathogènes. Dans le système digestif, ils aident aussi à la digestion en cassant des nutriments spécifiques. Il s'agit d'une interaction mutualiste entre l'hôte et les bactéries commensales, qui se nourrissent des nutriments ingérés par l'hôte. Bien sûr, les bactéries sont aussi responsables de nombreuses pathologies : une souche virulente peut parvenir à supplanter les autres populations bactériennes, perturber le fonctionnement de l'écosystème et déclencher une inflammation (causant par exemple des caries, des diarrhées, des orgelets, etc). Parfois, une telle souche pénètre l'organisme et se répand dans les différents organes, causant de graves infections systémiques. Les infections bactériennes sont responsables de millions de morts chaque année : selon l'Organisation Mondiale de la Santé, elles contribuent largement à trois des dix premières causes de mort à l'échelle mondiale en 2016, avec la pneumonie, les maladies diarrhéiques et la tuberculose [3]. Les infections bactériennes sont d'autant plus un enjeu de santé publique que la résistance aux antibiotiques semble maintenant s'étendre plus rapidement que les nouveaux traitements ne sont conçus [4], alors même que les antibiotiques constituent le principal outil pour lutter contre elles. C'est pourquoi il est essentiel de développer notre connaissance des populations bactériennes et des infections bactériennes.

Différentes approches coexistent pour étudier les populations bactériennes. Une première classe d'approches pourrait être qualifiée d'*ascendante* : l'idée est de partir du système le plus simple possible *in vitro*, et d'introduire progressivement des niveaux de complexité de manière maîtrisée afin d'imiter les conditions naturelles. [5] [6] [7] [8]. A l'inverse, une autre classe d'approches complémentaires aux précédentes pourrait être qualifiée de *descendante* : l'idée est cette fois de

partir de systèmes complexes en conditions naturelles et d'essayer d'analyser et de séparer les différents facteurs qui en régissent les comportements. Ces approches sont essentielles, car le niveau de complexité en conditions naturelles et/ou dans des systèmes vivants est très difficile, voire impossible, à reproduire artificiellement. D'une part, afin de recréer les contraintes mécaniques ressenties dans un environnement naturel, ou *in vivo* pour des populations bactériennes vivant à l'intérieur d'un hôte, il faut que ces contraintes puissent être correctement caractérisées, ce qui n'est pas toujours le cas. Par exemple, dans l'appareil digestif, des mouvements péristaltiques complexes visant à mélanger le digestat de manière optimale afin de favoriser l'absorption des nutriments imposent des contraintes complexes sur le contenu extrêmement non-Newtonien[9, 10]. D'autre part, des composants chimiques ou enzymatiques, des bactériophages ou d'autres microorganismes, pourraient jouer un rôle important dans les comportements observés sans être correctement identifiés. Et même lorsque ces composants sont correctement identifiés, certains composants moléculaires peuvent être difficiles à reproduire artificiellement ou à isoler en quantités suffisantes pour être ajoutés à des systèmes *in vitro*. C'est pourquoi de nombreux modèles animaux ont été développés pour étudier les infections bactériennes directement dans l'hôte. Le modèle de l'infection de la souris par la salmonelle [11], utilisé dans les études expérimentales qui ont initialement motivé ce travail, présente de nombreux avantages. D'abord, le système immunitaire de la souris présente un degré de complexité similaire à celui des humains, avec en particulier un système immunitaire adaptatif développé. En outre, plusieurs aspects de l'infection peuvent être contrôlés. Les souris peuvent être élevées en environnement stérile de sorte que la composition de leur microbiote puisse être contrôlée et présenter une complexité moindre que dans des conditions naturelles. Les souris peuvent également être manipulées génétiquement de manière à présenter des déficiences immunitaires particulières. Tous ces outils permettent de mieux séparer les différentes effets étudiés.

L'étude des populations bactériennes en conditions naturelles ou dans des systèmes vivants complets nécessite le recours à la modélisation pour de nombreuses questions, et, inversement, ces systèmes représentent tout un champ de possibles applications pour les modélisateurs et les physiciens. Par exemple, les progrès des techniques de séquençage permettent maintenant l'analyse de microbiomes entiers, que l'on peut confronter à des modèles de réseaux [12], de théorie des graphs et théorie des jeux [13], ou encore de génétique des populations [14]. Un autre exemple est la contribution que la physique a déjà apporté à l'immunologie, notamment en utilisant le concept d'information, en particulier en ce qui concerne la reconnaissance des agents pathogènes qui déclenchent la sécrétion de nouveaux effecteurs, et la constitution de répertoires de récepteurs à antigènes d'une grande diversité [15, 16].

En ce qui concerne la colonisation d'un hôte, à part dans des configurations très particulières [17, 18], il est impossible de suivre le processus infectieux en détail sans avoir recours à des mesures invasives et donc perturbatrices. Des me-

sures indirectes sont plutôt privilégiées, et c'est là que la modélisation devient un outil essentiel pour comprendre les observations expérimentales. L'organisme animal peut être vu comme une "boîte noire", et le rôle du modélisateur consiste alors à inférer ses règles internes à partir d'observations extérieures. Par ailleurs, dans le contexte des infections, les outils de physique statistique sont nécessaires pour étudier les dynamiques de populations. En effet, une infection peut démarrer d'un petit nombre de microorganismes, ce qui requière une modélisation stochastique. Ces microorganismes se répliquent ensuite en grands nombres, ce qui nécessite une description de type champ moyen.

Du point de vue du physicien, le système digestif est particulièrement intéressant. D'une part, en raison des forces exercées par le flux qui mélange et transporte le contenu des intestins, et d'autre part, parce qu'il n'y a qu'un nombre restreint d'effecteurs immunitaires à prendre en compte, ce qui rend l'étude de leurs mécanismes physiques plus aisée. En effet, le système digestif est topologiquement "à l'extérieur" de l'organisme hôte, les effecteurs immunitaires sécrétés sont donc pour ainsi dire "perdus" pour l'organisme une fois dans le lumen intestinal. Cela pourrait expliquer pourquoi, sauf dans des cas particuliers, seul un petit nombre de cellules immunitaires sont sécrétées dans les intestins. Les cellules de l'immunité présentent plusieurs états, réagissent à de nombreux signaux, et sont donc complexes à modéliser. Au contraire, et bien que les mécanismes menant à sa sécrétion soient complexes eux aussi, l'immunoglobuline A (IgA), un type d'anticorps et le principal effecteur de la réponse immunitaire adaptative sécrété dans les intestins, est une molécule dont la concentration peut être mesurée, et dont les effets sont plus faciles à caractériser. Par ailleurs, ces anticorps reconnaissent une souche spécifique de bactéries, ce qui signifie que la population bactérienne avec laquelle ils interagissent est essentiellement homogène, ce qui facilite encore la modélisation biophysique.

Cette thèse porte sur plusieurs aspects des dynamiques d'une population bactérienne et de son interaction avec le système immunitaire dans les intestins. Elle se compose de deux parties distinctes et essentiellement indépendantes. Dans une première partie, je présente mon travail portant sur les dynamiques de colonisation d'une population bactérienne au début d'une infection. Je développe des modèles stochastiques (car les nombres initiaux de bactéries d'intérêt sont faibles) qui visent à déterminer à partir de données expérimentales indirectes des paramètres biologiquement pertinents, comme des taux de réplication et d'élimination, et la probabilité pour une bactérie de s'établir dans l'organisme et de participer à l'infection. Dans un premier chapitre 1.1, je présente brièvement les données expérimentales qui ont motivé cette étude (et qui reposent principalement sur du marquage bactérien), et je présente les principales méthodes de physique statistique qui sont utilisées dans la suite – à la fois analytiques (avec principalement des processus de branchement) et numériques (avec principalement des simulations de type Gillespie). Le second chapitre 1.2 est ensuite consacré à l'étude d'une première classe de modèles dont la population bactérienne est initialisée par un tirage Poissonien puis suit un processus de mort et de naissance

de Markov en temps continu. Cette étude permet également une réflexion plus large sur le choix optimal d'une observable pour caractériser la variabilité dans la distribution d'un ensemble de variables aléatoires suivant la même dynamique temporelle mais avec des nombres initiaux différents. Je montre que dans certains cas, les paramètres estimés à l'aide des différentes observables à partir d'un même jeu de données ne sont pas clairement cohérents. Cela m'a menée à l'étude d'une autres classe de modèles au chapitre 1.3, avec deux sous-populations distinctes suivant les mêmes types de dynamiques mais avec des paramètres différents. Plusieurs facteurs biologiques pourraient en effet être à l'origine de telles disparités au sein d'une population d'une même souche bactérienne. Je montre que si ces modèles permettent de plus grandes possibilités de combinaisons des différentes observables, qui permettent d'expliquer certaines des expériences considérées ici, on ne peut pas tirer de conclusion générale pour l'ensemble (restreint) de données exploitées dans ce travail.

La seconde partie de ma thèse porte sur les mécanismes physiques de la réponse immunitaire dans les intestins. Elle s'appuie sur les résultats d'une étude récente, à laquelle j'ai contribué, montrant que le principal effecteur du système immunitaire dans les intestins, l'immunoglobuline A (un anticorps spécifique), enchaîne les bactéries filles en agrégats clonaux lors de la réplication [19]. Dans un premier chapitre 2.1, je reprends les résultats de cette étude et présente en particulier ma contribution, avec un modèle qui montre la diminution de diversité dans la population bactérienne résultant de ce processus de croissance enchainée, tout en expliquant comment ce mécanisme suffit à protéger l'organisme de l'infection. Dans les deux chapitres suivants (2.2 et 2.3), j'explore les conséquences de ce phénomène, d'abord à l'échelle de l'hôte (chapitre 2.2), puis à l'échelle d'une population d'hôtes, en terme d'évolution de la résistance aux antibiotiques dans la population bactérienne (chapitre 2.3). A l'échelle de l'hôte, dans le chapitre 2.2, je montre que dans les individus immuns produisant de l'immunoglobuline A, l'interaction entre croissance bactérienne et dislocation des agrégats pourrait être un moyen pour le système immunitaire de contrôler la composition du microbiote en impactant préférentiellement les bactéries à croissance rapide, plus susceptibles de déséquilibrer la flore intestinale [20]. En effet, si les bactéries se répliquent plus vite que les agrégats ne cassent, alors elle finissent piégées dans ces agrégats, ce qui les empêche d'approcher les parois de l'intestin et de coloniser le reste de l'organisme, c'est-à-dire de commencer une infection systémique. En outre, si les bactéries sont présentes sous forme d'agrégats clonaux au sein de l'hôte, il est aussi plausible qu'elles soient transmises sous cette forme. Dans le chapitre 2.3, je montre qu'à l'échelle de la population d'hôtes, si tous les autres paramètres sont inchangés, la probabilité qu'une infection émerge est réduite dans une population immunisée (où les bactéries s'agrègent) par rapport à une population naïve (où les bactéries restent libres). En effet, dans le cas où les bactéries sont transmises sous forme d'agrégats clonaux (soit totalement résistants, soit totalement sensibles), si le nombre moyen de bactéries résistantes transmises est conservé, la proportion de transmissions contenant au moins une bactérie résistante est plus faible par rapport au cas où il n'y a pas d'agrégation.

Les deux parties de cette thèse ont été motivées par l'étude de données quantitatives d'expériences d'infections de la souris par la salmonelle. Cependant, les résultats présentés ici dépassent le cadre de la simple interprétation de données. Les mécanismes d'immunité que je présente ici sont en effet de portée très générale : le système immunitaire de la souris est proche de celui de nombreux autres vertébrés (dont les humains), et la croissance enchaînée n'est pas un processus qui se limite à la salmonelle, puisqu'il a déjà été mis en évidence notamment chez *E. coli* [19]. En outre, dans la première partie, les outils de physique statistique qui sont développés pourraient être appliqués à de plus grands ensembles de données sur les infections bactériennes de l'intestin dans divers animaux et avec diverses souches bactériennes, mais pourraient aussi être adaptés facilement à d'autres systèmes en écologie, pourquoi pas à des échelles différentes. Dans la deuxième partie, l'étude de la croissance et de la fragmentation des agrégats présente en soi un problème général de physique statistique, qui s'est déjà révélé utile dans d'autres contextes [21][22]. Ainsi, l'objectif est d'identifier des mécanismes génériques, et les ingrédients minimaux nécessaires à la compréhension de la portée de ces mécanismes.

1 Dynamiques d'une population bactérienne lors d'une infection des intestins

Introduction

Les maladies infectieuses sont souvent étudiées sous l'angle des processus moléculaires impliqués. L'identification des molécules et des cellules clés est utile, mais ce point de vue peut être complété par d'autres approches. En particulier, la contribution des méthodes de type dynamiques de populations est essentielle : on ne peut pas imaginer concevoir des stratégies optimales pour combattre une infection sans savoir par quelle voie l'agent pathogène pénètre l'organisme hôte, quels organes il colonise, à quelle vitesse il se réplique, migre, et se fait potentiellement éliminer. L'identification de points faibles du processus infectieux à l'échelle de toute la population pourrait amener à concevoir de nouveaux vaccins et thérapies. De telles approches ont été développées dans la communauté des virus [23, 24] et se sont déjà montrées prometteuses en ce qui concerne les infections bactériennes [25, 26]. Inversement, les populations bactériennes fournissent des systèmes modèles pour l'étude des dynamiques de population et d'évolution, en particulier en raison des taux de croissance élevés.

Dans cette partie, on développe des outils génériques, et on les applique à un cas spécifique : les infections bactériennes intestinales résultant d'une intoxication alimentaire. L'objectif est en particulier de caractériser les dynamiques de colonisation des bactéries dans les premiers stades de l'infection, en déterminant des paramètres biologiquement pertinents, comme des taux de réplication, d'élimination, ou la probabilité pour une bactérie de s'établir et de prendre part au processus infectieux. Pour un même nombre final de bactéries issues d'un nombre initial de bactéries donné, différents scénarios sont possibles. Par exemple, des taux élevés

de réplication et d'élimination pourraient mener au même nombre final qu'un scénario avec des taux de réplication et d'élimination faibles, bien que le premier scénario présente un taux de renouvellement bien plus important. La modélisation sert donc à déchiffrer les données indirectes auxquelles on a accès et à déterminer lequel de ces scénarios est le plus probable.

D'un point de vue plus général, cette partie de mon travail vise également à traiter des questions plus théoriques. Le but est de développer des modèles génériques d'une ou de plusieurs sous-populations dans un système ouvert et de déterminer les meilleurs observables pour extraire un maximum d'information sur la dynamique. Le recours à différentes observables doit permettre soit d'inférer un plus grand nombre de paramètres, soit de vérifier la cohérence des résultats.

Dans la suite, je commence par présenter les données expérimentales qui ont motivé mon travail, en particulier les différents marqueurs bactériens utilisés afin d'obtenir plus d'informations quantitatives sur les nombres initiaux et finaux de la population bactérienne, et je dresse une liste non exhaustive des méthodes analytiques et computationnelles qui sont utilisées dans la suite. Dans un deuxième temps, je décris un premier modèle à une population, ainsi que toutes les questions soulevées en ce qui concerne le choix des observables. Enfin, je présente un modèle à deux sous-populations inspiré d'arguments biologiques. Je finis en discutant les résultats obtenus avec ces différents modèles et leurs limitations, et présente les perspectives de développement de ce travail.

1.1 Données expérimentales et méthodes

1.1.1 Données expérimentales

Le travail de cette thèse s'inspire de données expérimentales produites dans l'équipe du Dr. Emma Wetter-Slack, immunologiste à l'ETH de Zürich. Ces données portent sur l'infection orale de la souris par la souche bactérienne *Salmonella enterica* serovar *enterica* Typhimurium [11]. Cette bactérie cause entre autre de sévères diarrhées chez l'humain, responsables de plusieurs millions de morts chaque année [27]. On s'intéresse principalement aux premiers stades de l'infection, lorsqu'elle n'est pas encore systémique et qu'il n'y a pas encore d'inflammation. On se concentre sur le contenu du cæcum, une poche située au début du gros intestin de dimension importante chez les rongeurs, et où les bactéries s'établissent et se répliquent pendant les premiers stades de l'infection.

Comptage direct des bactéries Les bactéries sont rendues résistantes à un antibiotique. Plusieurs dilutions (par exemple à la fin de l'expérience, dans le contenu cæcal) sont faites, et déposées dans des boîtes de Pétri. Des colonies bactériennes visibles se forment et sont comptées [28].

Plasmides et nombres de générations Des plasmides (brins d'ADN circulaires) qui ne se répliquent plus une fois dans l'organisme de la souris sont ajoutés aux bactéries. Ainsi, lorsque les bactéries se divisent dans la souris, les plasmides doivent choisir entre l'une ou l'autre des deux bactéries-filles. Toutes les bactéries

sont initialement porteuses du plasmides, puis la proportion de bactéries porteuses est divisée par deux à chaque cycle de réplication. Des expériences *in vitro* permettent de calibrer précisément la correspondance entre dilution des plasmides et nombre de cycles de réplication. De cette manière, on a accès au nombre moyen de cycles de réplication qu'a connu la population bactérienne au cours de l'expérience.

WITS Des souches isogéniques marquées (WITS pour *Wild type Isogenic Tagged Strains* [29]) sont également utilisées. Ces bactéries se comportent identiquement aux bactéries non-marquées et peuvent être identifiées par PCR quantitative grâce à une séquence d'ADN spécifique non-codante ajoutée à leur génome (dans les données exploitées ici, 7 WITS différents sont utilisés). Seul un petit nombre de WITS sont inoculés à la souris, de sorte que les effets de la stochasticité des processus impliqués soient observables sur leur distribution [29, 25, 26]. Par exemple, un scénario avec de forts taux de réplication et d'élimination mènera à une plus grande variabilité dans la distribution de WITS qu'un scénario avec de faibles taux de réplication et d'élimination.

1.1.2 Méthodes

Dans cette partie on considère de manière abstraite un modèle "zéro" dont les composantes seront justifiées dans la section suivante, où la taille initiale de la population résulte d'un tirage Poissonien de moyenne βN_0 et où la dynamique après $t = 0$ suit un processus Markovien en temps continu avec un taux de réplication r et un taux d'élimination c.

Méthodes analytiques L'équation maîtresse sur la distribution de probabilité $P(n,t)$ que n bactéries soient présentes dans le cæcum au temps t s'écrit :

$$\frac{\partial P(n,t)}{\partial t} = (-cn - rn)P(n,t) + c(n+1)P(n+1,t) + r(n-1)P(n-1,t)$$

En sommant cette équation sur tous les n, on obtient une équation aux dérivées partielles sur la fonction génératrice de cette distribution, qu'on résout. On trouve l'expression suivante pour la fonction génératrice :

$$g(z,t) = \exp\left[\frac{\beta N_0 (r-c)(z-1)e^{(r-c)t}}{rz - c - (z-1)re^{(r-c)t}}\right]$$

On peut ensuite extraire les moments de la distribution de probabilité à l'aide des dérivées de la fonction génératrice.

Méthodes computationnelles Afin de simuler la dynamique de ce modèle, on utilise des algorithmes de type Gillespie. Les transitions (dans notre cas : naissance ou mort d'un individu) sont prises en compte les unes après les autres. A chaque pas de temps, des nombres aléatoires sont générés afin de déterminer

l'intervalle de temps ainsi que la prochaine transition qui va se produire. Pour ce travail, j'ai utilisé l'outil *adaptivetau* du langage R, qui est une version approchée de l'algorithme de Gillespie permettant de réaliser plusieurs transitions à la fois et ainsi d'accélérer le temps de calcul.

1.2 Modèles à une population

Les bactéries inoculées à la souris sont prélevées d'une solution, ce que l'on modélise par un tirage Poissonnien de moyenne N_0. Il est ensuite possible que l'acidité de l'estomac représente une barrière que toutes les bactéries ne parviennent par à franchir vivantes, on considère donc que chaque bactérie a une probabilité β de la traverser et de s'établir dans le cæcum. Le nombre initial de bactérie résulte donc d'un tirage Poissonnien de moyenne βN_0. On considère ensuite un taux fixe de réplication plutôt qu'un temps fixe de division pour des raisons de simplicité (cette hypothèse est discutée plus en détail en annexe). Dans les simulations, on considère une saturation de la réplication lorsque la capacité maximum du cæcum est atteinte, par le biais d'un terme de croissance logistique. Dans l'étude analytique en revanche, on néglige cette saturation (on verra plus loin que les observables que l'on considère dépendent surtout de la dynamique au tout début de l'infection, ce qui permet de négliger la saturation).

On peut déterminer le taux maximal initial de réplication (qui est une quantité robuste dépendant peu des spécificités de l'expérience) grâce notamment au taux de réplication maximal *in vitro*. La proportion d'étiquettes génétiques perdues au cours de l'expérience peut s'écrire comme le zéro de la fonction génératrice. Cette expression dépend à la fois de β et de c. On cherche donc une seconde observable pour séparer les effets de β et de c.

1.2.1 Une nouvelle observable

On cherche une observable afin de caractériser la variabilité dans la distribution des étiquettes génétiques. L'evenness a d'abord été privilégiée pour des raisons de continuité avec des travaux précédents de nos collaborateurs. Cependant, cette observable est difficilement manipulable analytiquement. La variance est une quantité plus naturelle pour le calcul. Nous avons montré qu'en fait, elle dépend de la même combinaison des paramètres β et c que la probabilité de perte. Elle ne contient donc *a priori* pas d'information différente par rapport à la probabilité de perte. Elle peut néanmoins avoir une autre utilité : comparer les estimations des paramètres possibles en utilisant la perte des WITS versus leur variabilité et vérifier que ces estimations sont compatibles. Par ailleurs, afin de prendre correctement en compte la variabilité initiale dans la distribution des WITS, on regardera plutôt la variance sur les taux de croissance des populations individuelles de WITS.

1.2.2 Résultats

Stratégie pour l'estimation des paramètres Dotés de cette nouvelle observable, on peut maintenant explorer l'espace paramétrique (β, c). En chaque point, on peut calculer ou réaliser une simulation pour obtenir la valeur attendue dans le cadre du modèle étudié pour les trois observables suivantes : la proportion de perte des étiquettes génétiques, la variance sur le taux de croissance, et le taux de croissance moyen (qui vaut $\beta 2^G e^{-ct}$ avec G le nombre moyen de cycles de réplications estimé par la donnée de la dilution des plasmides). Puis les valeurs expérimentales de ces trois observables définissent des courbes de l'espace paramétrique qui correspondent aux ensembles de paramètres autorisés par chacune des observables. Dans certains cas, des incertitudes sur ces courbes de niveaux peuvent être estimées (voir figure 1).

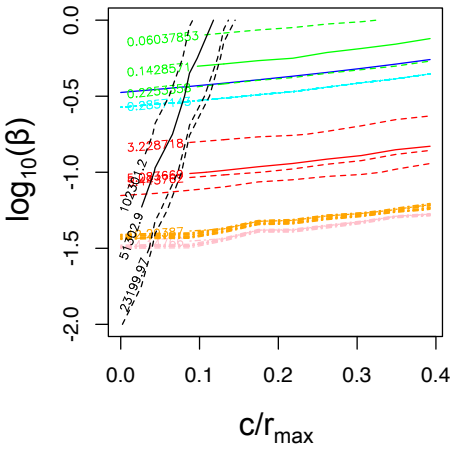


FIGURE 1 – Courbes de contour des valeurs expérimentales des différentes observables dans l'espace des paramètres (β, c) pour une expérience démarrant avec un inoculum de taille 10^3. *Pour la variance* : rouge pointillé : contour de la variance pour trois souris et rouge trait plein : pour la valeur moyenne. Rose pointillé et orange pointillé : incertitudes sur ce contour estimées de diverses manières, voir détails dans le texte principal. *Pour la perte des WITS* : bleu et vert : courbes de contour pour la proportion de WITS expérimentalement perdue, estimées de diverses manières (voir détails dans le texte principal). Vert pointillé et cyan : incertitudes sur ces contours, estimées de diverses manières (voir détails dans le texte principal). *Sur le taux de croissance* : **noir** : en utilisant l'expression $\beta 2^G e^{-ct}$, contour pour la valeur expérimentale du taux de croissance dans trois souris (traits pointillés) ainsi que la valeur moyenne (trait plein).

Discussion On observe tout d'abord que bien que l'on ne prenne pas en compte la saturation dans l'étude analytique, elle correspond parfaitement aux résultats des simulations. En effet, en ce qui concerne les observables de la proportion de WITS perdus et de la variance, les effets importants se produisent au tout début, lorsque les populations de WITS sont encore de petite taille. Par ailleurs, les

contours pour la variance et la probabilité de perte sont parallèles, ce qui était attendu puisque ces deux observables dépendent de la même combinaison des paramètres β et c. Quant au contour du taux de croissance moyen, il restreint l'espace des paramètres possibles à de faibles valeurs de c, ce qui était également attendu. Dans certaines expériences, on ne peut pas dire clairement si l'espace entre les contours de la variance et de la probabilité de perte peuvent être compris dans le bruit attendu ou non. Afin d'explorer de plus grandes possibilités de combinaison des observables, on se tourne donc vers des modèles à plusieurs sous-populations.

1.3 Modèles à deux sous-populations

1.3.1 Arguments en faveur d'un modèle à deux sous-populations

Plusieurs facteurs biologiques pourraient expliquer la coexistence au sein d'une population d'une même souche bactérienne de plusieurs sous-populations suivant des dynamiques différentes. On pourrait par exemple imaginer que la spatialité joue un rôle : par exemple, si les nutriments ne sont pas bien répartis dans le cæcum, il pourrait y avoir des zones où les bactéries se répliquent plus que d'autres. Il peut aussi exister plusieurs états phénotypiques dans la population bactérienne : une partie de la population peut par exemple exprimer un facteur de virulence, ou un flagelle, tandis que l'autre non. Ces différences pourraient mener à des dynamiques différentes, et notamment des taux de réplications différents.

Par ailleurs, comme les nombres initiaux de WITS sont petits, si deux sous-populations se répliquent à des taux différents, alors il peut arriver que toutes les bactéries marquées avec un premier WITS soient à croissance rapide, tandis que toutes les bactéries marquées avec un autre WITS soient à croissance lente. Dans ce cas, on arriverait à la fin de l'expérience à une plus grande variance dans la distribution des WITS que si toutes les bactéries se répliquaient au même taux moyen : ainsi, pour un même nombre de WITS perdus, on devrait avoir accès à une plus grande plage de valeurs pour la variance.

1.3.2 Résultats

Le calcul des différentes observables peut être adapté au cas avec deux sous-populations. On note α le ratio entre les deux taux de réplication (on définit $\alpha > 1$) et q la proportion initiale de bactéries à croissance rapide. Ici, on fixe tous les paramètres sauf α et q grâce à la donnée des valeurs expérimentales pour deux observables : le taux de croissance et la probabilité de perte. Une seule observable reste libre : la variance. On explore donc les valeurs attendues (analytiquement et par des simulations) pour la variance dans l'espace des paramètres (α, q), et l'on compare ces valeurs à la valeur expérimentale. Deux exemples de cartes paramétriques sont présentés figure 2.

Plusieurs expériences présentent des mesures de dilution de plasmides assez peu compatibles avec la mesure des tailles initiale et finale des populations, ou des problèmes expérimentaux. Deux expériences semblent cohérentes avec le modèle

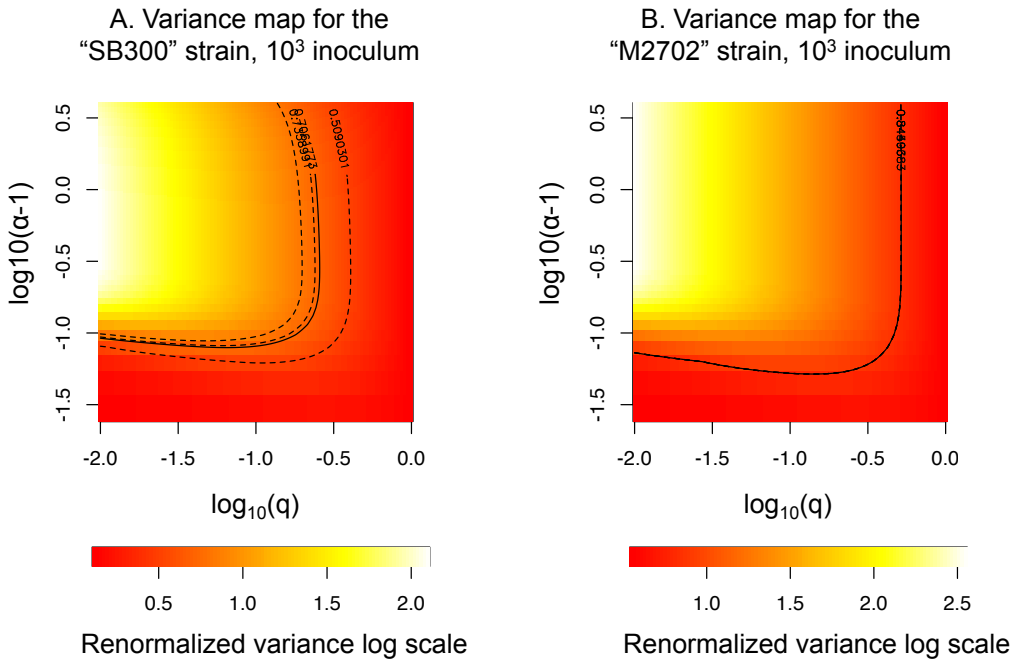


FIGURE 2 – Cartes de la variance attendue dans le cadre du modèle à deux sous-populations, pour deux expériences commençant avec un inoculum de taille 10^3. En abscisse on a $\log_{10}(q)$, avec q la proportion initiale de bactéries à croissance rapide ; en ordonnées, on a $\log_{10}(\alpha-1)$, avec α le ratio entre les deux taux de réplication. Les contours montrent les points de l'espace paramétrique où l'on retrouve la valeur expérimentale de la variance (pointillés pour trois souris et trait plein pour la moyenne). Le simple fait que ces lignes apparaissent sur la carte montre que certaines combinaisons des paramètres (q, α) permettent de retrouver les valeurs expérimentales des trois observables à la fois, ce qui n'était pas le cas dans le cadre du modèle à une seule population.

à deux sous-populations, bien que pour l'une d'entre elles les paramètres estimés soient plus faibles que l'ordre de grandeur attendu. Cependant, un plus grand nombre de données serait nécessaire afin de conclure plus généralement sur le sujet, comme il sera discuté dans la conclusion qui suit.

Conclusion

Dans cette partie, j'ai présenté la partie de mon travail qui porte sur les dynamiques de colonisation d'une population bactérienne au début d'une infection intestinale. J'ai développé des modèles stochastiques de dynamiques de population en systèmes ouverts, qui visent à déterminer des paramètres biologiquement pertinents de l'infection (comme les taux de réplication et d'élimination, et la probabilité pour une bactérie de s'établir dans l'organisme et de prendre part à l'infection) à partir de données indirectes. Dans une première section 1.1, j'ai présenté les données quantitatives sur l'infection de la souris par la salmonelle

qui ont motivé cette étude (essentiellement les nombres initiaux et finaux de bactéries, ainsi que les distributions initiales et finales de marqueurs génétiques), ainsi que les méthodes générales utilisées par la suite (à la fois analytiques, avec principalement des processus de branchement, et computationnelles, avec principalement des simulations de Gillespie). Dans une deuxième section 1.2, j'ai étudié un premier modèle à une population, dont la taille de la population est initialisé par un tirage Poissonnien et qui suit ensuite un processus de naissance et de mort de Markov en temps continu. Dans ce cadre, j'ai cherché l'observable optimale pour caractériser la variabilité dans la distribution des étiquettes génétiques, qui ont la particularité de partir de tailles de populations inégales, et ai montré que le variance du taux de croissance renormalisée était une observable adéquate. J'ai vérifié la cohérence des estimations de paramètres s'appuyant sur les différentes observables et ai montré que dans certains cas, on ne peut pas conclure clairement quant à cette cohérence. A partir d'arguments biologiques qui viennent soutenir cette hypothèse, et en me basant sur l'idée que cela pourrait mener à de plus grandes possibilités de combinaison des observables, j'ai ensuite développé dans une troisième section 1.3 un modèle avec deux sous-populations suivant la même dynamique, mais avec des taux de réplication différents. J'ai montré que ce type de modèle explique très bien certaines expériences, mais pas toutes, et en raison de la faible quantité de données considérées ici, on ne peut pas conclure clairement quant à la coexistence réelle ou non de plusieurs sous-populations.

Dans ces modèles, on a toujours considéré que les bactéries se répliquaient avec un taux fixe, pour des raisons de simplicité. Cependant, des temps fixes de réplication sont plus proches de la réalité. En appendice D j'ai étudié un modèle simple à une population, identique au modèle simple à une population étudié en section 1.2, mais où toutes les bactéries se répliquent au bout d'un temps fixe τ à la place de se répliquer à un taux fixe r. Les nouveaux paramètres sont ajustés afin que les tailles moyennes de populations soient identiques dans les deux modèles au bout du même temps. J'ai déterminé les nouvelles expressions pour les observables considérées (variance sur le taux de croissance renormalisée et probabilité de perte), qui sont toutes les deux modifiées de manière complexe : les effets sur l'estimation des paramètres ne sont pas triviaux, et dépendent des expériences considérées, on ne peut donc pas tirer de conclusion claire quant à l'effet de l'hypothèse "taux fixe de réplication" dans nos modèles.

Une autre piste concerne une étape des expériences qui a été identifiée comme étant source de variabilité additionnelle : avant les mesures de q-PCR des fréquences des différents WITS, il y a une étape préalable d'amplification de la population bactérienne. Des arguments théoriques indiquent que, sauf cas extrêmes, cette étape ne devrait pas contribuer significativement, mais ses effets devraient être étudiés expérimentalement plus en détail.

Une autre question est celle des plus grands nombres de sous-populations. Des arguments théoriques et des simulations préliminaires indiquent qu'on ne peut pas atteindre de plus grande variance avec 3 sous-populations ou plus que dans le cas avec deux sous-populations.

Pour conclure sur l'ensemble de données sur lequel on a testé nos méthodes,

une des plus grandes difficultés réside dans le fait que l'observable qui contient le plus d'information, la proportion de WITS perdus dans les expérience, est aussi celle qui est expérimentalement mesurée avec un échantillonnage insuffisant pour en avoir une valeur suffisamment fiable. Un bon moyen de contourner ce problème sans avoir recourt à des nombres d'expériences trop importants serait de développer un plus grand nombre de marqueurs génétiques (dans les données considérées ici, seuls sept marqueurs génétiques différents sont utilisés) de manière à ce que la probabilité de perte et les distributions de WITS soient mesurées de manière plus statistiquement significative. Cependant, même si l'on ne peut pas tirer de conclusion générale quant à ce jeu de données particulier, les méthodes développées dans cette partie sont suffisamment générales pour être applicables dans d'autres cas de dynamiques de populations.

2 Mécanismes d'action de l'immunoglobuline A lors de la réponse immunitaire

Introduction

La surface du système digestif est très grande [30][31], couverte d'une couche de cellules épithéliales essentielles à l'absorption des nutriments, mais qui constitue aussi une porte d'entrée pour de nombreux pathogènes. Contrairement à l'intérieur de l'organisme, où la présence de n'importe quelle bactérie est anormale, le lumen de l'appareil digestif abrite un microbiote important : il y a au moins autant de bactéries dans l'appareil digestif humain qu'il n'y a de cellules humaines composant le corps humain [1]. Le microbiote est donc important en taille, mais aussi par sa fonction : les bactéries sont nécessaires pour casser et absorber certains nutriments, et peuvent agir contre de potentiel agents pathogènes par effets de compétition [2]. Pour ces raisons, dans le système digestif, le système immunitaire de l'hôte doit à la fois lutter contre les bactéries dangereuses et préserver celles qui sont bénéfiques.

Le principal effecteur du système immunitaire adaptatif dans les intestins est un anticorps spécifique, l'Immunoglobuline A (IgA). Cet anticorps s'attache à des antigènes-cibles précédemment rencontrés (lors d'une infection ou grâce à une vaccination), et permet de protéger de l'infection [32]. Si les mécanismes moléculaires complexes permettant la sécrétion de l'anticorps ont été étudiés en détail [33], on commence seulement à développer notre compréhension des mécanismes d'action qui les rendent efficaces une fois dans le lumen intestinal. Dans cette partie, on vise à éclaircir certains aspects de ces mécanismes. Dans une première section, le concept de croissance enchaînée sera exposé ; je reprendrai l'élément de preuve par lequel j'ai contribué à l'étude de Moor *et al.* [19]. Dans une deuxième section, on explorera les conséquences de ce phénomène à l'échelle de l'hôte, et l'on montrera qu'il peut permettre au système immunitaire pour réguler la composition du microbiote. Enfin, dans une dernière section on s'intéressera aux conséquences de la croissance enchaînée sur l'évolution de la résistance aux antibiotiques à l'échelle d'une population d'hôtes.

2.1 Une idée nouvelle en immunologie : la croissance enchaînée

Nos collaborateurs ont développé un protocole de vaccination de la souris basé sur l'inoculation de grandes quantités de bactéries tuées [34]. Cette vaccination déclenche la sécrétion en grandes quantités d'IgA dans le lumen intestinal, et les souris ainsi vaccinées sont ensuite protégées d'infections ultérieures. Cependant, l'IgA n'a pas de pouvoir bactéricide : en fait, dans les souris vaccinées, les nombres de bactéries sont inchangés par rapport aux souris non-vaccinées. En revanche, on observe des agrégats de bactéries, trop gros pour approcher la surface de l'intestin et interagir avec, étape essentielle aux bactéries pour déclencher les étapes suivantes de l'infection. Les IgA possédant plusieurs sites de fixation, il était connu que ces anticorps pouvaient agglomérer les bactéries entre elles. Ce processus d'agglutination avait cependant toujours été pensé comme résultant de la rencontre aléatoire des bactéries et des IgA diffusant dans le contenu intestinal. Or, dans les expériences, avec une concentration initiale de bactéries dans la limite haute de ce qui peut provoquer une intoxication alimentaire, le temps typique de rencontre d'une bactérie avec une autre est d'environ 30 heures ; cette hypothèse ne permet donc pas d'expliquer comment les souris sont protégées. En fait, ce que montre l'étude de Moor *et al.* [19], c'est que les bactéries-filles restent collées entre elles par les IgA lors de la division. Les agrégats bactériens qui en résultent sont donc clonaux.

Plusieurs éléments de preuve ont été nécessaires. J'ai contribué à l'un d'eux en développant un modèle simple de croissance enchaînée basé sur la donnée des distributions de WITS. En effet, si les bactéries se développent et sont éliminées sous forme d'agrégats clonaux, tout se passe essentiellement comme si la taille effective de la population était réduite (voir figure 3). Mon modèle a ainsi permis de prédire la réduction de diversité observée dans la population bactérienne.

2.2 La croissance enchaînée comme moyen de régulation de la composition du microbiote

Introduction

Dans les intestins, le système immunitaire doit pouvoir protéger l'organisme de l'intrusion de bactéries pathogènes, tout en protégeant celles qui sont bénéfiques. Cette distinction peut s'avérer d'autant plus délicate que certaines bactéries pourtant très apparentées présentent des comportements tout à fait différents dans les intestins. Et comme la sur-croissance de n'importe quelle source bactérienne peut mettre en danger l'équilibre de la flore intestinale, il est nécessaire que le système immunitaire puisse maintenir un homéostat sur la composition du microbiote.

Or, les agrégats bactériens résultants de la croissance enchainée (voir section 2.1) peuvent casser, vraisemblablement sous l'effet des contraintes mécaniques imposés par les flux. L'interaction entre la cassure des agrégats et la croissance bactérienne pourrait avoir des conséquences importantes. Considérons un modèle simplifié à l'extrême où toutes les bactéries se répliquent au bout d'un temps τ_{div} et où tous les liens formés par les IgA se cassent au bout d'un temps τ_{break}

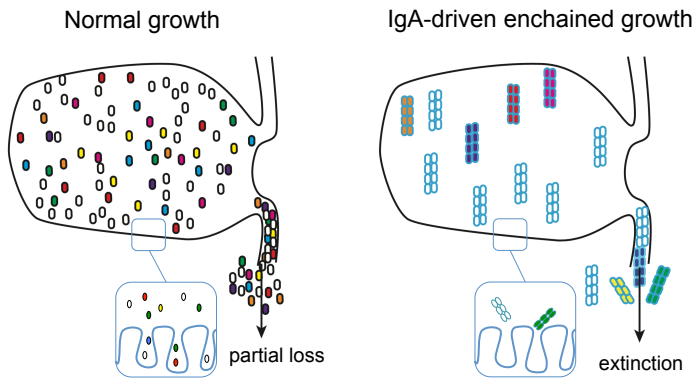


FIGURE 3 – Croissance bactérienne dans une souris naïve (gauche) ou dans une souris vaccinée (droite) (source : figure 2a in [19]). Les couleurs représentent les différentes souches de WITS, les contours bleus représentent les IgA qui recouvrent la surface. Les encadrés zooment sur la surface et montrent que les agrégats ne peuvent pas interagir avec (soit parce qu'ils sont trop gros pour pénétrer les cryptes, soit parce qu'ils sont trop gros pour parvenir à pénétrer la couche de mucus).

(voir figure 4). Alors si $\tau_{div} > \tau_{break}$, les liens se cassent avant que les bactéries aient le temps de se diviser, et il ne se forme pas d'agrégats de taille supérieure à 2. En revanche, si $\tau_{div} < \tau_{break}$, alors les bactéries se divisent toujours avant que les liens ne cassent, des clusters de taille toujours plus grande se forment et il n'y a plus aucune bactérie libre. Dans ce modèle, ce sont les bactéries qui se répliquent le plus vite, donc les plus susceptibles de déséquilibrer l'équilibre de la flore intestinale, dont l'organisme se protège en les piégeant sous forme d'agrégats. Dans la suite, on élabore cette idée à l'aide de modèles plus réalistes.

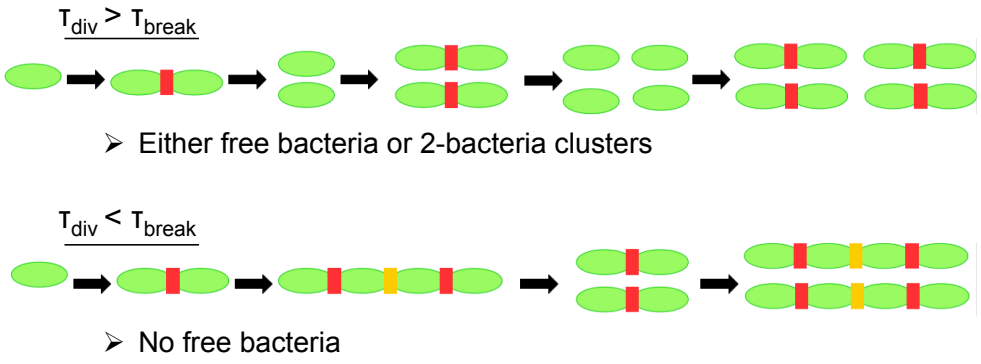


FIGURE 4 – Modèle simplifié de croissance et dislocation des agrégats, où toutes les bactéries se divisent après un temps τ_{div} et où tous les liens se cassent après τ_{break}.

2.2.1 Modèle de base

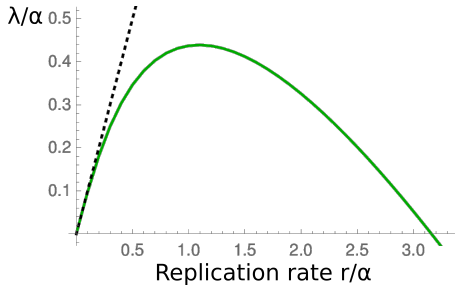
On considère d'abord un modèle de base, où les bactéries se répliquent à taux constant r et sont parfaitement enchaînées à la réplication, et où les liens se cassent tous avec un taux α. On considère qu'au début de l'infection, on peut négliger les rencontres aléatoires de bactéries ou d 'agrégats indépendants dus à la diffusion. On considère que les agrégats sont initialement formés de chaînes linéaires de bactéries (comme cela a été observé expérimentalement), sauf lorsqu'une chaîne casse ailleurs qu'aux extrémités : dans ce cas, on considère qu'elle se recombine de manière plus complexe (par les côtés, comme cela a également été observé), et que comme dans ces clusters complexes, les bactéries ont en moyenne plus d'attaches, elles sont moins susceptibles de se détacher. On considère donc que seuls les agrégats linéaires participent à la dynamique des bactéries libres, les agrégats plus complexes ne contribuent plus au système.

En partant de ces hypothèses, on peut écrire l'ensemble d'équations différentielles déterministes sur les nombres moyens n_i de chaînes de bactéries linéaires de taille i. Ces systèmes d'équations sont ensuite étudiés à la fois à l'aide de résolutions numériques du système complet (en coupant le système arbitrairement à une taille maximale de chaîne n_{max} qui n'impacte pas la dynamique), et à l'aide d'approximations qui permettent d'obtenir des expressions approchées de la distribution des tailles de chaînes. On se concentre en particulier sur le taux de croissance de la population de bactéries libres ; en effet, seules les bactéries planctoniques peuvent approcher la paroi de l'intestin et déclencher les étapes suivantes de l'infection.

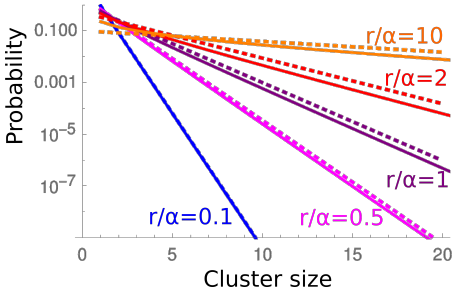
Les résultats pour le modèle de base sont présentés figure 5. Le taux de croissance des bactéries libres est drastiquement réduit par rapport au cas où il n'y a pas de croissance enchaînée, et devient non-monotone : plus r est élevé, plus on peut potentiellement produire de bactéries libres, mais lorsque r devient grand par rapport à α, les agrégats se forment, puis se cassent pour former des agrégats plus complexes dont aucune bactérie libre ne peut s'échapper. En ce qui concerne la distribution des tailles de clusters, celle ci décroit en loi de puissance, et décroit d'autant plus rapidement que r est petit devant α (auquel cas très peu d'agrégats se forment).

2.2.2 Variantes de modèle de base

Des variantes du modèle de base sont ensuite étudiées, afin de vérifier la robustesse des résultats par rapport aux hypothèses faites dans le modèle de base. En utilisant toujours les mêmes méthodes, on étudie ainsi : un modèle où l'enchaînement n'est pas parfait lors de la réplication, où les bactéries-filles ont donc une probabilité non-nulle de s'échapper ; un modèle avec un temps fixe au lieu d'un taux fixe de réplication ; un modèle où les chaînes lorsqu'elles cassent ailleurs qu'aux extrémités ont une probabilité non-nulle de donner deux chaînes linéaires au lieu d'un amas plus complexe qui ne contribue plus au système ; et un modèle où le taux de cassure des liens augmente avec la force exercée sur les liens par les contraintes hydrodynamiques.



A. Taux de croissance de la population de bactéries libres λ en fonction du taux de réplication r, avec α pour unité. Résultat numérique (en vert), et limite sans agrégation ($\lambda = r$ en pointillés noirs).



B. Distribution de la taille des clusters pour différents r. Traits pleins : résultats numériques, pointillés : approximations (confondue avec les résultats numériques pour $r/\alpha = 0.1$).

FIGURE 5 – Modèle de base. Pour les résolutions numériques, $n_{max} = 40$.

2.2.3 Comparaison aux données et discussion

Sauf dans le cas très particulier où les deux chaînes résultant d'une cassure au milieu d'une chaîne peuvent toujours s'échapper, le taux de croissance de la population de bactéries libres est toujours réduit par rapport au taux de réplication bactérien. De plus, dans la plupart des modèles étudiés (sauf si plus de la moitié des chaînes s'échappe après cassure ou si la probabilité de s'échapper au moment de la réplication est grande), le taux de croissance de la population de bactéries libre est une fonction non-monotone du taux de réplication, et il existe un taux de réplication fini qui maximise le taux de croissance. A plus grands taux de réplication, les bactéries restent piégées dans des agrégats complexes qui ne contribuent plus à la dynamique générale du système.

En ce qui concerne les distributions des tailles d'agrégats, celles-ci présentent toutes une décroissance qualitativement similaire, à l'exception près du modèle avec des temps fixes de réplications, qui présente des pics correspondants aux tailles qui sont des puissances de 2. L'analyse d'images du contenu cæcal de souris vacciné semble indiquer une plus grande compatibilité avec ce modèle, ainsi qu'avec le modèle où le taux de cassure dépend des forces, qui présente une distribution plus étroite.

2.3 Conséquences de la croissance enchaînée pour l'évolution de la résistance aux antibiotiques

Introduction

Depuis la découverte de la pénicilline, la conception d'un nouveau traitement antibiotique a systématiquement été suivie de l'apparition d'une résistance à cet antibiotique [4]. Les antibiotiques représentent un outil essentiel pour la médecine, et sont très largement utilisés : un quart des Français en prennent chaque année [35, 36], et le bétail des exploitations agricoles est souvent traité de manière rou-

tinière. Or, le traitement n'est efficace que sur les bactéries sensibles. Au contraire, si certaines bactéries sont résistantes, la prise du traitement ne peut qu'augmenter la proportion de bactéries résistantes au sein d'un hôte. Par ailleurs, la prise d'antibiotiques contre une bactérie pathogène peut favoriser l'apparition de résistance chez d'autres bactéries du microbiote, en particulier dans les intestins. Dans cette partie, on développe des modèles multi-échelle qui visent à comprendre l'interaction entre prise d'un traitement antibiotique et dissémination de la résistance, en prenant en compte l'aspect d'enchaînement clonal dû à l'immunité.

En effet, on a vu à la section précédente que les bactéries connaissaient une croissance enchaînée dans les hôtes immunisés. Or si les bactéries sont présentes sous forme d'agrégats clonaux au sein de l'hôte, il est probable qu'elles soient également transmises sous cette forme à d'autres individus via la route fécale-orale, réduisant ainsi la diversité des populations transmises. Le but de cette section est d'étudier les effets de ce phénomène sur l'émergence de la résistance aux antibiotiques à l'échelle d'une population d'hôtes.

2.3.1 Modèle

On utilise un modèle multi-échelle où la dynamique au sein de l'hôte est décrite de manière déterministe (les nombres de bactéries sont typiquement très grands), alors que la transmission entre les hôtes est décrite de manière stochastique, à l'aide d'un processus de branchement (les nombres transmis peuvent être faibles). On considère que les individus sont toujours initialement infectés d'un total N de bactéries, et qu'il s'agit exactement de la taille d'un agrégat au sein des hôtes immunisés.

Dynamique au sein de l'hôte Les hôtes sont soit immunisés (les bactéries qui le colonisent forment des agrégats) soit naïfs (elles n'en forment pas). Ils peuvent être traités avec des antibiotiques, auquel cas si l'hôte n'était colonisé que de bactéries sensibles toutes les bactéries sont tuées, et sinon l'infection au sein de l'hôte ne se fait plus qu'avec des bactéries résistantes. Des mutations peuvent se produire au sein de l'hôte, avec une probabilité μ_1 (resp. μ_2) pour une bactérie-fille de se transformer de sensible à résistant (resp. de résistant à sensible) lors de la réplication. Les bactéries résistantes le sont au coût d'un taux de croissance réduit d'un facteur $(1-s)$. Ainsi on écrit les équations différentielles régissant la dynamique au sein de l'hôte :

$$\frac{dS}{dt_g} = (1-\mu_1/\log(2))S + (1-s)R\mu_2/\log(2),$$

$$\frac{dR}{dt_g} = (1-s)(1-\mu_2/\log(2))R + \mu_1 S/\log(2),$$

que l'on peut résoudre exactement afin de connaître les proportions finales des différents types bactériens au sein de l'hôte au bout de G générations, avant l'étape de transmission.

Transmission Quand il n'y a pas de croissance enchaînée, les bactéries sont transmises indépendamment. Mais dans les hôtes immunisés, elles sont transmises sous forme d'agrégats. En l'absence de mutations, ou lorsqu'elles sont négligeables (notamment lorsque l'inoculum est mixte), les agrégats transmis sont principalement d'un seul type, soit totalement sensibles, soit totalement résistants. Lorsque l'inoculum est d'un seul type, il devient nécessaire de prendre en compte les mutations, et il est parfois nécessaire de prendre en compte la présence d'agrégats mixtes (notamment lorsque le nombre de générations G au sein de l'hôte n'est pas très grand devant le nombre de générations g nécessaire à l'élaboration d'un cluster de taille $N = 2^g$). Chaque individu naïf (resp. immunisé) transmet l'infection à un nombre d'hôtes qui suit une distribution de Poisson de moyenne λ (resp. λ').

2.3.2 Méthodes et équations

On utilise les processus de branchement pour décrire le début de l'infection [37, 38, 39], et on écrit donc les fonctions génératrices, avec $\wp_{i,\{n_0,n_1,....,n_N\}}$ la probabilité qu'un hôte initialement infecté avec i bactéries résistantes et $N-i$ bactéries sensibles infecte n_0 nouveaux individus avec 0 résistants et N sensibles, n_1 nouveaux individus avec 1 résistante et $N-1$ sensibles, etc. On suppose que le nombre de transmissions est distribué de manière Poissonienne et que les transmissions sont indépendantes les unes des autres. En notant f_{ij} la probabilité qu'une transmission d'un hôte initialement infecté de i bactéries résistantes contienne j bactéries résistantes, on obtient :

$$g_i(z_0,...,z_n) = \sum_{\{n_0,n_1,....,n_N\}} \wp_{i,\{n_0,n_1,....,n_N\}} z_0^{n_0}....z_N^{n_N}$$

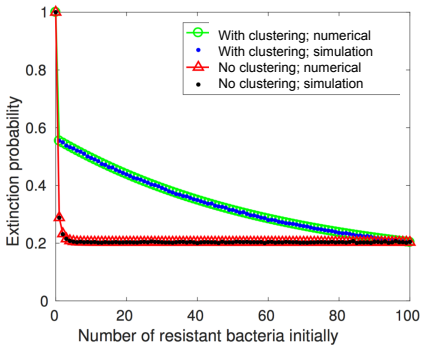
$$= \exp\left(-\tilde{\lambda}\sum_{j=0}^{N} f_{i,j}(1-z_j)\right)$$

avec $\tilde{\lambda} = \lambda$ pour un individu naïf et λ' pour un individu immunisé. Les probabilités d'extinction e_i (pour un "patient zéro" infecté de i bactéries résistantes) sont les points fixes de ces fonctions génératrices. Les f_{ij} doivent donc être explicités dans chacun des cas, puis l'on obtient un système d'équations linéaires que l'on peut résoudre numériquement et étudier à l'aide d'approximations analytiques.

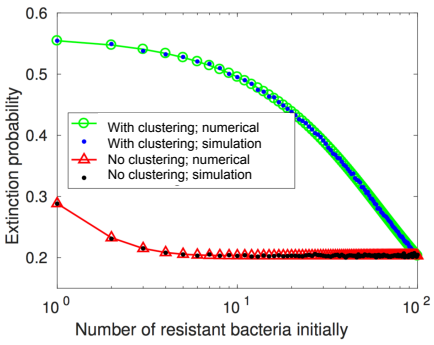
2.3.3 Résultats

En l'absence de mutations En l'absence de mutations, la résolution du système complet d'équations pour $\lambda = \lambda'$ et $q = q'$ permet de montrer que dans une population immunisée, la probabilité d'extinction de l'infection est plus grande (voir figure 6).

Effet de porteur sain Il est vraisemblable qu'un individu immunisé ne se sente pas malade. Dans ce cas, on peut imaginer qu'il reste moins isolé, et a donc



Échelle linéaire



Échelle logarithmique

FIGURE 6 – Probabilité d'extinction de l'épidémie en fonction du nombre de bactéries résistantes n ayant infecté le "patient zéro". Ici on a pris $\mu_1 = \mu_2 = 0$, $s = 0$, $N = 100$, $\lambda = \lambda' = 2$ et $q = q' = 0.55$, et n varie de 0 à $N = 100$.

la possibilité de transmettre l'infection à plus d'individus ($\lambda' > \lambda$). De plus, il est aussi possible qu'étant moins malade, il soit moins traité ($q' < q$). Toujours en l'absence de mutations, avec $\lambda' > \lambda$, l'effet de l'immunité sur la propagation peut être inversé, notamment lorsque le "patient-zéro" n'était initialement infecté que de bactéries résistantes.

Avec des mutations On considère en particulier le cas où le "patient-zéro" n'est initialement infecté que de bactéries sensibles, et où la résistance ne peut émerger que par une nouvelle mutation. On compare la probabilité de survie de l'épidémie dans une population entièrement naïve par rapport à une population entièrement immunisée en considérant la quantité :

$$\frac{1 - e_{0,naive}}{1 - e_{0,immune}}.$$

Dans ce cas il est important de distinguer le cas où les clusters mixtes sont importants à prendre en compte des cas où l'on peut les négliger. En écrivant les équations complètes que l'on résout numériquement, mais aussi par des approximations pour des cas limites, on montre que dans deux régimes différents (celui d'un petit nombre de génération avec s suffisamment petit, et celui d'un grand nombre de générations avec s pas trop petit), lorsque la souche bactérienne ne peut pas se propager en l'absence de mutations, la probabilité de survie de l'épidémie est réduite dans la population immunisée par rapport à une population naïve, d'un facteur qui dépend des détails de l'infection.

Discussion

Ainsi, dans la plupart des cas, on observe une réduction de la probabilité d'émergence d'une épidémie dans une population immunisée par rapport à une population naïve. L'idée principale qui explique cet effet est que pour un individu immun

les transmissions sont d'un seul type, et ainsi si le nombre moyen de bactéries transmises est conservé, le nombre moyen d'hôtes auquel on n'a transmis aucune bactérie résistante est plus élevé dans la population immunisée.

Conclusion

Dans cette thèse, j'ai présenté dans un premier temps la partie de mon travail qui porte sur les dynamiques de colonisation des populations bactériennes au début d'une infection de l'intestin. J'ai développé des modèles stochastiques de dynamiques de population en système ouvert, m'aidant à la fois de méthodes analytiques – comme les processus de branchements – et de méthodes numériques – comme les simulations de Gillespie. Le but de cette approche est d'inférer des paramètres de l'infection pertinents d'un point de vue biologique (par exemple des taux de réplication et d'élimination, et la probabilité pour une bactérie de s'établir dans l'organisme et de participer à l'infection) à partir de données indirectes (la dilution de plasmides qui ne se répliquent pas dans l'organisme de la souris, les nombres initiaux et finaux de bactéries, ainsi que les distributions initiales et finales d'étiquettes génétiques marquant les bactéries). Dans un premier temps, j'ai étudié un modèle à une population : le nombre initial est tiré d'un processus Poissonnien, puis la population suit un processus de Markov de naissance et de mort en temps continu. Dans ce contexte, j'ai cherché l'observable optimale afin de caractériser la variabilité de la distribution finale des étiquettes génétiques, qui ont la particularité d'avoir une distribution initiale inégale. J'ai vérifié si les paramètres estimés à partir des observables du taux de croissance moyen, de la variance renormalisée du taux de croissance et de la proportion d'étiquettes génétiques perdue étaient cohérents, et ai montré que pour certaines expériences, on ne peut pas conclure clairement quant à la cohérence du modèle. En m'appuyant sur des arguments biologiques et sur l'idée qualitative qu'un tel modèle devrait permettre de plus grandes possibilités pour combiner les différentes observables, j'ai ensuite développé des modèles à deux sous-populations suivant chacune le même type de dynamique, mais avec des taux de réplications différents. J'ai montré que ce type de modèles permet bien d'expliquer certaines expériences, mais pas toutes ; et en raison de la faible quantité de données, on ne peut pas tirer de conclusion claire quant à la coexistence de plusieurs sous-populations.

La deuxième partie de ma thèse porte sur les mécanismes qui permettent à la réponse immunitaire d'être efficace. J'ai d'abord repris les résultats de l'étude de Moor *et al.* [19] à laquelle j'ai contribué. Elle montre que le principal effecteur du système immunitaire dans les intestins, l'immunoglobuline A (un anticorps produit en grandes quantités après une infection ou une vaccination), enchaîne les bactéries-filles en agrégats clonaux lors de la réplication, agrégats que l'on pensait auparavant le résultat de rencontres aléatoires de bactéries dans les intestins (rencontres qui sont en fait assez rares en raison des faibles concentrations initiales en bactéries typiques lors d'une intoxication alimentaire). Ce mécanisme appelé *croissance enchaînée* suffit à protéger la souris en empêchant les bactéries d'interagir avec l'épithélium et de le traverser pour coloniser le reste de l'organisme.

J'ai contribué à mettre en évidence ce phénomène en construisant un modèle simple permettant de prédire la diminution de diversité au sein de la population bactérienne qui en résulte. J'ai ensuite cherché à déterminé les conséquences de ce processus. A l'échelle de l'hôte, j'ai étudié l'interaction entre la croissance des agrégats et leur fragmentation au sein des individus immuns, à l'aide de modèles basés sur des équations différentielles étudiées à la fois numériquement et avec des approximations analytiques. J'ai montré que la croissance enchaînée touche vraisemblablement d'avantage les bactéries à croissance rapide, c'est à dire les plus susceptibles de perturber l'équilibre de la flore intestinale, et que ce mécanisme pourrait donc permettre au système immunitaire de réguler la composition du microbiote. A l'échelle de la population d'hôtes, et dans le contexte de l'évolution de la résistance aux traitements antibiotiques, si les bactéries sont transmises sous forme d'agrégats clonaux (soit complètement résistants soit complètement sensibles) plutôt que sous forme de paquets aléatoires de bactéries sensibles et résistantes, la probabilité qu'une bactérie résistante soit transmise à un individu particulier est diminuée. A l'aide de modèles multi-échelles basés sur des processus de branchement, j'ai quantifié les variations dans la probabilité d'émergence d'une infection entre une population immunisée et une populations naïve.

Les deux parties de cette thèse ont été motivées par l'étude de données quantitatives d'expériences d'infections de souris par la salmonelle. Les mécanismes immunitaires présentés dans cette thèse sont pourtant de portée bien plus générale : en effet, le système immunitaire de la souris est proche de celui de nombreux autres vertébrés, notamment de celui des humains, et la croissance enchainée concerne *a priori* de nombreux autres micro-organismes (et a par exemple déjà été observée pour *E. coli* [19]). Bien sûr, des expériences complémentaires seraient nécessaires afin de conclure plus généralement quant à la portée exacte de ce phénomène. A long terme, la croissance enchaînée pourrait être maitrisée grâce à la vaccination orale, et constituer un moyen de lutte pour réduire l'utilisation des antibiotiques et ralentir l'évolution et la propagation de souches résistantes. Par ailleurs, les outils de dynamiques de populations développés dans la première partie pourraient être appliqués à de plus grands ensembles de données sur des infections bactériennes des intestins avec divers souches et divers animaux hôtes, mais également à des systèmes différents en écologie, pas nécessairement à la même échelle. Dans la deuxième partie, l'étude de la croissance et de la fragmentation des agrégats présente en soi un problème général de physique statistique, qui s'est déjà révélé utile dans d'autres contextes [21][22]. Ainsi, dans cette thèse on utilise des outils de physique statistique pour identifier des mécanismes génériques en biologie, et les propriétés essentielles pour en comprendre la portée.

Contents

Introduction

The human body is host to an important microbiota. Any anatomical compartment in direct contact with the environment, like the skin, the dental plaque, the saliva, the lungs, are home to many microorganisms. It has been estimated[1] that there are about 10^{13} bacteria in a human body, *i.e.* approximately as many as the number of human cells composing the body. Most of them ($\simeq 99\%$) are located in the digestive track. The presence of these ecosystems is essential to the good health of an individual, because they fulfill several important functions [2]: through competition effects, they protect the host against other potentially pathogenic invaders. In the digestive tract, they also help digestion by breaking specific nutrients, in a mutualistic interaction between the host and its commensal bacteria (which feed on nutrients ingested by the host). But bacteria are of course also responsible for many pathologies, when a virulent strain manages to overgrow the other populations, disrupts the functioning of the ecosystem and triggers inflammation (causing dental cavities, diarrhea, styes, etc), and sometimes, enters the organism and spreads to various organs, in serious systemic infections. Bacterial infections are responsible for millions of deaths every year: according to the World Health Organization, they contribute largely to three of the top ten causes of death worldwide in 2016 with pneumonia, diarrheal diseases and tuberculosis [3]. Bacterial infections are all the more a public health issue that antibiotic resistance seems now to increase faster than the rate of new drugs conception [4], while antibiotics constitute the main available tool to fight against them. Thus developing our knowledge of bacterial populations and bacterial infections is crucial.

To study bacterial populations, various approaches exist. A first range of approaches could be qualified as *bottom up*, starting from the simplest possible systems *in vitro* and introducing progressively mastered levels of complexity to mimic natural conditions. These microbiology approaches have developed in parallel to the emergence of more sophisticated experimental techniques. Thus bacteria were first extensively studied in suspensions, because it is the most convenient way to culture them. Then many studies focused on surface effects, typically on Petri dishes [5]. The development of microfluidics now allows to investigate the behavior of bacterial colonies both at the scale of interacting mixed communities (for example to study biofilm formation [6]) and at the single

cell level (for example to study the accumulation of new mutations[7]), in highly controlled environments [8].

Conversely, complementary approaches that could be qualified as *top down*, start from complex systems in natural conditions and try to analyze and separate the various factors driving the system's behavior. These approaches are essential because the level of complexity found in natural conditions and/or in living systems is very hard, or even impossible, to recreate artificially. First, reproducing the mechanical constraints felt in a natural environment, or *in vivo* for bacterial populations inside a host, supposes that these constraints are well characterized, which is not always the case. For example in the digestive tract, complex peristaltic motions aiming at the well-mixing of the digesta to favor nutrients absorption impose complex constraints on the very non-Newtonian content [9, 10]. Secondly, there could be chemical or enzymatic components, phages or other microorganisms, that play a role in the observed bacterial behavior but that are not necessarily well identified. And even when they are well identified, some molecular components which intervene in the immune response for example might be extremely hard to reproduce artificially or to isolate properly to add them in *in vitro* systems. Thus, various animal models have been developed to study bacterial infections directly in the host. The mice model for Salmonella diarrhea [11] used in the experimental studies that motivated this present work presents many advantages. First, mice immune system presents a similar level of complexity to the human one, and in particular harbors a developed adaptive immune system. Then several aspects of the infection can be controlled, like the stomach acidity. Mice can be grown in sterile environment so that the composition of their microbiota can be designed with a lower complexity than in wild-type animals. Mice can also be genetically engineered with specific immunology deficiencies, and all these tools allow a better separation of the various effects studied.

The study of bacterial populations in natural conditions and possibly in complete living animals requires modeling for a wide range of questions, and these systems represent a huge field of possible applications and interest for modelers and physicists.

A first example is linked to the progresses in sequencing technologies that now allow to sequence and analyze whole complex microbiomes. To infer the different types of interactions between the different species in presence, the contribution from the networks community is absolutely essential [12]. In a reverse way, bacterial complex systems can also represent an interesting way to confront networks and game theory models to the real world, for example in the context of evolution of structured populations [13]. Modeling can also decipher to which extent the composition of a microbiota is the product of an active selection or a neutral random process [14].

A second example is the contribution of physicists to new insights in immunology using the concept of information, in particular regarding pathogens

recognition that triggers the secretion of new effectors, and the building of diverse repertoirs of antigen receptors [15, 16].

Regarding host colonization, except in very particular configurations [17, 18], it is impossible to follow the infectious process in details without invasive, and hence, disruptive measurements. Indirect measures are favored instead, and this is where modeling becomes an essential tool to the understanding of the experimental observations. The animal organism can be seen as a "black box", and the role of the modeler is to infer its internal rules from external observations. Then, in the context of infections, statistical physics is needed to study microbial population dynamics. Indeed, an infection can start from a single or very few microorganisms, thus requiring full-fledged stochastic modeling. These microorganisms then replicate to very large numbers, thus requiring coarse-grained descriptions.

From the point of view of the physicist, the digestive system is particularly interesting. First because of the hydrodynamics (there is a flow, the fluid is non-Newtonian), and also because it is still topologically "outside" the host body, thus there are fewer different immune effectors to take into account, and studying their physical effects seems more tractable. Indeed, immune effectors secreted in the gut lumen are essentially "lost" for the body. This may be a reason why, except in particular circumstances, few immune cells are secreted in the gut. Immune cells can have different states, react to many cues, and thus are complex to model. In contrast, although the mechanisms leading to its secretion are complex, Immunoglobulin A (IgA), a type of antibody and the main effector of the adaptive immune response secreted in the gut, is a molecule which concentration can be measured, and which effect is more easily characterized. Furthermore, these antibodies recognize one specific strain of bacteria, meaning that the bacterial population they interact with should essentially be homogeneous, which facilitates physical modeling. Thus the action of IgA can be more easily studied by biophysical models.

This thesis focuses on several aspects of bacterial population dynamics and interactions with the immune system in the gut. It is constituted of two distinct parts largely independent. In the first part, I present my work regarding the colonization dynamics of bacterial populations in early infection. I develop stochastic models (because the initial numbers of bacteria of interest are small) which aim at determining biologically relevant parameters, such as replication and elimination rates, and the probability for one bacteria to settle in the organism and participate in the infection. I start in the first chapter (1) with a brief presentation of the experimental data – relying in particular on bacteria labeling – that motivated this study, along with a brief review of the general statistical physics methods used thereafter – both analytical (mostly branching processes) and numerical (mostly agent-based Gillespie simulations). I then dedicate the next chapter (chapter 2) to the study of a first class of models with Poissonian

initialization followed by continuous time Markovian birth-death process. This study is also the occasion for a broader reflection on the optimal choice of observable to characterize the variability of a distribution for a set of random variables following the same dynamics but starting with different initial values. I show that in some cases, the parameters estimates obtained by using different observables on the same experimental data are not clearly consistent. This led me to the study of another class of models in chapter 3, with two distinct subpopulations following the same type of dynamics but with different parameters. Various biological factors can account for the coexistence of different bacterial dynamics among a population of same strain. I show that these models allow for a broader range of combinations of the different observables values which can very well explain some of the experiments, although considering the whole set of data, there is not enough evidence to get a clear conclusion.

The second part focuses on the physical mechanisms underlying the immune response in the gut. It starts with a recent finding that IgA, a specific kind of antibody and main effector of the immune system in the gut, enchains daughter bacteria in clonal clusters upon replication [19]. In a first chapter (chapter 4), I summarize the findings of this study and focus on my contribution to it with a model showing the reduced diversity in the bacterial population caused by this *enchained growth* phenomenon, and show how this mechanism suffices to protect the organism from infection. Then the two following chapters (chapter 5 and chapter 6) explore the consequences of this phenomenon, at the scale of a host (chapter 5) and at the scale of a host population in terms of bacterial evolution of antibiotic resistence (chapter 6). At the scale of the host, in chapter 5, I propose to show that in vaccinated individuals producing IgA, the interplay between bacterial growth and cluster breaking could be a way for the immune system to maintain microbiota homeostasis by discriminating against fast-growing bacteria susceptible to disrupt the gut flora equilibrium [20]. Indeed, if bacteria replicate faster than bacterial aggregates break, then they end up trapped in cluster form, which prevents them from approaching the epithelium, a necessary step to colonize the rest of the organism, *i.e.* to start a systemic infection. Then if bacteria are agglomerated in clonal clusters, it is plausible that they are also transmitted under this form. In chapter 6, I show that at the scale of the host population, if other parameters do not change, the probability of infection emergence is reduced in immune populations (where bacteria are agglomerated) compared to naive ones (where bacteria grow freely). The main reason for that is that, as in the case of immune hosts bacteria are transmitted via clonal clusters (being either completely resistant or completely sensitive), then, with the same average number of transmitted resistant bacteria, the proportion of transmissions with at least one resistant bacteria transmitted is lower for immune donor hosts.

Both parts of this thesis were motivated by the study of quantitative data on *Salmonella* infection in the mice gut. However, the results presented here go beyond the scope of mere data interpretation. It is essential to understand that the

immune mechanisms we propose here are of very general scope: the mice immune system is indeed close to the one of many other vertebrates (including humans), and enchained growth is not a process limited to *Salmonella* but has already been evidenced in *E. coli* [19] for example. Then in the first part, the population dynamics tools developed could be applied to larger sets of data concerning bacterial infection of the gut with various strains in various animals, but could also easily be translated to different systems in ecology, not necessarily at the same scale. In the second part, the study of clusters growth and fragmentation is a more general statistical physics problem which had already proved to be useful in other contexts and at other scales (for example to the study of a specific kind of algae, see [21], or to explore reproduction modes [22]). We thereby aim to identify some generic mechanisms, and the minimal ingredients needed to understand the range of situations where these mechanisms may be important.

Part I

Population dynamics of a bacterial gut infection

Introduction

Infectious diseases are often studied from a molecular point of view. The identification of key molecules or key cells is very useful, but this vision can be completed by other approaches. In particular, the contribution from population dynamics is essential: one cannot design optimal strategies to block the infectious process without knowing where the pathogen enters the host, which organs it colonizes, how fast it replicates, migrates, and potentially gets eliminated. Identifying weak points of the infectious process at the whole pathogen population level could lead the way to the design of new vaccines and therapies. Such approaches have been developed in the virus community [23, 24], and have already provided promising insights regarding bacterial infections [25, 26]. Additionally, bacterial populations also provide model systems for the study of population dynamics and evolution, in particular because of their high growth rate.

In this part, we develop generic tools, and apply them to a specific case: gut infections following food poisoning. We aim at understanding them in terms of population dynamics, and in particular at characterizing the colonization dynamics of bacteria in the first stages of infection. In this perspective, modeling is a key tool to reconstruct what has happened in the black box that is the organism, since only indirect data are available: the initial data on bacteria inoculated and the final data retrieved after dissection. The goal is to estimate biologically relevant parameters, such as the replication rate, elimination rate (due either to feces production or immune response), and the probability for one bacteria to establish in the gut and not get killed before by the acidity in the stomach or being directly evacuated. Many different scenarios could account for exactly the same final numbers starting from the same initial numbers. For example, a high replication and elimination rates scenario could lead to the same final numbers as a low replication and elimination rates scenario, only that in the first scenario the turnover would be more important. Modeling is needed to decipher these data and determine which scenario is the most likely to have happened.

From a broader perspective, this part of the work is also motivated by more general and theoretical questions. It aims at developing generic models of one or several subpopulations in open system and determining which observables are best to extract maximum information on the dynamics. Using different observables would either enable to infer more parameters, or to infer the same parameters and check the consistency of the results.

In the following, in a first chapter I will present the experimental data that motivated this work, in particular the different markers used to get additional quantitative information on the initial and final bacterial populations, and will then review the different analytical and computational methods used thereafter. In a second chapter, I will describe a first one-population model, along with all the questions it raised, including the more theoretical ones concerning the choice of observables. In a third chapter, I will argue that the study of several sub-

populations models is relevant in our context and present a two-subpopulations model. I will finally discuss the results obtained with these different models and their limitations, and present some further thoughts and perspectives for future developments.

Chapter 1

Experimental data and methods

Contents

In this introductory chapter, I present in a first section the experimental system and related data of *Salmonella* colitis that motivated this study, and in particular the different bacterial labels developed to get additional quantitative data on the initial and final bacterial populations. In a second section, I review and re-derive some of the essential methods from statistical physics that will be of use in the following chapter, both analytical (mostly branching processes) and computational (Gillespie-based simulations).

1.1 Experimental data

Our work was motivated and inspired by data collected in the group of Dr. Emma Wetter Slack, immunologist at ETH Zürich. They study intestinal microbiota using the streptomycin mouse model for Salmonella diarrhea [11], whereby mice are orally infected with *Salmonella enterica* serovar *enterica* Typhimurium (referred to as S. Typhimurium in the following). This rod-shaped, gram-negative potentially flagellated, potentially anaerobe, non-typhoidal strain of bacteria, can cause acute diarrhea to humans and is responsible for several millions of death per year [27]. In normal conditions, an inoculation of this bacterium does not necessarily trigger an infection: the overgrowth of this particular strain can indeed

be regulated by competition effects with the multitude of other microorganisms composing the microbiota. In order to be sure to trigger an infection even starting with a low inoculum, mice were pre-treated with broad-spectrum antibiotics before the beginning of each experiment. Two different strains were used in the experiments under consideration in this part: a wild type strain producing virulence factors favoring the immune response ("SB300"), and an attenuated non-virulent strain avoiding inflammation ("M2702"), which can allow better observations in specific cases. As we focus on the beginning of the infection, before inflammation, the two strains should be equivalent.

We are mostly interested in the early stages of the infection (24-48h post inoculation), when it is not systemic yet and the immune system has had no time to react strongly yet, *i.e.* before inflammation is triggered [25]. We will focus exclusively on the content of the cecum, a pouch situated at the beginning of the large intestine and of important dimensions for rodents (see figure 1.1), where the bacteria colony settles and develops during the first stages of infection. The following sections present the different types of quantitative data collected during these infection experiments: the final numbers of bacteria, the dilution of a specific plasmid giving information on the number of generations, and the final numbers of genetically labeled bacteria.

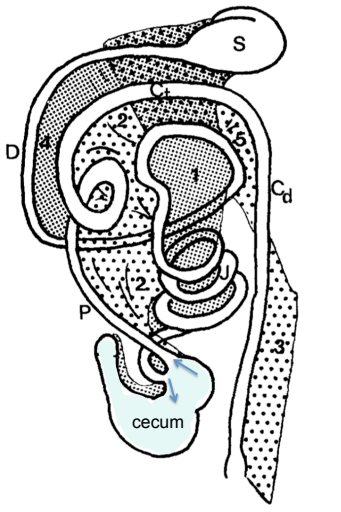


Figure 1.1 – Drawing of the digestive track of a laboratory mouse. The digesta enters through the stomach (S), passes the duodenum (D) and enters the cecum through the ileum. It then leaves it to the colon. The cecum is important for cellulose digestion in rodents. Figure adapted from [40]

1.1.1 Direct count of bacteria

Bacteria are genetically engineered to be resistant to an antibiotic (streptomycin). The technique to count them consists in making several different dilutions of the

cecum content and put them on Petri dishes with streptomycin. An adequate dilution will allow to count visually the number of colonies on the plate (not too diluted to obtain a large enough number of colonies and limit the noise but diluted enough so that each colony is distinct from one another on the plate). The unit to count bacteria this way is hence called CFU for Colony Forming Unit (it is considered that each colony results from a single bacterium if the dilution was appropriate). This counting technique is estimated to be accurate with a factor 2 [28].

Most of the data are from the cecum, measured after killing the mice. Something that could be done is to count the bacteria in the feces to get additional longitudinal points. But very little is know about the dynamics of microbiota composition along the gut (both in the cecum and downstream), which could be shaped by physical factors (complicated hydrodynamics of non-Newtonian fluid with peristaltic motions [41]), chemical factors (pH gradient shaping the bacterial landscape [42]), and ecological factors (competition between species). Comparing preliminary data from the feces and data from the cecum after euthanasia, it was unclear to which extent the feces composition is representative of the cecum content; therefore only cecum data are considered in the following.

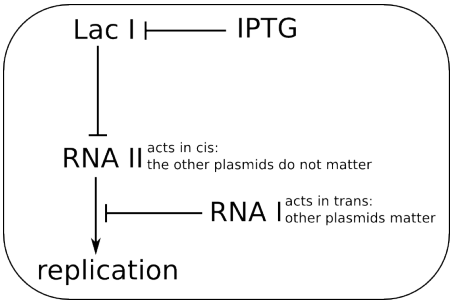
1.1.2 Plasmids and mean number of generations

The type of parameters we want to estimate depends on the type of model we choose (and the more complex the model, the more parameters are to be estimated). In any case, the mean number of generations will always be a relevant information, and it can be directly estimated from the dilution of a specific type of plasmid, as will be explained below.

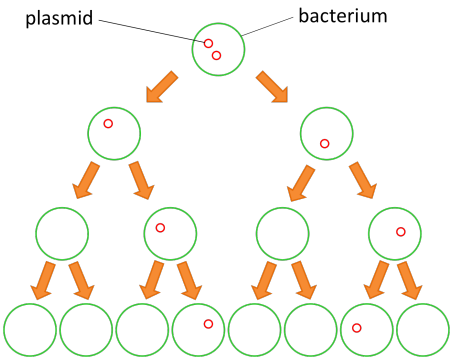
A plasmid is a strand of circular DNA separate from the chromosomal DNA and which can replicate independently. The E.coli plasmids pAM34 [43] used in the experiments we consider have two particular properties. First, these plasmids have been engineered to need IPTG to replicate (see fig. 1.2 A). IPTG is present in the *in vitro* culture but absent from the mouse system. Secondly, these plasmids have been engineered to carry the resistance to a second antibiotic (Ampicillin). This antibiotic resistance allows to measure the final proportion of bacteria still carrying a plasmid at the end of the experiment. And since a plasmid does not replicate once inside the mouse, it has to "choose" between one of the two daughter-bacteria at each step of replication (see fig. 1.2 B). Thus, if one supposes that initially, each bacteria carries a single plasmid, the link between the total number of replication cycles G and the final proportion of plasmid-carriers $y(t_f)$ can simply be written as:

$$y(t_f) = 1/2^G$$

In practice, it is slightly more complicated, since plasmids are usually found in more than one copy in each bacteria initially. There is thus an initial phase of replication during which each bacteria still carries at least one plasmid, before plasmid-carrier bacteria actually start diluting with replications. We also con-



A. Regulation of plasmids replication: in presence of IPTG the inhibition of RNA II production is inhibited [44]. As RNA II is needed for replication, absence of IPTG suppresses replication.



B. Dilution of the proportion of bacteria carrying plasmids: here dilution starts only after a first step of replication ($\Delta_d = 1$). At the end $1/4 = 1/(2^2)$ bacteria carry a plasmid, there were $G = 2$ replication steps after the begining of dilution.

Figure 1.2 – Plasmids which do not replicate *in vivo* allow the estimation of the mean number of replications through the measure of the final proportion of antibiotic-resistant bacteria

sider the possibility that there is a residual replication of the plasmids once in the mice, which would slow the dilution process. Calibration experiments are carried on *in vitro* (see the data in appendix A table A.1), in conditions for which there is no bacterial death. For each experiment i, let N_{0i} be the initial number of bacteria. After G replication cycles, the total final number of bacteria is $N_{0i}2^G$. Let us write 2^{Δ_d} the mean initial number of copies of the plasmid present in each bacteria at the beginning of the experiment. The initial number of plasmids is thus $2^{\Delta_d}N_{0i}$, and considering that a residual number of replications ϵG happens for the plasmids after the beginning of the experiment, the final number of plasmids is $N_{0i}2^{\Delta_d}2^{\epsilon G}$. Thus the final proportion of resistant bacteria measured as y_i in experiment i writes:

$$y_i = \frac{N_{0i}2^{\Delta_d}2^{G\epsilon}}{N_{0i}2^G} = 2^{\Delta_d - G(1-\epsilon)} = \frac{2^{\Delta_d}}{x_i^{(1-\epsilon)}} \tag{1.1}$$

with $x_i = 2^G$ the measured ratio between the final and initial numbers of bacteria in experiment i. The experimental points (x_i, y_i) are thus fitted (see fig. 1.3) through quadratic error minimization to the best equation of the following form:

$$\log_2(y) = \Delta_d - (1-\epsilon)\log_2(x)$$

corresponding to eq. (1.1) taken in base log_2.

During actual experiments in the mice, bacteria die or are carried out of the cecum, and this is why the final number of bacteria will only give a lower

bound for the number of replications having happened. But bacteria carrying a plasmid or not will die indifferently, leaving the proportion of plasmid-carriers unchanged. Thus, the data of the two best-fitting parameters ϵ and Δ_d for the proportion of plasmid-carrying bacteria directly allows to estimate the mean number of replications G.

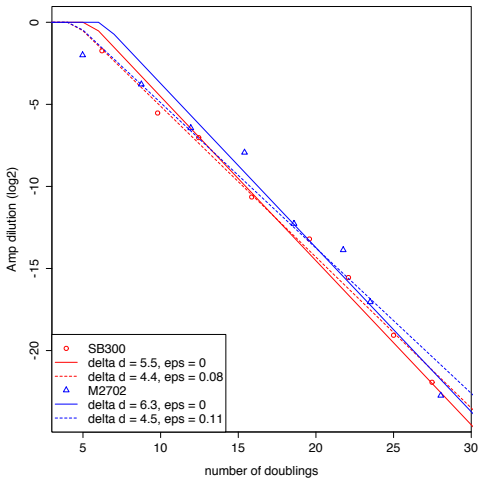


Figure 1.3 – Calibrating data from *in vitro* experiments, for the virulent (in red) and avirulent (in blue) strains: $\log_2$ of the fraction of antibiotic-resistant bacteria in function of $\log_2$ of the ratio between final and initial numbers. For both strains, a fit with ϵ constrained to zero and one with ϵ free are tested, leading to solutions close to one another.

1.1.3 Wild type Isogenic Tagged Strains

Some Wild type Isogenic Tagged Strains [29] (WITS in the following) are included in the inoculum given to the mice in the experiments. A specific sequence has been added to the genome of these bacteria, as well as a gene coding for the resistance to a third antibiotic (Kanamycin), so that, again, it is possible to tell apart those bacteria from the others and count them (as seen in section 1.1.1). They are otherwise identical to the unlabeled strain, and in particular they do not differ in fitness [29]. In the experiments considered in this part, seven different genetically tagged strains have been used, with seven different specific sequences. The proportion in which each strain is present can be determined through quantitative-Polymerase Chain Reaction (q-PCR). The use of these genetic tags gives an additional data: the initial and final distribution of these seven strains.

To extract relevant information from these distributions, the initial number of bacteria labeled with these genetic tags needs to be low enough so that the effects of the stochasticity of the processes involved can be observed on the distribution

[29, 25, 26], but still high enough so that we do not lose all the tags in the process. Typically if we go back to the example given in the introduction of this part, a scenario "high replication and elimination rates" leads to a higher variability in the WITS distribution than a scenario "low replication and elimination rates".

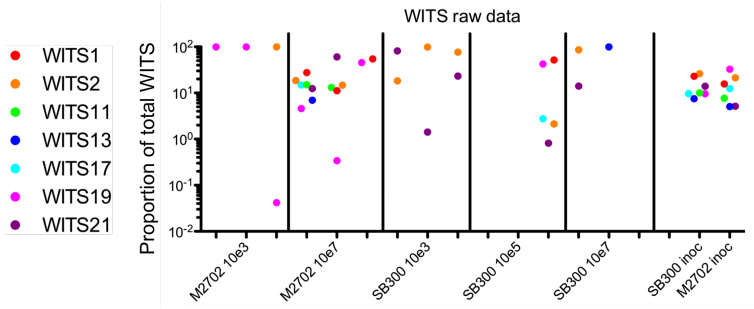


Figure 1.4 – An example of raw data from experiment ES15-010: wild type mice, $n = 3$ per group, were infected with approximately $10^3, 10^5$ or 10^7 wild type (SB300) or avirulent (M2702) Salmonella containing 5 to 10 CFU of each WITS. The mice were euthanised for cecal content at 24h post infection. In many cases, some WITS are present in much higher proportions than the others. This could indicate that very small proportions should actually be seen as experimental artifacts and considered as 0. In general we will thus define a cutoff under which WITS numbers are put to zero.

1.1.4 Summary of experimental data

In the end, we have seen in this section the three types of data we will be analyzing in the following, measured both at the beginning (in the inoculum) and at the end of the experiments (from the cecal content): the total number of bacteria, the plasmids dilution which gives a direct estimate of the mean number of generations, and the WITS distributions from which we will retrieve information on the stochasticity of the processes involved. We will study in this part I a simple data set (no vaccination), to test our methods.

1.2 General methods

In this section, I am going to review the main methods that will be used in the following. In a first section, I present analytical methods and re-derive results that can be found in [37, 45]: starting from the master equation of a Markovian birth-death process, I solve the equation for the generating function of its probability distribution and show how to obtain the moments of the distribution from it. I also introduce the method of log-likelihood maximization [46]. In a second section, I present computational methods, first the Gillespie simulation [47], then the tau-leaping procedure which derives from it [48]. I finally recall Bessel's correction [49] for the estimate of distribution parameters from a finite number of measures.

1.2.1 Analytical methods

In this section I present some general analytical methods through the study of a null model. For this abstract study, no data will be considered; all the issues of observables calculations and parameters estimations will thus be left to the next sections along with data interpretation questions. Likewise, all the biological arguments justifying the choice of the type of model will be treated in chapter2.

Null model: Let us consider a stochastic population dynamics model where the initial population size is drawn from a Poisson distribution of mean βN_0. At time $t = 0^+$ starts a continuous time Markov process with a constant replication rate r, and a constant elimination rate c.

1.2.1.1 Master equation

Let us define $P(n,t)$ as the probability density that the system contains n individuals at time t. At time $t + dt$, the new probability distribution only depends on the one at time t and thus one can write the following:

$$P(n,t+dt) = (1-(c+r)ndt)P(n,t)+c(n+1)dtP(n+1,t)+r(n-1)dtP(n-1,t) \tag{1.2}$$

with the following Poissonian initial condition:

$$P(n,t=0) = \frac{(\beta N_0)^n}{n!}e^{-\beta N_0} \tag{1.3}$$

Making dt small, one can re-write eq. (1.2) as the following master equation:

$$\frac{\partial P(n,t)}{\partial t} = (-cn-rn)P(n,t)+c(n+1)P(n+1,t)+r(n-1)P(n-1,t) \tag{1.4}$$

1.2.1.2 Generating function

This whole stochastic process is a branching process in continuous time. It can be treated by studying a generating function $g(z,t)$ over the probability distribution $P(n,t)$ [38]. This generating function corresponds to

$$g(z,t) = \sum_{n=0}^{n=+\infty} P(n,t)z^n \tag{1.5}$$

Summing eq. (1.4) multiplied by z^n over n, and having in mind that

$$\frac{\partial g}{\partial z} = \sum_{n=0}^{n=+\infty} nP(n,t)z^{n-1},$$

one gets:

$$\frac{\partial g}{\partial t} = -(c+r)z\frac{\partial g}{\partial z} + c\sum_{n=0}^{\infty}(n+1)P(n+1,t)z^n + r\sum_{n=1}^{\infty}(n-1)P(n-1,t)z^n.$$

The summation index can be shifted in the second term of the right hand side ($n' = n + 1$) without additional term since the $n' = 0$ term added to the sum is zero; we thus get $c\frac{\partial g}{\partial z}$. Similarly, the last summation index can be shifted easily (this time $n' = n - 1$). Hence, this term transforms into $rz^2\frac{\partial g}{\partial z}$ and one finally gets the following partial differential equation on g:

$$\frac{\partial g}{\partial t} + (1 - z)(rz - c)\frac{\partial g}{\partial z} = 0 \tag{1.6}$$

1.2.1.3 Characteristics method

This equation can be solved using the method of characteristics [50], which allows to reduce it to a collection of ordinary differential equations. This method consists of seeing the first member of eq. (1.6) as the full derivative of g with respect to a new parameter x, and the two parameters z and t as functions of x parametrizing the curves of space (the "characteristic lines") on which the equation $\frac{Dg}{Dx} = 0$ is true indeed. Since

$$\frac{Dg(z(x), t(x))}{Dx} = \frac{\partial g}{\partial t}\frac{dt}{dx} + \frac{\partial g}{\partial z}\frac{dz}{dx}$$

by identification, one can chose

$$\begin{cases} \frac{dt}{dx} &= 1 \\ \\ \frac{dz}{dx} &= -(z-1)(rz-c) \end{cases}$$

And integrating this system, one gets:

$$\begin{cases} t &= x + t_0 \\ \\ x &= \int \frac{-dz}{(z-1)(rz-c)} = \frac{1}{r-c}\ln\left(\frac{rz-c}{z-1}\right) + x_0 \end{cases} \tag{1.7}$$

One can choose $x_0 = 0$. Then, only one constant of integration is needed. Along $(t(x), z(x))$, $\frac{Dg}{Dx} = 0$, and hence g is a constant that only depends on the constant of integration t_0. Thus

$$g(z,t) = K(t_0) = K\left(t - \frac{1}{r-c}\ln\left(\frac{rz-c}{z-1}\right)\right)$$

K can then be found with the initial condition on P eq. (1.3). Thus:

$$g(z, t=0) = \sum_{n=0}^{\infty} P(n, t=0)z^n = e^{-\beta N_0}\sum_{n=0}^{\infty}\frac{(\beta N_0 z)^n}{n!} = e^{\beta N_0(z-1)}$$

Now,

$$g(z, t=0) = K\left(-\frac{1}{r-c}\ln\left(\frac{rz-c}{z-1}\right)\right) = K(-x)$$

And if one inverts the relation between x and z in eq. (1.7), $z = \frac{e^{(r-c)x}-c}{e^{(r-c)x}-r}$, one gets:

$$K(-x) = e^{\beta N_0(z-1)} = \exp\left[\beta N_0\left(\frac{e^{(r-c)x}-c}{e^{(r-c)x}-r} - 1\right)\right] = \exp\left[\frac{\beta N_0(r-c)e^{-(r-c)x}}{1-re^{-(r-c)x}}\right]$$

Now that we have the expression of the function K, one just needs to remember that $g(z,t) = K(t_0)$ and replace $-x$ with the expression of t_0 deduced from eq. (1.7), $t_0 = t - \frac{1}{r-c}\ln\left(\frac{rz-c}{z-1}\right)$. *In fine*, one gets the following generating function:

$$g(z,t) = \exp\left[\frac{\beta N_0(r-c)(z-1)e^{(r-c)t}}{rz-c-(z-1)re^{(r-c)t}}\right] \tag{1.8}$$

1.2.1.4 Moments of the probability distribution

The generating function eq. (1.8) is a powerful tool which allows the calculation of the different moments of the distribution $P(n,t)$, and from them, the calculation of relevant observables, as it will be presented in the next sections. Here are a few identities which will be useful in the following:

Mean population size By definition,

$$\langle n\rangle(t) = \sum_{n=0}^{+\infty} nP(n,t) = \left.\frac{\partial g}{\partial z}\right|_{z=1,t} \tag{1.9}$$

Higher order moments

$$\begin{aligned}\langle n^2\rangle(t) = \langle n(n-1)\rangle(t) + \langle n\rangle(t) &= \sum_{n=0}^{+\infty} n(n-1)P(n,t) + \langle n\rangle(t)\\ &= \left.\frac{\partial^2 g}{\partial z^2}\right|_{z=1,t} + \left.\frac{\partial g}{\partial z}\right|_{z=1,t}\end{aligned}$$

By induction, higher moments can similarly be expressed as combinations of the derivatives of g.

Variance In particular, the variance on the number of individuals can be expressed as:

$$\begin{aligned}\sigma^2(t) = \left\langle (n-\langle n\rangle)^2\right\rangle(t) &= \langle n^2\rangle(t) - \langle n\rangle^2(t)\\ &= \left.\frac{\partial^2 g}{\partial z^2}\right|_{z=1,t} + \left.\frac{\partial g}{\partial z}\right|_{z=1,t} - \left(\left.\frac{\partial g}{\partial z}\right|_{z=1,t}\right)^2\end{aligned}$$

Extinction probability The extinction probability is the probability that there is no individual left at a certain time t: $P(n = 0, t)$. This quantity can be obtained from the generating function by noticing that [12]

$$P(n = 0, t) = g(z = 0, t) \tag{1.10}$$

1.2.1.5 Log-likelihood maximization principle

The log-likelihood maximization principle [46] allows to get the best fitting parameters that fit a data set with an assumed probability distribution function. Let us suppose we have at our disposal the data of a set of draws $x_1, ..., x_N$ from a random variable X. We know the shape of its distribution – let us say for example that X is normally distributed – but the parameters of the distribution function (in our example, the variance and/or the mean which characterize the normal distribution) are unknown. The maximization of the log-likelihood allows for an estimation of these parameters from the data of $x_1, ..., x_N$. Let us write $f_\mu(x)$ the probability distribution of X, with μ the parameter we want to estimate. The likelihood can be defined as the quantity $L(\mu)$, with:

$$L(\mu) = \prod_{i=1}^{N} f_\mu(x_i)$$

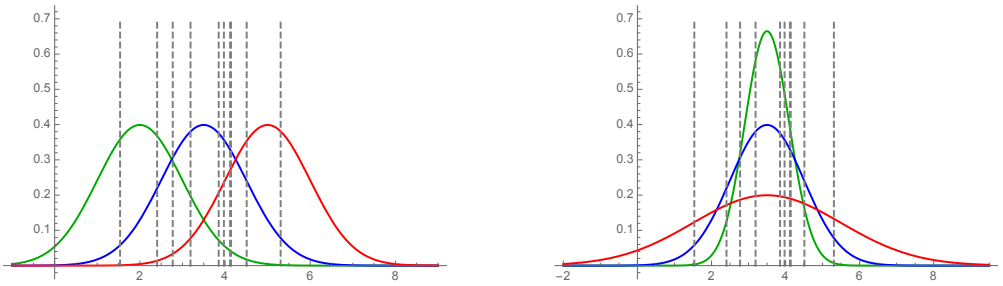
The best fit of the data will be obtained for the value of μ maximizing the quantity L. Figure 1.5 provides a qualitative explication of this principle. Thus, in order to find this best-fitting parameter, one must solve the equation $\frac{\partial L}{\partial \mu} = 0$ in μ. In many cases, the log-likelihood $\ln(L)$ is a quantity easier to manipulate; it does not change anything for the maximization, since a probability distribution function is always positive, and that the logarithm is a strictly monotonically increasing function.

1.2.2 Computational and numerical methods

In this section, I present computational methods which allow to simulate the stochastic birth-death process presented in the previous part. I first present the exact Gillespie algorithm, and then present in a second section an approached version which allows faster updates. I finish with a comment on the numerical estimate of the variance from a finite data set.

[1] In this context, it is agreed that $0^0 = 1$. The equality 1.10 can otherwise be obtained by considering the uniqueness of the development in power series with the Taylor expansion $g(z,t) = \sum_{n=0}^{\infty} \frac{\partial^n g}{\partial z^n}\Big|_{z=0} z^n$

[2] Note that although $n = 0$ is an absorbing state of this Markov process (once reached, it is impossible to escape), the extinction probability does not tend to 1 when $t \to \infty$. The reason is that there is an infinite number of states and that the rate to move away from the absorbing state is higher than the rate to come back. For more details, see [51].



A. Mean value μ determination

B. Spread σ determination

Figure 1.5 – Log-likelihood maximization principle applied to a normal distribution $f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{x-\mu}{\sigma}\right)^2}$. The vertical dashed lines represent the data of the draws $x_1, ...x_N$ from the random variable X. $L(\mu)$ (resp. $L(\sigma)$) is the product of the heights of those bars under the curve corresponding to μ (resp. σ). **A.** Here the variance is known already, we try to determine the mean. The parameter of the blue curve is better adapted. **B.** Here the mean is known, but not the variance. The parameter of the blue curve is the best adapted.

1.2.2.1 Gillespie algorithm

The Gillespie stochastic simulation algorithm (SSA) was first introduced in the context of chemistry, with microscopic systems of reactant molecules [47]. This algorithm generates possible trajectories for a stochastic equation. Let us write $X(t)$ the vector of the different population sizes $(X_1(t), ..., X_N(t))$ we are following. As will be explained in the next sections, in our case the different populations will correspond to the different types of labeled bacteria. Let us note M the number of possible transitions, $a_j(X(t))$ the rate (probability per unit of time) at which the transition j occurs (it can depend on the state of the system at time t $X(t)$), and $\nu_j = (\nu_{1j}, ..., \nu_{Nj})$ the state change vector of transition j (meaning that ν_{ij} gives the change in the X_i population induced by the transition j). We follow N populations[3] that can only replicate or die by increasing or decreasing their population size by 1. For example if the transition j is a replication event in the population X_k, then $\nu_{kj} = +1$ and $\nu_{ij} = 0$ for all $i \neq k$. Bellow are the steps followed by the SSA algorithm:

1. **Initialization:** Choose initial values for $X(0)$. In our case they are drawn from a Poisson distribution

2. **Monte Carlo Step:** Two random numbers r_1 and r_2 are generated from the uniform distribution on the interval between 0 and 1, boundary values excluded. The time for the next transition to occur is exponentially

[3] We will see in the next section that in our case, the only interaction we will contemplate between populations will be through the consideration of logistic growth, in which case the transition rates of replication events decrease with the sum of X_i

distributed and calculated as:

$$\tau = \frac{1}{\sum_{j=1}^{M} a_j(X(t))} \log\left(\frac{1}{r_1}\right)$$

and the selected transition is the smallest index j satisfying:

$$\sum_{l=1}^{j} a_l(X(t)) > r_2 \sum_{l=1}^{M} a_l(X(t))$$

3. **Update:** Increase the time step by the τ calculated at step 2 and actually perform the transition $X(t+\tau) = X(t) + \nu_j$

4. **Iterate:** Go back to step 2 unless the final time has been reached, or if all the rates are equal to zero (in our system, it is equivalent to all the populations being extinct).

1.2.2.2 Tau-leaping procedure

Some approximations were later made to optimize the computational time of the Gillespie algorithm, which is slow because it processes reactions one at a time. For my simulations I used the *adaptivetau* package of the R language [48]. Many sophistications are included in this package to optimize the algorithm, but we will review here only the main ones, of interest for our study. The tau-leaping principle is to choose a time step τ so that there is little change in the rates of the transitions over time step, and perform simultaneously several transitions at a time. The parameter controlling how strictly we want the rates not to change over the selected τ (and thus how accurately the algorithm converges) is ϵ. With $\Delta a_j(x)$ the change in rate a_j from t to $t+\tau$, the rates have to obey a leap condition of the form:

$$\Delta a_j(x) < \epsilon f(a_1(x), ..., a_M(x)). \tag{1.11}$$

If this condition cannot be met, then the algorithm switches back to the exact Gillespie procedure.

There are different tau-selection formulas (*i.e.* different f in eq. (1.11)) for each possible scheme. The updates can be done either following an explicit scheme – where the state of the system at time $t+\tau$ is calculated based only on the state of the system at time t, like in the standard Gillespie algorithm – or implicit scheme – where the state of the system at time $t+\tau$ is similarly written but with the rates of transition taken at time $t+\tau$ instead of t – the advantage of implicit schemes being that they allow bigger tau leaps. There are also other special cases when some specific conditions are verified by the system, for example of partial equilibrium. The tau-selection formulas have themselves been approximated and simplified so that the calculus of the biggest τ value allowed by a scheme is optimized in time. At each step, the largest possible τ is calculated for each possible scheme, and then the scheme allowing the biggest tau-leap is automatically selected (that is

why the algorithm is called *adaptive*). In particular, the parameter controlling the switch between the implicit/explicit schemes is N_{stiff}, which specifies how much larger $\tau^{implicit}$ needs to be relative to $\tau^{explicit}$ to make it worth enough to choose the implicit scheme over the explicit scheme (the implicit scheme is favored if $\tau^{implicit} > N_{stiff}\tau^{explicit}$, with $\tau^{implicit}$ the tau-leap allowed with the implicit scheme and $\tau^{explicit}$ with the explicit one).

The standard value for ϵ is 0.05. But is that small enough for our case? Not all the time, especially when high replication rates are taken. I ran some tests with the "null model" (see fig. 1.6 and fig. 1.7), checking if the mean size of the final population (average over many realizations of the procedure) actually matches the size we are analytically expecting, and came to the conclusion that we should pay special attention to the choice of this parameter in the next sections.

A special care is also taken of the populations close to extinction: if a population gets below a critical size n_c then the algorithm goes back to the standard Gillespie algorithm. If a population had a size above this limit but still accidentally ends up with a negative population size after a tau leap, the procedure goes back one step and divides the chosen τ by two before trying again. On the other hand, if two populations are on partial equilibrium then higher values of τ are allowed.

1.2.2.3 Bessel's correction

Let us suppose that we could reproduce the same stochastic experiment corresponding to our null model as defined at the beginning of section 1.2, a certain number h of times. For each realization of the experiment i – which can either be a simulation or an actual biological experiment if the model is adapted to the situation – let us denote the measure of the size of the population at the end of the experiment m_i, which are then different realizations of the random variable m. Let μ be the exact average of m (as put in the simulation for example), and σ^2 its exact variance.

We can estimate a mean population size from this finite number of realizations:

$$\bar{m} = \frac{1}{h}\sum_{i=1}^{h} m_i$$

If we then want to estimate the variance of the distribution of m from the *same* set of $(m_1, ..., m_h)$, the correct estimate is [49]:

$$s^2 = \frac{1}{h-1}\sum_{i=1}^{h}(m_i - \bar{m})^2 \tag{1.12}$$

Differing from $s_h^2 = \frac{1}{h}\sum_{i=1}^{h}(m_i - \bar{m})^2$ by a correcting factor $h/(h-1)$. This correcting factor comes from the fact that we try to estimate both the average and the variance with the same data set, and there is thus one degree of freedom less than what one could think. Let us demonstrate eq. (1.12) is the right estimate

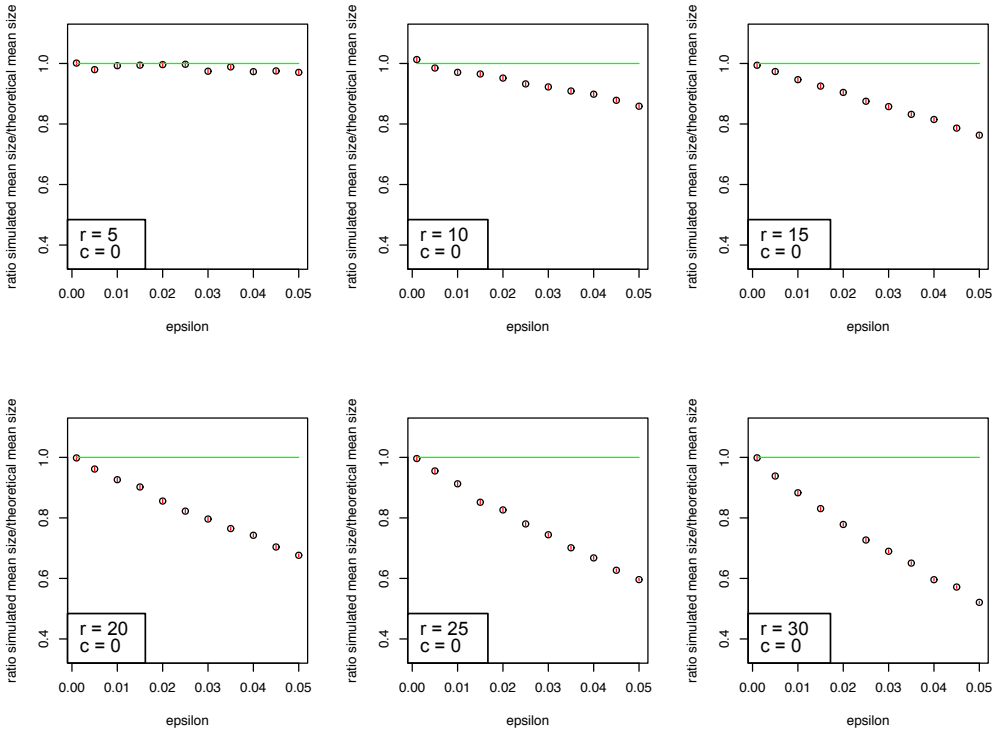


Figure 1.6 – Ratio between the size of the population as a result of the simulation (averaged over 10000 realizations) and the mean size of the population calculated analytically (see eq. (1.9)), in the context of the null model defined at the beginning of section 1.2. Initial population size is 5, final time is one day, probability to establish is $\beta = 0.8$. From left to right and from top to bottom, different values of the replication rate $r = 5, 10, 15, 20, 25, 30 day^{-1}$, with no loss. For the high values of r we get quickly far from the right population size as soon as ϵ is non-zero. The error bars in red (contained within the points) are given by the standard error renormalized by the expected mean population size.

by calculating the expectation value of s^2:

$$\begin{aligned}
\mathbb{E}(s^2) &= \mathbb{E}\left(\sum_{i=1}^{h}\frac{(m_i-\bar{m})^2}{h-1}\right) = \frac{1}{h-1}\mathbb{E}\left(\sum_{i=1}^{h}\left[m_i-\mu-(\bar{m}-\mu)\right]^2\right) \\
&= \frac{1}{h-1}\mathbb{E}\left(\sum_{i=1}^{h}\left([m_i-\mu]^2-2[m_i-\mu][\bar{m}-\mu]+[\bar{m}-\mu]^2\right)\right) \\
&= \frac{1}{h-1}\mathbb{E}\left(\sum_{i=1}^{h}[m_i-\mu]^2-2h[\bar{m}-\mu]^2+h[\bar{m}-\mu]^2\right) \\
&= \frac{1}{h-1}\mathbb{E}\left(\sum_{i=1}^{h}[m_i-\mu]^2-h[\bar{m}-\mu]^2\right) \\
&= \frac{1}{h-1}\left[\sum_{i=1}^{h}\sigma^2-h\frac{\sigma^2}{h}\right] \\
&= \frac{1}{h-1}(h-1)\sigma^2 = \sigma^2
\end{aligned}$$

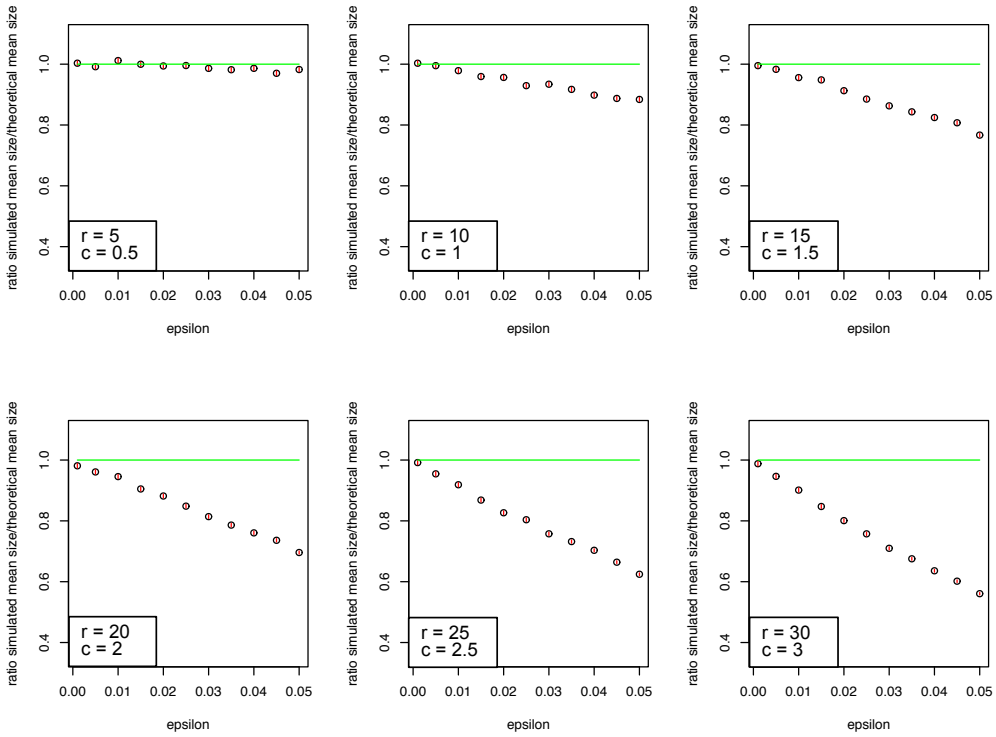


Figure 1.7 – Everything as in fig. 1.6, except that c is non zero ($c = 0.1r$). Same conclusion on the need to be careful with ϵ

1.2.3 Summary of general methods

In this section, I presented and rederived part of the general methods I will use thereafter. As much as possible, I will try to obtain analytical expressions, writing generating functions for the branching processes. I will also use numerical methods, in particular, Gillespie-like algorithms in agent-based simulations.

1.3 Symbols for Part I

Plasmids dilution	
G	Number of generations estimated from plasmids dilution
N_{0i}	Initial number of bacteria in the *in vitro* calibration experiment i
Δ_d	"Delay" to dilution: each bacteria carries initially 2^{Δ_d} plasmids
x_i	Ratio between final and initial numbers of bacteria in *in vitro* calibration experiment i
y_i	Final proportion of resistant bacteria in *in vitro* calibration experiment i
One-population model	
$P(n,t)$	Probability that n bacteria are in the cecum at time t
$g(z,t)$	Generating function for the probability distribution $P(n,t)$
β	Probability for one bacteria to establish initially in the cecum
N_0	Total initial number of bacteria
r_{mean}	$= G \ln 2/t$, mean replication rate, another way to express the plasmids dilution data
r_{max}	Initial replication rate (before saturation) for the bacterial population
K	Carrying capacity at which the growth saturates
c	Loss rate for the bacterial population
m_i	Final number of WITS i
n_i	Initial number of WITS i
n_0	The mean of the initial numbers of WITS, taken in first approximation of an equal inoculum
$B(t)$	Number of bacteria in the cecum at time t (deterministic description for the total population)
B_f	Total final number of bacteria
u	$= \frac{r}{\beta n_0 (r-c)}$ Indicator of stochasticity
$\tilde{u}$	$= \frac{r}{\beta (r-c)}$ Indicator of stochasticity normalized by the initial population size
Two-populations model	
$r_{max,1}$	Initial growth rate of population 1 (fast-replicating)
$r_{max,2}$	Initial growth rate of population 2 (slow-replicating)
α	$= \frac{r_{max,1}}{r_{max,2}}$ ratio between the replication rates of the two sub-populations
q	Initial proportion of the fast-replicating subpopulation 1

Chapter 2

One-population models

Contents

In this chapter I will consider different variations of a model where all the bacteria follow the same dynamics, replicating at the same rate and getting eliminated at the same rate. I will first review the biological ground which justify the choice of model (section 2.1). Then I will present the principles underlying our approach for parameters estimation through the study of an oversimplified approximation (section 2.2). I will then discuss the best strategy to extract information from the data through the choice of an appropriate observable (section 2.3) and will then show how the null model can be refined to best take into account the specificity of the data (section2.3.4). Finally, I will present the results and discuss the appropriateness of these models to describe the data (section 2.5).

2.1 Biological grounds to the one-population model

In the experiments, bacteria are taken from a suspension at a certain concentration $\mathcal{C}_0$ to be orally inoculated to mice. Thus we will consider in our model that the sample of bacterial solution of volume V_0 given to a mouse contains a number of bacteria k Poisson-distributed of average of $N_0 = \mathcal{C}_0 V_0$. Once inside the organism, the bacterial population may undergo a first bottleneck, best taken into account by an identical probability β for each bacteria to settle in the cecum, instead of being killed by the acidity in the stomach before reaching it or being directly carried out of the niche . We will consider this phenomenon to be instantaneous and to happen at $t = 0$, while in practice, inoculated bacteria reach the cecum over a couple of hours. Writing $P(n,t)$ the probability to find a *total* number of bacteria n at time t in the cecum:

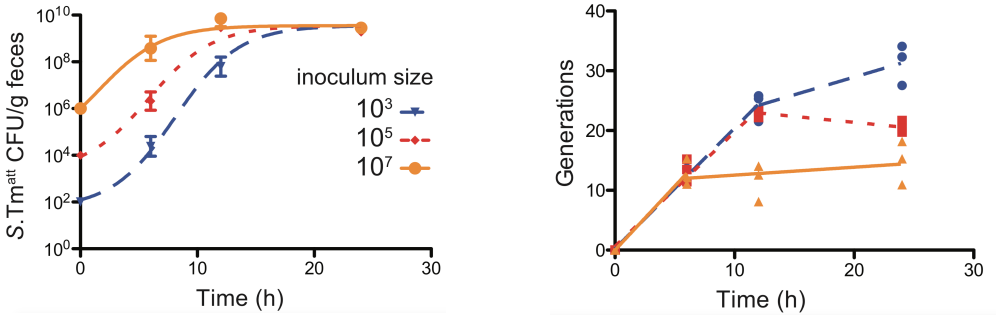
$$\begin{aligned} P(n, t = 0^+) &= \sum_{k=n}^{\infty} \frac{N_0^k}{k!} e^{-N_0} \binom{k}{n} \beta^n (1-\beta)^{k-n} \\ &= e^{-N_0} \sum_{k'=0}^{\infty} \frac{N_0^{k'+n}}{(k'+n)!} \frac{(k'+n)!}{k'!n!} \beta^n (1-\beta)^{k'} \\ &= e^{-N_0} \frac{(\beta N_0)^n}{n!} \sum_{k'=0}^{\infty} \frac{(N_0(1-\beta))^{k'}}{k'!} \\ &= \frac{(\beta N_0)^n}{n!} e^{-\beta N_0} \end{aligned}$$

The binomial selection combined with the Poisson sampling is thus equivalent to a Poisson sampling of average βN_0.

At time $t = 0$ then starts the replication and death of bacteria. For simplicity reasons we will consider that replication happens at constant rate r, *i.e.* that division times are exponentially distributed[1]. In practice, division times are typically more narrowly distributed; the case of a model with fixed division time will therefore be discussed in appendix D. Data show (see fig. 2.1) that the number of replication cycles depends on the size of the inoculum: even with different bacterial inoculum sizes, after one day the final bacterial population reaches the same size K, the carrying capacity. When the inoculum is large, only a few replication events take place. We estimate the size of the niche to be around 10^9 bacteria in the cecum (the maximum number of *Salmonella* experimentally reached). Thus the model should include a saturation term. Taking the classic model of logistic growth, the replication rate at time t is $\frac{K-B(t)}{K} r$ with r the replication rate far

[1]Demonstration: let us consider a population $B(t)$ replicating exponentially so that $\frac{dB}{dt} = rB(t)$. Let us write $S(t)$ the fraction of the population which did not undergo a replication event yet at time t. The probability not to have known such an event at $t+dt$ is the probability not to have known it until time t minus the probability to have known it during dt, so that $S(t+dt) = S(t) - rdtS(t)$ and $S(t) = e^{-rt}$ (because the counters of replication are set to zero at time $t = 0$: $S(t = 0) = 1$). Then the probability to replicate exactly between τ and $\tau + d\tau$ is the probability not to have replicated until τ but to have replicated before $\tau + d\tau$: $P(\tau)d\tau = -dS = re^{-r\tau}d\tau$. Thus $\langle\tau\rangle = \int_{\tau=0}^{\infty} \tau P(\tau)d\tau = 1/r$. τ is thus exponentially distributed, of mean $1/r$, with $P(\tau) = re^{-r\tau}$.

from saturation and $B(t)$ the number of bacteria in the cecum at time t. In practice, since the data we will analyze concerns the beginning of the infection, or rather because it is the beginning of the infection that matters regarding the WITS distribution, it will almost always be justified to consider we are far from saturation and make the approximation $\frac{K-B}{K} r \simeq r$ for the analytical analysis. We will however take this saturation into account in the simulations. The elimination of bacteria can be due to evacuation through feces production. As we will see in the following, feces production alone accounts for a loss rate of the order of $c \simeq 10 day^{-1}$, much smaller than the initial replication rate (around $33 day^{-1}$ as will be seen in the following). It has been shown that after a few days, inflammation may cause bacterial death [25], but that early in infection bacterial death is negligible.



A. Bacterial load per gram of feces

B. Number of generations

Figure 2.1 – 3 mice per group were fed with different sizes of inoculum of the attenuated strain of S.Tm one day after receiving a Streptomycin treatment. A. The bacterial load is fitted with four-parameter sigmoids. B. Corresponding numbers of generations are calculated from plasmids dilution data as explained in section 1.1.2. The final bacterial concentrations are similar, and the bigger the inoculum, the smaller the number of replications. Figures from [19]

We use the term *one-population models* since in this model, all bacteria follow the same dynamics. But in practice, and in order to study the WITS data, we are going to follow several independent populations (only potentially interacting via the saturation term which is global), one for each WITS. We will note $n_1, ..., n_h$ the initial numbers of the h WITS populations ($h = 7$ in the data considered), and will first consider all of them to be equal ($n_i = n_0$ for all i). The fig. 2.2 represents the dynamical steps of the model.

2.2 Approximation $c = 0$

A first approximation one can make is to consider that $c = 0$ since we focus on the beginning of the infection, when it is not systemic yet and the immune system has had no time to react strongly yet, *i.e.* before inflammation is triggered [25]. This approximation will allow us to present simply in this section the principles underlying our approach, because then the determination of the other parameters

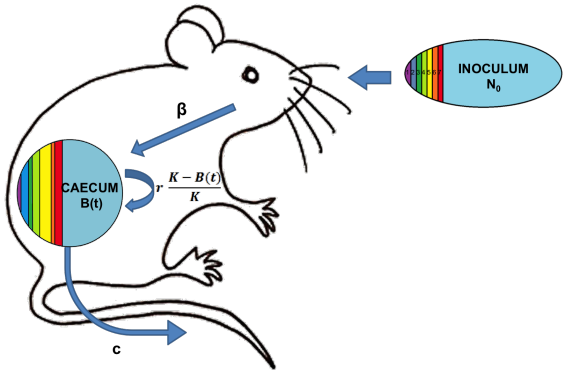


Figure 2.2 – Illustration of the one-population model under consideration: Poissonian inoculum of mean size N_0 including a small proportion of seven different genetically labeled strains (WITS, represented here with rainbow colors), probability to settle in the gut β followed by a Markov birth and death process with constant elimination rate c and saturating replication rate $r\frac{K-B(t)}{K}$, $B(t)$ being the number of bacteria in the cecum at time t

β and r is simple. In the following sections, we will study the case $c > 0$ but the reasoning will remain similar.

2.2.1 Replication rates

In our model, for simplicity reasons, we model bacterial growth at a constant rate: each bacteria has the probability r to replicate per unit of time. Depending on the quantity we look at, we will either choose to use a mean replication rate over the considered period of time, r_{mean}, or the initial replication rate r_{max} (corresponding to the initial slope on figure 2.1B). For example, to estimate the mean final population size, r_{mean} will be more relevant. But for the quantities linked to the WITS distribution (variance and loss probability, as will be exposed in the next sections), then r_{max} is more relevant, because it is the early stages of the dynamics which will shape the WITS distribution: the saturation starts having a significant effect only when the WITS population sizes are already large enough so that the effect of stochasticity is negligible and the deterministic mean evolution describes well what happens next (*i.e.* only the absolute numbers change, but the distribution remains quite unchanged).

We have seen in section 1.1.2 how to get the mean number of generations G from the data of plasmids dilution (the plasmid dilution data for the experiments under consideration can be found in appendix A table A.3). The link between G and r_{mean} can be found by writing the final number of bacteria, starting from βN_0, in both cases, with only replication:

$$\beta N_0 e^{r_{mean} t} = \beta N_0 2^G$$

Thus r_{mean} can be estimated with $r_{mean} = \frac{G \ln(2)}{t}$, with G estimated from the

plasmids data as explained in section section 1.1.2.

r_{max} can be inferred from the data of figure 2.1: bacteria replicate approximately every half hour, so that the corresponding replication rate is $r_{max} = 48 \ln(2) day^{-1}$. This quantity has been observed to be very robust, and does not depend on the type of strain studied or the size of the inoculum. It corresponds also to the replication rate *in vitro* in favorable conditions.

2.2.2 Establishment probability with the WITS loss

In the case where $c = 0$, no WITS can be lost during the Markov process, and all the WITS that are lost are lost during the Poissonian initial sampling. The probability to lose the tag i during the first step writes as the probability to draw a zero in the Poisson distribution of mean βn_i:

$$\text{Probability to lose strain i} = \frac{(\beta n_i)^0}{0!} e^{-\beta n_i} = e^{-\beta n_i}$$

Since we suppose all WITS are all inoculated in the same proportions, then the probability to lose a strain is the same for each strain, and equals the mean number of strains lost, which is available in our data set (it can *a priori* be averaged on several mice at once):

$$\text{Proportion of WITS lost} = e^{-\beta n_0}$$

2.2.3 Limit of the approximation $c = 0$

With r_{max} and the β estimated from the WITS loss, we ran some preliminary simulations. By choosing arbitrary values of c remaining small (of the order of $0.1 r_{max}$) we noticed that the proportion of WITS lost after the Poissonian initialization was not a negligible portion of the totality of the WITS lost. We thus looked for a more accurate way of estimating the parameters β and c when c is not zero. The way to estimate r (taken either equal to r_{max} or to r_{mean}) remains unchanged.

2.2.3.1 WITS loss when $c \neq 0$

First of all, let us look at the expression taken by the proportion of WITS lost when $c \neq 0$. Since we consider each WITS to be inoculated in the same number n_0, each WITS population gets exactly the same generating function eq. (1.8) as the one calculated with the null model, with n_0 instead of N_0. We can then directly use the extinction probability exposed with eq. (1.10), and thus:

$$\text{Proportion of WITS lost}(t) = g(s = 0, t) = \exp\left[-\frac{\beta n_0 (r - c) e^{(r-c)t}}{-c + r e^{(r-c)t}}\right] \tag{2.1}$$

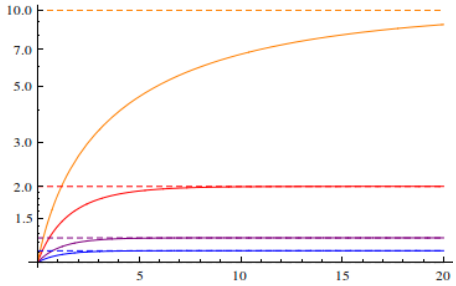
If we compare this β estimate with the previous one assuming that $c = 0$, that we will now denote β_0 for convenience

$$\frac{\beta}{\beta_0} = \frac{r e^{(r-c)t} - c}{(r - c) e^{(r-c)t}} = \frac{1 - e^{-(1-c/r)rt} c/r}{1 - c/r}. \tag{2.2}$$

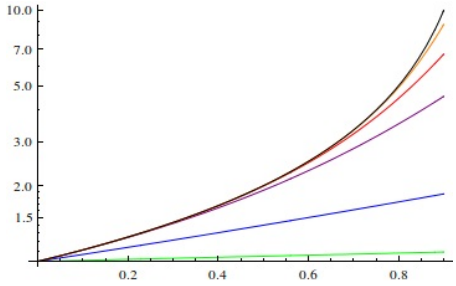
We can check that for $c = 0$ (and any t), as well as for $t = 0$ (and any c), this ratio is 1 as expected. The limit when $rt \to \infty$ is:

$$\frac{\beta}{\beta_0} \to \frac{1}{1 - c/r}. \tag{2.3}$$

The correction coefficient (2.2) depends on two parameters, rt and c/r. For values of interest for us, rt is typically around 10 or more, and c/r is believed to be small, likely smaller than 0.5. Thus as can be seen on fig. 2.3, the correcting factor is close to its limit when rt is large, i.e. $1/(1 - c/r)$. It is interesting because then, it depends only on one parameter.



As a function of rt. $c/r = 0.1$, 0.2, 0.5, 0.9, solid lines for the exact expression (2.2), dashed lines for the limit expression (2.3).



As a function of c/r. Exact expression (2.2) : coloured lines with $rt = 0.1$,1, 5, 10, 20, limit (2.3): black line

Figure 2.3 – Correction factor β/β_0 as a function of c/r and rt.

2.2.3.2 The need for a new observable to disentangle β and c

We have thereby seen that when c cannot be approximated to zero, the data of the proportion of WITS lost only gives a constraint linking β and c. In order to separate the effects of β and c on the WITS distribution, additional information must be extracted from the data. A possibility to do so is to look for an additional observable, to complete the data of the proportion of zeros in the distribution. Another possibility would simply be to apply numerically the log-likelihood maximization principle (as explained in section 1.2.1.5) on the whole WITS distribution. However, this method is just a way to obtain the best fit possible by introducing a black box providing no actual understanding on the data. On the other hand, the use of another observable can provide a feedback on the model we choose: either it provides totally new information and allows to disentangle the roles of β and c, or at least we can check if the informations provided by both observables on the parameters are coherent. This is the direction we will follow in the next section, by looking for an observable to characterize the variability of the WITS distribution.

2.3 The quest for another observable

We have seen in the previous section that the data of the number of WITS lost provided an information linking the parameters β and c, but was not sufficient to disentangle their effects. In this section, I look for a second observable on the WITS distribution that could provide complementary information on the parameters, and that could be measured in the experiments. I thus present different possible observables to characterize the variability in the WITS distribution. The evenness index was first chosen for continuity reasons with previous work of our collaborators, and proves to be convenient for simulations, but is hard to handle analytically. To overcome this difficulty, two other observables will be considered in the following: the variance on the final numbers of WITS, and the variance restrained on surviving WITS. Then the variance over the growth rate will be introduced in order to take into account the initial variability in the inoculum.

2.3.1 Evenness

The evenness that we use here is the Gini index, or relative evenness as defined in [52]. It is an index of diversity, traditionally used to measure the biodiversity of an environment (taking into account both the number of species, or richness, and the equitability of the proportional abundances of the species). Let $\{x_1, ...x_h\}$ be the normalized distribution ($\sum_{i=1}^{h} x_i = 1$) giving the final frequencies of the h different WITS we want to characterize. Here are the steps to calculate the evenness E:

1. Order the values so that $x_1 \leq x_2 ... \leq x_h$
2. Replace the ordered distribution with its cumulative sum distribution: $\{x_1, x_1 + x_2, ..., x_1 + x_2 + ... + x_h\}$
3. Take the sum of the cumulative sum distribution:
$$S(\{x_i\}) = \sum_{i=1}^{h} (h + 1 - i)\ x_i$$
4. Subtract the same quantity calculated with the most uneven distribution $\{0, ...0, 1\}$: $S_{uneven} = 1$
5. Renormalize by the difference between this quantity for the most even distribution $\{\frac{1}{h}, ..., \frac{1}{h}\}$: $S_{even} = \sum_{i=1}^{h} \frac{i}{h} = \frac{h+1}{2}$ and this quantity for the most uneven distribution S_{uneven}

In the end the evenness writes:

$$E = \frac{S(\{x_i\}) - S_{uneven}}{S_{even} - S_{uneven}} = \frac{2}{h-1} \left[\sum_{i=1}^{h} (h + 1 - i)\ x_i - 1 \right]$$

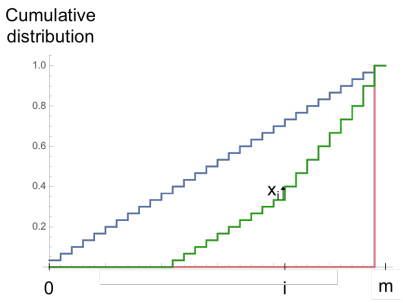


Figure 2.4 – The blue line is the cumulative of the most even distribution, its sum is S_{even}. The red line is the cumulative of the most uneven distribution, its sum is S_{uneven}. The green line is some cumulative distribution to characterize with the evenness

and can be interpreted as a difference of areas below the curves shown on figure 2.4. Note that in this expression, the x_i need to be ordered, which is the main source of difficulties in the analytical manipulation of evenness.

Since all WITS populations follow the same dynamics, and assuming that the initial distribution is even, then each WITS should have an average final frequency of $1/h$. However, that does not mean that the average final evenness should be equal to one. Let us illustrate this point, with an example in the simplest case of $h = 2$. Let us note $\{m_1, m_2\}$ the absolute numbers of WITS corresponding to the WITS frequencies $\{x_1, x_2\}$, with always $x_1 < x_2$. In the most uneven case, $x_1 = 0$ and $x_2 = 1$, thus $S_{uneven} = 1$. In the most even case, $x_1 = x_2 = 1/2$ and thus $S_{even} = 3/2$. For any distribution $\{x_1, x_2\}$, $S(\{x_i\}) = 2x_1 + x_2 = 1 + x_1$. Let us assume for example that x_1 and x_2 are identically distributed, with a Poisson distribution of mean λ, and try to calculate the mean evenness:

$$\langle x_1 \rangle = \left\langle \frac{m_1}{m_1 + m_2} \right\rangle = \sum_{j=0}^{+\infty} e^{-\lambda} \frac{\lambda^j}{j!} \sum_{i=0}^{j} e^{-\lambda} \frac{\lambda^i}{i!} \frac{i}{i+j}$$

I found no simple exact expression. Ordering the distribution necessarily implies to introduce partial sums, even for the simplest case; thus the evenness is not suitable for analytical analysis.

Entropy A more natural observable to characterize variability for a physicist is entropy. The Shannon entropy is expressed as:

$$H(X) = \sum_{i=1}^{\infty} P_i \log(P_i)$$

with P_i the probability that the random variable X takes the value i, *i.e.* the probability that the final size of a WITS population is i in our case. Although this quantity should in principle be less difficult to manipulate analytically in general, it is not ideal either, since we do not have enough data at our disposal to make relevant estimates of the values of the P_i.

Comparison on real data We will see in the following that the variance is a much more suitable observable both to handle analytically and to estimate from the data. On figure 2.5 we check that variance and evenness are correlated, meaning variation is similarly translated in both.

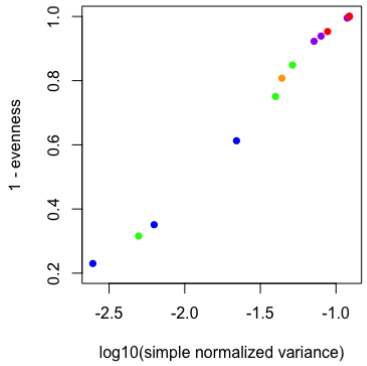


Figure 2.5 – Comparison of the evenness index and the simple normalized variance (see section 2.3.2) on empirical data. Each color corresponds to a combination of an inoculum size and a strain.

2.3.2 Variance

A natural observable to characterize the variability of the WITS distribution is a simple variance on the final numbers of each of the h different WITS strains $\{m_1, ...m_h\}$. We will see that the mean variance expected from the model is a much more natural quantity to derive analytically.

2.3.2.1 Mean expected variance

With Bessel's correction 1.2.2.3 the expression of the variance estimated from the final numbers of WITS of one experiment or one run of simulation is:

$$var = \frac{1}{h-1}\sum_{i=1}^{h}\left(m_i - \frac{1}{h}\sum_{j=1}^{h} m_j\right)^2 \tag{2.4}$$

Since we consider all WITS are equivalent, it is straightforward to check that the mean expected value of var is indeed $\langle m^2\rangle - \langle m\rangle^2$, which is the variance of the distribution. As it was shown in section 1.2.1.4, this variance can be expressed with derivatives of the generating function 1.5. In the end the mean expected variance writes:

$$\langle var\rangle = (\beta n_0 \exp((r-c)t))^2 \frac{1}{\beta n_0}\left(\frac{2r}{r-c} - \frac{r+c}{r-c}\exp(-(r-c)t)\right), \tag{2.5}$$

which in the limit $(r-c)t \gg 1$, and with $u = r/((r-c)\beta n_0)$ tends to:

$$\langle var \rangle_l = (\beta n_0 \exp((r-c)t))^2 \, 2u \tag{2.6}$$

i.e. the mean population size value 1.9 to the square multiplied by $2u$, with u a measure of the stochasticity: when the initial population is small (βn_0 small) and the loss rate is close to the replication rate ($(r-c) \ll r$), then u is large and stochasticity will have a large effect. On the other hand, if the initial population is large (βn_0 large) and the loss rate is small compared to the replication rate ($c \ll r$) then u is small and the effects of stochasticity are negligible.

2.3.2.2 Comparison with the WITS loss

We have also seen in section 1.2.1.4 how to calculate the WITS loss using the generating function 1.8. It writes:

$$\begin{aligned} P(n=0,t) = g(s=0,t) &= \exp\left(-\frac{n_0\beta(r-c)}{r - c\exp(-(r-c)t}\right) \\ &\xrightarrow[(r-c)t \gg 1]{} \exp\left(-\frac{n_0\beta(r-c)}{r}\right) = e^{-\frac{1}{u}} \end{aligned} \tag{2.7}$$

We see from expressions 2.6 and 2.7 that in the limit $(r-c)t \gg 1$ the variance and the WITS loss depend on the same combination u of the parameters of the model. So, fundamentally, the variance and the loss probability contain similar information on the parameters, and they cannot separate β and c, since the parameters are similarly combined. However, if the model matches properly the data, it also means that we should be able to check that those two observables actually convey the same information.

2.3.2.3 Variance of the variance

The variance calculated in the previous section is only a mean expected value: for each realization of the same experiment (a run of the simulation or one mouse), the final numbers of WITS $\{m_1, ...m_h\}$ will give a different estimate of the variance with expression 2.4. Thus, providing that more than one experiment/simulation is completed, it is possible to estimate a variance on the variance observable, and also to get an expression for its expected value in the model. It is useful to estimate the expected variance on this observable. If we measure the variance on a single mouse, the variance of the variance can help us estimate a lower bound on the expected accuracy of our estimate. When measuring the variances on several mice, the experimental variance of the variance could be compared with the theoretical one (and likely be larger, as in the theoretical one, we assume that there is no inter-mice differences).

The expected variance of the variance writes as follows:

$$\langle var_{var} \rangle = \left\langle \left(\frac{1}{h-1} \sum_{i=1}^{h} \left(m_i - \frac{1}{h} \sum_{j=1}^{h} m_j \right)^2 - \left(\langle m^2 \rangle - \langle m \rangle^2 \right) \right)^2 \right\rangle$$

In appendix B.1 we derive the whole expression (which might have already been derived elsewhere) and find:

$$\begin{aligned}\langle var_{var}\rangle = &\left(e^{(r-c)t}n_0\beta\right)^4 \frac{2r-(r+c)e^{-(r-c)t}}{h(h-1)(r-c)^2(n_0\beta)^2}\times \\ &\left(\frac{(h-1)(c^2+10cr+r^2)e^{-2(r-c)t}}{(r-c)\beta n_0} - \frac{12(h-1)r(c+r)e^{-(r-c)t}}{(r-c)\beta n_0}\right. \\ &\left. -2h(r+c)e^{-(r-c)t} + 4hr + \frac{12r^2(h-1)}{n_0\beta(r-c)}\right)\end{aligned} \tag{2.8}$$

In the limit $(r-c)t \gg 1$,

$$var_{var} \to \left(\beta n_0 e^{(r-c)t}\right)^4 \frac{8u^2(h+3(h-1)u)}{h(h-1)}.$$

with $u = \frac{r}{\beta n_0(r-c)}$. As expected, it scales with the mean population size to the power four, and the higher the stochasticity (*i.e.* the higher u), the higher the variance. We also notice that, in this limit:

$$\frac{var_{var}}{\langle var\rangle^2} \propto \frac{2h(h+3(h-1)u)}{h(h-1)}.$$

Also, in the limit of h large, var_{var} scales as $1/h$.

2.3.3 Variance conditioned on WITS survival

One idea to try to decorrelate the information obtained from the WITS loss and the information obtained from the variability, is to restrain the study of the variability to the surviving WITS. The mean number of bacteria per WITS, conditioned on WITS survival, with n_s the number of surviving WITS types is:

$$\langle m\rangle_s = \left\langle \frac{1}{n_s}\sum_{i\in s} m_i \right\rangle = \frac{\sum_{q=1}^{\infty} p(q)q}{\sum_{q=1}^{\infty} p(q)} = \frac{\sum_{q=0}^{\infty} p(q)q}{1-p(0)} = \frac{\partial_s g|_{s=1}}{1-g(0)}.$$

With eq. (1.8),

$$\langle m\rangle_s = \frac{\beta n_0 e^{(r-c)t}}{1-e^{-X}},$$

with

$$X = \frac{\beta n_0(r-c)}{r-ce^{-(r-c)t}}.$$

In the limit $(r-c)t \gg 1$, $X \to 1/u$, and

$$\langle m\rangle_s \to \frac{\beta n_0 e^{(r-c)t}}{1-e^{-1/u}}.$$

Again, the variance used is the variance with the Bessel's correction:

$$var_s = \frac{1}{n_s - 1} \sum_{i \in s} (m_i - \langle m \rangle_s)^2.$$

The expected mean value of the variance is:

$$\langle var_s \rangle = \langle m^2 \rangle_s - \langle m \rangle_s^2$$

$$\langle var_s \rangle = \frac{1}{1 - g(0)} \left(\left. \frac{\partial^2 g}{\partial s^2} \right|_{s=1} + \left. \frac{\partial g}{\partial s} \right|_{s=1} - \frac{1}{1 - g(0)} \left(\left. \frac{\partial g}{\partial s} \right|_{s=1} \right)^2 \right).$$

With eq. (1.8), we obtain:

$$\langle var_s \rangle = \frac{e^{2(r-c)t} n_0 \beta}{(r-c)(1 - e^{-X})} \left(n_0 \beta (r-c) + 2r - (r+c) e^{-(r-c)t} + \frac{(r-c) n_0 \beta}{1 - e^{-X}} \right).$$

In the limit $(r-c)t \gg 1$,

$$\langle var_s \rangle \to \left(\frac{\beta n_0 e^{(r-c)t}}{1 - e^{-1/u}} \right)^2 (2u(1 - e^{-1/u}) - e^{-1/u}) = \langle m \rangle_s^2 (2u(1 - e^{-1/u}) - e^{-1/u}).$$

Again, all is function of $\langle m \rangle$ and u, thus it does not help more in distinguishing $n_0\beta$ and $(r-c)/r$. Another issue with this observable is that if there is an error in the presence/absence of a WITS (WITS present in very small proportion), then it bias considerably the result.

2.3.4 Variance over the growth factor

Among all the indicators of variability for the WITS distribution seen in the previous sections, the simple variance over the absolute final numbers is the only quantity that can easily be analytically predicted. However, until now, we have always made the assumption that the initial numbers of WITS were all equal, which meant that the h final numbers could be seen as h independent draws from the same random variable. But in the experiments, there is always some initial variability in the concentrations of the solutions of the different WITS prepared to make the inoculum. This translates into different n_i values, and this variability is sometimes non-negligible compared to the final variability. It is thus essential that this initial variability can be taken into account, and we will see in this section how to do so. The solution we propose is to look at the variance over the ratio between the final (m_i) and initial (n_i) numbers of WITS m_i/n_i. These expressions might have been derived in the literature, but we rederive them, building on the previous section.

2.3.4.1 Mean growth rate

It is straightforward that the mean expected value of the random variables m_i/n_i are the same for all i and equal to:

$$\left\langle \frac{m_i}{n_i} \right\rangle = \beta \exp((r-c)t) \tag{2.9}$$

2.3.4.2 Variance

With Bessel's correction:

$$\langle var \rangle = \left\langle \frac{1}{h-1} \sum_{i=1}^{h} \left(\frac{m_i}{n_i} - \frac{1}{h} \sum_{j=1}^{h} \frac{m_j}{n_j} \right)^2 \right\rangle = \frac{1}{h-1} \sum_{i=1}^{h} \left\langle \left(\frac{m_i}{n_i} - \frac{1}{h} \sum_{j=1}^{h} \frac{m_j}{n_j} \right)^2 \right\rangle$$

Expanding the square:

$$\langle var \rangle = \frac{1}{h-1} \sum_{i=1}^{h} \left(\left\langle \left(\frac{m_i}{n_i} \right)^2 \right\rangle - \frac{2}{h} \left\langle \frac{m_i}{n_i} \sum_{j=1}^{h} \frac{m_j}{n_j} \right\rangle + \frac{1}{h^2} \left\langle \left(\sum_{j=1}^{h} \frac{m_j}{n_j} \right)^2 \right\rangle \right)$$

As the WITS are independent, separating the term $i = j$ from the others in the second right-hand side term:

$$\begin{aligned} \langle var \rangle = & \frac{1}{h-1} \sum_{i=1}^{h} \left(\left(1 - \frac{2}{h} \right) \left\langle \left(\frac{m_i}{n_i} \right)^2 \right\rangle - \frac{2(h-1)}{h} \left\langle \frac{m_i}{n_i} \right\rangle^2 \right. \\ & \left. + \frac{1}{h^2} \sum_{j=1}^{h} \left\langle \left(\frac{m_j}{n_j} \right)^2 \right\rangle + \frac{h(h-1)}{h^2} \left\langle \frac{m_i}{n_i} \right\rangle^2 \right) \end{aligned}$$

$$\langle var \rangle = \frac{1}{h-1} \sum_{i=1}^{h} \left(\left(1 - \frac{2}{h} \right) \left\langle \left(\frac{m_i}{n_i} \right)^2 \right\rangle + \frac{1}{h^2} \sum_{j=1}^{h} \left\langle \left(\frac{m_j}{n_j} \right)^2 \right\rangle - \frac{h-1}{h} \left\langle \frac{m_i}{n_i} \right\rangle^2 \right)$$

The second sum does not depend on i. Thus $\sum_{i=1}^{h} \sum_{j=1}^{h} \left\langle \left(\frac{m_j}{n_j} \right)^2 \right\rangle = h \sum_{j=1}^{h} \left\langle \left(\frac{m_j}{n_j} \right)^2 \right\rangle$ and the two first terms can be regrouped:

$$\langle var \rangle = \frac{1}{h-1} \left(-(h-1) \left\langle \frac{m_i}{n_i} \right\rangle^2 + \sum_{i=1}^{h} \left[\left(1 - \frac{1}{h} \right) \left\langle \left(\frac{m_i}{n_i} \right)^2 \right\rangle \right] \right)$$

$$\langle var \rangle = - \left\langle \frac{m_i}{n_i} \right\rangle^2 + \frac{1}{h} \sum_{i=1}^{h} \left\langle \left(\frac{m_i}{n_i} \right)^2 \right\rangle \tag{2.10}$$

We now study:

$$\left\langle \left(\frac{m_i}{n_i} \right)^2 \right\rangle = \left\langle \frac{m_i(m_i - 1)}{n_i^2} \right\rangle + \left\langle \frac{m_i}{n_i^2} \right\rangle = \frac{1}{n_i^2} \left. \frac{\partial^2 g_i}{\partial s^2} \right|_{s=1} + \frac{\beta e^{(r-c)t}}{n_i}$$

Using the definition of the generating function,

$$\left\langle \left(\frac{m_i}{n_i} \right)^2 \right\rangle = \beta^2 e^{2(r-c)t} \left(1 + \frac{2r(1 - e^{-(r-c)t})}{n_i \beta (r-c)} \right) + \frac{\beta e^{(r-c)t}}{n_i}, \tag{2.11}$$

which put back in the equation 2.10 for the variance leads to:

$$\langle var \rangle = \frac{1}{h}\sum_{i=1}^{h}\left(\beta^2 e^{2(r-c)t}\left(1+\frac{2r(1-e^{-(r-c)t})}{n_i\beta(r-c)}\right)+\frac{\beta e^{(r-c)t}}{n_i}\right)-\beta^2 e^{2(r-c)t}$$

$$\langle var \rangle = \beta^2 e^{2(r-c)t}\frac{2r-(r+c)e^{-(r-c)t}}{\beta(r-c)}\frac{1}{h}\sum_{i=1}^{h}\frac{1}{n_i} \tag{2.12}$$

We notice that in the limit $(r-c)t \gg 1$, this variance tends to:

$$\langle var \rangle_l = (\beta \exp((r-c)t))^2\left[\frac{1}{h}\sum_{i=1}^{h}\frac{1}{n_i}\right]\frac{2r}{\beta(r-c)} \tag{2.13}$$

and there is thus always a dependence in the same combination of the parameters $\tilde{u} = \frac{r}{\beta(r-c)}$. We also check that if all the n_i are equal to n_0, the results from the simple variance 2.5 and 2.6 are recovered.

2.3.4.3 Variance on the variance

Following the same reasoning as for the simple variance, a mean expected variance over this new variance can be calculated. We show in appendix section B.2 that:

$$\begin{aligned}\langle var_{var} \rangle = \left(\beta^4 e^{4(r-c)t}\right)\Bigg[&\frac{2\left((c+r)e^{-(r-c)t}-2r\right)^2}{\beta^2(h-1)^2h^2(r-c)^2}SN^2\\ &+\frac{2(h-2)\left((c+r)e^{-(r-c)t}-2r\right)^2}{\beta^2(h-1)^2h(r-c)^2}SN2\\ &-\frac{e^{-3rt}}{\beta^3h^2(r-c)^3}\Big\{c^3e^{3ct}+c^2r\left(11e^{3ct}-14e^{t(2c+r)}\right)\\ &+r^3\left(-14e^{t(2c+r)}+36e^{t(c+2r)}+e^{3ct}-24e^{3rt}\right)\\ &+cr^2\left(-44e^{t(2c+r)}+36e^{t(c+2r)}+11e^{3ct}\right)\Big\}SN3\Bigg].\end{aligned} \tag{2.14}$$

with $SN = \sum_{i=1}^{h}\frac{1}{n_i}$, $SN2 = \sum_{i=1}^{h}\frac{1}{n_i^2}$ and $SN3 = \sum_{i=1}^{h}\frac{1}{n_i^3}$. Note that if all the n_i are replaced by n_0, equation 2.8 divided by n_0^4 is recovered.

In the limit of $(r-c)t \gg 1$,

$$\langle var_{var} \rangle \to \left(\beta e^{(r-c)t}\right)^4\frac{8u^2}{(h-1)^2h^2}\left(SN^2+h(h-2)SN2+3u(h-1)^2SN3\right)$$

with $u = \frac{r}{\beta(r-c)}$. As expected, it scales with the mean growth factor to the power four, and the higher the stochasticity (*i.e.* the higher u), the higher the variance. Also, in the limit of large h, it scales as $1/h$.

2.3.5 Summary of the quest for a new observable

To characterize the variability in the WITS distribution, the evenness index was first chosen for continuity reasons with previous work of our collaborators. It proves to be convenient for simulations, but it is hard to handle analytically. To overcome this difficulty, the variance on the final numbers of WITS was first considered: it is indeed easier to derive but provides no new information on the parameters, and neither does the variance restrained on surviving WITS that was considered next. More importantly, these observables did not take into account the initial variability of the WITS distribution. We thus introduced the variance over the growth factor, which is the observable we will keep for our study.

2.4 Strategy for parameters estimation

Now that we have settled for the choice of our second observable for the variability in the WITS distribution, with the variance over the growth factors, let us go back to the question of parameters estimation. Using the different observables, let us look at the constraints imposed on the parameters by the data.

2.4.1 Constraint on β and c from the mean growth rate

In addition to the data on the WITS distribution, we can also exploit the initial and final total count of bacteria measured experimentally combined to the data of the plasmids dilution for another observable, the one of the mean growth rate. As seen in equation 2.9, it expresses as:

$$\left\langle \frac{B_f}{N_0} \right\rangle = \beta \exp((r-c)t)$$

In this expression, r can be replaced by $r_{mean} = \frac{G \ln 2}{t}$ given by the data of the plasmids dilution (see section 1.1.2 and 2.2.1). This gives a first constraint on the parameters (β, c).

2.4.2 Constraint on β and c from the tags loss

As seen in the previous section, for a WITS population i starting with n_i bacteria, the probability that this population dies during the process modeled is given by the following probability distribution function:

$$P(0,t|n_i) = \exp\left[-\beta n_i \frac{(r-c)e^{(r-c)t}}{re^{(r-c)t} - c}\right]$$

This probability is thus different for each WITS. Let us write Θ_i the random variable taking the value 1 if the population of WITS i got extinct in the process, and 0 otherwise. The probability to get the value θ_i for Θ_i writes as the binomial (either WITS i gets lost, or not):

$$P(\theta_i) = (P(0,t|n_i))^{\theta_i}(1 - P(0,t|n_i))^{(1-\theta_i)}$$

Thus we can apply the method described in section 1.2.1.5 on this probability distribution and can define the likelihood as:

$$L(\beta) = \prod_{i=1}^{n} P(\theta_i)$$

and taking the derivative of the logarithm of this quantity, assuming all the other parameters well known, one can find β as the root of the following equation:

$$\begin{aligned}\frac{d\ln(L(\beta))}{d\beta} = 0 \ &= -\sum_{i=1}^{h} n_i\theta_i \left[\frac{(r-c)e^{(r-c)t}}{re^{(r-c)t}-c}\right] \\ &+ \sum_{i=1}^{h}(1-\theta_i)n_i \left[\frac{(r-c)e^{(r-c)t}}{re^{(r-c)t}-c}\right] \frac{\exp\left(-\beta n_i \left[\frac{(r-c)e^{(r-c)t}}{re^{(r-c)t}-c}\right]\right)}{1-\exp\left(-\beta n_i \left[\frac{(r-c)e^{(r-c)t}}{re^{(r-c)t}-c}\right]\right)}\end{aligned}$$

2.4.3 Constraint on β and c from the renormalized variance over the growth factor

It has been seen in the previous section that the variance can be expressed as a function of both β and c (see eq. 2.12). Thus, the value of the experimental variance gives a constraint linking β and c.

2.4.4 Summary of the strategy for parameters estimation

The strategy to extract the parameters of the model from the data will therefore be the following:

1. Depending on the case, use either r_{mean} or r_{max} for r as defined and explained in section 2.2.1.

2. Explore all the possible sets of parameters β and c: for each set, calculate the expected value for the three observables (the growth factor, the variance on the growth factor and the tags loss) and see which region of the parameters space is compatible with their experimental values.

3. Each observable will result in a $curve(\beta, c)$ of plausible parameters. First, we will check that the estimates are compatible, *i.e.* that these curves either cross or are close enough in at least one region of the parameter space. Then if this region is smaller than each curve, that will give some constrained estimate of (β, c).

The same strategy can be adopted using simulations instead of analytics, as it will be shown in the next section.

2.5 Simulations and results

In this section, we will see how the reasoning and techniques developed in the previous sections applies to the data set that drove our study (the data set can be found in appendix A), and what conclusions can be drawn from them. Of course the same techniques can be applied to larger amounts of the same kind of data, or even adapted to other systems behaving similarly.

2.5.1 Determining the carrying capacity

In order to run the simulations, one needs to determine one additional parameter compared to the set of parameters used for the analytical study, where we neglected the growth saturation: we need to estimate K, the carrying capacity. The easiest way to do so, which also allows for the determination of a K value that does not vary much with the other parameters of the model, is to consider that at the end of the experiment (24h post-infection), equilibrium of the population size is reached (which is also coherent with the data shown on figure 2.1). Let us consider $B(t)$ the total number of bacteria in the cecum at time t. At the end of the experiment, the size of the bacterial population is very large (of the order of 10^9 at least), and thus we take the deterministic equation:

$$\frac{\partial B}{\partial t}(t) = \left(\left(1 - \frac{B(t)}{K}\right) r_{max} - c\right) B(t)$$

Then when equilibrium is reached, $\frac{\partial B}{\partial t} = 0$ and thus

$$K = \frac{B_f r_{max}}{r_{max} - c}$$

with B_f the total final number of bacteria (which can be averaged over several repetitions of the same experiment, in our case three mice per experiment). Thus, since we are going to explore different values of the parameters β and c, the value of K put in the simulations will have to be recalculated for every new value of c, but since c should be relatively small compared to r_{max}, there will not be a large range of values for K.

2.5.2 Results

For each set of parameter (β, c), a simulation can be launched, with all the other parameters being known : we start with the initial numbers of WITS n_i and the initial number of untagged bacteria N_0 experimentally measured (see the data concerning the inoculum in the appendix A, table A.2). We use an effective replication rate $r_{max}\left(1 - \frac{B(t)}{K}\right)$, recalculated at each new time step with the new value of $B(t)$ (the total number of bacteria in the system at time t), with r_{max} and K determined as previously explained (section 2.5.1). Many runs of the same simulation are repeated, so that we can average on the different iterations all the observables of interest : the proportion of WITS lost, the renormalized variance

over the growth rate (that we will simply call "variance" hereafter), and also the mean growth rate. The full source program in .R is shown in appendix C. Simultaneously, the theoretical expected value for the same observables can be calculated, using either r_{max} or r_{mean} for r, as exposed in section 2.2.1:

- r_{max} for the two observables – WITS loss and variance – linked to the WITS distribution that mostly depends on the early dynamics, and
- r_{mean} for the mean growth rate, which writes $\left\langle \frac{m_i}{n_i} \right\rangle = 2^G e^{-ct} = \beta e^{(r_{mean}-c)t}$ with G the number of replications estimated from the plasmids dilution,

to compensate the fact that we do not take saturation into account in our analytical study. We thus obtain parametric maps of those observables, as presented for example for the variance on figures 2.6. On those heat maps can be added contour lines, corresponding to the experimental values of the observables under consideration (calculated from the data in appendix A table A.3). These curves indicate the region of the parameter space which is coherent with the data, in the model we chose, for each observable considered.

Then these different contour lines can be superposed on the (β, c) grid, this is what is shown on figures 2.7 and 2.8. To the contour lines of the variance (in red) can be added the ones for the WITS loss. There are several possibilities for the WITS loss :

1. The log-likelihood maximization method can be used as seen in section 2.4 to estimate a value of β for each value of c in order to take properly into account the uneven inoculum.
2. The contour curve of the experimental loss can be traced on the landscape of the lost proportion of WITS in the simulations. This method takes into account the initial unevenness in the simulation but not in the way the final lost proportions is counted (with simply the number of WITS lost over h the number of WITS, averaged over the three mice, or over the simulation runs, regardless of the initial WITS proportions).
3. The same experimental contour line can be taken, but on the landscape of the analytical loss probability (expression 2.1), making this time the approximation of an equal inoculum (taking all the n_i to be $\langle n_{0i} \rangle$).

Figures 2.7 and 2.8 show (in blue and green) that these three options give very similar results.

Last but not least, to the variance and the WITS loss is added the mean growth factor: its theoretical value on the (β, c) grid is $\beta 2^G e^{-ct} = \beta e^{(r_{mean}-c)t}$ with G the number of replications estimated from the plasmids dilution. On this theoretical landscape we draw the contour lines of the experimental growth factor, calculated as the ratio between the final and initial whole population sizes, both experimentally measured[2].

[2]In the simulations, B_f/N_0 is already used to constraint K. We could then look at the final average number of replications G_{simu} and compare it with its experimental value.

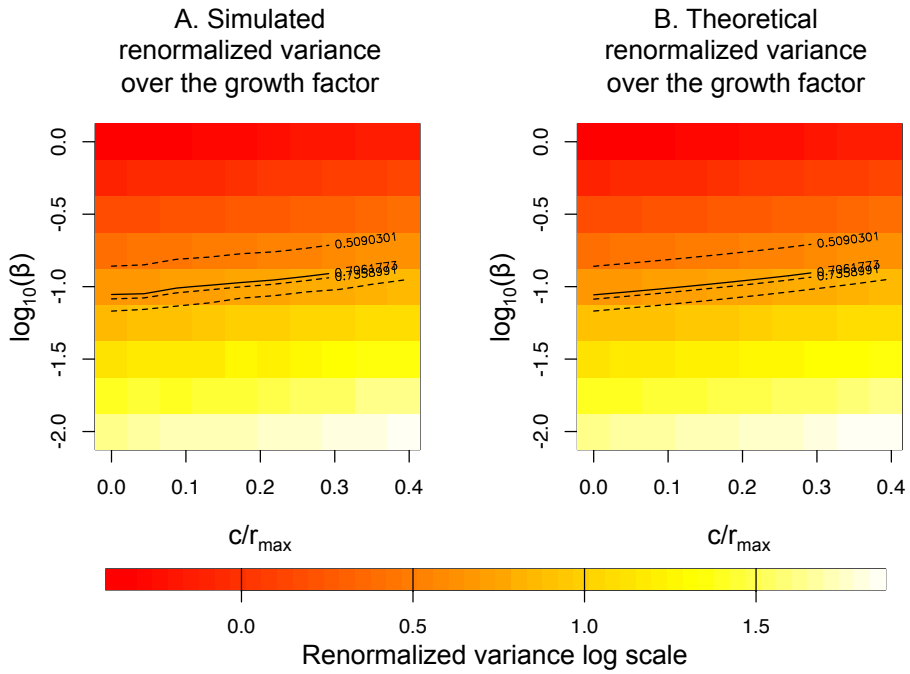


Figure 2.6 – Parametric maps for the renormalized variance over the growth factor: "x" coordinate is the loss rate c in units of the initial replication rate r_{max}, "y" coordinate is the logarithm in base 10 of the establishment probability β. The heat colors correspond to the variance in log scale (see the color scale), either as resulting from the simulations (A, ratio of the variance on the growth factor averaged on 5000 iterations on the square of the final growth factor averaged on the same iterations) or from analytics (B, ratio of expression 2.12 on the square of expression 2.9 with $r = r_{max}$). The variance is higher when the numbers of bacteria are lower, especially the initial numbers: it is thus higher for the lower values of β and the higher values of c. The initial numbers of bacteria put in the model are from three repeats of an experiment with the strain SB300, starting from an inoculum of size 10^3, containing approximately 5 CFU of each of the seven WITS (see data tables A.2 andA.3). The final experimental values for the variance after one day is shown on the three dashed contour curves, and the average on the three mice is represented in solid line. It corresponds to the sets of parameters (β, c) authorized by the variance.

Then some uncertainties on the estimates can be added to the picture. For the variance, we can get the coordinates of the points on the contour line, and calculate the expected variance on the variance as in 2.14 and renormalize it with the expected growth rate 2.9 to the power four, each of them calculated with r_{max}. For each point on the variance contour line we thus have a confidence interval of $+/-$ the square root of this variance of the variance on the analytical figure. Since we do not know *a priori* the shape of the variance distribution, on the figure with the simulations we also add the 10% and 90% quantiles of

the distribution of the variance resulting from additional simulation runs for the (β, c) coordinates of the variance contour curve. Then, let us write p the probability for one WITS to be lost (which we estimate by the proportion of WITS lost experimentally). The number of WITS lost follows a binomial law (each WITS being seen as an independent draw of the same Bernouilli variable) which standard deviation writes $\sqrt{hp(1-p)} \simeq \sqrt{hp}$, and the standard deviation on the proportion thus writes $\sqrt{p/h}$. We thus add the contour lines corresponding to the experimental lost proportion $+/-$ this standard deviation on the landscape maps of the lost proportion. On the simulations figure we also add the 10% and 90% quantiles of the distribution of the proportion of WITS lost resulting from additional simulation runs for the (β, c) coordinates of the experimental loss contour curve. All of this is shown on figures 2.7 and 2.8.

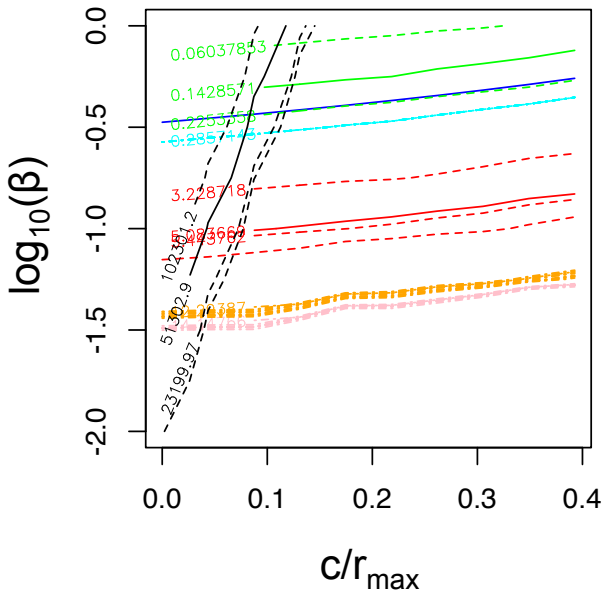
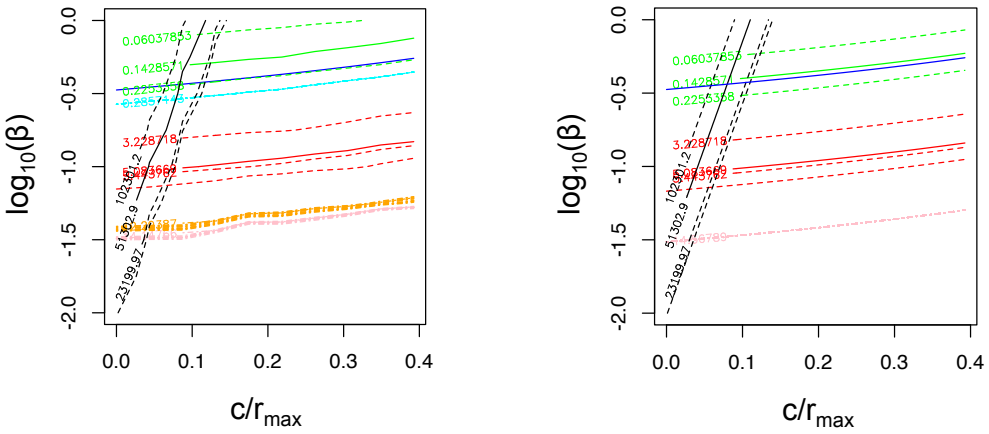


Figure 2.7 – Contour curves for the experimental values of the observables in the parameter space (β, c). Same experiment as in figure 2.6.
On the simulated variance landscape: red dashed: contour for the final experimental variance in three mice, and red solid line: for the mean value. Pink two-dashed lines: contour for the mean experimental variance +/− the square root of the theoretically expected value for the renormalized variance of the variance at this place of the parameter space (one line for each coordinate extracted from the contour line). The upper line does not show because for these values of the parameters the variance is so high that the average is smaller than the square root of the variance. Orange two-dashed lines: contour for the 10% and 90% quantiles of the variance distribution resulting from additional runs of simulations (5000 iterations for each point) for the coordinates extracted from the variance contour line (one line for each coordinate). The upper line does not show because its corresponding value does not appear in this range of parameters.
Relative to the WITS loss: Blue line: $\beta(c)$ as calculated from the log-likelihood maximization (option 1). Green line: contour of the experimental proportion of WITS lost (combining the data on the three mice) on the simulated proportion of WITS lost landscape (option 2). Green doted lines: on the same landscape, contour lines for the proportion of WITS lost +/− the expected binomial standard deviation. Cyan: on the same landscape, contour for the 10% and 90% quantiles of the proportion of WITS lost distribution resulting from additional runs of simulations (5000 iterations for each point) for the coordinates extracted from the experimental WITS loss contour line (in green). The upper line does not show, because its corresponding value is not found on the map.
On the growth factor landscape: **Black**: using the analytical expression $\beta 2^G e^{-ct}$, contour for the growth factor experimental value in the three mice (dashed lines) and for the average (solid line).



A. Contour lines on the simulated landscapes

B. Contour lines on the theoretical landscapes

Figure 2.8 – A. is a repetition of figure 2.7, repeated to ease visual comparison. B. Everything as in A, except that the contour curves for the variance and the WITS loss are taken on the landscapes analytically inferred, instead of the simulated ones, and that the cyan and orange lines corresponding to simulated confidence intervals (see figure 2.7) are not shown here. Note that the grid pattern is finer (since analytics do not require high calculation times), which accounts for the smoother aspect of the lines (in particular, note that in both figure the contour for the growth factor is on the analytical landscape, the only difference for the black lines is thus the grid pattern).

2.5.3 Discussion

In this chapter, I have presented a one-population model of early infection of the gut. I first exposed the biological grounds to this model, and made a first oversimplified approximation that allowed me to present simply the principles underlying my approach. I then looked for an observable to characterize the distribution of the genetic tags to complement the tags loss, and settled for the renormalized variance over the growth factor, which is derivable analytically and allows to take properly into account the initial variability in the tags distribution. I then reviewed how the experimental data of the three observables (growth factor, variance and WITS loss) constrains the parameters of our model, and presented parametric maps of these constraints materialized in contour curves in the last section.

What we observe first is that simulations and analytics give very similar results, despite the fact that in the analytic study we neglect the saturation of growth. What counts for the WITS distribution is the initial dynamics, when the absolute numbers are low enough so that the effects of stochasticity are important, and when the carrying capacity is, indeed, far from being reached (the inoculum size varies from 10^3 to 10^7 and the carrying capacity is of the order of 10^9). Thus we bypass successfully the problem risen by this no-saturation analytic approximation by using either a replication rate corresponding to the initial rate, r_{max}, when it comes to the observables on the WITS distribution, and the mean replication rate r_{mean} (or, equivalently, $(G \ln 2)/t$ with G the mean number of replications estimated by the plasmids dilution data) when it comes to calculating the growth factor. The only noticeable difference between analytics and simulation results is for the WITS loss contours (green curves), but this is explained by the fact that in the simulations, we can start from an uneven distribution, while for the theoretical parametric map, the approximation that we start from an even distribution of same mean was made.

We also observe that the data of the experimental growth factor constrains the range of possible values for c to small values in front of r_{max}, which is coherent to what we expected, regarding the fact that we focus on the beginning of the infection. The fact that these contour lines are not parallel to the others shows very well that this observable contains another information on the parameters, through a dependence in $\beta e^{(r-c)t}$ (see equation 2.9) rather than in $\tilde{u} = \frac{r}{\beta(r-c)}$ like the two others (see equations 2.13 and 2.7). On the contrary, the curves for the variance and the WITS loss are parallel, which was expected since we showed that the two observables depend on the same combination of the parameters $\tilde{u}$. They are not superposing either, however. The question thus remains to know if the gap between the two curves is coherent with the level of noise expected in the framework of our model or if it surpasses it. The fact is that the answer to this question does not seem so clear for some experiments (see figure 2.9). It might indicate that the frame imposed by the choice of our model is too restricted to actually explain what is observed *in vivo*. In particular, we would like to consider models that allow for a higher variability in the WITS distribution. We

thus turned to models with several subpopulations, which are presented in the following chapter.

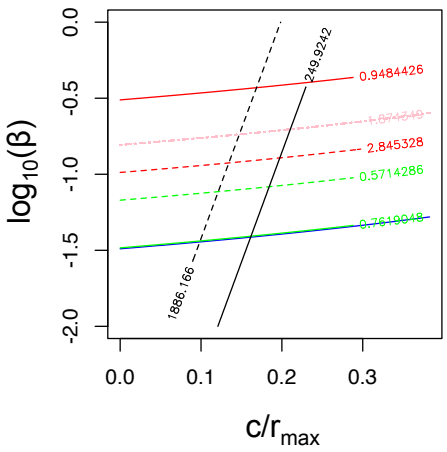


Figure 2.9 – Contour curves for the experimental values of the observables in the parameter space (β, c) for the experiment with the strain "SB300" and the inoculum size 10^5. Same color code: on the theoretical variance landscape, red dashed: contour for the final experimental variance in three mice (only one appears, because all the WITS were lost in the two other mice and a null variance does not appear on this grid of parameters), and red solid line: for the mean value. Pink two-dashed lines: contour for the mean experimental variance +/− the square root of the theoretically expected value for the renormalized variance of the variance at this place of the parameter space (one line for each coordinate extracted from the contour line). The upper line does not show because for these values of the parameters the variance is so high that the average is smaller than the square root of the variance. Blue line: $\beta(c)$ as calculated from the log-likelihood maximization (option 1). Green line: contour of the experimental proportion of WITS lost (combining the data on the three mice) on the simulated proportion of WITS lost landscape (option 2). Green doted lines: on the same landscape, contour lines for the proportion of WITS lost +/− its square root. **Black**: contour for the growth factor experimental value in the three mice (dashed lines) and for the average (solid line). In this experiment, the contour lines for the variance are situated above the ones for the WITS loss on the parametric map. With our confidence interval estimates, the two contours of the WITS loss and the variance seem to be further appart than the noise allowed in the framework of the one-population model.

Chapter 3

Two-subpopulations models

Contents

We have seen in the previous chapter that in some experiments, there was an incompatibility in the parameters estimates inferred from the experimental values of the different observables, in the framework of the one-population model. In this chapter, we will consider bacterial populations composed of several subpopulations, each following the same type of dynamics as in the null model exposed at the beginning of section 1.2: a Poissonian selection corresponding to the sampling in the inoculum and the possible bottleneck of the establishment in the cecum (other bacteria being directly eliminated or not surviving the acidity of the stomach), followed by a Markovian process with death and growth. What distinguishes the different subpopulations are the parameters, *e.g.* rates, associated to these processes. In a first subsection (3.1), we will see the various reasons why studying such models. Then we will see how the analytical study presented in the last section adapts to the new configuration of two subpopulations (section 3.2). We will see how to infer a maximum of parameters from the data and what strategy to adopt to determine the remaining free parameters, both analytically and with simulations (section 3.3). Finally, we will present and discuss the results obtained with this class of models (section 3.4).

3.1 Arguments for a two-subpopulations model

3.1.1 Biological arguments

Various biological factors could account for the simultaneous existence of bacteria of a same strain undergoing different dynamics in the cecum of a same mouse. I do not intend to discuss an exhaustive list of those factors here, but only to present a few ideas on the subject.

The first idea that comes to a physicist's mind is the idea of spatiality. In our models, we never take space into account, but always make the approximation that the cecum is a well-mixed, homogeneous environment. But various factors could easily contradict this hypothesis. First of all, nutrients could be inhomogeneously spread, and since it is well established that bacteria division rates highly depend on nutrients concentration [53], this could account for the coexistence of two different replication rates. Then, the anatomy of the cecum itself is not homogeneous either: there is an entry and a way out ([40], see also figure 1.1). One could thus imagine that bacteria located closer to the way out undergo a higher clearance rate. However, since the hydrodynamics of the cecum is not very well known – complicated peristaltic contractions of the wall muscles playing an important role in the mixing of the highly non-Newtonian digesta, in order to favor nutrients absorption [9, 10] – one can hardly speculate more on that matter.

Another possibility is that bacteria are in different phenotypic states (*i.e.* non-genetic diversity, mostly due to stochastic gene expression). For instance, a study from Sturm *et al.* [54] shows that under certain conditions *in vitro*, there is a bistable expression of a gene coding for a virulence factor: although all bacteria are isogenic, a part of the population expresses this gene and the rest does not. It is also established in this paper that the expression of this virulence factor comes at a cost: a reduced replication rate for those bacteria[1]. This is typically the type of phenotypic bi-expression which could explain different dynamics, but it is only one possibility among many others. Stewart *et al.* [55] reviews some of these other possibilities. For example, it has also been shown that not all *Salmonella* Tyhimurium express a flagella during infection[2]. We can speculate that there may be a link between replication rate and motility, either through an enhanced nutrients search for motile bacteria, or on the contrary, that flagella expression comes at a cost. In any case, non genetic diversity can effectively be responsible for the coexistence of different subpopulations of isogenic bacteria with different dynamics.

3.1.2 Qualitative argument

The other argument motivating our interest for two sub-populations models is our analysis of the data in the previous section. For instance on figure 2.7,

[1] On the whole population level however, this expression favors virulence by triggering the inflammation which clears the niche from a part of the *Salmonella* population, but more importantly, other species: it is thus a case of bet-hedging or of cooperation.

[2] It might be that the bacterial population divides labor, so that the more motile bacteria can better approach the epithelium to trigger systemic infection.

the contour curve of the proportion of WITS lost is situated above the contour curve of the variance on our parametric map (it is the case in almost all the experiments we consider), which as we can check on figure 2.6, is an area of lower variance than the experimental one. That means that in the framework of our one-population model, at fixed WITS loss proportion, the model does not allow for a high enough variance in the WITS distribution to get perfectly coherent parameters estimate. We are thus looking for models allowing for a higher variance in the WITS distribution for the same WITS loss. A model where two subpopulations replicate at different rates should allow it, providing that the initial WITS population sizes are low enough, as it is qualitatively discussed on figure 3.1.

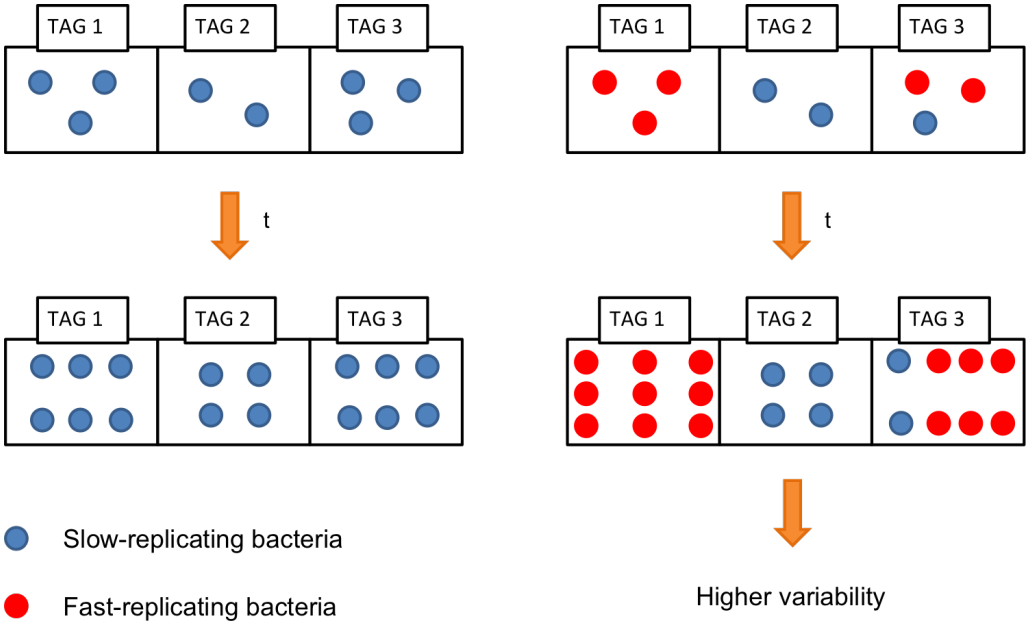


Figure 3.1 – As the initial WITS populations are taken to be low enough so that the effects of stochasticity can be observed on their distribution, and since bacteria are indifferentially genetically tagged, it might happen that the distribution of fast-replicating bacteria versus slow-replicating ones is specifically uneven in one of the WITS population, resulting after some replicating time in very different final sizes depending on the initial ratio, providing that the dynamics are otherwise identical.

3.1.3 Summary of the arguments for a two-subpopulations model

There are biological reasons that could make that genetically identical bacteria replicate at different rates. Qualitatively, having such subpopulations could increase the variance in the WITS for a given probability of WITS loss.

3.2 Analytical approach

In this section, we are going to see how the analytical study of the one-population model can be adapted to the case with two sub-populations: we will calculate the new generation function, and derive the expressions for the three observables (growth factor, WITS loss and variance).

3.2.1 Generating function

First of all, we need to write a generating function describing the whole population composed of the two sub-populations at once. Each of the subpopulations can be described by the generating function 1.8, providing that we replace the parameters (N_0, β, r, c) by the corresponding ones. We will write q (resp. $(1-q)$) the initial proportion of bacteria of type 1 (resp. type 2), so that $(qN_0, \beta_1, r_1, c_1)$ are the parameters to describe the dynamics of population 1 (resp. $((1-q)N_0, \beta_2, r_2, c_2)$ to describe population 2). To write a single generating function describing both populations at once, one needs two silent variables z_1 and z_2 and needs to use the simultaneous probability that n_1 bacteria of type 1 and n_2 bacteria of type 2 are present in the cecum, $P(n_1, n_2)$:

$$g(z_1, z_2) = \sum_{n_1, n_2} P(n_1, n_2) z_1^{n_1} z_2^{n_2}$$

Now $P(n_1, n_2) = P(n_1)P(n_2)$ if we assume the populations to be evolving independently. In our model, if we neglect the growth saturation as we have done already in our analytical study, this is the case. Thus,

$$\begin{aligned} g(z_1, z_2) =& g(z_1)g(z_2) \\ =& \exp\left[\frac{\beta_1 q N_0 (r_1 - c_1)(z_1 - 1)e^{(r_1 - c_1)t}}{r_1 z_1 - c_1 - (z_1 - 1)r_1 e^{(r_1 - c_1)t}} + \frac{\beta_2 (1-q) N_0 (r_2 - c_2)(z_2 - 1)e^{(r_2 - c_2)t}}{r_2 z_2 - c_2 - (z_2 - 1)r_2 e^{(r_2 - c_2)t}}\right] \end{aligned}$$

3.2.2 Observables calculations

For this class of models, we will use the same observables as developed in the one-population model case, adapted to the fact that we start with an uneven distribution of WITS: the mean growth factor, the WITS loss and the renormalized variance on the growth factor. In order to calculate them, we will use the same relations between the moments of the distribution and the derivatives of the generating function as in sections 1.2.1.4 and 2.3.4, but using the new generating function 3.2.1. Note that this new generating function can be adapted either to the whole population, or to each of the h WITS population by replacing N_0 by n_i (as was already done for the one-population model).

3.2.2.1 Mean growth factor

With m_i the final number of WITS of type i and m_{i1} (resp. m_{i2}) the final number of WITS of type i from population 1 (resp. 2), so that $m_i = m_{i1} + m_{i2}$:

$$\begin{aligned}\left\langle \frac{m_i}{n_i} \right\rangle &= \frac{1}{n_i}\left(\langle m_{i1}\rangle + \langle m_{i2}\rangle\right) = \frac{1}{n_i}\left(\left.\frac{\partial g}{\partial z_1}\right|_{z_1=1,z_2=1} + \left.\frac{\partial g}{\partial z_2}\right|_{z_1=1,z_2=1} \right) \\ &= q\beta_1 e^{(r_1-c_1)t} + (1-q)\beta_2 e^{(r_2-c_2)t}\end{aligned}$$

3.2.2.2 WITS loss

The extinction probability of WITS i now writes:

$$g(0,0) = \exp\left(-\beta_1 q n_i \frac{(r_1-c_1)e^{(r_1-c_1)t}}{r_1 e^{(r_1-c_1)t} - c_1} - \beta_2(1-q)n_i \frac{(r_2-c_2)e^{(r_2-c_2)t}}{r_2 e^{(r_2-c_2)t} - c_2} \right) \quad (3.1)$$

3.2.2.3 Variance on the growth rate

In our analytical study, m_{i1} and m_{i2} are independent random variables (no saturation). Thus the variance of $\frac{m_i}{n_i} = \frac{m_{i1}}{n_i} + \frac{m_{i2}}{n_i}$ is the sum of the variance for each subpopulation. We thus use expression 2.12 and replace the parameters accordingly:

$$\begin{aligned}\langle var\rangle = &\frac{1}{h}\left[q\beta_1^2 e^{2(r_1-c_1)t} \frac{2r_1 - (r_1+c_1)e^{-(r_1-c_1)t}}{\beta_1(r_1-c_1)} \right. \\ &\left. + (1-q)\beta_2^2 e^{2(r_2-c_2)t} \frac{2r_2 - (r_2+c_2)e^{-(r_2-c_2)t}}{\beta_2(r_2-c_2)} \right] \left(\sum_{i=1}^{h} \frac{1}{n_i} \right)\end{aligned}$$

One can check that taking $q = 0$ or 1, one recovers the one-population expression 2.12.

3.3 Parameters search strategy

In the following, we will focus on the models where the two subpopulations differ only by their replication rates $r_1 \neq r_2$. The replication rate is indeed biologically the most likely parameter to differ between subpopulations, for the various reasons exposed in section 3.1.1. In comparison, it is less likely that the parameter β, which corresponds to the probability for one bacteria to cross the acidity barrier of the stomach and settle in the cecum, should take different values for different individuals of the same isogenic bacterial population. This possible bottleneck corresponds indeed to a physical barrier that all bacteria have to pass, and known possible phenotypic differences are very unlikely to impact β. Different c could also be plausible (for example with the argument of spatiality presented in section 3.1.1) and have their importance, but since the c values are expected to remain small, this should have a negligible effect compared to changing the replication

rates. We will thus study a model of two subpopulations of respective parameters (qN_0, β, r_1, c) and $((1-q)N_0, \beta, r_2, c)$ (with N_0 the total number of bacteria in the inoculum). We will write the ratio between the two replication rates $\alpha = r_1/r_2$ and denote the faster-growing population 1, so that we will always have $\alpha \geq 1$.

Compared to the one-population case, we have thus added two dimensions to the parameter space, adding the two parameters α and q. In the former case, we fixed only the parameter r (either with r_{mean} estimated by the plasmid dilution or with r_{max} experimentally measured), and, for the simulation, the additional parameter K with the final number of bacteria. The two other parameters (β and c) remained free, and none of the three observables (mean growth rate, WITS loss and variance) were fixed. Here, we propose to fix all the parameters but α and q by enforcing the values of the experimental WITS loss and the experimental mean growth rate (we will see how below), so that only the variance remains free. We will thus explore the parameter space (q, α). The values $\alpha = 1$, $q = 0$ or $q = 1$ recover the one population case. Thus for these values, it should be impossible to recover the experimental variance when there was a gap in our parameters estimate from the different observables. But if other sets of parameters (q, α) allow the recovery of the experimental variance level, then it will mean that it is plausible that there actually are two subpopulations. Now let us see how to fix the other parameters values with the experimental values for the WITS loss and the mean growth rate.

3.3.1 Carrying capacity and replication rates

We will arbitrarily take $K = 10^9$ in the following (the order of magnitude of all the values for K calculated as exposed in section 2.5.1 in the one-population case), since we have seen that the exact value for K did not have much impact on the observables. Then we fit the initial growth rates $r_{max,1}$ and $r_{max,2}$ so that the growth of the whole population is on average of rate r_{max}. Since at the beginning of the experiment growth is exponential, this condition writes:

$$e^{r_{max}t} = qe^{r_{max,1}t} + (1-q)e^{r_{max,2}t}. \tag{3.2}$$

Since we have parametrized $r_{max,1} = \alpha r_{max,2}$, the upper equation has only one unknown at fixed (q, α) and can thus be numerically solved for each new point of the parameter space.

3.3.2 β and c with plasmid dilution, mean growth factor and WITS loss

Then, to fix the parameters β and c, we will use a combination of three informations: the mean replication rate given by the plasmid dilution, the mean growth factor measured in the experiments and the WITS lost in the experiments. The mean growth factor indeed writes as

$$\frac{B_f}{N_0} = \beta 2^G e^{-ct} = \beta e^{(r_{mean}-c)t}$$

with G the number of replications from the plasmids (see section 2.2.1), B_f and N_0 the final and initial experimental total numbers of bacteria, which are all the three experimental measures we have at our disposal apart from the WITS data. Consequently, this gives a first relation between the parameters β and c, that we can inject in the maximization of the log-likelihood (presented in sections 1.2.1.5 and 2.4) which now writes, with the loss probability of WITS i expressed above (3.1):

$$\begin{aligned}\frac{d\ln(L(\beta))}{d\beta} = 0 \ &= -\sum_{i=1}^{n} n_0^i \theta_i \left[q\frac{(r_1-c)e^{(r_1-c)t}}{r_1e^{(r_1-c)t}-c} + (1-q)\frac{(r_2-c)e^{(r_2-c)t}}{r_2e^{(r_2-c)t}-c} \right] \\ &+ \sum_{i=1}^{n}(1-\theta_i)n_0^i \left[q\frac{(r_1-c)e^{(r_1-c)t}}{r_1e^{(r_1-c)t}-c} + (1-q)\frac{(r_2-c)e^{(r_2-c)t}}{r_2e^{(r_2-c)t}-c} \right] \\ &\times \frac{\exp\left(-\beta n_0^i \left[q\frac{(r_1-c)e^{(r_1-c)t}}{r_1e^{(r_1-c)t}-c} + (1-q)\frac{(r_2-c)e^{(r_2-c)t}}{r_2e^{(r_2-c)t}-c} \right]\right)}{1-\exp\left(-\beta n_0^i \left[q\frac{(r_1-c)e^{(r_1-c)t}}{r_1e^{(r_1-c)t}-c} + (1-q)\frac{(r_2-c)e^{(r_2-c)t}}{r_2e^{(r_2-c)t}-c} \right]\right)}\end{aligned}$$

with θ_i taking the value 1 if the population of WITS i got extinct in the process and 0 otherwise, and taking $r_1 = r_{max,1}$ and $r_2 = r_{max,2}$ as the WITS loss depends mainly on the early dynamics. Those two constraints allow for an estimate of parameters β and c, which completes the set of parameters we wanted to fix with the experimental data. We can now switch to the exploration of the variance in the parameter space (q, α).

3.4 Simulations and results

In this section I will review the results obtained when applying a two-subpopulations model (with two subpopulations replicating at different rates), both analytically and with simulations, to the data set we are interested in. The full code for simulations is printed in annexe appendix C. As explained in the previous section, we fix the values of the parameters $(\beta, r_{max,1}, r_{max,2}, c, K)$ with the experimental values of the plasmids dilution, the growth factor and the proportion of WITS lost, and will explore the values allowed for the variance in the parameter space (q, α). The values fixed for β and c also depend on those parameters, so for each set of parameter (q, α) we will also check their estimated values for consistency with the range of values that is biologically expected. We will see in the next chapter (4) that c can be mechanically estimated if we consider the feces production, that are of 10% of the cecum content every 15min. This corresponds to a survival probability of $(0.9)^{4\times24}$ for a bacteria in the cecum after a day, which with our parameters expresses as e^{-ct}, and thus allows an estimate of $c = -96\ln(0.9) \simeq 10day^{-1}$. Then an underestimate of β can be obtained considering that all the WITS were lost during the Poisson process, as explained in section 2.2, which gives a range of estimated lower bounds for β between $0.01 - 0.4$.

For $q = 0$, $q = 1$ or $\alpha = 1$, the same value for the variance is found everywhere, as expected because we recover the one population case. For intermediate values

of the parameters, we obtain higher ranges of variance, as predicted. We first check that simulations and analytics give similar results (see for example figure 3.2). There is an excellent concordance between the variance calculated on the final numbers resulting from the Gillespie algorithm and the analytical variance, despite the approximation of no saturation we made in our analytical study. As we explained in the one-population model, this is because the variance depends mostly on the initial dynamics: when the WITS populations have reached a large enough size, then the distribution cannot change much. In the following we will thus only present analytical maps of the variance, since this allows a sharper exploration of the parameter space (without requiring a high calculating time).

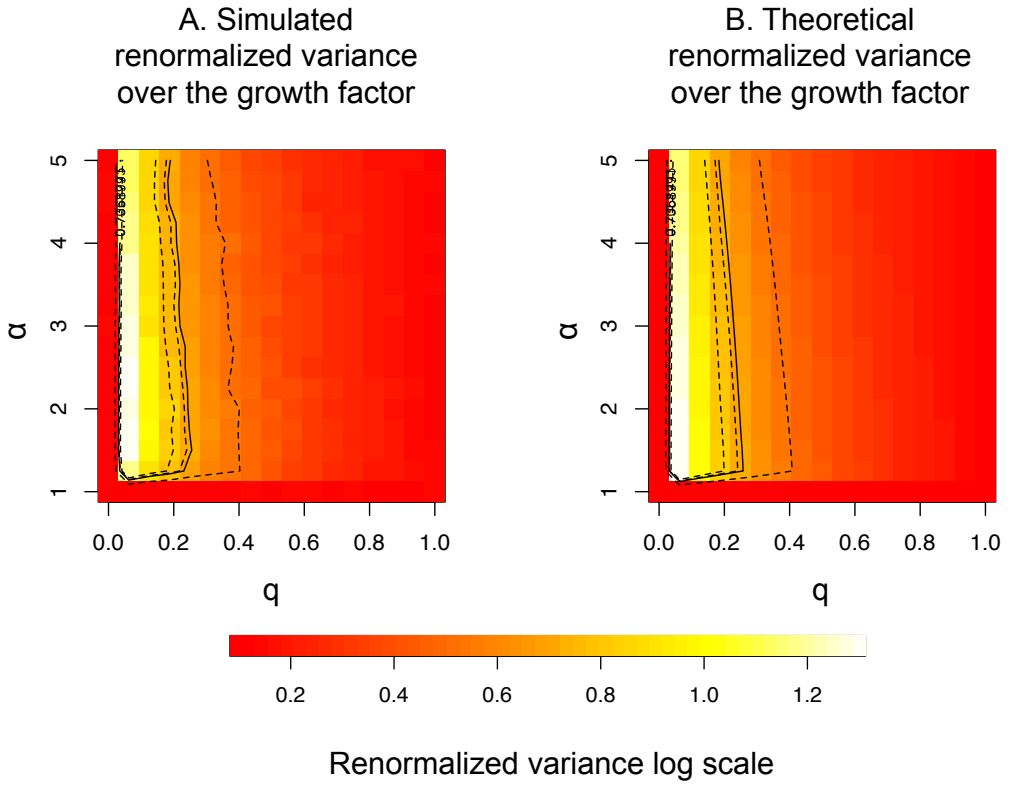


Figure 3.2 – Maps of the expected renormalized variance over the growth factor (color scale), in the framework of the two-subpopulations model, using the data from the experiments with "SB300" strain and 10^3 inoculum size to constraint the values of β and c. "x" axis is q, the initial proportion of subpopulation 1 that replicates faster. "y" axis is α, the authorized ratio between the two replication rates. The contour curves show the points of the parameter space where the experimental value of the variance is met (dashed lines for the three mice and solid for the mean). The fact that these lines appear on the map indicates that some combinations of the parameters (q, α) allow to recover all the experimental values for the three observables at once. The variance takes the same value for $q = 0$, $q = 1$ or $\alpha = 1$ because we recover the one-population case.

Note that there is a sharp transition of the variance when q is small (there is no discontinuity). Infinitesimal values of q in equation 3.2 give rise to very unlikely values for $r_{max,1}$ (essentially, if q is small enough, then $r_{max,2}$ equals r_{max} and $r_{max,1} = \alpha r_{max}$). From the biological point of view, it is very unlikely that a subpopulation might replicate more than 50% faster than the maximum replication rate observed in the experiments. We could check by zooming on these very low q values that this part of the contour corresponds to areas where the values of $r_{max,1}$ exceed by far this limit. Thus this part of the contour is an artifact of our model coming from the fact that we put no bound in the values of the replications rates, and should be ignored.

Now, we do not expect the two-subpopulations model to apply well in the 6 types of experiments we are considering (two different strains and three different inoculum sizes). First of all, the experiment with strain "M2702" and inoculum size 10^5 had a defect in the WITS dilution, and thus we lack the information of a proportion of lost WITS, which prevents us from exploiting it. Then in the experiment with strain "SB300" and inoculum size 10^5, the experimental variance is lower than the variance allowed if the WITS loss is fixed (because the contour curve for the variance is situated above the one for the loss in the parametric map, see figure 2.9). Since the two-subpopulations model generates higher variances than the one-population model, it will not help bridging the gap between the estimates in this particular experiment. Finally, regarding the two experiments starting with the inoculum size 10^7 with strain "SB300" or "M2702", the model does not achieve to find positive values for c and a value of β comprised between 0 and 1 using the WITS loss and the growth factor. Looking further into these data, it appears that in these experiments, the value of the experimental growth factor measured by the ratio between the final and initial total numbers of bacteria B_f/N_0 is close in order of magnitude to the maximum possible growth rate allowed by the data of the plasmids dilution $e^{r_{mean}t} = 2^G$. If we take arbitrarily a c value of one third of the lower bound estimated from the mechanical loss in the feces, we even have $2^G e^{-ct} < \frac{B_f}{N_0}$, which means that no value of β comprised between 0 and 1 will allow to recover the experimental value for the growth rate.

Then, in the two last experiments (starting with inoculum size 10^3), positive values of c and β comprised between 0 and 1 are found, and the parametric map of the variance displays the contour curve for the value of the experimental variance (see figure 3.3). It means that some combinations of the parameters (q, α) allow to recover all the experimental values for the three observables at once. In both cases, a value of α around 1.2 with values of q smaller than 0.5 seem to allow an optimal recovery of the experimental variance. Higher values of α also give some possibilities, but then very high values of α would probably not be very realistic from the biological point of view. For that matter, in the experiment with "SB300", too high values of α (around 6, 7) combined to low values of q lead to estimate of β over 1, indicating that the model is not relevant anymore in this region of the parameter space. In these experiments, the β estimated are around 0.5 in the "SB300" case and 0.05 in the "M2702" case (the

values of the parameters do not vary much over the part of the parameter space explored). If both estimates are plausible, since we estimated a lower bound of 0.12 on average, we would however rather expect similar estimates from those two experiments, because the β selection has a priori few reasons to depend on the type of strain inoculated. Then the c values are around $0.01day^{-1}$ in the "M2702" strain and $3day^{-1}$ in the "SB300" case. If the later recovers the order of magnitude we expected (we estimated a loss of around $c = 10day^{-1}$ just from mechanical loss in the feces), the first is surprisingly low. There might be some biological reasons that would explain a lower loss than the one expected from the mechanical loss, for example bacteria might stick to the wall and not go along with the feces. But then again, nothing would explain why so different estimates would be retrieved from two experiments supposedly very similar (and moreover, virulent bacteria from strain "SB300" would be more likely to interact more with the epithelium). Thus only one of the six experiments under consideration here could be explained by the two-subpopulations model. However, more data would be required to conclude more generally on the subject, as will be further discussed in the following conclusion of this part.

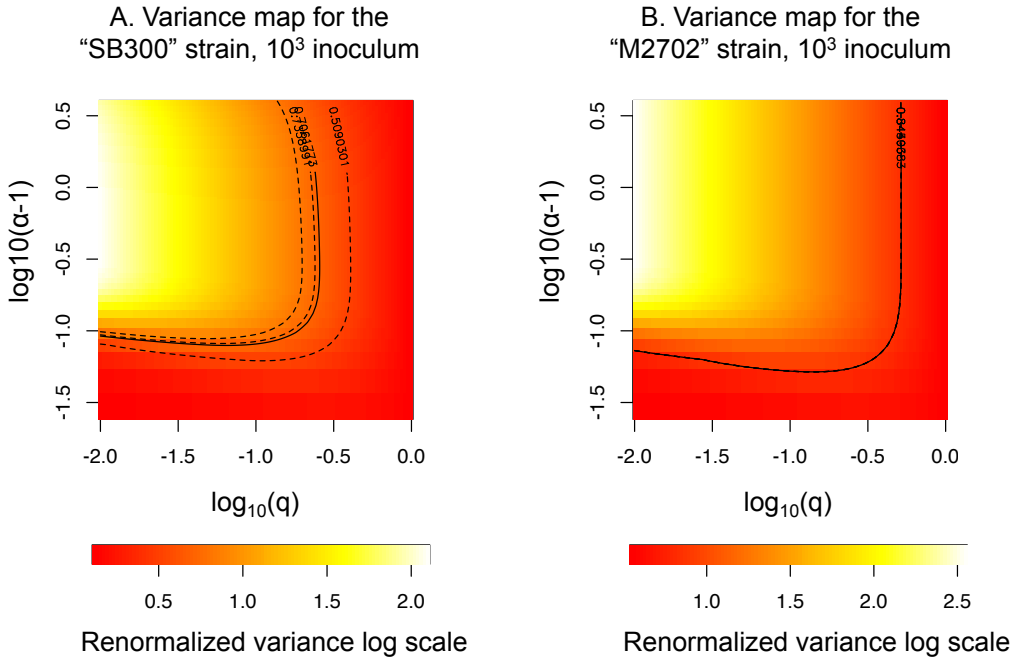


Figure 3.3 – Maps of the expected renormalized variance over the growth factor (color scale), in the framework of the two-subpopulations model, using the data from the experiments with 10^3 inoculum size and "SB300" strain (A) or "M2702" strain (B). "x" axis is $\log_{10}(q)$, with q the initial proportion of subpopulation 1 that replicates faster. "y" axis is $\log_{10}(\alpha - 1)$, with α the authorized ratio between the two replication rates. The contour curves show the points of the parameter space where the experimental value of the variance is met (dashed lines for the three mice and solid for the mean). The fact that these lines appear on the map indicates that some combinations of the parameters (q, α) allow to recover all the experimental values for the three observables at once. Note that the two color scales are slightly different (the variance is globally higher on the "M2702" map, which is coherent with the fact that the estimated β is smaller, see detailed values in the text).

Conclusion

In this part, I presented my work concerning the colonization dynamics of bacterial populations in early infection of the gut. I developed stochastic models of population dynamics in an open system, which aim to infer biologically relevant parameters of the infection (such as replication and elimination rates, and the probability for one bacteria to settle in the organism and participate in the infection) from indirect data. In a first chapter, I presented the quantitative data on *Salmonella* colitis in mice that motivated the study (essentially initial and final numbers of bacteria, as well as initial and final distributions of genetic tags), along with the general methods used subsequently (mostly branching processes and agent-based Gillespie simulations). In a second chapter, I studied one-population models initialized with a Poissonian draw and following a continuous-time birth-death Markovian process. In this framework, I looked for the optimal observable to characterize the variability in the distribution of genetic tags, which have the particularity of starting from unequal population sizes, and showed that the renormalized variance on the growth factor was a suitable measure of variability. I checked for consistency between the parameter estimates based on the observables of the mean growth rate, the renormalized variance over the growth rate and the proportion of genetic tags lost, and showed that in some cases, it is not clear whether these estimates are truly coherent. Based on biological arguments and the qualitative idea that it could lead to broader possibilities of observables combination, and in particular, to a higher variance, I then developed in a third chapter models with two subpopulations following the same kind of dynamics, but with different replication rates. I showed that this kind of model explains very well some experiments, but not all of them, and due to the small quantity of data no clear conclusion can be drawn as to the coexistence of several subpopulations.

In these models, we have always considered a fixed replication rate for the sake of simplicity. However, fixed replication times are closer to the reality of bacterial replication. In appendix D I study a simple model with one population, identical to the null model defined at the beginning of section 1.2, but with all the bacteria that divide after a fixed replication time τ instead of at a fixed rate r. I scale the new parameters so that the population size is identical to the constant replication rate case, and I derive the new expressions for the renormalized variance over the growth factor and the loss probability, which are both modified: in the limit of large time and small elimination rate compared to the growth, the variance is reduced by a factor two, and the loss probability is also reduced in a more complex fashion. However, the effects on parameters estimation on the data set are hard to predict, and depend on the experiment: in some cases, the new observables expressions pull the estimates further apart, while in some others, it gets them closer together. Thus, no clear conclusion can be drawn on the impact of the

fixed replication rate hypothesis on our study.

Another track of investigation concerns a step of the experiments that has been identified as a source of additional variability. Before the q-PCR measurements of the relative abundances of WITS, there is a previous state of amplification: the cecum content is diluted and nutrients are added. The aim is to allow the detection of WITS that were initially present in very small quantities, that would otherwise not be sequenced properly. The problem is that the measure of variability is different whether the q-PCR is actually done right away (just after dilution) or after this enrichment phase. This could be at the origin of a part of the inconsistency observed between the parameters estimates in the one-population model at section 2.5.2. One hypothesis could be that a part of the population is "dormant" when it is extracted from the cecum, and that it experiences a delay before starting to replicate again. However, at the end of the experiment, the WITS populations before enrichment are already of important sizes. Therefore the effects of stochasticity are negligible, and the amplification step should not have any effect on the distribution of WITS. The only case for which we could observe an increase of variability over the amplification process is the case where almost all the bacteria are dormant, and only a tiny proportion of bacteria (rapid simulation allowed me to estimate a proportion of $\simeq 10^{-5}$) replicate with a very high rate to compensate for the other bacteria not replicating. But this situation is extremely unlikely from the biological point of view. If there was a bias in the q-PCR measurements, it should be systematic, and thus should not be a source of additional noise. The effect of amplification would therefore deserve to be further investigated.

Then, a question that comes naturally is the question of a higher number of subpopulations. Intuitively, in terms of variability of the WITS distribution, the limit case must be the one with two subpopulations, with one that does not replicate at all, and the other replicating with a high rate to compensate and keep the final number of bacteria unchanged. This should be the most efficient way to obtain the most unequal distributions: if a WITS population (starting with a small initial number) is all taken from the pool of non-replicating bacteria, while another WITS population is entirely taken from the pool of fast-replicating bacteria, that will allow the higher difference possible in final population sizes. Some very preliminary simulations with three subpopulations indeed showed that no higher variance could be reached compared to the case with two subpopulations, but the question would deserve to be resolved more formally.

To conclude on the data set on which we tested our methods, one of the biggest difficulty is that it seems that the observable that conveys the more information, *i.e.* the proportion of WITS lost in the experiment, is not measured with a sampling important enough so that it can be completely reliable. A good way to overcome this issue without requiring a higher number of experiments (and thus, mice), would be to develop a higher number of genetic tags (in the data we

studied, there are only seven different labels), so that the loss probability and the distribution measured become more statistically relevant. However, even if there is no clear conclusion regarding the data set under examination here, there is no doubt that the kind of models developed are sufficiently general to be applicable in other situations of population dynamics.

Part II

Mechanisms of the IgA immune response in the gut

Introduction

The digestive system has a large surface area[30][31], covered by a single layer of epithelial cells, essential for nutrient absorption, but also a gateway for many pathogens. Contrary to the inside of the body, where the presence of any bacteria is abnormal, the lumen of the digestive system is home to a very important microbiota : there are at least as many bacteria in the human digestive system as there are human cells composing the body [1]. The microbiota is thus important in numbers, but also in function: bacteria are necessary to break down and absorb certain nutrients, and can compete against potentially pathogenic intruders[2]. Inside the organism, the immune system can and must be able to eliminate generically any bacteria. On the other hand, in the digestive system, the host has to find alternative ways to fight dangerous bacteria while sparing beneficial ones.

The adaptive immune response is the only strong handle that the host has on directly controlling microbiota composition at the species level[56][57]. The main effector of the adaptive immune response in the digestive system is secretory IgA, an antibody. sIgA specifically bind to targets (*i.e.* antigens) that the organism has already encountered via infection or vaccination. It was observed more than 40 years ago that this prevents infection[32]. Many studies have focused on the complex molecular and cellular pathways that trigger an immune response on the host side of the digestive surface[33]. However, we are only just beginning to understand by which physical mechanisms the immune effectors act once secreted into the intestinal lumen. For example, the influence on bacteria dynamics of abiotic factors such as the flow in the gut has only recently started being quantitatively studied [42, 58]. The aim of this part of my thesis is to enlighten some aspects of these physical mechanisms.

In the following, I will first expose a new concept in immunology named *enchained growth*, the process through which dividing bacteria remain clumped in clonal aggregates, preventing them to approach the gut epithelium and colonize the rest of the organism. I will review the different elements produced to evidence this phenomenon in the study from Moor *et al.* [19], and in particular my contribution with a model based on genetically labeled bacteria data. In a second chapter, we will explore the consequences of this phenomenon at the host level, and see how this phenomenon could be a way for the immune system to discriminate dangerous bacteria from the others by targeting preferentially fast replicating bacteria. Finally, in a last chapter, we will explore the evolutionary consequences of enchained growth at the population level, namely in the context of antibiotic resistance spread, through the study of cross scale models of infection propagation.

Chapter 4

A new idea in immunology: *enchained growth*

Contents

In this chapter, I will review the different elements produced to evidence the phenomenon of *enchained growth* in the study from Moor *et al.* [19]. I will first explain how the production in high quantities of Immunoglobulin A can be triggered by vaccination. Then I will show that random encounters of bacteria are rare in the gut at the beginning of infection, so that another mechanism than classical agglomeration must account for the protection of the vaccinated mice. I then present my contribution with a model predicting the loss of variability resulting from enchained growth and conclude.

4.1 Vaccination triggers sIgA production

Our collaborators from the Wetter-Slack group have developed a vaccination protocol based on the oral inoculation of large quantities of bacteria killed by peracetic acid [34]. In this process, the organism is put in contact with large quantities of specific antigens marking the surface of bacteria, without risking an infection. It triggers the production in large quantities of sIgA in the gut lumen, and when mice are later inoculated with active S. Thyphimurium, it is observed that they do not become sick and there is no inflammation. However, the mechanisms through which this sIgA production protects the mice are not immediately obvious. Indeed, sIgA are not able to kill bacteria. In fact, the dynamics of the infection in the gut remain quite unchanged in vaccinated mice compared to naive ones the first day. In particular, the intestinal bacterial load is very similar in both cases (see fig. 4.1). Less bacteria however are crossing the epithelium barrier and migrating to the lymph nodes.

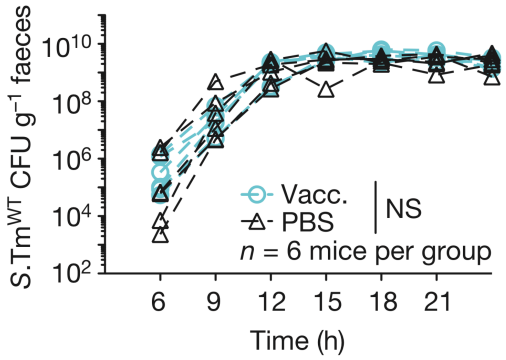


Figure 4.1 – Six mice per group were either vaccinated with peracetic-inactivated bacteria or mock-vaccinated with a buffer solution before being pre-treated with antibiotic and later inoculated with 10^5 CFU of wild type *Salmonella* Typhimurium. The bacterial load per gram of feces is shown during the 24h following inoculation. There is no significant difference between vaccinated and naive mice (from fig. 1h of [19]). Note that samples were systematically agitated with bead-beating before plating and CFU count (see section 1.1.1). This allows to break the potential IgA-mediated clumps (see the following) and permits a correct count in any case.

Most bacteria are actually observed to be agglomerated together in the gut lumen, far from the epithelium. It has been known for a long time that high-avidity sIgA are able to bind specific target-bacteria together. This is due to their structure with several binding sites (see fig. 4.2). However, agglutination was classically thought as a process driven by the diffusion of bacteria in the gut: there are random encounters of bacteria of the same strain. If the concentration of sIgA targeting this bacterial strain is high enough, then those bacteria will be coated by sIgA and will remain attached to each other upon encounter. But at the realistic initial concentrations of bacteria following food poisoning, how frequent are those encounters?

4.2 Limit of the classical agglomeration idea

Order of magnitude of the encounter time between two bacteria According to [59], 10^5 is a realistic estimate of the number of bacteria ingested by a human victim of typical food poisoning. The corresponding number should be much smaller for mice, but we will keep this number as a higher limit. The typical time to find one target of radius a in a sphere of radius b by diffusion is of the order of $b^3/(Da)$, with D the diffusion coefficient, so the typical time for one bacteria to find another when there are N bacteria in a volume V is of the order of $V/(NDa)$. For bacteria, a is in the micrometer range. Bacteria typically swim at $10\mu m/s$, and change direction every second, which gives a diffusion coefficient of the order of $10^{-10}m^2/s$ (the peristaltic motions of the digesta are large scale

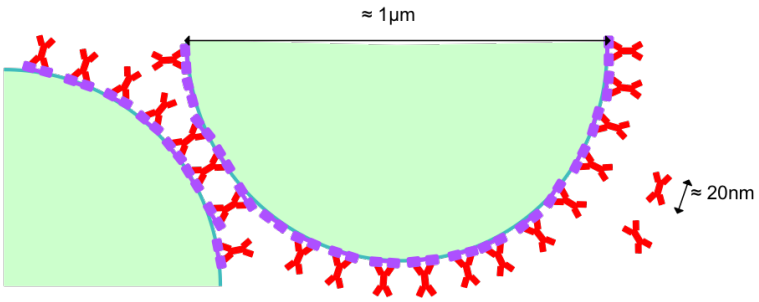


Figure 4.2 – Schematic diagram of IgA-mediated bounds between bacteria (scale is not respected). It has been observed (fluorescent imaging [19]) that IgA (here represented in red) are present in high enough concentrations, so that they can coat the surface of bacteria by attaching the specific antigen (here in purple). The structure of IgA allows the cross-binding of two bacteria together.

movement rather than local diffusion, so we assume they have a smaller effect on diffusion). The mouse's cecum has a volume of the order of $1cm^3$. The smallest inoculum in the experiments is $N_0 = 10^5$ bacteria. With these numbers, the typical encounter time is of the order of $10^5 s$, *i.e.* $30h$, about ten times longer than the typical digestion time in mice. Mice are nevertheless protected in these experiments, thus, classical agglomeration cannot be the mechanism through which mice are protected after vaccination, because it is inefficient at the initial low bacterial densities.

Other hypothesis Other hypothesis were tested. It was checked that sIgA does not kill bacteria nor prevents them from growing. Another hypothesis is that IgA could act on flagella. However, experiments with a bacterial strain missing flagella show that they still make naive mice sick, while vaccinated mice are protected.

Enchained growth hypothesis Bacteria may remain stuck together upon division. Vaccinated mice were infected with a 1:1 mixture of GFP-tagged (green fluorescence) and mCherry-tagged (red fluorescence). In mice inoculated with small quantities of bacteria (10^5), only monochromatic clumps of bacteria were observed (see fig. 4.3). When larger quantities of bacteria are inoculated (from 10^8 on), then the initial densities are high enough so that classical diffusion-driven encounters play an important part, and randomly-colored clusters are observed. For intermediate inoculum sizes, at first the dominating effect is enchained growth, and when high enough densities are reached, monochromatic clusters start to regroup randomly, forming oligoclonal structures.

Modeling showed that enchained growth can account for the observed ratio of clumped bacteria versus planktonic ones. In the experiments with vaccinated mice, it has been observed that only planktonic bacteria approach the epithelium and can interact with it; agglomerated bacteria, on the contrary, remain far from

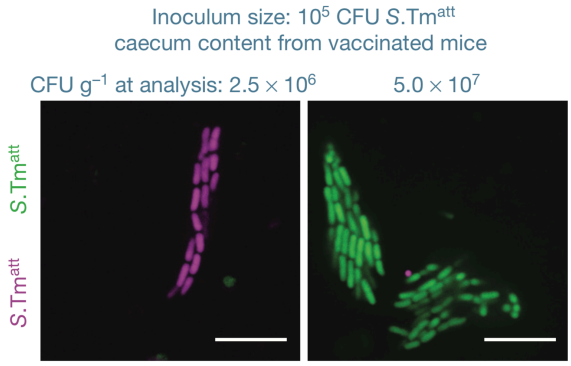


Figure 4.3 – Vaccinated mice were infected with 10^5 CFU inoculum of attenuated strains (attenuated mutant used to avoid inflammation in mock-vaccinated controls). Live microscopy shows that most bacteria are trapped in monochromatic clusters, particularly when bacterial densities remain below 10^8 (fig. 1i from [19])

the wall, deep in the lumen. Modeling has also shown that the observed ratio of clumped bacteria versus planktonic ones can account for the reduced level of lymph node colonization[19].

4.3 Modeling clonal loss with enchained growth

To confirm the existence of enchained growth, I developed a model based on the data of the genetically labeled bacteria called WITS (see section 1.1.3). Indeed, if bacteria develop and are eliminated in clonal clusters, then everything happens as if the effective population size were reduced. A higher clonal loss should thus ensue, as schematically explained in figure 4.4. Our model aims at predicting the amplitude of this increased clonal loss.

To estimate the effect of enchained growth on clonal loss, we simulated a simple scenario on the basis of the following experiment : vaccinated and naive mice were infected with 10^5 CFU of the attenuated strain, spiked with an average of 10 copies of each of the seven WITS strains, as in the experiments. WITS frequencies were determined by plating, enrichment culture and qPCR as described in section 1.1.3. The evenness of the final distribution was calculated as described in section 2.3.1. Only tagged bacteria need to be tracked in the simulation[1]. It starts from the numbers of the seven barcoded bacteria strains in the inoculum. These bacteria establish in the cecum following a Poisson process with a probability β (as in the *null model* in section 1.2). We reasonably assume that the established bacteria become uniformly spatially distributed in the cecal content by peristaltic mixing and that the untagged bacteria have no effect on the bar-

[1]contrary to what was done in the previous part, here we will not include a saturating term, which would have required to keep track of the whole population size, but will rather consider two successive phases for the growth, which will be described in the following

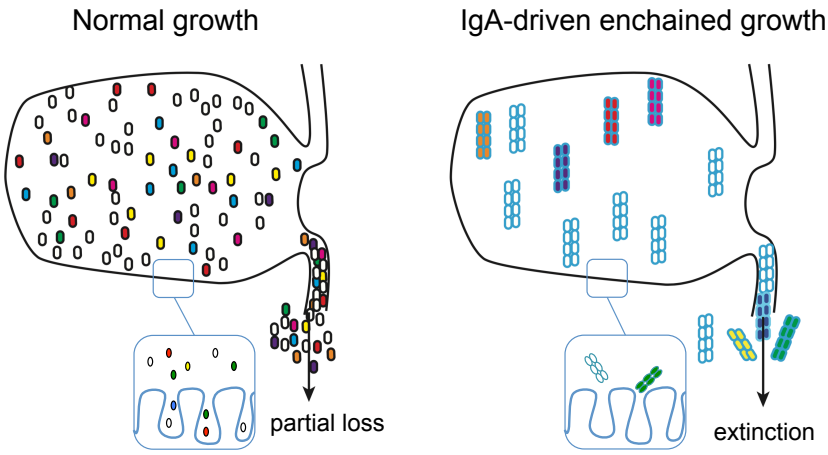


Figure 4.4 – Bacterial growth in naive versus vaccinated mice. Colours represent the different strains of WITS. Blue lines represent sIgA coating. From figure 2a in [19]. The zoom-in boxes show how clustered bacteria are prevented from interacting with the epithelium, one of the hypothesis to explain this being steric: clusters might be too big to penetrate the crypts forming the surface. Another hypothesis is that the mucus covering the epithelium prevents the bacterial clusters, too large compared to the mucus mesh size, to approach the wall.

coded clonal distribution. The growth is taken as deterministic, with replication every τ during the first phase of fast replication, until the carrying capacity is reached and growth only compensates the loss so that the population size remains stable. Bacterial loss is taken to be random : at each time step, a proportion p_c of the population is lost on average. I simulated the two extreme cases, which differ only in the way bacteria are lost:

- The "normal growth" case, with no enchained growth and where all bacteria remain independent (mimicking what should happen in naive mice). In this case, we simply track the numbers of bacteria, and at each time step, each bacteria has the probability p_c to be eliminated.

- The "perfect enchained growth" case, where individual clones never segregate, meaning that bacteria get eliminated in the feces in perfect clonal clusters. In this case, we track the numbers of clusters, and at each time step, each cluster has the probability p_c to get eliminated.

In these simulations, there are thus three important parameters for the dynamics of the bacterial population that need to be inferred from experimental data: β, τ and p_c.

1. The probability to seed the cecum β is estimated using WITS loss in the mock-vaccinated mice data. Indeed, in this case clonal loss can be mainly attributed to this initial Poisson bottleneck, as subsequent bacterial loss is a

small effect compared to the division rate (because of the first phase of fast replication). Thus as if in section 2.2.2, we consider that the experimental proportion of WITS lost equals $\exp(-\beta\langle n_0\rangle)$, with a cutoff on the final WITS numbers (the frequencies below 10^{-2} are considered to actually be zero, based on a separation of the frequencies). With this approximation we find $\beta = 0.115$.

2. The growth kinetics in the cecal lumen is based on the plasmid dilution data from figure 2.1B (see also extended data figure 1 from [19]). When streptomycin pretreated mice are infected with 10^5 CFU of *S.* Typhimurium, bacteria divide twice per hour until 12h after infection, when they reach the cecal carrying capacity of approximately 10^{10} and net growth stops (that is, the growth rate equals the clearance rate).

3. The kinetics of clearance is only thought here as mechanic: to maintain feces production, approximately 10% of the cecal content is cleared to the colon and lost in the feces every 15 min.

The results of these simulations are shown on figure 4.5. Note that in the normal growth case (black line), the evenness stabilizes after 1h. The final value estimated by the simulation is close to the experimental value for unvaccinated mice[2]. On the other hand, evenness continues to drop in the perfect enchained growth case (cyan line), indicating on-going clonal extinction. The experimental evenness values of the tags distribution at 18h post-infection in the vaccinated mice (cyan points) are on average located between the simulated final values of evenness with no enchained growth and with perfect enchained growth. This is qualitatively consistent with the idea that enchained growth *in vivo* might be imperfect : IgA may be not perfectly sticky, and clusters may break at some point (the idea will be elaborated in the following chapter).

4.4 Conclusion

Thus, we have seen in this chapter that enchained growth is the key element which allows a vaccinated mice to be protected from infection. Even at low initial concentrations, bacteria remain bound together by IgA-mediated links upon replication, resulting in clonal clusters prevented from interacting with the epithelium to trigger inflammation and invade the lymph nodes : this explains immune exclusion. Moreover, clumped bacteria are likely to be more easily evacuated in the feces [19], and the absence of inflammation preserves the rest of the intestinal flora and thus favors elimination through competition with other pathogens. This mechanism is thought to be quite generic: the immune system of a mice is quite similar to the ones of other animals such as humans, and the sIgA-binding is valid for a large panel of bacteria (for instance, *E. coli* has also been tested and found to behave likewise in [19]). It therefore constitutes an

[2]In this dataset with non-vaccinated mice, WITS loss (used to calculate β) and WITS variability (here the evenness) are compatible.

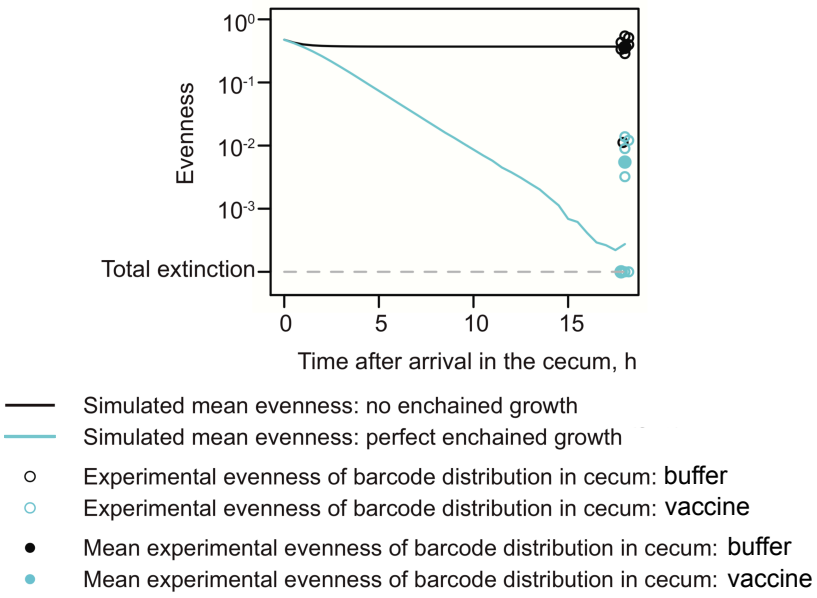


Figure 4.5 – Simulation of the two extreme cases ("normal growth" and "perfect enchained growth") and comparison with the experimental data described at the beginning of this section. The simulated mean evenness is an average over 300,000 realizations of the same process. Note that the final simulated evenness with normal growth matches the experimental value in naive mice, as expected. The experimental values in naive mice are situated somewhere between the two extreme cases, indicating that *in vivo*, enchained growth may not be perfect. From extended data figure 8 in [19].

important mechanism of protection of the intestinal ecosystem. The dynamics of these clusters may be important for their functions, this will be the object of chapter 5. Moreover, the structure of the bacterial population decreases diversity, which may have evolutionary consequences. This will be the object of chapter 6.

Chapter 5

Enchained growth as a way to regulate microbiota homeostasis

Contents

We have seen in the previous chapter that in immune animals, daughter bacteria remain enchained by sIgA upon replication. In this chapter (which, in a modified form, is submitted [20]) we argue that this enchained growth process can be a way for the immune system to regulate the microbiota composition, through the interplay between clusters growth and fragmentation. In a first section, I expose this qualitative idea in more details. I a second section, I present different plausible models of bacteria clusters dynamics, and the methods to study them. Then I give, for each model, the resulting dynamics and cluster size distribution, before putting these results in perspective with experimental data. Eventually, I discuss the results.

5.1 Interplay between clusters growth and fragmentation

As closely related bacteria (for example, *Salmonella* spp. and commensal *E. coli*) can show highly variable behaviors in the intestine, the task that falls to the immune system in the gut of identifying which bacteria are noxious among the beneficial ones is challenging. Besides, the overgrowth of any type of bacteria, even those that do not cause acute pathology, can actually impair the functionality of the microbiota. Thus the host needs mechanisms to maintain gut microbiota composition homeostasis.

We have seen in the previous chapter that mice vaccinated with inactivated *Salmonella* Typhimurium do produce specific sIgA which bind to *S.* Typhimurium, but this neither kills them nor prevents them from reproducing[60][19]. These mice are nevertheless protected against pathogen spread from the gut lumen to systemic sites like lymph nodes, liver or spleen. We have contributed to show that the main effect is that upon replication, daughter bacteria remain attached to one another by sIgA, driving the formation of clusters derived from a single infecting bacterium[19]. This "enchained growth" process is effective at any bacterial density. Clustering has physical consequences: the produced clusters do not come physically close to the epithelial cells, and as interaction with the epithelial cells is essential for *S.* Typhimurium virulence, this is sufficient to explain the observed protective effect.

Now if sIgA were perfectly sticky, we would expect all bacteria to be in clusters of ever increasing size. In these experiments, despite observing *S.* Typhimurium clusters in the presence of sIgA, there are still free planktonic bacteria, and clusters of small sizes. One possible explanation would be that the concentrations of sIgA are insufficient, and that not all bacteria are coated with it. But in these experiments, it has been demonstrated (with IgA fluorescent staining and flow-cytometry analysis) that they are (see extended figure 2c of[19]). This was expected. Indeed, a gram of digestive content contains at most 10^{11} bacteria, and typically 50 micrograms or more of sIgA[61], of molecular mass of about $385kD$. This leads to about 800 sIgA per bacteria. sIgA may not be all bound to bacteria, and sIgA for different specific antigens may be produced in proportions not matching the proportions of antigens present in the digestive system, so that not all bacteria are coated with 800 sIgA. Nevertheless, most bacteria already encountered by the organism will be coated with many sIgA, and thus the clusters sizes are not limited by the number of available sIgA. The other possible explanation is that the sIgA-mediated links break. Before plating bacteria to count them, they are subjected to bead beating, and this has indeed been observed to break the clusters (see extended data figure 3 of [19]). In related systems, the breaking of such links has been demonstrated to be dependent on the applied forces [62][63]. As there is shear in the digestive system, because mixing is needed for efficient nutrients absorption, it is plausible that links break over time.

Moreover, it has been observed that the small clusters are linear chains of bacteria, bound by sIgA. As bacteria are similar to each other, it is, at another

scale, analogous to other physical systems[64], such as polymers breaking under flow[65]. The main difference is that these chains grow by bacterial replication. Growth and fragmentation are competing effects, and the modelling of these clusters can be addressed as a statistical physics problem, where one wants to predict their size distribution, whether there is a typical cluster size, or if large clusters of ever-increasing size dominate the distribution, and how the growth in the number of free bacteria depends on the bacterial replication rate.

This could also have very important biological consequences. To illustrate this point, let us consider a simplified model (see figure 5.1): bacteria remain enchained by sIgA when they grow (replication time τ_{div}), and the sIgA-link between 2 bacteria breaks exactly after a time τ_{break} (although this latter hypothesis is not realistic, we make it for now for the sake of simplicity in getting the general idea). If $\tau_{div} > \tau_{break}$, then when a bacterium divides, it forms a 2-bacteria cluster, which dislocates into two free bacteria before the next replication step, so that bacteria remain in the state of free or 2-bacteria clusters and there are no larger clusters. If $\tau_{div} < \tau_{break}$, when a bacterium divides, it forms a 2-bacteria cluster, which becomes a 4-bacteria cluster before the first link breaks, so that there cannot be free bacteria anymore. In this model, the fast-growing bacteria are selectively targeted by the action of the immune system. The immune system does not need to sense which bacteria are growing faster, it only has to produce sIgA targeted to all the bacteria it has ever encountered, and bacteria with $\tau_{div} > \tau_{break}$ are unaffected, whereas bacteria with $\tau_{div} < \tau_{break}$ are trapped in clusters. That could be a simple physical mechanism to target the action of the immune system to the fast-growing bacteria which are destabilizing the microbiota, and thus to preserve microbiota homeostasis.

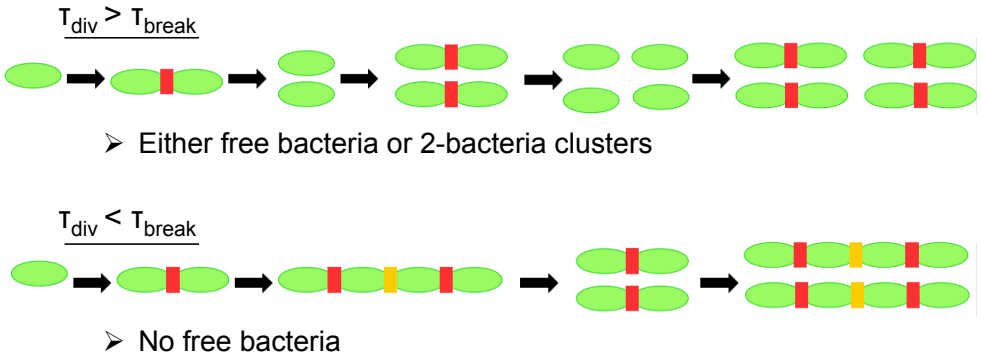


Figure 5.1 – Simplified model of clusters dislocation, where all bacteria divide at τ_{div} and all link break after τ_{break}. Larger clusters form only if bacteria have time to divide before the links break, that is if $\tau_{div} < \tau_{break}$.

In the following, we present different plausible models of bacteria clusters dynamics, and the methods to study them. Then we give, for each model, the resulting dynamics and cluster size distribution, before putting these results in perspective with experimental data. Eventually, we discuss the results.

5.2 Models and methods

We consider low bacterial densities, so encounters between unrelated bacteria and thus classical agglomeration are negligible. Thus, we consider each free bacteria and each cluster of bacteria independently of the others. *Salmonella* are rod-shaped bacteria, which divide at the middle of the longitudinal axis. Thus if the daughter bacteria remain enchained, they are linked to each other by their poles. With further bacterial replications, the cluster will then be a linear chain. This is consistent with experimental observations, in which clusters are either linear chains, with bacteria attached to one or two neighbors by their poles, or larger clusters which seem to be formed as bundles of such linear clusters (see pannel A figure 5.2). Our aim is to model the dynamics of these clusters.

5.2.1 Elements of the various models

A first element is bacterial replication (see figure 5.2 C). One way to model it is to assume that bacteria replicate every τ_{div}. Another way, that we will generally use, less realistic but easier for calculations, is to assume that there is a fixed replication rate r.

A second element is that when bacteria replicate, they may be able to escape enchainment (see figure 5.2 B), but likely with low probability (see discussion in the next section 5.2.3). In most cases, we will take the limit with perfect enchainment upon replication.

A crucial element is the possibility for the links between bacteria to break. We usually assume that the breaking rate α is the same for all links and over time. But we will also explore the case where the links breaking rate is force-dependent, in which case not all the links have the same breaking rate.

Another crucial element, is to model what happens when the chain breaks (see figure 5.2 D). If the subparts come in contact again at the same poles and get linked again, then this could simply be modeled by an effectively lower breaking rate. More likely, if the subparts come in contact again, they do so laterally, forming larger clusters of more complex shapes. Because in these clusters, most bacteria have more than two neighbors, and more contact surface, they are much less likely to escape. To simplify, we will consider that these clusters do not contribute anymore to releasing either free bacteria or linear chains. Thus when a link breaks, either the two subparts move sufficiently away and become two independent chains (probability q); or they collide and become a more complex cluster which does not contribute anymore to either free bacteria or linear chains dynamics (probability $1-q$). For simplicity, we consider that when an outermost link breaks, the single bacteria, more motile, always escapes ($q_{outermost} = 1$), but that else q is size independent. We will take $q = 0$ for the base model.

As digestive content leaves the digestive system, or the part of the digestive system under consideration, due to flow, we define c the loss rate of free bacteria, and c' the loss rate of clusters. We assume no death (which could break clusters). As free bacteria have more autonomous motility, enabling them to swim towards the epithelial cells, it is likely that $c' \geq c$. We will usually take $c = c'$. Crucially,

in this latter case, free bacteria and all clusters are lost at the same rate. The c value has a complex effect on stochastic quantities, such as the probability to have at least one cluster of a given size. However, here we study only the mean numbers of free bacteria and clusters of different sizes, so that the case with $c = c'$ is equivalent to $c = c' = 0$, with all numbers of bacteria and clusters multiplied by $\exp(-ct)$.

We start with the most basic model, with a replication rate r, bacteria perfectly bound upon replication, a fixed breaking rate per link α, and bacterial chains always binding into a more complex cluster when a link breaks (except for the outermost links) ($q = 0$). We then study variations of this base model to test the robustness of the results: with an non-zero escape probability upon replication and $c \neq c'$; with a replication time τ instead of a replication rate r; with the possibility for chains to escape when an inner link breaks ($q > 0$); with a force-dependent breaking rate.

5.2.2 Methods

We consider the beginning of the process, early enough so that the carrying capacity is far from reached, and thus the replication rate is constant. We do not consider generation of escape mutants which are not bound by IgA. We consider only the average numbers of free bacteria and linear clusters of different sizes, and we do not count more complex clusters, as they do not contribute to free bacteria dynamics in our model.

For each model, we write the equations for the derivative of these numbers with respect to time. With N the vector of the mean number of free bacteria and linear clusters of higher sizes (the i^{th} component being the mean number of clusters of size i) these equations give the coefficients of the matrix M, such that $dN/dt = MN$. The results are obtained in part via analytical derivations and in part via numerical studies. The latter are obtained in Mathematica by numerically solving the eigensystem written for clusters up to an arbitrary size n_{max}, chosen large enough not to impact the results. In the large time limit, $N(t) \to e^{\lambda t}X$, with λ the largest eigenvalue of M and X the corresponding eigenvector. For each model, we study how the growth of the free bacteria population size – the ones which are capable of causing systemic infection[19] – *i.e.* λ in the steady state, depends on the bacterial replication rate. Besides, we obtain distributions of the cluster sizes, which could be compared to experimentally observed distributions.

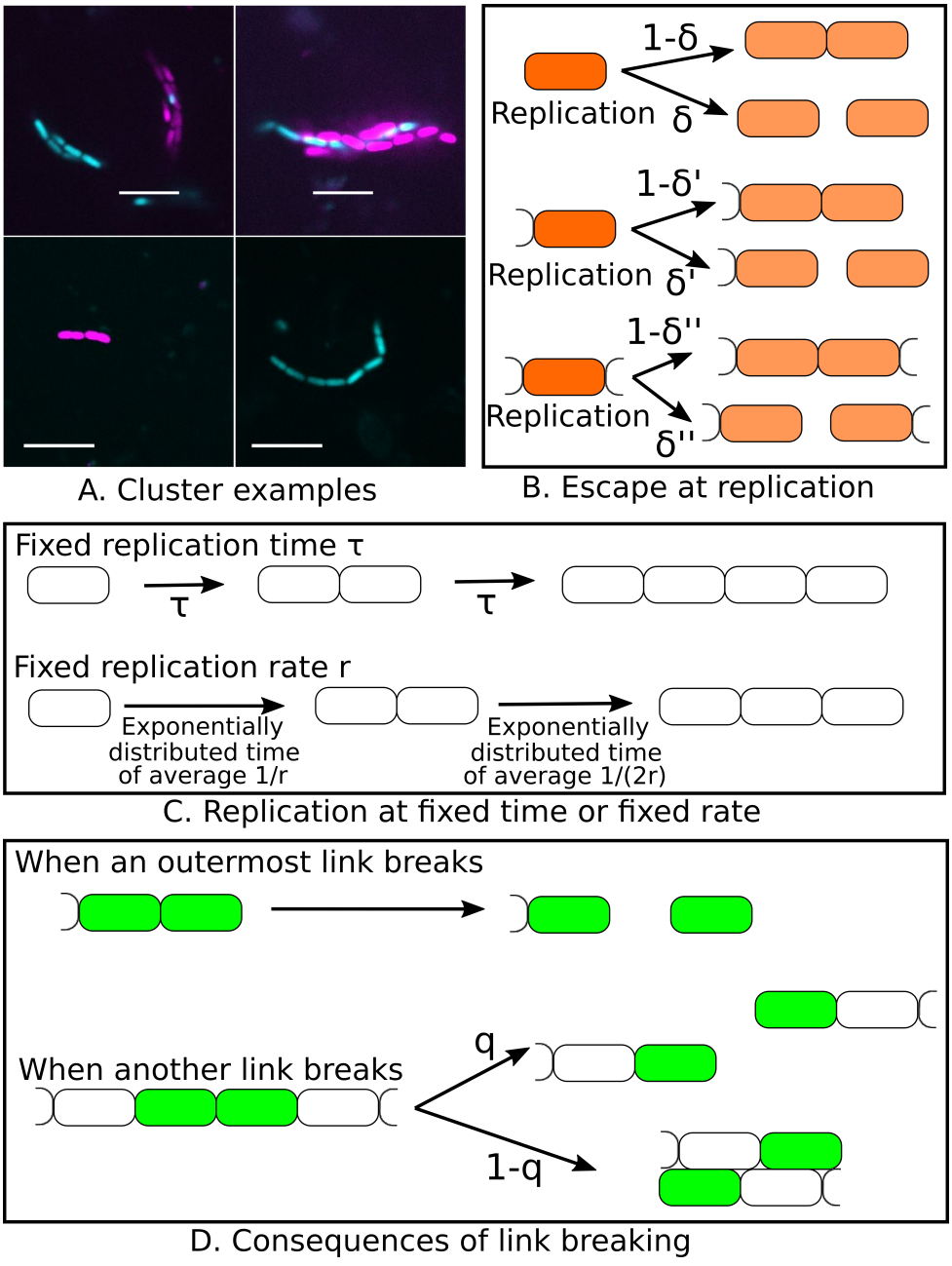


Figure 5.2 – Bacterial clusters modeling. A. Examples of experimental images of bacterial clusters in cecal content of vaccinated mouse at 5h post-infection with isogenic GFP and mCherry expressing *S.* typhimurium. The scale bar is $10\mu m$. Top Figures: complex clusters made from bundles of linear clusters, which could be re-attached single chains (left) or formed from at least two independent clones (indicated by fluorescence, right). Bottom Figures: linear clusters which dynamics we aim to model. B. Potential bacterial escape upon replication (in the base model, $\delta = \delta' = \delta'' = 0$). C. Fixed replication time or fixed replication rate (the latter is chosen for the base model). D. Consequences of link breaking. In the base model, $q = 0$.

5.2.3 Argument for a low escape probability

When a bacterium replicates, the time for septation is of the order of a few minutes. We intuitively think that this time is much larger than the time required for bacteria to stick when they randomly meet. The aim of this section is to give an overestimate of the typical time τ_k it takes for a bacterium to stick to another when they meet.

We use the data on figure 1k of [19] about non-dividing bacteria (so the only sticking is from random encounters). The majority of them are aggregated after a few hours for a concentration of $10^7 - 10^8$ bacteria. As we will calculate an overestimate of τ_k, we take the highest concentration and the maximum time, i.e. $N = 10^7$ bacteria in $V = 1cm^3$ (cecum volume) and $\tau_{exp} = 8$ hours.

If the diffusion coefficient is high enough, the time for bacteria to cluster will be limited by the rate k at which bacteria stick to each other when they are in close vicinity. k is the inverse of τ_k. If the diffusion coefficient is smaller, then the time to first encounter will also play a role, but as we calculate an overestimate of τ_k, we can neglect this.

Note that this is a large overestimate. Indeed, when bacteria get clumped to each other, their effective concentration decreases, thus it takes longer for the last bacteria to meet others, and thus the time for most bacteria to be clumped will be significantly larger than the inverse of the early clumping rate.

The bacteria typical size is of a few micrometers; let us take $3\mu m$ as an overestimate of the maximum bacterial size. To be in close contact, two bacteria must be at at most $a = 3\mu m$ away. Let us assume that then the volume of possible contact is $4/3\pi a^3$, which is an overestimate, because only certain orientations will allow bacteria to touch each other. Then, the proportion of time spent in close contact will be of the order of $(N4\pi a^3)/(3V)$, and the typical time to stick to each other will be $\tau_{exp} = \tau_k 3V/(N4\pi a^3)$. Then $\tau_k = \tau_{exp} N4\pi a^3/(3V)$. Numerically, we obtain about 5 minutes as an overestimate of τ_k.

With all these highly conservative estimates, we find τ_k at the very most of the same order of magnitude as the septation time, and very likely much smaller. Hence the probability for bacteria to escape enchainment is small, which justifies that we take in general the limit of no escape.

5.2.4 Table of the symbols used

Base model	
r	Bacterial replication rate
α	Breaking rate of the link between two bacteria
$n_i(t)$	Number of linear clusters of length i at t (n_1: free bacteria)
λ	Largest eigenvalue of the matrix, which is the growth rate of the free bacteria in the steady state
Model with bacterial escape and bacterial loss (all these parameters are taken as 0 in the base model)	
δ	When a free bacteria replicates, the probability that this will lead to 2 free bacteria
δ'	When a bacterium at the tip of a cluster replicates, the probability that the daughter cell at the exterior side escapes
δ''	When a bacterium replicates within a cluster, the probability that the daughter bacteria will not be bond to each other, resulting to the cluster breaking in two
c	Loss rate for the free bacteria
c'	Loss rate for the clusters
Model with fixed replication time	
τ	Time between one bacterial division and the next (the bacterial growth rate is $r_{eff} = \log(2)/\tau$)
$\mathcal{N}$	Largest eigenvalue of the matrix in this model. $\mathcal{N} = \exp(\lambda\tau)$
Model with linear chains independent after breaking	
q	Probability that when an inner link of a cluster breaks, the two subparts become independent linear clusters. In the base model, $q = 0$.
Model with force-dependent breaking rates	
β	A constant expressing the strength of the coupling between hydrodynamic forces and link breaking. In the base model, $\beta = 0$.

5.3 Clusters dynamics and distributions of sizes

5.3.1 Base model

5.3.1.1 Equations

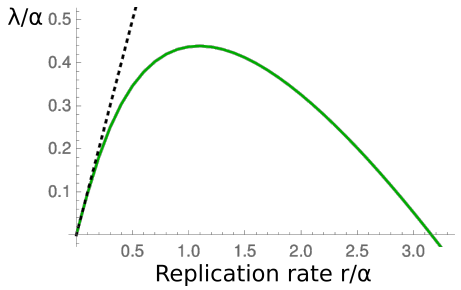
In the base model, bacteria have a replication rate r, daughters are perfectly bound upon replication, each link has a breaking rate α, and when a link which is not at a tip breaks, the resulting two chains of bacteria always bind into more complex clusters and thus do not contribute to free bacteria dynamics anymore ($q = 0$). With $n_i(t)$ the number of linear clusters of size i as a function of time, (n_1 is the number of free bacteria),

$$\frac{dn_1}{dt} = -rn_1 + \sum_{i=2}^{\infty} 2\alpha n_i$$

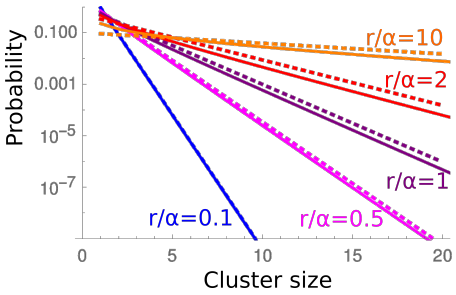
and for $i \geq 2$,

$$\frac{dn_i}{dt} = rn_{i-1}(i-1) - irn_i - (i-1)n_i\alpha + 2\alpha n_{i+1} \tag{5.1}$$

5.3.1.2 Free bacteria growth rate as a function of the bacterial replication rate



A. Growth rate λ as a function of the replication rate r, both in units of α. Numerical result (green solid line), and limit with no clusters ($\lambda = r$, black dotted line).



B. Distribution of the cluster sizes, for different r. Solid lines: numerical results, dotted lines: approximation (5.3) (almost overlaid with the numerical results for $r/\alpha = 0.1$).

Figure 5.3 – Base model. For the numerical calculations, $n_{max} = 40$.

Even for this simple version, the system of equations is hard to solve in the general case. We start by studying numerically the growth rate in the long term (the maximum eigenvalue λ of the matrix of coefficients $m_{i,j} = r(i-1)\delta_{i-1,j} - ir\delta_{i,j} - (i-1)\alpha\delta_{i,j} + 2\alpha(\delta_{i+1,j} + \delta_{i,1}(1-\delta_{j,1}-\delta_{j,2}))$), as a function of the replication rate (see figure 5.3 A). The growth rate has a maximum for a finite replication rate, of the order of α (the link breaking rate): the higher the replication rate, the higher the potential for growth in the number of free bacteria, but when the replication rate becomes too large compared to the breaking rate, the bacteria get trapped in clusters, which break and re-attach in more complex clusters from which independent bacteria cannot escape.

5.3.1.3 Chain length distribution

In the long time limit, the number of clusters of size i is of the order of $b_i \exp(\lambda t)$, with λ the largest eigenvalue. Equation (5.1) simplifies to:

$$\lambda b_i = -irb_i + rb_{i-1}(i-1) - (i-1)b_i\alpha + 2\alpha b_{i+1}$$

Assuming that i is large,

$$b_i \simeq \frac{r}{r+\alpha} b_{i-1} \tag{5.2}$$

is required. Using this approximation for all i, the probability that a randomly chosen chain is of size k is:

$$p_k = \left(1 - \frac{r}{r+\alpha}\right)\left(\frac{r}{r+\alpha}\right)^{k-1} \tag{5.3}$$

(note that to renormalize this probability, the sum of the geometric progression is taken from 1 to infinity). This approximation works relatively well, especially for smaller r/α values (see figure 5.3 B). Part of the discrepancy is that equation (5.2) is an approximation for large i, and thus does not hold at small clusters sizes.

5.3.2 Model with bacterial escape and differential loss

5.3.2.1 Equations

This is similar to the base model presented before, except that we take into account that upon replication, bacteria may not be perfectly bound, and may escape (pannel B of figure 5.2). We note δ the probability for the two daughter bacteria to remain free after the replication of a free bacteria. We note δ' the probability that when a bacterium at the tip of a cluster replicates, the daughter cell on the outside of the cluster escapes the enchainment. We note δ'' the probability that when a bacterium at the interior of the cluster divides, the daughter cells will not be enchained, effectively clipping the cluster in two. As free bacteria are more motile than clusters, then $\delta \geq \delta' \geq \delta''$. We also add here the possibility that the loss rate c for free bacteria and c' for clusters are different. Then the base equations are:

$$\frac{dn_1}{dt} = r(-1+2\delta)n_1 + \sum_{i=2}^{\infty} 2r\delta' n_i + \sum_{i=2}^{\infty} 2\alpha n_i - cn_1$$

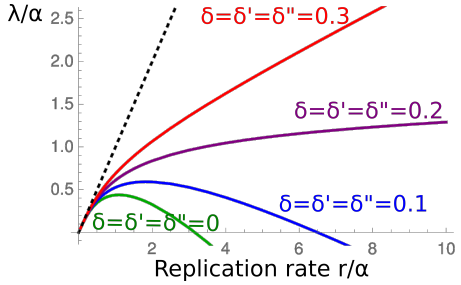
$$\frac{dn_2}{dt} = r(1-\delta)n_1 - 2r(1-\delta')n_2 - \alpha n_2 + 2n_3\alpha - c'n_2$$

and for $i \geq 3$,

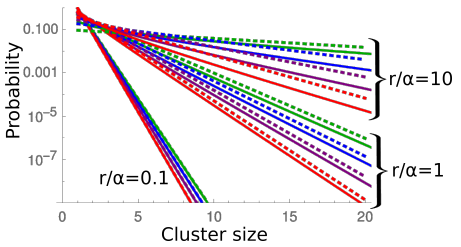
$$\frac{dn_i}{dt} = r(2\delta'-i)n_i + rn_{i-1}(i-1-2\delta'+3\delta''-i\delta'') - (i-1)n_i\alpha + 2\alpha n_{i+1} - c'n_i. \quad (5.4)$$

5.3.2.2 Free bacteria growth rate as a function of the bacterial replication rate

Similarly to the base model, we study numerically the growth rate as a function of the replication rate (see figure 5.4 A). The larger the replication rate, the more the deviation between the growth rate and the replication rate, which would be its value in the absence of clusters. If δ, δ', δ'' are small enough, the qualitative behavior is similar to the base model. But for larger δ, δ' and δ'', the growth rate continues to increase monotonously with the replication rate. The same is true when δ, δ' and δ'' are different (see figure 5.5). If $c = c'$, the growth rate is simply offset by minus the loss rate (see figure 5.5), and if $c \neq c'$, the effect is more complex, but for small r/α values it corresponds to an offset of $-c$.

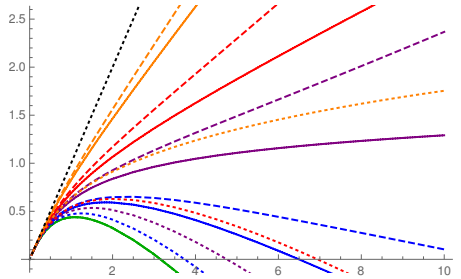


A. Growth rate λ as a function of the replication rate r, both in units of α. Numerical results (colors), with $\delta = \delta' = \delta''$. Black dotted line: limit with no clusters ($\lambda = r$).

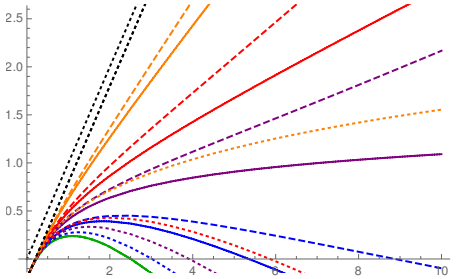


B. Cluster size distribution, for different r. The solid lines are the numerical results (for $\delta = \delta' = \delta''$.), the dotted lines are the approximation (5.5).

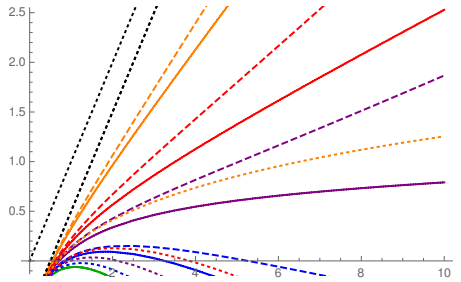
Figure 5.4 – Model with bacterial escape ($\delta > 0$). $\delta = \delta' = \delta'' =$ 0, 0.1, 0.2, 0.3. $c = c' = 0$. For the numerical calculations, $n_{max} = 40$.



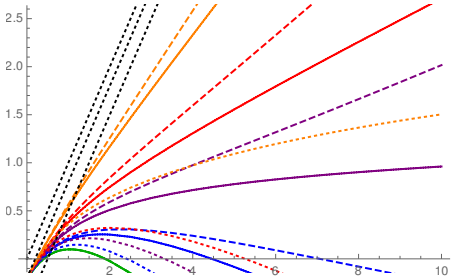
$c = c' = 0$



$c = c' = 0.2\alpha$



$c = c' = 0.5\alpha$



$c = 0.2\alpha$, $c' = 0.5\alpha$

Figure 5.5 – Growth rate λ as a function of the replication rate r, both in units of α. Numerical results (colors), with $\delta = \delta' = \delta''$ (solid lines), $\delta = \delta'$, and $\delta'' = 0$ (dashed lines), $\delta' = \delta'' = 0$ (dotted lines). $\delta =$ 0, 0.1, 0.2, 0.3, 0.5. The black dotted lines are either r/α, $(r - c)/\alpha$ or $(r - c')/\alpha$. As expected, if $c = c'$, the resulting growth rate are the same than when $c = c' = 0$, minus c. If $c \neq c'$, the results are closer for small r/α to the results if both c and c' had the c value. For the numerical results, $n_{max} = 40$.

5.3.2.3 Chain length distribution

We can reason similarly to the base model to obtain the approximation for the cluster size distribution. In the long time limit, the number of clusters of size i is

of the order of $b_i \exp(\lambda t)$, with λ the largest eigenvalue. Equation 5.4 simplifies to:

$$\lambda b_i = r(2\delta' - i)b_i + rb_{i-1}(i - 1 - 2\delta' + 3\delta'' - i\delta'') - (i-1)b_i\alpha + 2\alpha b_{i+1} - c'b_i$$

Assuming that i is large,

$$b_i \simeq (1-\delta'')\frac{r}{r+\alpha}b_{i-1}$$

is required. Using this approximation for all i, the probability that a randomly chosen chain is of size k is:

$$p_k = \left(1 - (1-\delta'')\frac{r}{r+\alpha}\right)\left((1-\delta'')\frac{r}{r+\alpha}\right)^{k-1} \tag{5.5}$$

This approximation works relatively well (figures 5.4 B and 5.6). The approximation (5.5) depends on δ'', but neither on δ nor δ', but δ and δ' could actually matter when i is small, and indeed we observe (see figures 5.7 and 5.8) that the approximation (5.5) works slightly less well when δ'' is different from δ or δ'. If $c = c'$, the distribution does not change, and if $c \neq c'$, the distribution changes very little (see figure 5.9).

Free bacteria are released at a rate $2r\delta' + 2\alpha$ per cluster. This rate is independent of the cluster size. The direct contributions to the increase of free bacteria from clusters of size i compared to all the larger clusters will be (with $K = (1-\delta'')r/(r+\alpha)$):

$$\begin{aligned}\frac{\textit{contribution larger}}{\textit{contribution } i} &= \frac{(2r\delta' + 2\alpha)\sum_{j=i+1}^{\infty}(1-K)K^j}{(2r\delta' + 2\alpha)(1-K)K^i} \\ &= \sum_{j=1}^{\infty} K^j = \frac{K}{1-K} = \frac{(1-\delta'')r}{\alpha + r\delta''}\end{aligned}$$

If r is small compared to α (replication rate $\ll$ breaking rate), then this ratio is small. Thus the larger clusters are quickly negligible. Indeed, in this regime, clusters typically dislocate before new replications, so there are few larger clusters.

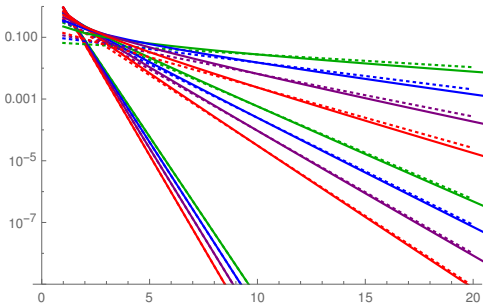
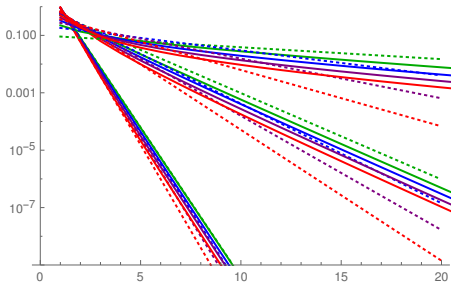
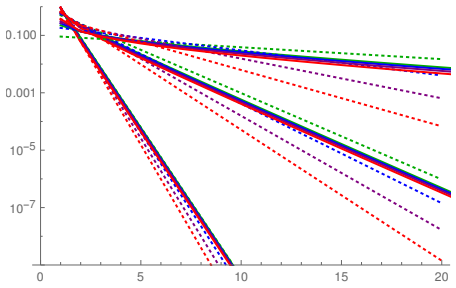


Figure 5.6 – Distribution of the cluster sizes. All as in figure 5.4 B, except that the approximation 5.5 is rescaled by the numerical value at $n = 10$. This shows that the approximation captures well the distribution of large clusters.

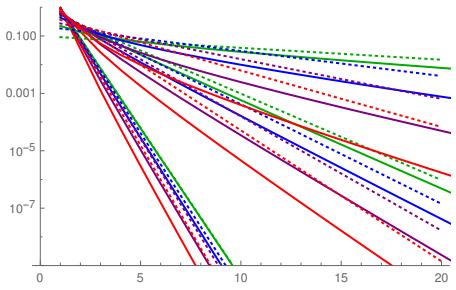


$\delta = \delta'$, $\delta'' = 0$

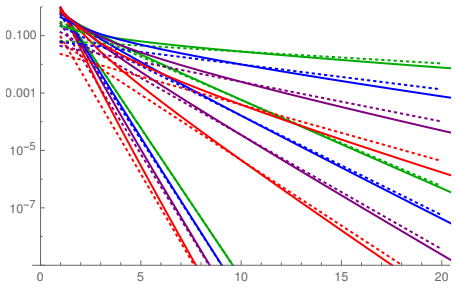


δ, $\delta' = \delta'' = 0$

Figure 5.7 – Distribution of the cluster sizes. All as in figure 5.4 B, except the values of δ' and δ''. The distribution is close to the result for $\delta = \delta' = \delta'' = 0$, which is in line with approximation 5.5 which is independent of δ and δ'.



Dotted lines : approximation 5.5



Dotted lines: approximation 5.5, rescaled by the numerical value at $n = 10$.

Figure 5.8 – Distribution of the cluster sizes, for $\delta = \delta' = 2\delta''$. Other parameters as in figure 5.4 B. The approximation does not work as well as when $\delta = \delta' = \delta''$.

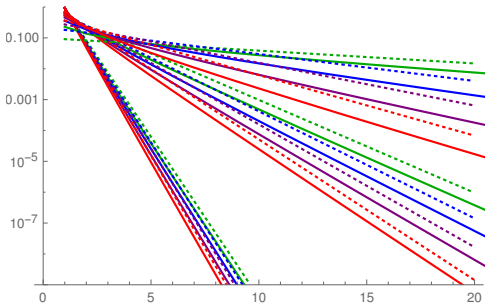


Figure 5.9 – All as in figure 5.4 B, except $c = 0.2\alpha$, $c' = 0.5\alpha$. There is very little change in the cluster size distribution.

5.3.3 Model with fixed replication time

In this variant of the base model, bacteria divide every τ, and there is no bacterial escape nor elimination. The effective growth rate is r_{eff} such that $\exp(r_{eff}t) = 2^{t/\tau}$, thus $r_{eff} = \log(2)/\tau$.

5.3.3.1 Equations

Let us start by considering a chain of n bacteria at $t = 0$, right after a replication event. During the time interval between two replication events, with $l(n, i, t)$ the probability that at t, the chain has lost i bacteria in total on the extremities, and consequently is of size $n - i$ at t (since we assume $q = 0$ as in the base model, if the chain breaks somewhere else, the subparts form a more complex cluster and thus are "lost" for the system):

$$\frac{dl(n,i,t)}{dt} = -\alpha(n-1-i)l(n,i,t) + 2\alpha l(n,i-1,t).$$

At $t = 0$, $l(n, 0, 0) = 1$ and for $0 < i < n - 1$, $l(n, i, 0) = 0$. The solution for any $0 \leq i \leq n - 2$ is:

$$l(n,i,t) = \frac{2^i}{i!}\exp(-\alpha t(n-1-i))(1-\exp(-\alpha t))^i \tag{5.6}$$

For any chain of size > 2, there are two outermost links, each breaking at rate α, liberating one free bacteria; and a chain of size 2 breaks at rate α, but liberates two free bacteria. Consequently, the average number of free bacteria generated during τ by this chain of n bacteria is:

$$l(n, free, \tau) = \int_0^\tau \sum_{i=0}^{n-2} l(n,i,t)2\alpha dt = 2\alpha \sum_{i=0}^{n-2} \int_0^\tau \frac{2^i}{i!}\exp(-\alpha t(n-1))(\exp(\alpha t)-1)^i dt.$$

A chain of length n right before replication becomes a chain of length $2n$ upon it, and will have contributed to chains of size k by $l(2n, 2n-k, \tau)$ and to the free bacteria by $l(2n, free, \tau)$ right before the next replication event. Writing $u(i, t)$ the number of chains of size i at time t (right before a replication event), we can write the matricial relation between the $u(i, t)$ and $u(i, t + \tau)$ (right before the next replication event) as follows:

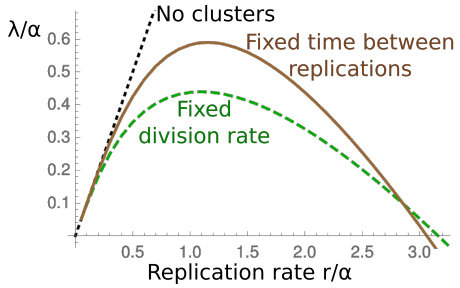
$$\begin{pmatrix} u(1,t+\tau) \\ u(2,t+\tau) \\ u(3,t+\tau) \\ u(4,t+\tau) \\ \vdots \end{pmatrix} = \begin{pmatrix} l(2,free,\tau) & l(4,free,\tau) & l(6,free,\tau) & \dots & \dots \\ l(2,0,\tau) & l(4,2,\tau) & l(6,4,\tau) & \dots & \dots \\ 0 & l(4,1,\tau) & l(6,3,\tau) & \dots & \dots \\ 0 & l(4,0,\tau) & l(6,2,\tau) & \dots & \dots \\ 0 & 0 & l(6,1,\tau) & \dots & \dots \\ 0 & 0 & \vdots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} u(1,t) \\ u(2,t) \\ u(3,t) \\ \vdots \\ \vdots \end{pmatrix}$$

The elements of this matrix can otherwise be written as:

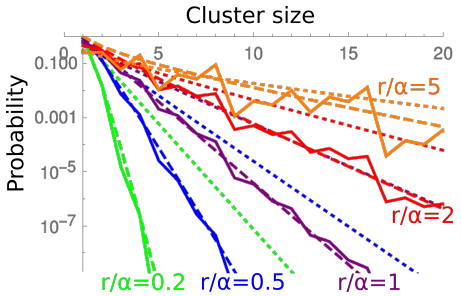
$$M_{ij} = \begin{cases} l(2j, free) & \text{for} \quad i = 1 \\ l(2j, 2j-i) & \text{for} \quad i > 1 \text{ and } 2j \leq i \\ 0 & \text{otherwise} \end{cases}$$

This matrix is then cut to size $n_{max} \times n_{max}$, and we numerically solve the eigensystem.

5.3.3.2 Free bacteria growth rate as a function of the bacterial replication rate



A. Growth rate, as a function of $r_{eff} = \log(2)/\tau$ (replication every τ, solid brown line) or as a function of r (base model with fixed replication rate r, green dashed line), all in units of α. The dotted black line is the expectation in the absence ofclusters.



B. Cluster size distribution, numerically solved (solid lines), compared with the analytical approximation (5.7) (dashed lines), and the numerical result in the base model (dotted lines). $r/\alpha = 0.2$, 0.5, 1, 2, 5.

Figure 5.10 – Fixed time between replications. For the numerical calculations, $n_{max} = 32$

The shape of the relation between the free bacteria growth rate and the (effective) replication rate is very similar in the fixed replication time versus fixed replication rate model (see figure 5.10 A, with a maximum of the growth rate for a finite value of the (effective) replication rate, at close values ($r_{eff} = 1.15\alpha$ versus $r = 1.09\alpha$ in the fixed replication rate model). When the replication is at fixed time intervals instead of a fixed replication rate, the maximum growth rate is higher, and it dips faster at increasing effective replication rate. Indeed, in the case of fixed replication rate, the distribution of durations between two replications is exponential, thus more spread. Close to the maximum, the presence of short replication intervals makes that there can be more cluster formation, and conversely, at higher replication rates, the presence of longer replication intervals results in more production of free bacteria.

5.3.3.3 Chain length distribution

We show here the assumptions and calculations to obtain an analytical approximation for the chain size distribution. We define $u(N,t)$ the number of chains of size N at time t. Assuming N even,

$$u(N, t+\tau) = \sum_{i=0}^{\infty} u\left(\frac{N}{2} + i, t\right) l(N + 2i, 2i, \tau).$$

In the long time, $u(N,t) = f(N)\exp(\lambda t)$, with λ the long term growth rate, that is such that $\exp(\lambda\tau) = \mathcal{N}$, with $\mathcal{N}$ the largest eigenvalue of the matrix. Then replacing $l(N+2i, 2i, \tau)$ by its expression 5.6 in the previous equation leads to:

$$\mathcal{N}f(N) = \sum_{i=0}^{\infty} f(\frac{N}{2}+i)\exp\left(-\alpha\tau\left(N-1+2i\right)\right)\left(\exp(\alpha\tau)-1\right)^{2i}\frac{2^{2i}}{(2i)!}.$$

We compare the 1st term of the sum to the rest of the sum. The first term is $f(\frac{N}{2})\exp\left(-\alpha\tau\left(N-1\right)\right)$, the rest of the sum is:

$$\sum_{i=1}^{\infty} f(\frac{N}{2}+i)\exp\left(-\alpha\tau\left(N-1+2i\right)\right)\left(\exp(\alpha\tau)-1\right)^{2i}\frac{2^{2i}}{(2i)!}.$$

We divide both by $\exp\left(-\alpha\tau\left(N-1\right)\right)$. Then this is equivalent of comparing $f(N/2)$ with:

$$S = \sum_{i=1}^{\infty} f(\frac{N}{2}+i)\exp\left(-2i\alpha\tau\right)\left(\exp(\alpha\tau)-1\right)^{2i}\frac{2^{2i}}{(2i)!}.$$

When $\alpha\tau$ is large, links typically break before the next replication, so there is little cluster formation, so it is expected that the chain length distribution decreases fast with N, so that for $i>0$, $f(\frac{N}{2}+i) \ll f(N/2)$. When $\alpha\tau$ is small, replication is slow compared to the typical time for one link to break. However, for a chain of length $N/2$, τ has to be compared to $(N/2-1)/\alpha$, the typical first link breaking time, thus for N large enough, we expect the number of large cluster to decrease, thus $f(\frac{N}{2}+i) \lesssim f(N/2)$ for $i>0$. We define B such as $f(\frac{N}{2}+i) \leq B, \forall i>0$. For $\alpha\tau$ large, $B \ll f(N/2)$, and for $\alpha\tau$ small, if N is large enough, $B \lesssim f(N/2)$. Then:

$$S \leq \sum_{i=1}^{\infty} B(1-\exp(-\alpha\tau))^{2i}\frac{2^{2i}}{(2i)!} = B\left(cosh\left(2(1-\exp(-\alpha\tau))\right)-1\right)$$

For $\alpha\tau$ large, $\left(cosh\left(2(1-\exp(-\alpha\tau))\right)-1\right) \simeq cosh(2)-1 \simeq 2.7$. For $\alpha\tau$ small, $\left(cosh\left(2(1-\exp(-\alpha\tau))\right)-1\right) \simeq 2(\alpha\tau)^2 \ll 1$. Thus in the case of $\alpha\tau$ large, S is small relative to $f(N/2)$ because S is smaller than a few units times B, with B much smaller than $f(N/2)$. In the case of $\alpha\tau$ small, S is small relative to $f(N/2)$ because S is of the order of $(\alpha\tau)^2 B$, with B of the order of $f(N/2)$. Then this justifies the assumption that only the first term of the sum matters:

$$f(N) \simeq \frac{1}{\mathcal{N}} f\left(\frac{N}{2}\right)\exp\left(-\alpha\tau\left(N-1\right)\right)$$

We assume $N = 2^k$, with k an integer. This is obviously true only for a very restricted set of N, but as we are interested on how the distribution depends on N for large N, looking at these specific points is good enough. Then, by recursion,

$$f(N) \simeq \frac{1}{\mathcal{N}^k} f(1)\exp\left(-\alpha\tau\left(N(1+1/2+1/2^2+...+1/2^k)-k\right)\right).$$

If N is large enough, $1 + 1/2 + 1/2^2 + ... + 1/2^k \simeq 2$. Remembering that k was defined as $N = 2^k$, the result is:

$$f(N) \simeq f(1) N^{\frac{\alpha\tau - \log(\mathcal{N})}{\log(2)}} \exp\left(-2\alpha\tau N\right).$$

When $\alpha\tau \gg 1$, links typically break before the next replication, thus there is little impact of the clustering on the growth, and thus the growth will be close to its value in the absence of clustering, i.e. doubling every τ, thus in this limit $\mathcal{N} = 2$:

$$f(N) \simeq f(1) N^{\frac{\alpha\tau}{\log(2)} - 1} \exp\left(-2\alpha\tau N\right). \tag{5.7}$$

This rough approximation allows to explain the core of the observed distribution (figure 5.10). There are bumps, due to the replication every τ (which in the absence of link breaking would results in clusters of size 2^k only), which makes that clusters of power-of-two length are over-represented. Compared to the case with fixed replication rate, the distribution is much narrower.

5.3.4 Model with linear chains independent after breaking ($q > 0$)

5.3.4.1 Limit case: subchains always independent after breaking

In this model, when a cluster breaks, the two resulting clusters remain independent and can thus continue to participate in the dynamics of the system:

$$\frac{dn_i}{dt}(t) = (i-1)r\ n_{i-1}(t) + (-\alpha(i-1) - ir)\ n_i(t) + 2\alpha \sum_{j=i+1}^{\infty} n_j(t)$$

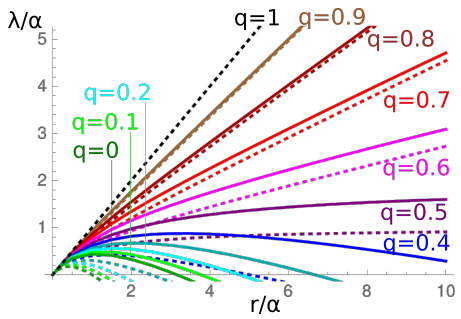
We recognize here the equation studied in [21], where they described chains of growing unicellular algae. As it has been shown, the steady state solution of the system is:

$$n_i(t) = C \exp(rt) \left(\frac{r}{\alpha + r}\right)^i.$$

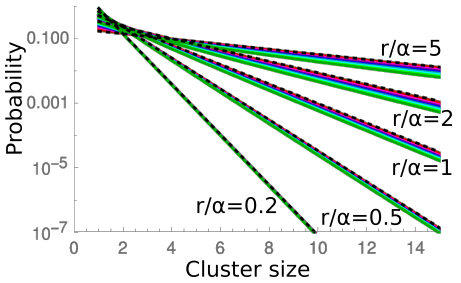
In the steady state, the growth rate is equal to the replication rate. The average cluster size is $1 + \frac{r}{\alpha}$, which shows that, as expected, if the link breaking rate is high compared to the replication rate ($r/\alpha \ll 1$), the average length is close to one as no cluster has the time to form: all the bacteria remain free.

5.3.4.2 Intermediate case: chains independent or trapped after breaking

More realistically, after breaking, chains will have some probability to either encounter each other and remain trapped in more complex clusters, or to escape and become independent. We will assume in the following that if a chain of size N breaks at a link at the extremity, releasing a cluster of size $N - 1$ and a free bacteria, then the free bacteria, smaller and likely more motile, will escape in all cases; but that if the link that breaks is elsewhere, the probability for the new clusters of sizes $N - k$ and k ($k > 1$) to escape and continue as two independent



A. λ/α as a function of r/α. The dotted black line is the case $q = 1$, for which $\lambda = r$, like in the absence of clusters. The colored dotted lines are the analytical approximation (5.10).



B. Distribution of cluster sizes. The dotted black lines are the approximate distribution (5.9) for each r/α, which is the exact distribution for $q = 1$. The colours represent the same q values than for the left panel. All curves are almost overlaid for small r.

Figure 5.11 – Model with linear chains independent after breaking. Numerical results: $n_{max} = 100$

linear clusters will be q, and the probability that they bind and form a more complex cluster will be $1 - q$, with q independent of k. We write the equations for the number $n_i(t)$ of cluster of i bacteria:

$$\frac{dn_1}{dt} = -rn_1 + 2\alpha \sum_{j=2}^{\infty} n_j$$

$$\frac{dn_i}{dt} = -rin_i + r(i-1)n_{i-1} - \alpha(i-1)n_i + 2\alpha n_{i+1} + 2\alpha q \sum_{j=2}^{\infty} n_{i+j}.$$

In the long time, $n_i \to f_i \exp(\lambda t)$ with λ the largest eigenvalue.

$$\lambda f_i = -rif_i + r(i-1)f_{i-1} - \alpha(i-1)f_i + 2\alpha f_{i+1} + \sum_{j=i+2}^{\infty} 2\alpha q f_j \tag{5.8}$$

This is valid for any i. We assume that f_i decreases fast enough with i so that the sum from $i+2$ to ∞ of the f_i is an order of magnitude less than if_i. Then, the largest elements of equation (5.8) when i is large enough are the terms multiplied by i, and consequently:

$$0 \simeq -rf_i + rf_{i-1} - \alpha f_i$$

Leading to:

$$f_i \simeq \frac{r}{\alpha + r} f_{i-1}$$

and then by recursion,

$$f_i \simeq C \left(\frac{r}{\alpha + r} \right)^i.$$

If this is valid for any i, the probability that a cluster taken at random is of size i is:

$$p_i = \frac{\alpha}{\alpha + r} \left(\frac{r}{\alpha + r} \right)^{i-1}. \tag{5.9}$$

We compare this approximation with the numerical results and they are in good agreement (figure 5.11 B), except when both q is small and r/α is large, and even in this case it gives a reasonable approximation.

Replacing f_i by $C \left(\frac{r}{\alpha+r} \right)^i$, equation (5.8) simplifies to:

$$\lambda \simeq -r \left(1 + \frac{\alpha}{r} \right) + \alpha + 2\alpha \left(\frac{r}{\alpha + r} \right) + \sum_{j=2}^{\infty} 2\alpha q \left(\frac{r}{\alpha + r} \right)^j$$

which after simplifications leads to:

$$\lambda \simeq r \frac{\alpha + (2q-1)r}{\alpha + r}. \tag{5.10}$$

This approximation does not work for $q < 0.5$, but it works well for q close to 1, and gives the right dependence for r/α large for $q > 0.5$ (figure 5.11 A). Intuitively, if $q > 0.5$, when a cluster breaks it leads to more than one independent linear cluster, thus the population of linear clusters and thus free bacteria may continue to increase with r/α, whereas if $q < 0.5$, clusters that break lead to less than one independent cluster on average, and thus, as the breaking rate increases with the size, the growth of the population is stunted when r/α increases.

5.3.5 Model with force-dependent breaking rates

5.3.5.1 Equations

What drives link breakage? The links could break if there was some process degrading the sIgA, but the sIgA are thought to be very stable[66]. Another possible explanation for link breaking is that the bound antigen can be extracted from the bacterial membrane, at a rate which may vary exponentially with the force[67][63]. The forces applied on the links are likely mostly due to the hydrodynamic forces exerted by the digesta flow on the bacterial chain. Taking the linear chain as a string of beads, as done for polymer chains, and in a flow with a constant shear rate, the force is predicted to be larger as the chain grows longer, and the largest at the center of the chain[65]. A more detailed discussion and the calculations can be found in section E in the appendix. Taking α as the breaking rate in the absence of shear, and β a constant expressing the strength of the coupling between hydrodynamic forces and link breaking, the resulting equations for this minimal model taking into account the forces are:

$$\frac{dn_1}{dt} = -rn_1 + 2\sum_{i=2}^{\infty} \alpha n_i \exp\left(\beta \frac{i-1}{2} \right)$$

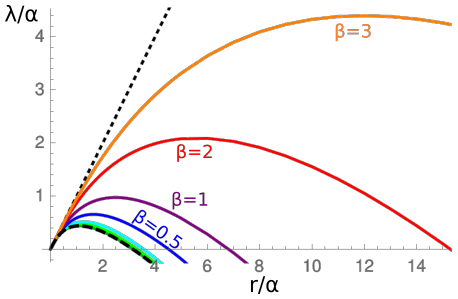
and for i even,

$$\frac{dn_i}{dt} = -rin_i - \alpha n_i e^{\beta i^2/8}\left(1 + 2\sum_{k=2}^{i/2} e^{-(k-1)^2\beta/2}\right) + r(i-1)n_{i-1} + 2\alpha n_{i+1}e^{\beta i/2} \tag{5.11}$$

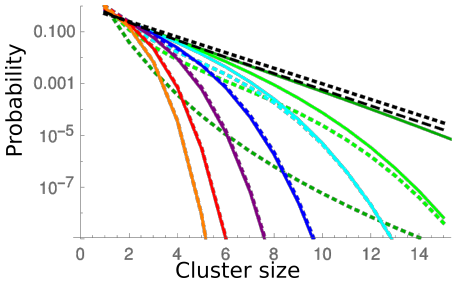
and for $i > 1$ odd,

$$\frac{dn_i}{dt} = -rin_i - 2\alpha n_i e^{\beta i^2/8}\sum_{k=1}^{(i-1)/2} e^{-(k-1/2)^2\beta/2} + r(i-1)n_{i-1} + 2\alpha n_{i+1}e^{\beta i/2} \tag{5.12}$$

5.3.5.2 Free bacteria growth rate as a function of the bacterial replication rate



A. λ/α as a function of r/α. The dotted black line is the case with no clusters.



B. Distribution of the cluster sizes. The solid lines are the numerical results, the colored dotted lines the analytical approximation (5.13), and the black dotted line the approximation for the base model (5.3).

Figure 5.12 – Model with force-dependent breaking rates. The solid colored lines represent the numerical results. Each color represents a different β : $\beta = 0.01$ ($n_{max} = 20$), $\beta = 0.1$ ($n_{max} = 15$), $\beta = 0.2$ ($n_{max} = 15$), $\beta = 0.5$ ($n_{max} = 15$), $\beta = 1$ ($n_{max} = 15$), $\beta = 2$ ($n_{max} = 10$), $\beta = 3$ ($n_{max} = 10$). The black dashed lines are the numerical results for the base model, equivalent to $\beta = 0$. The curves for $\beta = 0.01$ (dark green) are almost overlaid with the curves for $\beta = 0$.

The growth rate as a function of the replication rate has a qualitatively similar shape as for the base model (figure 5.12 A), with a finite replication rate maximizing the growth rate. The limit $\beta \to 0$ corresponds well to the base model, as expected. When β increases, the replication rate maximizing the growth rate increases, as the effective breaking rate is higher. Numerically, we find (see figure 5.13) that the replication rate maximizing the growth rate scales as $\alpha \exp(0.8\beta)$.

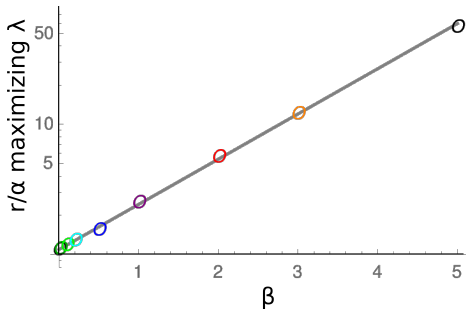


Figure 5.13 – Log of the value of r/α maximizing the growth rate in the force-dependent breaking rate model as a function of β. The points are numerical maximums, the line is $1.09 \times \exp(0.8\beta)$. 1.09 is the value of (r/α) maximizing the growth rate for the base model (i.e. for $\beta \to 0$).

5.3.5.3 Chain length distribution

Similarly to the other models, we start from equations 5.11 and 5.12, and assume that for t long enough, $n_i \simeq p_i \exp(\lambda t)$ (with λ the largest eigenvalue). Then,

$$\lambda p_i = -rip_i - \alpha p_i \exp(\beta i^2/8)X + r(i-1)p_{i-1} + 2\alpha p_{i+1}\exp(\beta i/2)$$

with $X = 1 + 2\sum_{j=1}^{i/2-1}\exp(-\beta j^2/2)$ (i even) or $X = 2\sum_{j=1}^{(i-1)/2}\exp(-\beta(j-1/2)^2/2)$ (i odd). For i large enough, $\lambda \ll ri$. X will tend to a finite number (converging sum) (to $Y = \theta_3(0, \exp(-\beta/2))$) for i even, $Z = \theta_2(0, \exp(-\beta/2))$) for i odd, and θ_i the Jacobi Theta functions), thus, because β is positive, for i large enough, $ri \ll \alpha \exp(\beta i^2/8)X$. Then we have to determine which of $r(i-1)p_{i-1}$ and $2\alpha p_{i+1}\exp(\beta i/2)$ dominates. If the second one dominates, $\alpha p_i \exp(\beta i^2/8)X \simeq 2\alpha p_{i+1}\exp(\beta i/2)$, thus $p_{i+1}/p_i \simeq \alpha\exp(\beta i(i/8 - 1/2))X$, which for i large enough means that the larger the cluster, the more of it, which would diverge and does not make sense in this system. Thus $\alpha p_i \exp(\beta i^2/8)X \simeq r(i-1)p_{i-1}$,

$$\frac{n_i}{n_{i-1}} \to \frac{p_i}{p_{i-1}} \simeq \frac{r}{\alpha}\frac{i-1}{X}\exp\left(-\beta\frac{i^2}{8}\right).$$

This approximation is valid for large i. Assuming that it is valid for any $i \geq 2$, and using $\sum_{i=2}^{n} i^2 = \frac{n+3n^2+2n^3}{6} - 1$:

$$p_{i,even} \simeq \left(\frac{r}{\alpha}\right)^{i-1}\frac{(i-1)!}{Y^{i/2}Z^{i/2-1}}\exp\left(-\frac{\beta}{8}\left(-1 + \frac{i+3i^2+2i^3}{6}\right)\right)$$

$$p_{i,odd} \simeq \left(\frac{r}{\alpha}\right)^{i-1}\frac{(i-1)!}{Y^{(i-1)/2}Z^{(i-1)/2}}\exp\left(-\frac{\beta}{8}\left(-1 + \frac{i+3i^2+2i^3}{6}\right)\right)$$

These two equations can be combined, and ultimately lead to:

$$p_i \simeq \left(\frac{r}{\alpha}\right)^{i-1}\frac{(i-1)!}{Y^{floor(i/2)}Z^{floor((i-1)/2)}}\exp\left(-\frac{\beta}{8}\left(-1 + \frac{i+3i^2+2i^3}{6}\right)\right) \quad (5.13)$$

This approximation works well, except for small β (figures 5.12 B and 5.14). Compared to the base model, the number of clusters decreases much faster with their size. Indeed, the breaking rates for each link increase importantly with the cluster size, thus larger clusters are much less stable than in the base model.

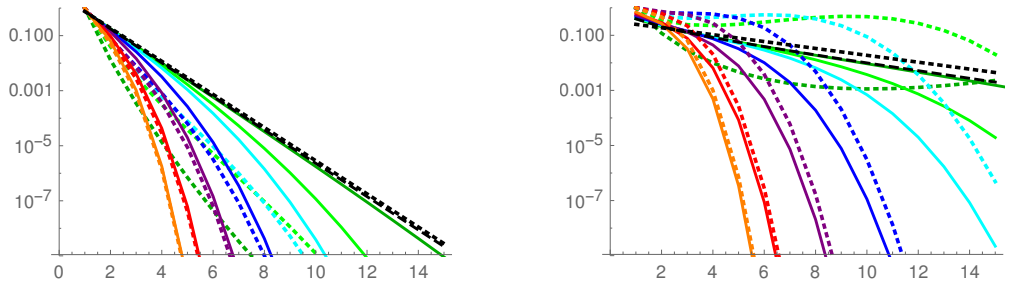


Figure 5.14 – Distribution of the cluster sizes, as in figure 5.12, except for the value of r/α

5.4 Comparison with experimental data

We analyzed experimental data from [19]: mice, which were previously vaccinated with a peracetic-acid inactivated S.Typhimurium strain (PA-S.Tm), were pretreated with 0.8g/kg ampicillin sodium salt in sterile PBS. 24h later, mice received 10^5 CFU of a 1:1 mix of mCherry-(pFPV25.1) and GFP-(pM965) expressing attenuated S. Tm M2702. For imaging, cecum content was diluted gently 1:10 w/v in sterile PBS containing $6\mu g/mL$ chloramphenicol to prevent growth during imaging. $200\mu L$ of the suspension were transferred to an 8-well Nunc Lab-Tek Chambered Coverglass (Thermo Scientific) and imaged at $100\times$ using the Zeiss Axiovert $200m$ microscope. To determine the distribution of bacteria in aggregates, $n = 25$ high power fields per mouse were randomly selected and imaged for mCherry and GFP fluorescence. For some mice, sequential sampling was done: these mice were terminally anaesthetised and artificially respirated. Cecum content was sampled by tying off parts of the cecum each hour for 3h. More details about the experimental procedures can be found in [19].

We analyzed all the images for the early data points (4 and 5 hours) of experiments starting from a low inoculum (10^5), to minimize the clustering from random encounters. Still, most clusters are large, and of complex shape. But only smaller linear clusters were counted. Figures are for the red and green fluorescence, so complex clusters with two colors were not counted. The data were analyzed manually. Below, the table of the linear clusters counted on the images from several experiments, either with mice sampled once (o), or with mice sequentially sampled (s).

cluster size	4h PI o (7 mice)	4h PI s (3 mice)	5h PI o (4 mice)	5h PI s (2 mice)	total
2	21	30	17	38	106
3	22	4	9	5	40
4	51	9	25	9	94
5	7	0	1	3	11
6	5	3	3	4	15
7	10	1	5	3	19
8	12	0	4	3	19
9	1	0	0	0	1
10	1	0	0	1	2
11	1	0	0	0	1
12	0	0	1	1	2
13	0	0	1	1	2
14	0	0	1	0	1

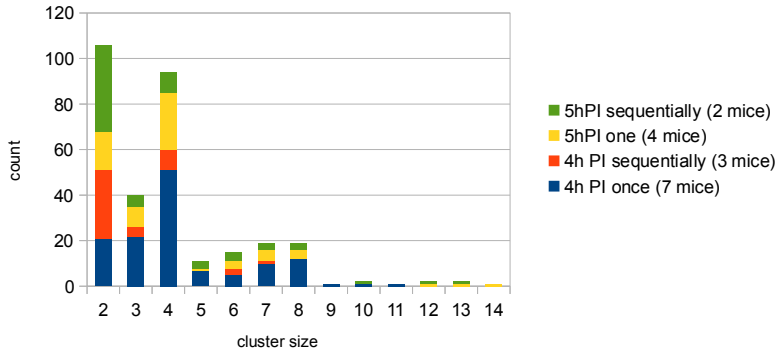


Figure 5.15 – Histogram of the linear cluster sizes 4 to 5 hours post infection

The data may be biased, as longer chains may not be fully in the focal plane. Because of gravity, they would fall close to the cover slip. The mass of one bacterium is about one pg, and its density is about 10% more than the water density[68, 69]. The thermal energy at ambient temperature is of the order of $4.10^{-21}J$, and gravity g is of the order of $10m/s^2$, thus thermal fluctuations will lift a bacterium by typically 4 μm higher than the bottom. Thus parts of the chains may be out of focus, as this is confocal microscopy, which typical optical section is less than $1\mu m$.

As there are not enough data points, we cannot quantitatively fit the data, in particular for larger chain lengths. We can nevertheless give some qualitative points. The larger value at 4 is in line with a fixed time between divisions. Clusters of uneven size could be evidences that linear chains do break. The distribution is relatively narrow, which could be compatible with force-dependent breaking rates.

5.5 Summary of the results and discussion

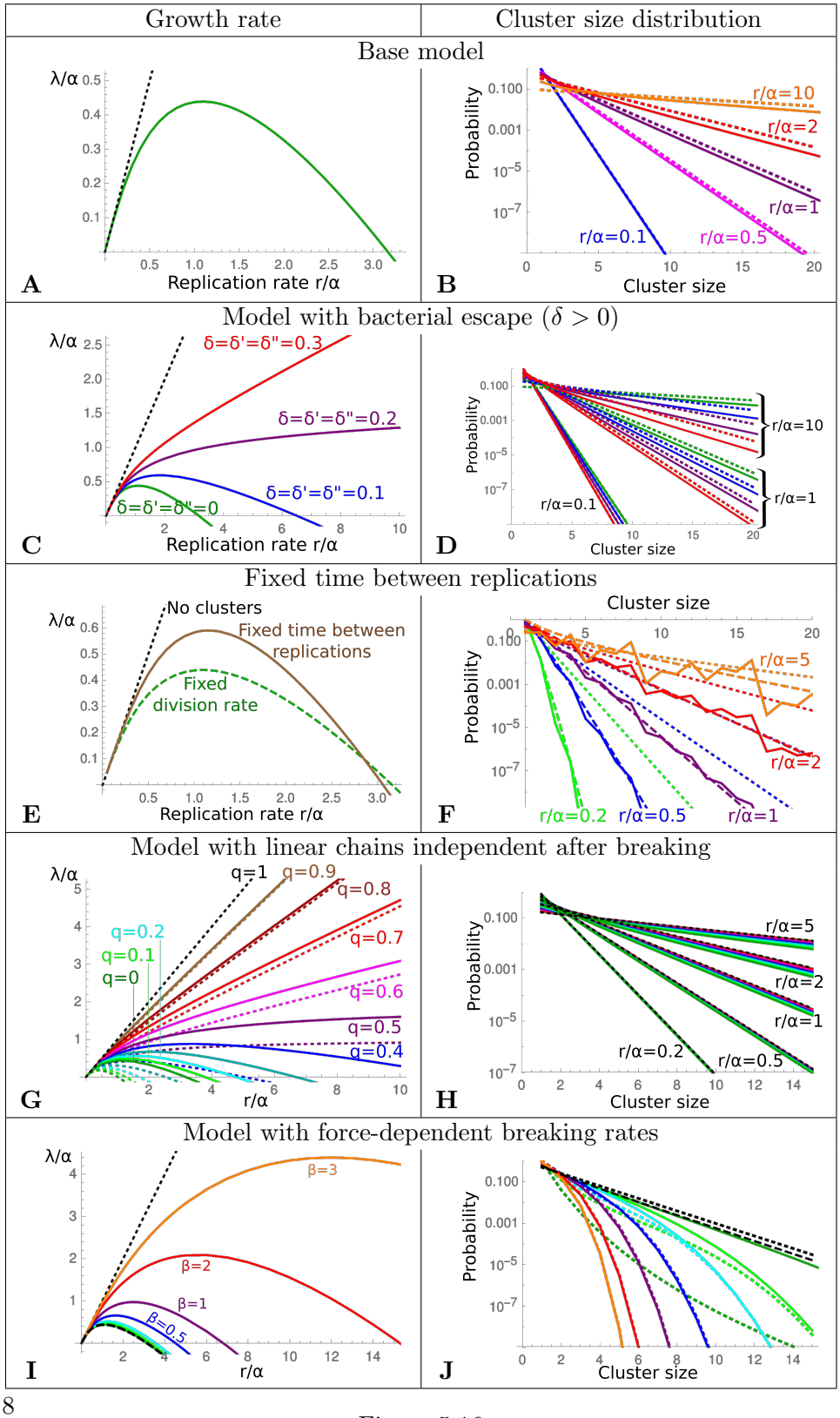


Figure 5.16

Figure 5.16: A,C,E,G,I: Growth rate λ of the free bacteria as a function of the bacteria replication rate r, both in units of α. Numerical results (solid colored lines), and limit with no clusters ($\lambda = r$) (black dotted line). B,D,F,H,J: Cluster size distribution. Solid lines: numerical results. A,B: Base model, $n_{max} = 40$. B. dotted lines: approximation (5.3) (almost overlaid with the numerical results for $r/\alpha = 0.1$). C,D: Model with bacterial escape. $\delta = \delta' = \delta'' = 0$, 0.1, 0.2, 0.3. $c = c' = 0$, $n_{max} = 40$. D. dotted lines: approximation (5.5). E,F: Fixed time between replications. $r_{eff} = \log(2)/\tau$. $n_{max} = 32$. F. approximation (5.7) (dashed lines), numerical result in the base model (dotted lines). $r/\alpha = 0.2$, 0.5, 1, 2, 5. G,H: Model with linear chains independent after breaking. $n_{max} = 100$. G. The dotted black line is the case $q = 1$, for which $\lambda = r$, like in the absence of clusters. The colored dotted lines are the analytical approximation (5.10). H. The dotted black lines are the approximate distribution (5.9) for each r/α, which is the exact distribution for $q = 1$. The colours represent the same q values than for the left panel. All curves are almost overlaid for small r. I,J: Model with force-dependent breaking rates. Each color represents a different β: $\beta = 0.01$ ($n_{max} = 20$), $\beta = 0.1$ ($n_{max} = 15$), $\beta = 0.2$ ($n_{max} = 15$), $\beta = 0.5$ ($n_{max} = 15$), $\beta = 1$ ($n_{max} = 15$), $\beta = 2$ ($n_{max} = 10$), $\beta = 3$ ($n_{max} = 10$). The black dashed lines are the numerical results for the base model, equivalent to $\beta = 0$. The curves for $\beta = 0.01$ (dark green) are almost overlaid with the curves for $\beta = 0$. J. Distribution of the cluster sizes for $r/\alpha = 1$. The colored dotted lines the analytical approximation (5.13), and the black dotted line the approximation for the base model (5.3).

We started from the recent finding [19] exposed in chapter 4 that the protection effect of sIgA, the main effector of the adaptive immune system in the gut, can be explained by enchained growth. Because sIgA are multivalent, they can stick identical bacteria together if they encounter each other. Early in infection, bacteria of the same type are at low density, thus typical encounter times are very long, but when a bacterium replicates, the daughter bacteria are in contact and thus can remain enchained to each other by IgA. Bacteria in clusters are less motile than individual bacteria, and in particular, are not observed close to the epithelial cells. In the case of wild type *S.* Typhimurium, only free bacteria which can interact with the epithelial cells contribute to the next steps of the infection process. Despite the presence of sIgA, some free bacteria are observed. It could be that they escape at the moment of replication. But, along with the observation that clusters do not grow indefinitely, it could also be a sign that the links between bacteria break. It is also physically expected that the links have some finite breaking rate. If the typical time between two bacterial divisions is much larger than the typical time for the link to break, then there would be no cluster. Conversely, in the inverse case, bacteria will be very likely to be trapped in large clusters. Then, even if sIgA are produced against all bacterial types, the bacteria dividing faster will be disproportionately affected.

We investigated whether this qualitative idea holds with more realistic models. We started from a base model in which bacteria replicate at a fixed rate and remain enchained upon replication, until the link between them breaks at a given fixed breaking rate, identical for all links. Considering that because of the way bacteria such as *Salmonella* or *E.coli* divide, the early clusters are linear chains of bacteria, it is also considered in the base model that when the chain breaks at an outermost link, the free bacteria will escape, while when the chain breaks elsewhere, the two resulting sub-chains encounter each other quickly and form clusters of more complex shapes from which individual bacteria do not escape. We studied this base model with a combination of analytical and numerical approaches. We also tested the robustness of our findings by studying separately several variations of the base model: a probability of escaping upon replication, loss rates, fixed replication time, non-zero probability for the subchains to escape, and force-dependent breaking-rates. For each model, we studied how the growth rate of the free bacteria varies with the replication rate (which would be equal if there were no clusters), and the distribution of cluster sizes. As a reminder, clusters seem unable to come close to the epithelial cells[19], thus only free bacteria interact directly with epithelial cells and may lead to systemic infections.

We find that, except in the very specific case in which subchains always escape upon link breaking, the growth rate of the free bacteria population is lower than the replication rate. And more spectacularly, in most of the models studied (but not if more than half of the subchains escape upon link breaking, or if there is a significant probability for bacteria to escape enchainement upon replication), the growth rate of the number of free bacteria is non-monotonous with the replication rate : there is a finite replication rate which maximizes the growth rate of non-clustered bacteria. At very high replication rates, bacteria get trapped

in more complex clusters and cannot contribute anymore to the free bacteria dynamics and thus to the next steps of the infection process. The replication rate maximizing the growth rate is of the order of the breaking rate, though its specific value depends on the details of the model.

The cluster size distribution is dependent on the model. In most cases, the probability for a linear cluster to be of size k decreases as γ^k, with γ some constant smaller than 1. When replication occurs at fixed time, and when breaking rates are force-dependent, the probability of larger clusters decreases faster. There are models with different cluster size distributions but qualitatively similar dependence of the growth rate on the replication rate, and the opposite is also true. This shows that large clusters have little importance for free bacteria production, what matters most is the small clusters dynamics. It is reassuring, as we did not consider buckling, which would make long linear chains fold on themselves and produce more complex clusters, and may bias the linear cluster distribution for very large sizes. It should also be noted that with fixed division time, not only is the distribution bumpy, as clusters comprising a power of two number of bacteria are more frequent than others, but the distribution is also narrower. Bacteria divide at approximately fixed division times, while replication is most often taken as occurring at fixed rates, because this makes calculations easier. Sometimes this modeling choice can lead to significant differences.

We analyzed experimental data on clusters of *S.* Typhimurium in the cecum of vaccinated mice. We have not enough data to quantitatively fit the cluster size distribution, but the distribution is qualitatively plausible with the fixed division time model (which is indeed more realistic for bacteria), and with force-dependent breaking rates. With more data, the shape of the distribution could be fitted to compare which model is the most plausible. To test the dependence of the growth rate with the replication rate, an ideal experiment would be to compare similar bacterial strains, but with differing replication rates, and compete them in the same individual. It is however very challenging to obtain bacteria that differ only by their replication rate, particularly *in vivo*.

sIgA-enchained bacterial clusters could be studied *in vitro* to measure how they break. However, using *in vitro* results to draw conclusions on *in vivo* systems is limited. First, there could be chemical or enzymatic components of the lumen that could facilitate or hinder link breaking, and the non-Newtonian viscosity of the digesta could play a role in the mechanic forces felt by the links, thus a simple buffer may not mimic well the real conditions. More crucially, the exact forces felt by particles of the size of bacterial clusters are not well characterized. Most studies of the flow characteristics in the digestive system rely either on external observations of the peristaltic muscles[10] or indirect measures of times for a marker to exit some section of the digestive track[70]. More quantitative study of the digestive flow at small scales is just beginning[9, 71, 72, 41, 73, 42, 58] and in the future it may give more clues to assess to which forces bacteria are subjected to in the digestive track.

The mechanism we propose is nevertheless plausible. The observation in vaccinated mice of the existence of single bacteria and small clusters, and particularly

small linear clusters with an odd number of bacteria, are pieces of evidence that clusters do break in these *in vivo* conditions. An alternative explanation could be that some bacteria escape enchainement upon replication. However, at higher bacterial densities, we have evidence of independent bacteria binding when they encounter, thus sIgA coated bacteria are adhesive. When two daughter bacteria divide, they are in contact, thus if sIgA is adhesive, escape is unlikely (see section 5.2.3). Importantly, even though our results show that specific conditions are needed for the growth rate to decrease with high replication rates, we almost always find that the higher the replication rate, the higher the proportion of bacteria trapped in clusters. Thus, even when it does not reverse the relationship between the growth rate of the free bacteria and the replication rate, it is at least dampening this relationship, and can be a tool both to control pathogenic bacteria, but also to maintain homeostasis of the gut microbiota. It is also interesting that there are other host effectors besides sIgA that bind bacteria together (neutrophil extracellular traps for instance[74]), and there could also be an interplay between replication rates and the breaking of the links mediated by these other effectors, as the mechanism we propose here is generic.

As for any mechanism to fight against bacteria, how easily resistance can be evolved is crucial. On the one hand, the replication rate could evolve. But a bacterial strain replicating slower would be less competitive with other bacteria in the absence of sIgA, and a slower growth leaves more time for further host response. On the other hand the typical link breaking time could evolve. On the host side, sIgA is thought to be mechanically very stable, and experiments about the bonding of cells by sIgA seem to point to the link failing because of the extraction of the antigen rather than because of sIgA breaking, and rather than the sIgA/antigen bond detaching[67][62]. In the case of IgA defficiency, there is more secretion of IgM, and microbiota is disturbed[75]: we may speculate that IgM being less powerful for microbiota homeostasis is related to these immunoglobulins being more protease-sensitive than IgA and thus cleaved on shorter time scales[76]. On the other side, bacteria could evolve surface antigens. It could be interesting to think that bacteria could produce decoy antigens with no functional value, but against which the immune system will mount an immune response, and that are more easily released from the bacteria, thus disabling the main sIgA mode of action (being easily evolvable would also be a benefit). Such decoys would however be a metabolic cost for the bacteria, and when breaking, may unmask other antigens corresponding to crucial functions of the bacteria. It could be argued that the capsule around bacteria such as *Salmonella* spp., and also common in pathogenic *E.coli*, may behave as a decoy, though it has also other functions. Along the same lines, we may speculate whether mechanical aspects could be a reason why sIgA against some antigens are not efficient for protection. For instance, while anti-flagella sIgA aggregate very well *Salmonella* Enteriditis together, they are not efficient for protection[77]. A main reason could be that as *Salmonella* can switch flagella production on and off, then some *Salmonella* will always escape these sIgA, and seed the infection[78]. An additional possibility could be that flagella may more easily break, especially as distance between

bacteria bound by flagella (long) is likely larger than for bacteria bound by O-antigens (on chains shorter than flagellas), and thus the shear forces would be larger. Further, the mechanical properties of the outer sugar layer of the gram negative bacteria could vary, and thus could be used to tune interactions. However, it would add another constraint on bacteria, and the general result that the growth rate compared to the replication rate is at least dampened by the cluster formation would remain.

In the crowded environment of the gut, it is hard for the host to identify the 'good" and the "bad" bacteria. That vaccination with dead bacteria is sufficient to produce sIgA and protection, shows that the host does not discriminate well against which bacteria they produce sIgA, as these dead bacteria do not harm. Linking the effect (here the clustering) of the immune effectors with a property directly relevant to the potential bacterial pathogenicity (here the replication rate) saves the immune system from having to make complex decisions about which bacteria to produce effectors against.

Chapter 6

Consequences of enchained growth on the evolution of antibiotic resistance

Contents

6.1 Introduction

Since the discovery of penicillin, every release of a new antibiotic has been followed a few years later by the emergence of bacteria resistant to it [4]. Antibiotics are an essential tool for medicine, thus the spread of resistance is problematic. If bacteria are sensitive, antibiotic treatment can kill them. But if some of them are resistant, then treatment will increase the proportion of antibiotic resistant bacteria in the host. Furthermore, the body is home to a diverse microbiota, most of it in the gut [1], important both in numbers and in function [57, 2]. Taking an antibiotic treatment against one pathogenic bacteria can favor the evolution

of drug resistance in other bacteria, in particular in the gut, and it can then be transmitted via the fecal oral route. Antibiotic use is widespread: for instance, about a quarter of French people are treated with antibiotics every given year [35, 36]. Besides, antibiotics are often routinely given to farm animals, and the drug resistance in bacteria they harbor may spread to humans [79, 80], though the magnitude of this effect is disputed [81]. The interaction between antibiotic use and spread of resistance in a population has been the subject of many models. Here, we develop a multiscale model, with more realistic within-host dynamics, integrating an important aspect of immunity.

Immunity could interfere with the spread of resistance. If the immune system in the gut were massively killing bacteria, it could destabilize the microbiota. Thus, it has to resort to other strategies. We have seen in chapter 4 that IgA neither kills its target bacteria nor prevents them from reproducing, but enchain daughter bacteria upon division [19], in a process we called *enchained growth*. Clusters of bacteria cannot come close to epithelial cells and thus this prevents systemic infection and protects the host. Besides, interaction of pathogenic bacteria with the epithelial cells can trigger inflammation, which can turn on the bacteria SOS response, enabling more horizontal gene transfer between bacteria. Enchained growth is thus one possible mechanism for the immunity (acquired either through previous encounters or vaccinations) to dampen horizontal transfer in the gut [82]. Furthermore, the simple fact that such IgA-mediated clusters of bacteria are mostly clonal, makes that even if there is horizontal transfer, it most likely occurs between bacteria close in space, and thus probably between very closely related bacteria, which will be inefficient for getting new genes. These effects will work unequivocally towards reducing the emergence of antibiotic resistance within the host. In this chapter, we look at another subtler effect: because bacteria will be in clonal clusters, there will be less effective genetic diversity within the host, and thus transmission will be less diverse too. We suspect that this will also decrease the probability of emergence of antibiotic resistance at the scale of the host population. The aim of this chapter is to study this effect.

New mutations occur upon replication within the host. What is crucial is whether resistant bacteria can spread among the host population. We thus need a multiscale model. We will use a minimal model, with deterministic within-host dynamics (as the number of bacteria within a host is very large), and with a simple stochastic branching process at the between-host scale. This is appropriate for the beginning of an epidemic, when very few hosts have been infected. The kind of scenario we have in mind is when there is an individual infected with a bacterial strain, similar enough to other circulating strains so that a portion of the population has immunity against it, but on which a mutation conferring resistance to an antibiotic can occur. Thus, what we will compute is the probability that, starting from an infected individual, the bacteria can invade the population or not.

This chapter presents a work in progress developed in collaboration with Loïc Marrec and Anne-Florence Bitbol from Laboratoire Jean Perrin, Paris. In a first section, we introduce all the components of the model : within-host dynamics,

transmission step, and between host transmission rules. Then we present the main methods, both analytical and numerical, and write the complete equations for the system. We then study the simplest model, in which naive and immune hosts only differ by how grouped are the bacteria they transmit, and no mutation is allowed. Then mutations and fitness cost of resistance are introduced, and we compare the probability of the emergence of infection in the whole immune and whole naive cases. We then introduce very briefly the case where the immune hosts may have a different number of contacts, and a different probability of treatment, *e.g.* if immune hosts are less sick and thus silent carriers of the bacteria. Finally, the results are discussed.

6.2 Model

6.2.1 Within-host dynamics

There is often a typical number of bacteria transmitted from one host to the next for successful infection, called the bottleneck size N_b. For instance, 10^5 is the typical number of *Salmonella* for food poisoning in humans [59]. Here we will assume that an infection within a host always starts with the same number N_b of infecting bacteria. Within the host, the number of bacteria is typically very large. For instance with a *Salmonella* infection, its density can reach 10^{10} bacteria per gram of gut content [19]. Then, stochastic fluctuations are likely small, which justifies using a deterministic model.

6.2.1.1 Types of bacteria

We assume that there are two types of bacteria, a sensitive type, growing at rate r within the host, and a resistant type, growing at rate $r(1-s)$ without antibiotics. It is often the case that resistance comes at a cost, with typical values for s from 0.005 to 0.3 [83, 84, 85, 86] (though it can sometimes be larger, see for instance [87]). Here we assume that the fitness difference only affects the within-host growth rate, but not the transmissibility [88, 89]. We assume that for each replication, the probability that each daughter bacteria is mutant is μ_1 from sensitive to resistant, and μ_2 from resistant to sensitive. Typical mutation rates for bacteria are in the range of $10^{-10} - 4.10^{-9}$ per base pair per replication [90]. There are often several mutations conferring resistance, so the sum of these different pathways will result in μ_1 of the order of $10^{-6} - 10^{-10}$ per replication [85], and μ_2 possibly smaller as the exact same mutation has to be reverted [83]. Thus we will assume that $1 \gg s \gg \mu_1, \mu_2$.

6.2.1.2 Treatment

When the host is treated with an antibiotic, we will assume that if it was initially infected with sensitive bacteria only, the treatment is very efficient and kills all bacteria before resistant bacteria appear via mutations, and before transmission to other hosts. But if there was at least one resistant bacteria at the beginning of

the infection, then the resistant strain will take over, and eventually the infection within the host will be made of resistant bacteria only.

When the host is not treated, there can be both resistant and sensitive bacteria. We write the differential equations for the mean numbers of resistant and sensitive bacteria in function of time and look at their relative proportions in the following section.

6.2.1.3 Within-host growth equations

Discrete vs. continuous time representation

Let us first consider the case without mutations, with S the number of sensitive bacteria, R the number of resistant bacteria, s the fitness cost of resistance. If there are G generations for the sensitive strain, there are $G(1-s)$ generations for the resistant strain. With r the growth rate of the sensitive bacteria, the representation in ordinary differential equations is:

$$\frac{dS}{dt} = rS$$

$$\frac{dR}{dt} = r(1-s)R$$

The solutions of these equations are $S(t) = S_0 \exp(rt)$, $R(t) = R_0 \exp(r(1-s)t)$. If we start from one bacteria of each type, after a time τ corresponding to G generations of the sensitive bacteria, there are $2^G = \exp(r\tau)$ sensitive bacteria, and $2^{G(1-s)} = \exp(r(1-s)\tau)$ resistant bacteria. Thus:

$$G\log(2) = r\tau. \tag{6.1}$$

Then, let us look at the case with mutations. When a sensitive bacteria divides, each of the daughter cells has a probability μ_1 to have mutated (and thus to have become resistant in this simplified system). When a resistant bacteria divides, each of the daughter cells has a probability μ_2 to have mutated (and thus to have become sensitive in this simplified system). Let us denote $\tilde{\mu_1}$ and $\tilde{\mu_2}$ the mutation rates for the system of differential equations, such that:

$$\frac{dS}{dt} = r(1-\tilde{\mu_1})S + \tilde{\mu_2}r(1-s)R$$

$$\frac{dR}{dt} = r(1-s)(1-\tilde{\mu_2})R + \tilde{\mu_1}rS \tag{6.2}$$

We then look for the relation between μ_i and $\tilde{\mu_i}$. The accumulation of mutants in the early dynamics has to be the same. Starting from sensitive bacteria only, neglecting back mutations, and taking the limit of s very small; when considering a bacteria after G generations, there was G opportunities for mutation, *i.e.* the proportion of resistant bacteria will be $G\mu_1$. If there were S_0 sensitive bacteria (and no resistant bacteria) at $t = 0$, and still neglecting back-mutations, $S(t) =$

$S_0 \exp(r(1-\tilde{\mu_1})t)$, and, replacing $S(t)$ by this expression in (6.2), and solving for $R(t)$ with $R(0) = 0$:

$$R(t) = S_0\tilde{\mu_1} \exp(r(1-\tilde{\mu_1})t)\frac{\exp(r(\tilde{\mu_1}+\tilde{\mu_2}(-1+s)-s)t)-1}{\tilde{\mu_1}+\tilde{\mu_2}(-1+s)-s}.$$

In the limit of small t,

$$R(t) \simeq S_0\tilde{\mu_1} rt \exp(r(1-\tilde{\mu_1})t)$$

and the proportion of resistant bacteria then reads:

$$p(t) = \frac{R(t)}{R(t)+S(t)} \simeq \frac{R(t)}{S(t)} = \tilde{\mu_1} rt$$

Thus $G\mu_1 = \tilde{\mu_1} r\tau = \tilde{\mu_1} G \log(2)$ (the latter because of (6.1)), and consequently we have to take $\tilde{\mu_1} = \mu_1/\log(2)$ for consistency.

Resolution of the continuous time system

When the host is not treated, the dynamics within the host can be complex. The growth could be limited by some carrying capacity and taken as logistic, there could be a loss term, etc. In the limit of small s, as we want to calculate the proportions of sensitive and resistant bacteria, the following equations will give the same results than equations with a carrying capacity:

$$\frac{dS}{dt_g} = (1-\mu_1/\log(2))S + (1-s)R\mu_2/\log(2), \tag{6.3}$$

$$\frac{dR}{dt_g} = (1-s)(1-\mu_2/\log(2))R + \mu_1 S/\log(2), \tag{6.4}$$

with t_g the time rescaled by the generations. The aim is to obtain the proportion of sensitive and resistant bacteria at the end of the infection within a host, depending on the initial composition of the infection. The total number of replications within a host G can vary. Typical minimal doubling time for bacteria is half an hour [53], but it can also be as large as a few hours [91]. Bacterial carriage can last several days or even more, but when close to carrying capacity, the growth rate decreases. As a portion of the bacteria will be lost in feces, there will be ongoing replication, though at a lower rate. Thus G can take a wide range of values. For instance, in experimental infection of mice by *Salmonella* starting at different inoculum sizes, the number of replications is typically 10 (inoculum of 10^7 bacteria) to 35 (inoculum of 10^3 bacteria) after 24h [19].

Solving the equations 6.3 and 6.4 with the initial conditions $S(0) = N - i$ and $R(0) = i$, we find for all i between 0 and N the following exact expression for the proportion $\frac{R}{R+S}$ of resistant bacteria after G generation, knowing that the infection was seeded with i resistant and $N - i$ sensitive bacteria at time $t = 0$:

$$p_i = \frac{(2^{\Delta G}-1)(2\mu_1 N + i(-\mu_1-\mu_2-s\log(2)+\mu_2 s)) + i\Delta\log(2)(2^{\Delta G}+1)}{N((2^{\Delta G}-1)(\mu_1+\mu_2+s\log(2)-2is\log(2)/N-\mu_2 s)+\log(2)\Delta(2^{\Delta G}+1))}$$

with

$$\Delta = \sqrt{s^2\left(1-\frac{\mu_2}{\log(2)}\right)^2 + 2s\left(-\frac{\mu_1}{\log(2)}+\frac{\mu_2}{\log(2)}-\frac{\mu_1\mu_2}{\log^2(2)}-\frac{\mu_2^2}{\log^2(2)}\right)+\frac{(\mu_1+\mu_2)^2}{\log^2(2)}}.$$

Let us look at some particular limits:

Case where there was no resistant initially: In this case, as $1 \gg s \gg \mu_1, \mu_2$, the final proportion of resistant bacteria can be approximated to:

$$p_0 \simeq \mu_1 \frac{1-2^{-sG}}{s\log(2)}. \tag{6.5}$$

Case with mixed inoculum: Let us distinguish two cases:

- If $sG \ll 1$, when starting with both strains, their relative proportion will have little time to change. If there is one resistant bacteria initially, then let us neglect mutations in both ways, as the mutation rate is small, and then the final proportion of resistant bacteria is:

$$p_1 \simeq \frac{1}{1+(N_b-1)(1+G\log(2)s)} \tag{6.6}$$

- If $sG \gg 1$, then if there was at least one resistant and one sensitive initially, the mutation selection balance is reached within the infected host. The final proportion of resistant bacteria p_{MSB} is obtained by writing the differential equation on $\frac{R}{R+S}$ and looking for its equilibrium. In the limits we are considering, p_{MSB} tends to $\mu_1/(s\log(2))$.

Case starting from resistant bacteria only : In this case, the final proportion of resistant bacteria will be:

$$p_N = 1 - \frac{2(2^{\Delta G}-1)\mu_2(1-s)}{(2^{\Delta G}-1)(\mu_1+\mu_2-s\log(2)-s\mu_2)+(1+2^{\Delta G})\Delta\log(2)}$$

with

$$\Delta = \sqrt{s^2\left(1-\frac{\mu_2}{\log(2)}\right)^2 + 2s\left(-\frac{\mu_1}{\log(2)}+\frac{\mu_2}{\log(2)}-\frac{\mu_1\mu_2}{\log^2(2)}-\frac{\mu_2^2}{\log^2(2)}\right)+\frac{(\mu_1+\mu_2)^2}{\log^2(2)}}.$$

For G not too large, as $1 \gg s \gg \mu_1, \mu_2$, then $\Delta \simeq s$. Consequently:

$$p_N \simeq 1 - \frac{(2^{sG}-1)\mu_2(1-s)}{2^{sG}\mu_2(1-s)+s\log(2)}. \tag{6.7}$$

For G very small, $p_N \simeq 1 - G\mu_2$ (simply the accumulation of mutations).

6.2.2 Transmission

We will consider that for each transmission, transmitted bacteria are chosen from the donor host using these probabilities, without correlation between two transmissions from the same host (we will notably suppose that selections are done with replacement).

For naive individuals, bacteria will remain independent from each other, whereas for hosts who are immune to this bacteria, they will be bound together by the secreted IgA. However, as forces grow significantly larger on larger clusters, or because there may be some lifetime of the bounds, clusters getting to a certain size will break [62, 63, 64, 65, 20]. Thus at the end of the infection, clusters will be of a typical size N_c. We will take the limit in which $N_c = N_b = N$. It may be more realistic that $N_c < N_b$, *i.e.* multiple clusters are transmitted to a recipient. However, we want to focus on the effect of clustering, and a more comprehensive model, taking into account N_b/N_c clusters transmitted, would be more complex without adding much to the comprehension of this effect. The limit $N_c = N_b = N$ gives an upper bound of the effect of clustering.

We assume that the concentration of the studied bacteria remains small in the gut, so that the typical encounter time between clusters is large, thus the clusters present are of the same lineage, as bacteria get enchained by IgA upon replication. Then, in the absence of mutations, or when they are negligible (when the initial inoculum was mixed), clusters are made of bacteria of the same type, either all sensitive or all resistant. A simplified view of the process of cluster growth and breaking is that clusters break once a certain size is reached, then grow, then break again, and so on. As daughter bacteria are physically close in the clusters, then the subclusters formed after a larger cluster breaks will be of related bacteria. Let us assume that there are G generations in total during a host infection, and that the maximal cluster size is $2^g = N_c = N$ (achieved in g generations; $g \leq G$).

Let us study what happens when bacteria are initially of one type and a mutation occurs: if a mutation occurs at one replication, the cluster will be of mixed bacterial types, until after g generations, when the cluster will be all made of descendants of this mutant (see figure 6.1). In the limit of $Gs \ll 1$, for a mutation rate μ for each daughter bacteria per replication, the proportion of clusters made of mutant bacteria only will be the probability for the bacteria that seeded this cluster to have been mutant, *i.e.* $\mu(G-g)$ (as $G-g$ is the number of replications this bacteria has gone through before seeding the cluster). The proportion of mixed clusters will be equal to the probability for a mutation to have occurred during the replications between this seeding bacteria and the final cluster, and thus $2\mu(2^g-1) = 2\mu(N-1)$ (see equation F.4). For a more detailed discussion, see appendix section F. We will consider two limit cases for when the host was initially infected with sensitive bacteria only: when the number of mixed clusters is small compared to the number of fully mutant clusters (then the probability to transmit a fully mutant cluster will simply be the mean proportion of mutants in the donor host p_0), and the case when most transmissions of mutant bacteria are through mixed clusters.

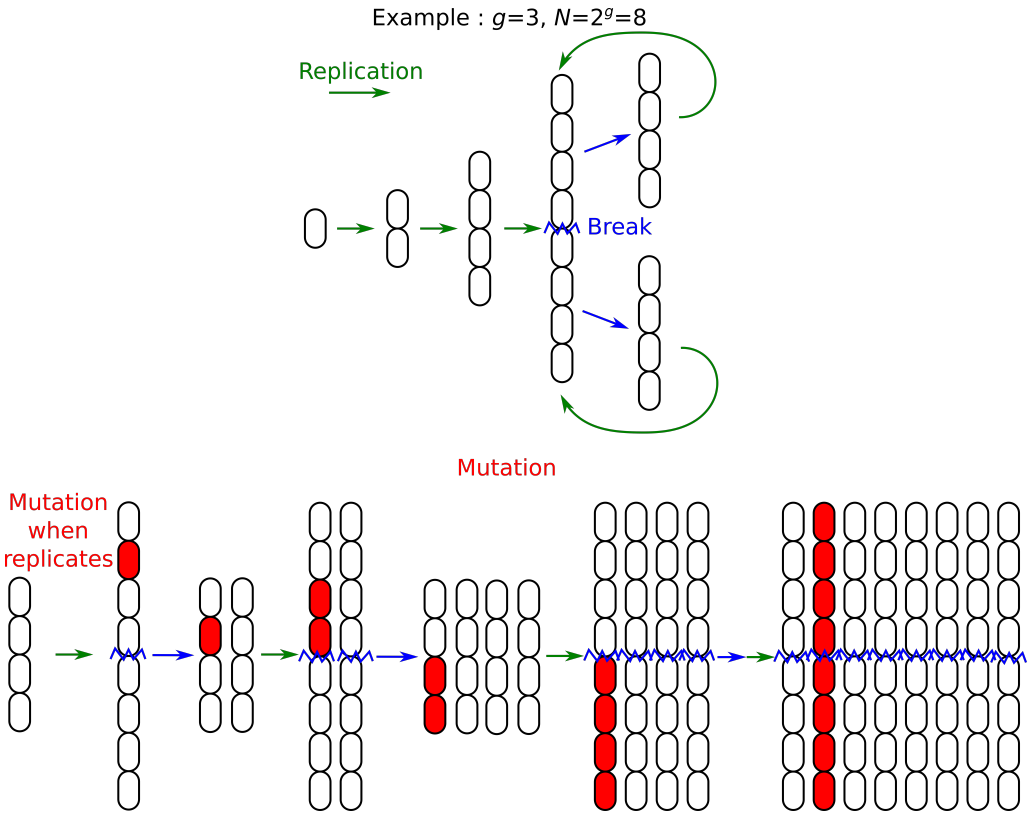


Figure 6.1 – Schematic of growth and mutations in clusters. Represented here is the case of simple linear clusters. For more complex clusters, it will remain true that more closely related bacteria will be close to each other, as daughter bacteria remain bound together after replication.

6.2.3 Between hosts

We focus on the effect of clustering on the apparition of antibiotic resistance. We thus take a simple model in which there is no population structure. Each host has a probability w to be immune to the bacteria studied (and thus $1 - w$ of being naive) (we will also compare the results for $w = 0$ and $w = 1$, *i.e.* fully naive vs. fully immune host population). The rationale is that we are interested in the spread of a strain with resistance risk, and, while this strain is new, it is similar enough to other strains present in the population so that there is some cross immunity. We look at the beginning of the spread of this strain, so that we neglect the effect that over time, w will increase, as infected hosts become immune to this new strain.

In addition to the clustering of bacteria, immune hosts may differ in other ways from naive hosts.

First, the mean number of other hosts infected by a donor host (that we will denote as λ for a naive donor host, and λ' for immune donor host) may differ. On

the one hand, because clustering prevents direct interaction with the epithelial cells and thus inflammation for some pathogenic bacteria, the infection may be cleared faster, by competition with other bacteria of the microbiota for instance. Thus an infected host would infect on average fewer contacts, and that would decrease the probability of resistance emergence. On the other hand, in the case of a pathogenic bacteria, an immune host may be less sick while still shedding bacteria, and be a silent carrier, thus its number of contacts may be increased in comparison to a visibly sick host. In this case, there would be competition between the different effects. We assume that the number of transmissions to recipient hosts from one donor host is Poisson distributed [92, 93, 94, 95, 89].

Secondly, immune hosts and naive hosts may not be treated at the same frequency (here the probability of antibiotic treatment is q for naive hosts, q' for immune hosts). If antibiotics are given for another reason (to fight another bacteria, or for growth-enhancement in farm animals), then the probability of treatment will be the same for all. But if antibiotics are given specifically when a host displays symptoms linked to the infection we model, then an immune host, less sick, will be less likely to be treated. Then an immune host could be a reservoir of sensitive bacteria, which can either increase the emergence of resistance just by enabling spreading to more hosts, or decrease the emergence of resistance due to a reduced use of antibiotics and thus less competitive advantage of resistant versus sensitive bacteria. We will consider both $q = q'$ and $q \neq q'$, with in all cases both q and q' finite and not too small.

6.3 Methods and equations

6.3.1 General methods

Since we focus on the beginning of the spread, we use the framework of branching processes [37, 38, 39]. The probability for a host initially infected by i resistant and $N - i$ sensitive bacteria to infect n_0 hosts with 0 resistant and N sensitive bacteria, n_1 hosts with 1 resistant and $N - 1$ sensitive bacteria, and so on, is $\wp_{i,\{n_0,n_1,....,n_N\}}$. We write the equations for the generating functions:

$$g_i(z_0, ..., z_n) = \sum_{\{n_0,n_1,....,n_N\}} \wp_{i,\{n_0,n_1,....,n_N\}} z_0^{n_0} ....z_N^{n_N}$$

As there is no correlation between transmissions, and as the number of infected contacts is Poisson distributed, of mean $\tilde{\lambda}$ (with $\tilde{\lambda} = \lambda$ for naive hosts and λ' for immune ones), then:

$$g_i(z_0, ..., z_n) = \sum_{k=0}^{\infty} \frac{\tilde{\lambda}^k e^{-\tilde{\lambda}}}{k!} \left(\sum_{j=0}^{N} f_{i,j} z_j \right)^k = \exp\left(-\tilde{\lambda} \left(1 - \sum_{j=0}^{N} f_{i,j} z_j \right) \right)$$

$$= \exp\left(-\tilde{\lambda} \sum_{j=0}^{N} f_{i,j} (1 - z_j) \right)$$

with $f_{i,j}$ the probability that when a host, initially infected with i resistant bacteria and $N-i$ sensitive ones, infects another host, it transmits to this other host j resistant bacteria and $N-j$ sensitive ones. We can write the equations for all g_i with i between 0 and N.

We calculate the probability that this new strain spreads in the population, *i.e.* one minus its extinction probability. Our model is adequate for the early steps of spread, more complex models are needed for studying the later stages of an epidemic. The extinction probabilities e_i, starting from an individual infected with i resistant and $N-i$ sensitive bacteria, are the fixed point of the generating functions, and thus solutions of $e_i = \exp(-\tilde{\lambda}(1-\sum_{j=0}^{N} f_{i,j}e_j))$ [38]. Either the bacterial strain will have a limited spread in the host population and go extinct, or it will transmit to an ever increasing number of hosts, acquiring resistance on the way, and spread resistance.

Our starting point will be these equations for the extinction probabilities.

Because when the host is treated, if it was initially infected with no resistant bacteria then it does not transmit anything, the general equations write:

$$\begin{aligned} e_i = {} & (1-w)(1-q)g_{naive,i}(e_0,e_1,...,e_N) \ \} \textit{ non immune non treated} \\ & +w(1-q')g_{immune,i}(e_0,e_1,...,e_N) \ \} \textit{ immune non treated} \\ & +(1-w)q(\delta(i,0)+(1-\delta(i,0))\exp(-\lambda(1-e_N))) \ \} \textit{ non immune treated} \\ & +wq'(\delta(i,0)+(1-\delta(i,0))\exp(-\lambda'(1-e_N))) \ \} \textit{ immune treated} \end{aligned}$$

We solve numerically the system of equations giving the values of e_i and look for analytical approximations.

6.3.2 Naive hosts

In naive hosts initially infected with i mutant bacteria, when there is transmission, the probability to transmit j resistant bacteria and $N-j$ sensitive ones will be:

$$f_{i,j} = \binom{N}{j} p_i^j (1-p_i)^{N-j}.$$

Then, for all i between 0 and N:

$$g_{naive,i}(e_0,e_1,...,e_N) = \exp\left(-\lambda\left(\sum_{j=0}^{N}\binom{N}{j}p_i^j(1-p_i)^{N-j}(1-e_j)\right)\right).$$

6.3.3 Immune hosts

In immune hosts, there is the question of whether the clusters are of one bacterial type, or mixed. As explained in appendix section F, the clusters are made of daughter cells enchained together, thus in the absence of mutations, clusters will be made of bacteria of one type only. As the mutation rates are very small, when the initial inoculum was mixed, the frequency of mixed clusters will be very small compared to the frequency of clusters of one type only. We thus neglect mixed clusters when the initial inoculum was mixed. Thus for all i between 1 and $N-1$,

$$g_{immune,i}(e_0,e_1,...,e_N) = \exp\left(-\lambda'\left((1-p_i)(1-e_0)+p_i(1-e_N)\right)\right).$$

6.3.3.1 Limit $G \gg N$

In this limit, we neglect mixed clusters in all cases (see section F). Thus for all i between 0 and N:

$$g_{immune,0}(e_0, e_1, ..., e_N) = \exp\left(-\lambda'\left((1-p_0)(1-e_0) + p_0(1-e_N)\right)\right).$$

$$g_{immune,N}(e_0, e_1, ..., e_N) = \exp\left(-\lambda'\left((1-p_N)(1-e_0) + p_N(1-e_N)\right)\right).$$

6.3.3.2 Limit $G \ll N$

In this limit, we will consider only the mutants in the mixed clusters (see section F). We assume s small enough so that $sg \ll 1$ and we can neglect the differences in growth time between the different types of clusters. A cluster at generation G was founded by one bacteria at generation $G-g$. Since we count only mixed clusters, at the earliest, a mutation occurred when this founding bacteria duplicated with probability 2μ, resulting in a cluster of size 2^g containing 2^{g-1} mutants. The probability for mutation will be $2^2\mu$ at next round, and so on. Thus, the probability for a cluster to include one mutant is $2^g\mu = N\mu$, 2 mutants is $2^{g-1}\mu = N\mu/2$, 4 mutants is $2^{g-2}\mu = N\mu/2^2$, ... , 2^{g-1} mutants is 2μ. Then:

$$g_{immune,0}(e_0, e_1, ..., e_N) = \exp\left[-\lambda'\left((1-2\mu_1(N-1))(1-e_0) + \sum_{j=0}^{g-1} N\mu_1 \frac{1-e_{2^j}}{2^j}\right)\right]$$

$$g_{immune,N}(e_0, e_1, ..., e_N) = \exp\left[-\lambda'\left((1-2\mu_2(N-1))(1-e_N) + \sum_{j=0}^{g-1} N\mu_2 \frac{1-e_{N-2^j}}{2^j}\right)\right]$$

General case We can take both cases into account. The probabilities p_0 and p_N should be computed at $G-g$ instead of G. We denote that p'_0 and p'_N. Then:

$$g_{immune,0}(e_0,e_1,...,e_N) = \exp\left[-\lambda'\left((1-2\mu_1(N-1)-p'_0)(1-e_0) + \sum_{j=0}^{g-1} N\mu_1 \frac{1-e_{2^j}}{2^j} + p'_0(1-e_N)\right)\right]$$

$$g_{immune,N}(e_0,e_1,...,e_N) = \exp\left[-\lambda'\left((p'_N-2\mu_2(N-1))(1-e_N) + \sum_{j=0}^{g-1} N\mu_2 \frac{1-e_{N-2^j}}{2^j} + (1-p'_N)(1-e_0)\right)\right]$$

The numerical resolution of the system will be used to explore its behavior, and will be compared to analytical approximations.

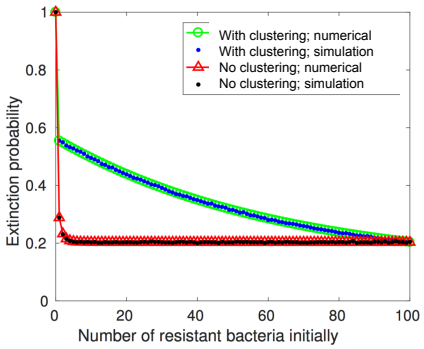
6.3.4 Table of the symbols used

Parameters specific to the infection	
G	Duration in number of replications of the intra-host infection
N_b	Typical bottleneck size, *i.e.* the number of bacteria seeding the infection in a new host
N_c	$= 2^g$ Maximum clusters size (when they reach it, they break in half before the next replication). In general we take $N_c = N_b = N = 2^g$
Intra-host dynamics	
$S(t)$	Number of sensitive bacteria within a specific host at time t
$R(t)$	Number of resistant bacteria within a specific host at time t
μ_1	Probability for one of the two daughter bacteria of a sensitive bacterium to have become resistant because of a mutation during the replication
μ_2	Probability for one of the two daughter bacteria of a resistant bacterium to have become sensitive because of a mutation during the replication
p_i	Proportion at transmission of resistant bacteria within a host that was initially infected with i resistant and $N - i$ sensitive
p_{MSB}	Proportion of resistant bacteria in the mutation-selection balance (*i.e.* when equilibrium in proportions is reached)
Immune vs. Naive hosts	
w	Probability for a host to be immune
q	Proportion of naive individuals in the host population who are antibiotic-treated
q'	Proportion of immune individuals in the host population who are antibiotic-treated
λ	Mean number of contacts a naive host transmits the infection to
λ'	Mean number of contacts an immune host transmits the infection to
Systems of equations	
e_i	Probability of extinction for an infection that was seeded in patient zero by i resistant bacteria and $N - i$ sensitive bacteria
$f_{i,j}$	Probability that when a host, initially infected with i resistant and $N-i$ sensitive bacteria, infects another host, it transmits j resistant bacteria and $N - j$ sensitive ones.

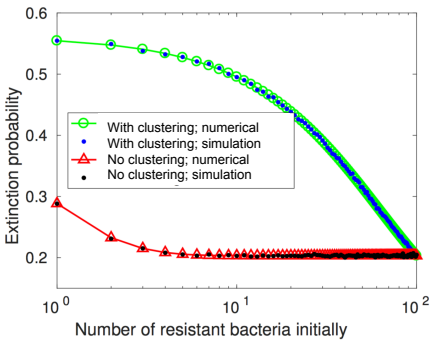
6.4 Results

6.4.1 Impact of clustering in the absence of mutations

We first study the simplest case, in which there is no mutation, and no fitness cost of the resistance (so the proportion of resistant bacteria is the same at the beginning and at the end of the infection within a host). If all the hosts in the



Linear scale



Log scale

Figure 6.2 – Extinction probability of the epidemic as a function of the number of resistant bacteria i in the first infected individual. Here $\mu_1 = \mu_2 = 0$, $s = 0$, $N = 100$, $\lambda = \lambda' = 2$ and $q = q' = 0.55$, and i is varied from 0 to $N = 100$. Results come from numerical resolution of the system of equations and from numerical simulation of the branching process (designed by Loïc Marrec and Anne-Florence Bitbol, not detailed here), with an averaging on 10^5 replicates.

population are naive, then:

$$e_0 = q + (1-q)\exp\left[-\lambda(1-e_0)\right], \tag{6.8}$$

$$e_i = q\exp\left[-\lambda(1-e_N)\right] + (1-q)\exp\left[-\lambda\sum_{j=0}^{N} f_{i,j}(1-e_j)\right] \quad \forall i \in [1, N-1],$$

$$e_N = \exp\left[-\lambda(1-e_N)\right], \tag{6.9}$$

with $f_{i,j} = \binom{N}{j}\left(\frac{i}{N}\right)^j\left(\frac{N-i}{N}\right)^{N-j}$.

If all the hosts in the population are immune, then we should solve:

$$e_0 = q' + (1-q')\exp\left[-\lambda'(1-e_0)\right], \tag{6.10}$$

$$e_N = \exp\left[-\lambda'(1-e_N)\right], \tag{6.11}$$

and then e_i can be obtained from e_0 and e_N via

$$\begin{aligned} e_i &= q'\exp\left[-\lambda'(1-e_N)\right] + (1-q')\exp\left[-\frac{\lambda' i}{N}(1-e_N)\right]\exp\left[-\frac{\lambda'(N-i)}{N}(1-e_0)\right] \\ &= q'e_N + \left[(1-q')e_N\right]^{\frac{i}{N}}(e_0-q')^{1-\frac{i}{N}}. \end{aligned}$$

after replacing the expressions from 6.10 and 6.11.

One can notice that if $\lambda = \lambda'$, e_N and e_0 do not depend on whether the population is naive or immune. Numerically (see figure 6.2), for the other values of i, extinction is more likely when the population is immune than when the population is naive.

6.4.2 Impact of clustering with mutations

We now take into account mutations, and the fitness cost of the resistance. We will compare the probability of emergence of an infection at the scale of the host population when we start from a host infected with sensitive bacteria only. We will compare the case of a whole naive vs. all immune population, *i.e.* study the ratio:

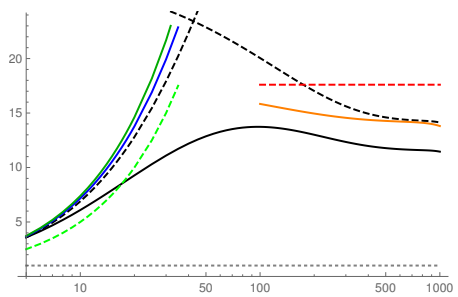
$$\frac{1 - e_{0,naive}}{1 - e_{0,immune}}.$$

Figure 6.3 shows this ratio calculated with the numerical solving of the complete systems of equation presented in the previous part, for different values of the parameters in function of G, the number of generations. The first observation is that this ratio is always bigger than 1 in the range of parameters chosen, meaning that emergence probability is reduced in the immune population compared to the naive one. We then make some approximations to grasp the general behavior of the system in certain limits, as will be shown thereafter, and compare them with the numerical solutions.

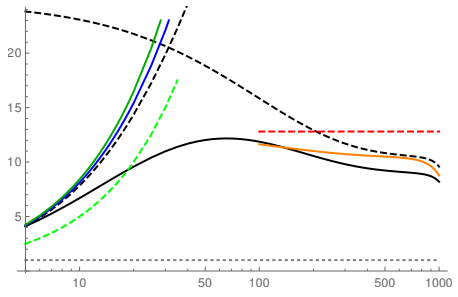
Regarding the fitness cost s, we can consider two limits. Either the number of generations within a host G is large enough ($sG \gg 1$) so that infections started with a mixed inoculum end up at proportions close to the mutation selection balance, or the number of generations is small ($sG \ll 1$) so that the final proportions of bacterial types is close to the initial values. Another matter is that, *starting from a fully sensitive infection,* mutations in an immune host could lead to either fully resistant clusters or mixed clusters. When comparing the number of these two types of transmissions, when $G \gg N$, then most transmissions containing resistant bacteria will be from clusters containing only resistant bacteria; and when $G \ll N$, then most transmissions containing resistant bacteria will be from mixed clusters. We present here the limit with few generations, with both $sG \ll 1$ and $G \ll N$, and the limit with many generations, with both $sG \gg 1$ and $G \gg N$, and give the main steps for the approximations.The details of these approximations can be found in appendix section G.

6.4.2.1 Small number of generations

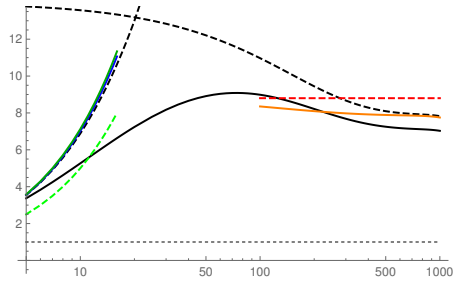
In the case of small number of generations (with both $sG \ll 1$ and $G \ll N$), starting from a host infected with only sensitive bacteria, the main path to resistance will be that, through mutations, resistant bacteria appear. But as their proportion will remain small, and the number of within-host generations is small, for both naive and immune hosts, in most cases transmission will be of either zero resistant and N sensitive or 1 resistant and $N - 1$ sensitive. The frequency of resistant will change little in hosts infected with one resistant bacteria, except if the host is treated, in which case the infection will become fully resistant. Untreated naive hosts will transmit either one or no resistant bacteria. Untreated immune hosts initially infected with one resistant bacteria will mostly transmit fully sensitive clusters of bacteria, with some occurrences of transmission of fully resistant clusters of bacteria. Most paths to resistance will thus involve princi-



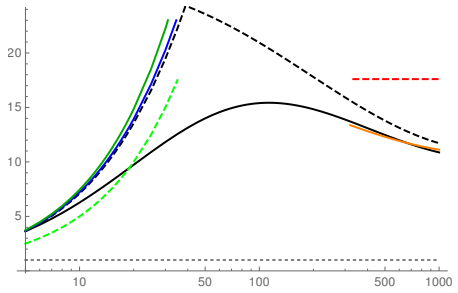
$\lambda = \lambda' = 1.7$, $q = q' = 0.55$, $s = 0.01$, $N = 32$



$\lambda = \lambda' = 1.4$, $q = q' = 0.4$, $s = 0.01$, $N = 32$



$\lambda = \lambda' = 1.7$, $q = q' = 0.55$, $s = 0.01$, $N = 16$



$\lambda = \lambda' = 1.7$, $q = q' = 0.55$, $s = 0.003$, $N = 32$

Figure 6.3 – Ratio of probability of emergence in a fully naive population relative to a fully immune population. Full system (black solid lines) (see section 6.3.3.2), system with only mixed clusters or only fully resistant clusters for transmissions from immune hosts initially infected with sensitive bacteria only (see section 6.3.3.1 and 6.3.3.2) (black dashed lines). Approximations for G small ($sG \ll 1$ and $G \ll N$): resolution of system (6.12) (6.13) (6.14) (solid blue line), expression (6.17) (solid dark green line) and limit $G/2$ (dashed green line). Approximations for G large ($sG \gg 1$ and $G \gg N$): system (6.18) (6.19) (6.20) (orange line) and limit Nq (red dashed line). The dotted gray line indicates 1, *i.e.* everything above this line shows that emergence is more likely in a naive population than in an immune population. $\mu_1 = 10^{-6}$, $\mu_2 = 10^{-8}$.

pally hosts initially infected with 0, 1, or N resistant bacteria, and taking into consideration only the most likely paths, we find that:

$$e_N = (1-w)\exp(-\lambda(1-e_N)) + w\exp(-\lambda'(1-e_N)). \tag{6.12}$$

$$\begin{aligned} e_1 =& (1-w)q\exp(-\lambda(1-e_N)) + wq'\exp(-\lambda'(1-e_N)) \\ &+ (1-w)(1-q)\exp\left[-\lambda(\exp(-Np_1)(1-e_0) + (1-\exp(-Np_1))(1-e_1))\right] \\ &+ w(1-q')\exp(-\lambda'((1-p_1)(1-e_0) + p_1(1-e_N))) \end{aligned} \tag{6.13}$$

$$\begin{aligned} e_0 =&(1-w)q + wq' + (1-w)(1-q)\exp(-\lambda((1-Np_0)(1-e_0) + Np_0(1-e_1))) \\ &+ w(1-q')\exp\left(-\lambda'\left(\left(1-\frac{2Np_0}{G}\right)(1-e_0) + \frac{2Np_0}{G}(1-e_1)\right)\right) \end{aligned} \tag{6.14}$$

Remarkably, equation (6.12) does not depend on neither q nor q'. If $\lambda = \lambda'$, it also does not depend on w. It also does not depend on either e_0 or e_1, and thus can be numerically solved independently. Then the result can be imported to equation (6.13) and the system of equations (6.13) and (6.14) can be solved numerically. It is still numerical, but less complex than the system of N equations.

When there is no mutation, and if an individual is infected only with sensitive bacteria, the mean number of recipients it infects is $R_{0,WT} = (1-w)\lambda(1-q) + w\lambda'(1-q')$. If this is smaller than one, then the probability to start an epidemic is zero. It is in this limit that we expect the most important effect of and immune population vs. a naive one. In this limit ($R_{0,WT} < 1$), then $1 - e_0$ is of the order of μ_1, and further simplifications can be made. Then, comparing the probability of spread of the strain in a fully naive relative to a fully immune population,

$$ratio = \frac{1-e_{0,naive}}{1-e_{0,immune}} = \frac{G}{2}\frac{\lambda(1-\lambda'(1-q'))}{\lambda'(1-\lambda(1-q))}\frac{(1-q)}{(1-q')}\frac{(1-e_{1,naive})}{(1-e_{1,immune})}$$

with

$$e_{1,naive} \simeq qe_{N,naive} + (1-q)\exp(-\lambda(1-e^{-Np_1})(1-e_{1,naive})).$$

$$e_{1,immune} = q'e_{N,immune} + (1-q')e_{N,immune}^{p_1}$$

$$e_{N,naive} = \exp(-\lambda(1-e_{N,naive})). \tag{6.15}$$

$$e_{N,immune} = \exp(-\lambda'(1-e_{N,immune})). \tag{6.16}$$

Then, taking further $q = q'$, and $\lambda = \lambda'$, then 6.15 and 6.16 are the same equations, thus in this regime $e_{N,naive} = e_{N,immune} = e_N$. In this regime we can further show that $e_{1,naive} < e_{1,immune}$, and that:

$$ratio = \frac{1-e_{0,naive}}{1-e_{0,immune}} = \frac{G}{2}\frac{1-e_{1,naive}}{1-e_{1,immune}} > \frac{G}{2} \tag{6.17}$$

6.4.2.2 Large number of generations

In the case of a large number of generations (with both $sG \gg 1$ and $G \gg N$), we consider that hosts can be categorized by inoculum counting 0, $0 < i < N$ or N resistant bacteria. Indeed, Because $sG \gg 1$, then mutation selection balance will be attained in hosts infected with a mixed inoculum. But as μ_1, μ_2 are very small, $\mu_i sG$ may not be large, thus infections starting with one type of bacteria

have to be considered separately. Then, considering the main paths:

$$\begin{aligned}1-e_N=&(1-w)(1-q)(1-\exp(-\lambda((1-e_i)(1-p_N^N-(1-p_N)^N)+p_N^N(1-e_N)+(1-p_N)^N(1-e_0))))\\&+(1-w)q(1-\exp(-\lambda(1-e_N)))\\&+wq'(1-\exp(-\lambda'(1-e_N)))\\&+w(1-q')(1-\exp(-\lambda'(p_N(1-e_N)+(1-p_N)(1-e_0))))\end{aligned} \tag{6.18}$$

$$\begin{aligned}1-e_i=&(1-w)q(1-\exp(-\lambda(1-e_N)))+wq'(1-\exp(-\lambda'(1-e_N)))\\&+(1-w)(1-q)(1-\exp(-\lambda((1-p_{MSB})^N(1-e_0)+(1-(1-p_{MSB})^N)(1-e_i)))\\&+w(1-q')(1-\exp(-\lambda'(1-e_0+p_{MSB}(1-e_N))))\end{aligned} \tag{6.19}$$

$$\begin{aligned}1-e_0=&(1-w)(1-q)(1-\exp(-\lambda((1-p_0)^N(1-e_0)+(1-(1-p_0)^N)(1-e_i))))\\&+w(1-q')(1-\exp(-\lambda'(1-e_0+p_0(1-e_N))))\end{aligned} \tag{6.20}$$

Though much simpler than the whole system (only 3 equations), this system has the disadvantage that all 3 equations depend on all the three extinction probabilities.

Then, taking similarly the limit when there is limited spread of the bacterial strain in the absence of mutations ($(1-w)\lambda(1-q)+w\lambda'(1-q') < 1$[1], then $1-e_0$ is of the order of μ_1, and further simplifications can be made. We also here take the assumption that p_N^N remains close to 1 even at the longest G considered, *i.e.* that μ_2 is very small. Then, comparing the probability of spread of the strain in a fully naive population relative to a fully immune population, and taking $q = q'$,

$$1 - e_N = 1 - (1 - w)\exp(-\lambda(1 - e_N)) - w\exp(-\lambda'(1 - e_N))$$

and

$$\frac{1 - e_{0,naive}}{1 - e_{0,immune}} \simeq Nq\frac{\lambda}{\lambda'}\frac{1 - e_{N,naive}}{1 - e_{N,immune}}$$

when $\lambda = \lambda'$, this ratio converges to Nq.

6.4.2.3 Conclusion

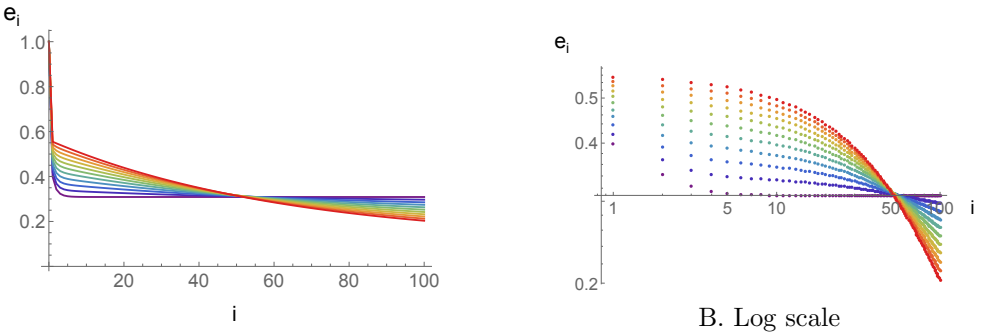
We have shown in two different regimes that when the bacterial strain cannot spread in the absence of resistant mutations, then clustering decreases the probability of spread of this strain and thus of resistance emergence, by a factor which depend on the details of the infection, but will typically be of several fold difference.

[1] without mutations all transmissions are of sensitive bacteria, which are wiped out when the recipient host is treated

6.4.3 But this effect can be countered by silent carrier effect

It is possible that if an immune host feels less sick, it may have more contacts, and thus $\lambda' > \lambda$; and it could be less likely to be treated, *i.e.* $q' < q$.

Let us first look at the results in the absence of mutations, comparing a fully immune with a fully naive population. When $\lambda = \lambda'$ and $q = q'$, e_0 and e_N have the same values for a naive and an immune population (and $e_0 > e_N$ if $q > 0$). If $\lambda' > \lambda$, then $e_{N,immune} < e_{N,naive}$ (see equations (6.11) and (6.9)). Also, if $q' \leq q$, then $\lambda'(1-q') > \lambda(1-q)$, leading to $e_{0,immune} < e_{0,naive}$ (see equations (6.10) and (6.8)). Numerically, we can see (figure 6.4) that despite immune population leading to more spread when the first infected host was infected with bacteria of only one type, for a wide range of parameters, the reverse is true when the first host was infected with a mix of different types of bacteria. Thus this is a case where there is a trade-off.



A. Linear scale

B. Log scale

Figure 6.4 – Probability of extinction e_i in function of the initial number of resistant bacteria i, for mixed populations of proportions of immune individuals $w = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$ (from purple to red). Numerical resolution for the whole system of equations 6.3.3.1 in the absence of mutations: the proportion of resistant bacteria within a host remains identical to the proportion with which it was inoculated, and thus, in immune hosts, only clusters of one type will be found. Other parameters: $N = 100$, $\lambda = 1.7$, $\lambda' = 2$, $q = q' = 0.55$.

Then let us look at the case with mutations, starting from an host infected with sensitive bacteria only. When both $\lambda(1-q) > 1$ and $\lambda'(1-q') > 1$, both types of bacteria may survive in the absence of mutations, so mutations will have little impact on the survival probability and we will be back to the case of the previous paragraph. When $\lambda(1-q) < 1$ and $\lambda'(1-q') > 1$, then immunity will make a large difference in survival. When all hosts are naive, the bacteria cannot spread in the absence of mutations. When there are enough immune hosts, the bacteria may spread to many individuals, mutations will eventually occur and lead faster to widespread resistance. The tipping point will be for w such as $(1-w)\lambda(1-q) + w\lambda'(1-q') = 1$ because 1 is the threshold of possible epidemic propagation. When both $\lambda(1-q) < 1$ and $\lambda'(1-q') < 1$, mutations are always needed for the epidemic to spread. We could study the details as in section 6.4.2.

6.5 Discussion

In this chapter, we have studied the consequences of enchained growth at the scale of a host population: in immune individuals, bacteria grow in clusters of enchained daughter-bacteria. In the context of the evolution of resistance, we propose a model with within-host deterministic dynamics, and compute the proportion of resistant bacteria in the host at the end of its infection, when it transmits bacteria to other recipient hosts, in function of the initial number of resistant bacteria it was seeded with. The transmission is stochastic and described in the framework of branching processes. We first study the simple case with no mutation allowed, and find that if the host is originally seeded with a mixed inoculum, the extinction is more likely in an all-immune population than in a all-naive population. Then we introduce mutations, and a fitness cost for resistance, and focus on the emergence of infection starting from a individual seeded with only sensitive bacteria. In this case, there are different regimes: in the immune individuals, when the number of generations is small, clusters are mostly of mixed composition. On the contrary, when the number of generation is large, then most clusters transmitted are of one type only, and within the host, the proportion of resistant bacteria reaches the mutation-selection balance. In both regimes, in the limit where sensitive bacteria cannot successfully seed an infection in the absence of resistant mutations, the probability of spreading the infection is reduced several folds in an immune population compared to a naive one. However, when taking into account other differences that might be linked to the status "immune" or "naive" of the host, and in particular, the fact that immune individuals may be silent carriers because they feel less sick, this effect can be dampened or even reversed.

In most cases though, we observe a reduction in the emergence probability when the host population is immune. The main reason for that is that, as in the case of immune hosts the transmissions are of bacteria of the same type, then, with the same average number of transmitted resistant bacteria, the proportion of transmissions with at least one resistant bacteria transmitted is lower for immune donor hosts. This system thus presents some similarities with the concept of bet-hedging, which could be summarized in the saying "don't put all your eggs in one basket"[96]. Indeed, microbial populations are typically very large, and experience high replication rates, even though the spread of mutants (like antibiotic-resistant ones) is less frequent. One reason is that upon transmission, not all the diversity of the bacterial population is transmitted, since it goes through important bottlenecks. Thus the details of the way bacteria are transmitted between the hosts really matter to understand properly the behavior at the scale of the population.

This chapter presents a work in progress that will be elaborated in the future. For example, we considered here that for each transmission, transmitted bacteria are chosen from the donor host using average probabilities, without correlation between two transmissions from the same host. In reality, there are hosts with similar starting bacterial composition, and different end composition, for instance

when a mutation occurred very early vs. no mutation occurred at all. Thus there would be correlations between two recipients of the same donor host. We want to check with stochastic simulations that these effects are negligible, and that the assumptions we made are valid. Then we could also study the case where a first mutation confers resistance but with a decreased fitness, and a second one restoring it. Integrating more realistic within-host dynamics and especially a more detailed description of the action of the immune system is essential to broaden our understanding of the interactions between antibiotics use and the spread of resistance in a population.

Conclusion

In this thesis, I presented a first part of my work focusing on the colonization dynamics of bacterial populations in early infections of the gut. I developed stochastic models of population dynamics in an open system, using analytical methods such as branching processes, and numerical methods such as agent-based Gillespie simulations. The aim is to infer biologically relevant parameters of the infection (such as replication and elimination rates, and the probability for one bacteria to settle in the organism and participate in the infection) from indirect data (dilution of plasmids that do not replicate inside the mice, initial and final numbers of bacteria, as well as initial and final distributions of genetic tags). First, I studied one-population models initialized with a Poissonian draw and following a continuous-time birth-death Markovian process. In this framework, I looked for the optimal observable to characterize the variability in the distribution of genetic tags, which have the particularity of starting from unequal population sizes, and showed that the renormalized variance on the growth factor was an adequate and suitable measure of variability. I checked for consistency between the parameter estimates based on the observables of the mean growth rate, the renormalized variance over the growth rate and the proportion of genetic tags lost, and showed that in some cases, it is not clear whether these estimates are truly coherent. Based on biological arguments and the qualitative idea that it could lead to broader possibilities of observables combination, and in particular, a higher variance, I then developed models with two subpopulations following the same kind of dynamics, but with different replication rates. I showed that this kind of model explains very well some experiments, but not all of them, and due to the small quantity of data no clear conclusion can be drawn as to the coexistence of several subpopulations.

The second part of my thesis concerns the mechanisms that make the immune response effective. I first reviewed the study from Moor *et al.* [19] to which I contributed. It shows that sIgA, a specific kind of antibody and main effector of the immune system in the gut produced in high quantities after vaccination, enchains daughter bacteria in clonal clusters upon replication, while this aggregation process was previously thought to happen through the random encounters of bacteria in the gut, which are actually rare at typical initial concentrations after food poisoning. This mechanism called *enchained growth* suffices to protect the mice by preventing bacteria to approach the epithelium and cross it to colonize the rest of the organism. I contributed to evidence the phenomenon with a model predicting the reduction of diversity in the bacterial population it causes.

I then investigated the consequences of this process. At the scale of the host, I studied the interplay of cluster growth and fragmentation in immune individuals, with models based on differential equations, studied by a mixed approach of analytical approximations and numerical solving. I showed that enchained growth likely preferentially targets fast-growing bacteria – which are the most susceptible to disrupt the gut flora equilibrium – and that this process could thus be a way for the immune system to regulate the microbiota composition. At the scale of the hosts population, in the context of evolution of antibiotic resistance, if bacteria are transmitted via clonal clusters (being either completely resistant or completely sensitive) rather than via random collections of resistant and sensitive bacteria, the probability that resistant bacteria are transmitted to a specific individual is decreased. With cross-scale models using branching process, I quantify how the probability of infection emergence is modified in immune population compared to naive ones.

Both parts of this thesis were motivated by the study of quantitative data on *Salmonella* infection in the mice gut. The immune mechanisms we propose here are of very general scope: the mice immune system is indeed close to the one of many other vertebrates (including humans), and enchained growth is not a process limited to *Salmonella* but has already been evidenced in *E. coli* [19] for example. Of course, complementary experiments should be carried out to draw more general conclusions on the actual impact of this phenomenon, but in the long run, enchained growth may be harnessed, through oral vaccination, to reduce the use of antibiotics and thus decelerate the evolution and spread of antimicrobial resistance. Besides, the results presented here go beyond the scope of mere data interpretation. In the first part, the population dynamics tools developed could be applied to larger sets of data concerning bacterial infection of the gut with various strains in various animals, but could also easily be translated to different systems in ecology, not necessarily at the same scale. Then in the second part, the study of clusters growth and fragmentation is a more general statistical physics problem which had already proved to be useful in other contexts and at other scales (for example to the study of a specific kind of algae, see [21], or to explore reproduction modes [22]). We thereby use statistical physics to identify some generic mechanisms, and the core properties needed to understand the range of situations where these mechanisms may be of importance.

Appendix

Appendix A

Experimental data tables

These are the data used in the part I. As it was a simple experiment (no vaccination, etc.), we used it as a test for our general methods.

Strain	Dilution from 10^7	final number	initial number	P_{AmpR}
SB300	10	3200000000	42880000	3.0E-003
SB300	100	3840000000	4288000	2.2E-002
SB300	1000	2400000000	428800	7.7E-005
SB300	10000	2560000000	42880	6.3E-006
SB300	100000	3360000000	4288	1.1E-006
SB300	1000000	1920000000	428.8	2.1E-007
SB300	10000000	1440000000	42.88	1.8E-008
SB300	100000000	800000000	4.288	2.5E-009
M2702	10	1280000000	40640000	2.5E-003
M2702	100	1760000000	4064000	7.3E-004
M2702	1000	1600000000	406400	1.2E-004
M2702	10000	1760000000	40640	4.1E-005
M2702	100000	1600000000	4064	2.0E-006
M2702	1000000	1440000000	406.4	6.7E-007
M2702	10000000	480000000	40.64	7.4E-008
M2702	100000000	1120000000	4.064	1.4E-009

Table A.1 – **Data for the plasmid dilution standard curve:** for each strain, several dilutions were prepared. The initial and final numbers of bacteria were estimated with plating and CFU count (see section 1.1.1), and the final proportions of ampicilin-resistant bacteria were measured with plating following antibiotic addition. In section 1.1.2 we take x_i as the ratio of the final number over the initial number, and y_i as the final proportion of Ampicilin-resistant bacteria (last column).

Strain	Inoculum	CFU count	n_1	n_2	n_3	n_4	n_5	n_6	n_7
SB300	10^3	1472	8.72	9.97	3.82	2.85	3.68	3.64	5.31
SB300	10^5	281600	13.32	15.22	5.84	4.36	5.62	5.55	8.10
SB300	10^7	42880000	16.53	18.90	7.25	5.41	6.97	6.89	10.05
M2702	10^3	4400	17.61	23.80	8.61	5.66	13.92	36.62	5.78
M2702	10^5	220800	1069	1444	522	343	845	2223	350
M2702	10^7	40640000	9.75	13.17	4.76	3.14	7.71	20.27	3.20

Table A.2 – **Inoculum data:** There are six different experiments (with two different strains and three possible inoculum sizes). The inoculum have been plated and the CFU load was counted as explained in 1.1.1 to get a more precise estimate. The initial WITS numbers $\{n_1, ...n_7\}$ were counted as explained in 1.1.3 (plating for the total number + q-PCR for the proportions) and correspond to an estimate of the numbers given to each of the 3 mice per experiment. WITS were given in too high numbers for the "M2702-10^5" experiment, consequently the WITS data on this experiment could not be used.

St.	I	total	m_1	m_2	m_3	m_4	m_5	m_6	m_7	P_{AmpR}
S	10^3	101184000	65.98	172561	1.60	1.86	5.05	21.82	769742	4.9E-09
S	10^3	56940800	321.17	6.95E6	5.79	14.47	20.25	114.29	100272	1.0E-06
S	10^3	68428800	2.28	861651	0.18	0.11	0.53	4.91	259813	1.7E-07
S	10^5	67328000	0	0	0	0	0	0	0	7.4E-09
S	10^5	73920000	0	0	0	0	0	0	0	6.8E-09
S	10^5	69888000	55638	2289	4E-03	0.02	2968	45751	874.26	4.8E-05
S	10^7	89523200	6E-05	18.50	2E-06	3E-05	8E-06	4E-04	3.02	2.4E-03
S	10^7	97216000	1E-04	4E-05	9E-06	37.20	4E-05	6E-04	3E-05	0.3E-02
S	10^7	42649600	0	0	0	0	0	0	0	5.1E-04
M	10^3	644160000	17.33	2.36	0.39	0.79	0.79	863737	1.77	7.1E-07
M	10^3	1380192000	0.0987	0.029	0.0058	0.0174	0.0058	10456	0.0348	4E-08
M	10^3	1431091200	118.95	1.18E7	14.42	54.07	14.42	4938	115.35	1.3E-08
M	10^5	1099929600	2.4E6	1.0E6	4.5E5	4.2E6	6.3E5	86129	393992	2.6E-05
M	10^5	1107302400	2.97E6	2.1E6	1.2E6	1.6E6	1.8E6	999479	2.5E6	8.5E-05
M	10^5	820838400	2.4E6	762835	601341	707152	2.6E6	1.0E6	1.7E6	5.9E-05
M	10^7	2005785600	5066	3391	2764	1279	2720	844.45	2282	1.9E-02
M	10^7	772608000	15.76	20.70	18.53	4E-04	5E-04	0.477	85.37	2.2E-04
M	10^7	3374771200	3200	0.0104	0.0057	0.0057	0.0062	2664	0.0073	1.3E-03

Table A.3 – **Final data for the experiments:** Data measured 24h post-infection, after mice euthanasia. The same experiment is repeated in three different mice. The strain names have been abbreviated ("S" for "SB300" and "M" for "M2702"), and "I" stands for the inoculum size (see the precise count in table A.2). $\{m_1, ...m_7\}$ are the final numbers of WITS (infered from a measure of their relative proportions with q-PCR plus a total count by plating, which explains that we can get numbers below 1, that probably correspond to noise in the q-PCR measurement and for which we define a cutoff), and P_{AmpR} is the final proportion of bacteria carrying a plasmid.

Appendix B

Variance of the variance

B.1 Variance of the simple variance

The expected variance of the variance from section 2.3.2.3 writes as follows:

$$\langle var_{var} \rangle = \left\langle \left(\frac{1}{h-1} \sum_{i=1}^{h} \left(m_i - \frac{1}{h} \sum_{j=1}^{h} m_j \right)^2 - \left(\langle m^2 \rangle - \langle m \rangle^2 \right) \right)^2 \right\rangle$$

To perform this calculation, one has to keep in mind that all WITS have the same distribution and are independent. Thus $\langle m_i \rangle = \langle m_j \rangle = \langle m \rangle$ for all i, j, and $\langle m_i m_j \rangle = \langle m_i \rangle \langle m_j \rangle = \langle m \rangle^2$ for $i \neq j$, but $\langle m_i m_i \rangle = \langle m^2 \rangle$. Developing the outermost square in the previous expression, one gets:

$$\begin{aligned} \langle var_{var} \rangle = & \frac{1}{(h-1)^2} \underbrace{\left\langle \left(\sum_{i=1}^{h} \left(m_i - \frac{1}{h} \sum_{j=1}^{h} m_j \right)^2 \right)^2 \right\rangle}_{\mathbf{A}} \\ & - \frac{2}{h-1} \left(\langle m^2 \rangle - \langle m \rangle^2 \right) \underbrace{\left\langle \sum_{i=1}^{h} \left(m_i - \frac{1}{h} \sum_{j=1}^{h} m_j \right)^2 \right\rangle}_{\mathbf{B}} \\ & + \left(\langle m^2 \rangle - \langle m \rangle^2 \right)^2 \end{aligned} \tag{B.1}$$

Let us first calculate the term B:

$$\begin{aligned}
B =& \left\langle \sum_{i=1}^{h} \left(m_i \frac{h-1}{h} - \frac{1}{h} \sum_{\substack{j=1 \\ j\neq i}}^{h} m_j \right)^2 \right\rangle \\
=& \left\langle \sum_{i=1}^{h} \left(m_i^2 \left(\frac{h-1}{h} \right)^2 + \frac{1}{h^2} \left(\sum_{\substack{j=1 \\ j\neq i}}^{h} m_j \right)^2 - 2\frac{h-1}{h^2} m_i \sum_{\substack{j=1 \\ j\neq i}}^{h} m_j \right) \right\rangle \\
=& \frac{(h-1)^2}{h} \langle m^2 \rangle + \frac{1}{h} \left\langle \left(\sum_{j=1}^{h-1} m_j \right)^2 \right\rangle - 2\frac{(h-1)^2}{h} \langle m \rangle^2
\end{aligned}$$

Now,

$$\left\langle \left(\sum_{j=1}^{h-1} m_j \right)^2 \right\rangle = \left\langle \sum_{j=1}^{h-1} m_j^2 + \sum_{\substack{j,k \\ j\neq k}} m_j m_k \right\rangle = (h-1)\langle m^2 \rangle + (h-1)(h-2)\langle m \rangle^2$$

thus

$$B = (h-1)\left(\langle m^2 \rangle - \langle m \rangle^2 \right)$$

Now for A:

$$\begin{aligned}
A =& \left\langle \sum_{i=1}^{h} \left(m_i - \frac{1}{h} \sum_{j=1}^{h} m_j \right)^4 + \sum_{\substack{k,l \\ k\neq l}} \left(m_k - \frac{1}{h} \sum_{j=1}^{h} m_j \right)^2 \left(m_l - \frac{1}{h} \sum_{j=1}^{h} m_j \right)^2 \right\rangle \\
=& h \left\langle \left(m_i \frac{h-1}{h} - \frac{1}{h} \sum_{\substack{j=1 \\ j\neq i}}^{h} m_j \right)^4 \right\rangle \\
&+ h(h-1) \left\langle \left(m_k \frac{h-1}{h} - \frac{1}{h} \sum_{\substack{j=1 \\ j\neq k}}^{h} m_j \right)^2 \left(m_l \frac{h-1}{h} - \frac{1}{h} \sum_{\substack{j=1 \\ j\neq l}}^{h} m_j \right)^2 \right\rangle \\
=& \frac{1}{h^3} \left\langle \left((h-1)m_i - SMI \right)^4 \right\rangle \\
&+ \frac{(h-1)}{h^3} \left\langle \left((h-1)m_k - m_l - SMJK \right)^2 \left((h-1)m_l - m_k - SMJK \right)^2 \right\rangle
\end{aligned}$$

with $SMI = \frac{1}{h}\sum_{\substack{j=1\\j\neq i}}^{h} m_j$ and $SMJK = \sum_{\substack{j=1\\j\neq k\\j\neq l}}^{h} m_j$. We calculate

$$\begin{cases} \langle SMI\rangle = \left\langle \sum_{\substack{j=1\\j\neq i}}^{h} m_j \right\rangle = (h-1)\langle m\rangle \\ \\ \langle SMI^2\rangle = \left\langle \sum_{\substack{j=1\\j\neq i}}^{h} m_j^2 + \sum_{\substack{j,a\\j\neq a}} m_j m_a \right\rangle = (h-1)\langle m^2\rangle + (h-1)(h-2)\langle m\rangle^2 \\ \\ \langle SMI^3\rangle = \left\langle \left(\sum_{\substack{j=1\\j\neq i}}^{h} m_j \right)^3 \right\rangle \\ \qquad = (h-1)\langle m^3\rangle + (h-1)(h-2)(h-3)\langle m\rangle^3 + 3(h-1)(h-2)\langle m^2\rangle\langle m\rangle \\ \\ \langle SMI^4\rangle = \left\langle \left(\sum_{\substack{j=1\\j\neq i}}^{h} m_j \right)^4 \right\rangle \\ \qquad =(h-1)(h-2)(h-3)(h-4)\langle m\rangle^4 + 6(h-1)(h-2)(h-3)\langle m^2\rangle\langle m\rangle^2 \\ \qquad\quad + (h-1)(h-2)\left[3\langle m^2\rangle^2 + 4\langle m^3\rangle\langle m\rangle\right] + (h-1)\langle m^4\rangle \end{cases}$$

Because in SMI^3, among the $(h-1)^3$ terms of the form $m_a m_b m_c$, there are $(h-1)$ terms where all the m are identical, $(h-2)(h-2)(h-3)$ where they are all different, and $3(h-1)(h-2)$ with two different indexes (there are $(h-1)(h-2)$ pairs of distinct indexes, and then $\binom{3}{1} = 3$ ways of attributing this pair of index to three elements). For the power four it works the same except that the count is slightly more complicated. And exactly the same for $SMJK$ and the higher powers (the sums of the powers of m_i where i is different from j and k), except with one term less in the sums, so one just needs to replace all the h with $h-1$.

Replacing everything in equation B.1, one finally gets the variance of the variance in function of the different moments of the final WITS numbers distribution:

$$\begin{aligned} var_{var} =& \frac{1}{h(h-1)}\left((h-1)\langle m^4\rangle - 2(2h-3)\langle m\rangle^4 - (h-3)\langle m^2\rangle^2\right. \\ & \left. -4(h-1)\langle m\rangle\langle m^3\rangle + 4(2h-3)\langle m\rangle^2\langle m^2\rangle\right) \end{aligned} \tag{B.2}$$

With eq. (1.8), it ultimately leads to:

$$\begin{aligned} \langle var_{var}\rangle = & \left(e^{(r-c)t} n_0\beta\right)^4 \frac{2r-(r+c)e^{-(r-c)t}}{h(h-1)(r-c)^2(n_0\beta)^2} \times \\ & \left(\frac{(h-1)(c^2+10cr+r^2)e^{-2(r-c)t}}{(r-c)\beta n_0} - \frac{12(h-1)r(c+r)e^{-(r-c)t}}{(r-c)\beta n_0} \right. \\ & \left. -2h(r+c)e^{-(r-c)t} + 4hr + \frac{12r^2(h-1)}{n_0\beta(r-c)} \right) \end{aligned}$$

B.2 Variance of the variance on the growth factor

The only difference from the simple variance case is that if the mean value $\left\langle \frac{m_i}{n_i} \right\rangle$ is the same for all i, it is not the case of the higher moments $\left\langle \left(\frac{m_i}{n_i} \right)^p \right\rangle$ (cf. eq. (2.11)) which depends on n_i. Thus the aim is to express everything in terms of sums of the moments. Let us note:

$$\begin{cases} M = \left\langle \frac{m_i}{n_i} \right\rangle = M \\ S[2] = \sum_{i=1}^{k} \left\langle \left(\frac{m_i}{n_i} \right)^2 \right\rangle \\ S[2^2] = \sum_{i=1}^{k} \left\langle \left(\frac{m_i}{n_i} \right)^2 \right\rangle^2 \\ S[3] = \sum_{i=1}^{k} \left\langle \left(\frac{m_i}{n_i} \right)^3 \right\rangle \\ S[4] = \sum_{i=1}^{k} \left\langle \left(\frac{m_i}{n_i} \right)^4 \right\rangle \end{cases}$$

And another identity which will prove extremely useful:

$$S[2]^2 - S[2^2] = \sum_{q=1}^{h} \sum_{\substack{l=1 \\ l \neq q}}^{h} \left\langle \left(\frac{m_l}{n_l} \right)^2 \right\rangle \left\langle \left(\frac{m_q}{n_q} \right)^2 \right\rangle$$

Writing directly that the variance is the mean of the squares minus the square of the mean:

$$\langle var_{var} \rangle = \frac{1}{(h-1)^2} \underbrace{\left\langle \left(\sum_{i=1}^{h} \left(\frac{m_i}{n_i} - \frac{1}{h} \sum_{j=1}^{h} \frac{m_j}{n_j} \right)^2 \right)^2 \right\rangle}_{\mathbf{A}} - \langle var \rangle^2$$

Developing the square in A and separating the terms in i and different from i:

$$A = \underbrace{\left\langle \sum_{i=1}^{h} \left(\frac{h-1}{h} \frac{m_i}{n_i} - \frac{1}{h} \sum_{\substack{j=1 \\ j \neq i}}^{h} \frac{m_j}{n_j} \right)^4 \right\rangle}_{\mathbf{B}} + \underbrace{\sum_{q=1}^{h} \sum_{\substack{l=1 \\ l \neq q}}^{h} \left\langle \left(\frac{h-1}{h} \frac{m_q}{n_q} - \frac{1}{h} \sum_{\substack{j=1 \\ j \neq q}}^{h} \frac{m_j}{n_j} \right)^2 \left(\frac{h-1}{h} \frac{m_l}{n_l} - \frac{1}{h} \sum_{\substack{j=1 \\ j \neq l}}^{h} \frac{m_j}{n_j} \right)^2 \right\rangle}_{\mathbf{C}}$$

Let us deal with B first. As for what was done for the simple variance, one must correctly count the terms when developing $(a-b)^4$ where there are zero, one, two, three or four a. The two first terms are easy, but the higher powers of the sum will be more delicate to deal with:

$$\begin{aligned} B = & \left(\frac{k-1}{k}\right)^4 S[4] - 4\left(\frac{k-1}{k}\right)^4 S[3] \\ & + 6\left(\frac{h-1}{h}\right)^2 \frac{1}{h^2}\sum_{i=1}^{h}\left\langle\left(\frac{m_i}{n_i}\right)^2\right\rangle\left\langle\left(\sum_{\substack{j=1\\j\neq i}}^{h}\frac{m_j}{n_j}\right)^2\right\rangle \Bigg\}\mathbf{B_3} \\ & - 4\left(\frac{h-1}{h}\right)\frac{1}{h^3}\sum_{i=1}^{h}\left\langle\frac{m_i}{n_i}\right\rangle\left\langle\left(\sum_{\substack{j=1\\j\neq i}}^{h}\frac{m_j}{n_j}\right)^3\right\rangle \Bigg\}\mathbf{B_4} \\ & + \frac{1}{h^4}\sum_{i=1}^{h}\left\langle\left(\sum_{\substack{j=1\\j\neq i}}^{h}\frac{m_j}{n_j}\right)^4\right\rangle \Bigg\}\mathbf{B_5} \end{aligned}$$

Then, as always, separating the terms where all the coefficient are two by two different from the others, one gets:

$$\begin{cases} B_3 = \frac{6(h-1)^2}{h^3}\left[(S[2])^2 - S[2^2] + (h-1)(h-2)M^2 S[2]\right] \\ B_4 = -\frac{4(h-1)}{h^4}M\left[(h-1)S[3] + h(h-1)(h-2)(h-3)M^3 + 3M(h-2)(h-1)S[2]\right] \\ B_5 = \frac{h-1}{h^4}S[4] + \frac{4(h-2)(h-1)}{h^4}MS[3] + \frac{3}{h^4}(h-2)\left(S[2]^2 - S[2^2]\right) \\ \qquad + \frac{6}{h^4}(h-3)(h-2)(h-1)M^2 S[2] + \frac{h}{h^4}(h-1)(h-2)(h-3)(h-4)M^4 \end{cases}$$

Then going back to C:

$$\begin{aligned} C = & \sum_{q=1}^{h}\sum_{\substack{l=1\\l\neq q}}^{h}\left\langle\left(\left(\frac{m_q}{n_q}\right)^2 - \frac{2}{h}\frac{m_q}{n_q}\sum_{j=1}^{h}\frac{m_j}{n_j} + \frac{1}{h^2}\left(\sum_{j=1}^{h}\frac{m_j}{n_j}\right)^2\right)\cdot\right. \\ & \left.\left(\left(\frac{m_l}{n_l}\right)^2 - \frac{2}{h}\frac{m_l}{n_l}\sum_{j=1}^{h}\frac{m_j}{n_j} + \frac{1}{h^2}\left(\sum_{j=1}^{h}\frac{m_j}{n_j}\right)^2\right)\right\rangle \end{aligned}$$

$$
\begin{aligned}
C = \sum_{q=1}^{h}\sum_{\substack{l=1\\ l\neq q}}^{h}\Bigg\langle & \underbrace{\left(\frac{m_l}{n_l}\right)^2\left(\frac{m_q}{n_q}\right)^2}_{C_1} - \underbrace{\frac{2}{h}\left(\frac{m_l}{n_l}\right)^2\left(\frac{m_q}{n_q}\right)\sum_{j=1}^{h}\frac{m_j}{n_j}}_{C_2} - \underbrace{\frac{2}{h}\left(\frac{m_l}{n_l}\right)\left(\frac{m_q}{n_q}\right)^2\sum_{j=1}^{h}\frac{m_j}{n_j}}_{C_3} \\
& + \underbrace{\frac{1}{h^2}\left(\frac{m_l}{n_l}\right)^2\left(\sum_{j=1}^{h}\frac{m_j}{n_j}\right)^2}_{C_4} + \underbrace{\frac{4}{h^2}\left(\frac{m_l}{n_l}\right)\left(\frac{m_q}{n_q}\right)\left(\sum_{j=1}^{h}\frac{m_j}{n_j}\right)^2}_{C_5} + \underbrace{\frac{1}{h^2}\left(\frac{m_q}{n_q}\right)^2\left(\sum_{j=1}^{h}\frac{m_j}{n_j}\right)^2}_{C_6} \\
& - \underbrace{\frac{2}{h^3}\left(\frac{m_l}{n_l}\right)\left(\sum_{j=1}^{h}\frac{m_j}{n_j}\right)^3}_{C_7} - \underbrace{\frac{2}{h^3}\left(\frac{m_q}{n_q}\right)\left(\sum_{j=1}^{h}\frac{m_j}{n_j}\right)^3}_{C_8} + \underbrace{\frac{1}{h^4}\left(\sum_{j=1}^{h}\frac{m_j}{n_j}\right)^4}_{C_9}\Bigg\rangle
\end{aligned}
$$

Now, by symmetry, $\sum_{q,l} C_2 = \sum_{q,l} C_3$, $\sum_{q,l} C_4 = \sum_{q,l} C_6$, $\sum_{q,l} C_7 = \sum_{q,l} C_8$. Then, still following the rules of terms separation:

$$
\begin{cases}
\sum_{\substack{l,q\\ l\neq q}} C_1 = S[2]^2 - S[2^2] \\
\sum_{\substack{l,q\\ l\neq q}} C_2 = -\frac{2}{h}\left[(h-2)(h-1)M^2S[2] + (h-1)MS[3] + S[2]^2 - S[2^2]\right] \\
\sum_{\substack{l,q\\ l\neq q}} C_4 = \frac{h-1}{h^2}\left[S[2]^2 - S[2^2] + S[4] + (h-1)(h-2)M^2S[2] + 2(h-1)MS[3]\right] \\
\sum_{\substack{l,q\\ l\neq q}} C_5 = \frac{4}{h^2}\left[5(h-2)(h-1)M^2S[2] + 2(h-1)MS[3] + 2(S[2]^2 - S[2^2])\right. \\
\qquad\qquad \left. + h(h-1)(h-2)(h-3)M^4\right] \\
\sum_{\substack{l,q\\ l\neq q}} C_7 = -\frac{2}{h^3}(h-1)\left[S[4] + 4(h-1)MS[3] + 6(h-2)(h-1)M^2S[2]\right. \\
\qquad\qquad \left. + 3(S[2]^2 - S[2^2]) + h(h-1)(h-2)(h-3)M^4\right] \\
\sum_{\substack{l,q\\ l\neq q}} C_9 = \frac{h-1}{h}\left[S[4] + 4M(h-1)S[3] + 3(S[2]^2 - S[2^2])\right. \\
\qquad\qquad \left. + 6M^2(h-2)(h-1)S[2] + h(h-1)(h-2)(h-3)M^4\right]
\end{cases}
$$

Finally summing all the terms of A and substracting the expression of the mean variance square, one finds:

$$
\begin{aligned}
\langle var_{var}\rangle = & \frac{4(2h-3)}{(h-1)h^2}M^2S[2] - \frac{4}{h^2}MS[3] + \frac{2}{(h-1)^2h^2}S[2]^2 \\
& - \frac{\left(h^2-2h+3\right)}{(h-1)^2h^2}S[2^2] + \frac{S[4]}{h^2} + \frac{2(3-2h)}{(h-1)h}M^4
\end{aligned}
$$

Notice that if one replaces all the n_i by n_0 one finds the same result as what was found for the variance over the simple variance in B.2, devided by n_0^4.

And finally, using the generating function, always paying attention to the fact that each population has now its own with the correct n_i:

$$\begin{aligned}\langle var_{var}\rangle = \left(\beta^4 e^{4(r-c)t}\right) \Bigg[& \frac{2\left((c+r)e^{-(r-c)t}-2r\right)^2}{\beta^2(h-1)^2h^2(r-c)^2} SN^2 \\ & + \frac{2(h-2)\left((c+r)e^{-(r-c)t}-2r\right)^2}{\beta^2(h-1)^2h(r-c)^2} SN2 \\ & - \frac{e^{-3rt}}{\beta^3h^2(r-c)^3}\Big\{c^3e^{3ct} + c^2r\left(11e^{3ct}-14e^{t(2c+r)}\right) \\ & + r^3\left(-14e^{t(2c+r)}+36e^{t(c+2r)}+e^{3ct}-24e^{3rt}\right) \\ & + cr^2\left(-44e^{t(2c+r)}+36e^{t(c+2r)}+11e^{3ct}\right)\Big\} SN3\Bigg]\end{aligned}$$

with $SN = \sum_{i=1}^{h} \frac{1}{n_i}$, $SN2 = \sum_{i=1}^{h} \frac{1}{n_i^2}$ and $SN3 = \sum_{i=1}^{h} \frac{1}{n_i^3}$.

Appendix C

Source code of the R simulations

I chose to run my agent-based simulations with the R-language, notably because it was a language used by our collaborators.

```
#####################################################################
#                          2 SUB-POPULATIONS MODEL                  #
#####################################################################


######################### VARIANCE FUNCTIONS ########################

####### For the results of simulations

newnew_variance <- function(final_numbers, initial_numbers)
{
  k <- length(final_numbers)
  if(sum(final_numbers) > 0)
  {
    (1/(k-1))*sum((final_numbers/initial_numbers-(1/k)*sum(final_
    numbers/initial_numbers))^2)
  }
  else 0 #If all the WITS populations got exctinct then there is no
    spread of the values
}


####### For the results of the experiment (normalization over one
    realization)
newnew_variance_norm <- function(final_numbers, initial_numbers)
{
  k <- length(final_numbers)
  if(sum(final_numbers) > 0)
  {
    newnew_variance(final_numbers, initial_numbers)/(((1/k)*sum(final_
    numbers/initial_numbers))^2)
  }
  else 0
}
```

```
mean_pop_size_exp <- function(final_numbers,initial_numbers)
{
  k<- length(final_numbers)
  if(sum(final_numbers) > 0)
  {
    (1/k)*sum(final_numbers/initial_numbers)
  }
  else 0
}

##### For the theoretical values

mean_size <- function(beta,r1,r2,c,t,k,q)
{
  beta*(q*exp((r1-c)*t)+(1-q)*exp((r2-c)*t))
}


newnew_variance_th <- function(beta, r1, r2, c, t, k, nombres_
    initiaux, q)
{
  if(sum(nombres_initiaux)!=0)
  {
    (beta^2)*(q*exp(2*(r1-c)*t)*((2*r1-(r1+c)*exp(-(r1-c)*t))/(beta*(
    r1-c)))+(1-q)*exp(2*(r2-c)*t)*((2*r2-(r2+c)*exp(-(r2-c)*t))/(beta
    *(r2-c))))*sum(1/nombres_initiaux)/k
  }
  else
  {
    0
  }
}

newnew_var_norm_th <- function(beta, r1, r2, c, t, k, nombres_
    initiaux, q)
{
  if(sum(nombres_initiaux)!=0)
  {
    newnew_variance_th(beta, r1, r2, c, t, k, nombres_initiaux, q)/((
    beta^2)*((q*exp((r1-c)*t)+(1-q)*exp((r2-c)*t))^2))
  }
  else
  {
    0
  }
}

#########################################################################
#                          SIMULATION GILLESPIE                         #
#########################################################################

#### Default values for the parameters

params.0 <- c(inoc=5e3, # Total inoculum size
              beta1=0.2,# Settlement probability subpopulation 1
              beta2=0.2,# Settlement probability subpopulation 2
```

```
                r1=16.6,  # Replication rate subpopulation 1
                r2=16.6,  # Replication rate subpopulation 2
                nwitsinoc=rep(1,7), # Initial WITS numbers
                                    # (vector of size n.wits)
                c=0,      # Elimination rate
                n.wits=7, # Number of different WITS
                K=5e9,    # Carrying capacity
                q = 1)    # Initial proportion of subpop 1

simulation <- function(params = params.0, tf.days=1, eps=0.05)
{
  require(adaptivetau)

  ###################### POISSON INITIALISATION ###################

  # INOCULUM B0

  # Two subpopulations for each WITS + 2 unlabelled followed:
  B0 <- vector("numeric",2*params["n.wits"]+2)

  # Poisson selection of unlabeled bacteria of type 1:
  B0[1] <- rpois(1,params["q"]*params["inoc"]*params["beta1"])

  # Poisson selection of WITS of type 1
  B0[2:(params["n.wits"]+1)] <- sapply(params[paste0("nwitsinoc",seq
    (1,params["n.wits"]))]*params["q"]*params["beta1"],rpois,n=1)

  # Poisson selection of unlabeled bacteria of type 2:
  B0[params["n.wits"]+2] <- rpois(1,(1-params["q"])*params["inoc"]*
    params["beta2"])

  # Poisson selection of WITS of type 2
  B0[(params["n.wits"]+3):(2*params["n.wits"]+2)] <- sapply(params[
    paste0("nwitsinoc",seq(1,params["n.wits"]))]*(1-params["q"])*
    params["beta2"],rpois,n=1)

  # Names: wt for unlabelled, Wij for a bacteria labelled with tag
    number j from the ith subpopulation
  names(B0)<-c("wt1",paste0("W1",1:params["n.wits"]), "wt2", paste0("
    W2",1:params["n.wits"]))


  # If no labelled bacteria passes the Poisson selection, the
    simulation ends here:
  if (sum(B0[c(2:(params["n.wits"]+1),(params["n.wits"]+3):(2*params[
    "n.wits"]+2))]) == 0)
  {
    sim = c(time = 0,B0)
    WITS.last <- sim[paste0("W1",1:params["n.wits"])] + sim[paste0("
    W2",1:params["n.wits"])]
    # (stores the final numbers of WITS)
    WITS.last1 <- sim[paste0("W1",1:params["n.wits"])]
    WITS.last2 <- sim[paste0("W2",1:params["n.wits"])]
    var = 0  # the variance is zero in this case
    moy = 0  # as well as the mean
  }
```

```
129
  # If some WITS have passed the Poissonian barrier,
131 # the simulation continues
  else
133 {
    ##################### TRANSITIONS ####################
135
    # The allowed transitions are for each population type to
137 # gain one element (through replication) or to lose one
    # (through death).
139 # All those transitions are independent from one another

141 transitions <- ssa.maketrans(names(B0),
                              rbind("wt1", +1),
143                           rbind("W11", +1),
                              rbind("W12", +1),
145                           rbind("W13", +1),
                              rbind("W14", +1),
147                           rbind("W15", +1),
                              rbind("W16", +1),
149                           rbind("W17", +1),
                              rbind("wt2", +1),
151                           rbind("W21", +1),
                              rbind("W22", +1),
153                           rbind("W23", +1),
                              rbind("W24", +1),
155                           rbind("W25", +1),
                              rbind("W26", +1),
157                           rbind("W27", +1),
                              rbind("wt1", -1),
159                           rbind("W11", -1),
                              rbind("W12", -1),
161                           rbind("W13", -1),
                              rbind("W14", -1),
163                           rbind("W15", -1),
                              rbind("W16", -1),
165                           rbind("W17", -1),
                              rbind("wt2", -1),
167                           rbind("W21", -1),
                              rbind("W22", -1),
169                           rbind("W23", -1),
                              rbind("W24", -1),
171                           rbind("W25", -1),
                              rbind("W26", -1),
173                           rbind("W27", -1))

175 ##################### RATES ##############################

177 # To each allowed transition a rate must be associated.
    # The rate can depend on the current state of the system (B)
179 # and directly of time

181 rates <- function(B, params, t)
    {
183   # When carrying capacity K is reached, replication stops:
      if(sum(B,na.rm=TRUE) > params["K"]) replication.rates <- rep
```

```
    (0,2*params["n.wits"]+2)

      # When carrying capacity is not reached, the replication
      # rates are proportional to 1-(sum(B)/K) and to the size
      # of the considered population:
      else
      {
        replication.rates <- vector("numeric",2*params["n.wits"]+2);
        replication.rates[1:(params["n.wits"]+1)] <- params["r1"] *
    (1-(sum(B)/params["K"])) * B[1:(params["n.wits"]+1)];
        replication.rates[(params["n.wits"]+2):(2*params["n.wits"]+2)
    ] <- params["r2"] * (1-(sum(B)/params["K"])) * B[(params["n.wits"
    ]+2):(2*params["n.wits"]+2)];
      }
      # The death rates are fixed
      # (proportional to the population size)
      killing.rates <- params["c"] * B
      return(c(replication.rates,killing.rates))
    }

    # Adaptivetau or exact SSA:
    if(eps>0)
    {
      sim <- ssa.adaptivetau(B0, transitions, rates, params, tf=tf.
    days,tl.params = list(epsilon=eps));
    }

    if(eps==0)
    {
      sim <- ssa.exact(B0, transitions, rates, params, tf=tf.days);
    }

    # Final variance and mean storage
    WITS.last <- sim[dim(sim)[1],paste0("W1",1:7)] + sim[dim(sim)[1],
    paste0("W2",1:7)]
    if(sum(WITS.last)>0)
    {
      var <- newnew_variance(WITS.last,params[paste0("nwitsinoc",seq
    (1,params["n.wits"]))])
      moy <- mean(WITS.last/params[paste0("nwitsinoc",seq(1,params["n
    .wits"]))])
    }
    else
    {
      var = 0
      moy = 0
    }

    WITS.last1 <- sim[dim(sim)[1],paste0("W1",1:7)]
    WITS.last2 <- sim[dim(sim)[1],paste0("W2",1:7)]
  } # Ends the part of the simulation to do only if not all
  # the WITS are lost at the first step

  WITS.loss = sum(WITS.last==0)
  # Output of the "simulation" function:
  list(sim=sim,final.variance=var,final.moy=moy,WITS.last = WITS.last
```

```
    ,WITS.loss = WITS.loss)
}
#########################################################################
#                                DATA TREATMENT                         #
#########################################################################

############################ DATA IMPORT ############################

# data on plasmid dilution in vitro:
datast=read.csv("ES15-010-standard.csv",header=TRUE)
# data on initial and final WITS counts:
dataw=read.csv("ES15-010-wits.csv",header=TRUE)
# data on initial numbers in the inoculum:
inoc=read.csv("ES15-010-inoc.csv",header=TRUE)

# Extracts the power of ten of the inoculum size from experience name
    (e.g. "SB300_10e5"):
inoc.size=function(x){substr(x,10,10)}

##### MEAN GENERATION NUMBER ESTIMATE FROM PLASMID DILUTION ####

######## Building the standard curve with in vitro data:

# Function giving the log2 of the plasmid dilution in function of the
    number of replication d:
ftheorysimple=function(deltad=0,eps=0,d)
  {min(0,deltad-d*(1-eps))}

# Quadratic error to minimize:
fntominimize.eps=function(para,data)
{
  deltad=para[1]
  eps=para[2]
  datad=data[,1] # log2 of the ratio final/initial numbers
  datap=data[,2] # log2 of the proportion of bacteria still carrying
    a plasmid
  sum((sapply(datad,ftheorysimple,deltad=deltad,eps=eps)-datap)^2)
}
stv=c()
for (i in 1:2)
{
  condi=(datast[,1]==c("SB300","M2702")[i])

  # log2 of the ratio between final and initial numbers:
  tempx=log2(datast[condi,]$CFU.per.ml.total/datast[condi,]$Starting.
    CFU.per.ml.total)
  # log2 of the proportion of bacteria still carrying a plasmid:
  tempy=log2(datast[condi,]$CFU.per.ml.AmpR/datast[condi,]$CFU.per.ml
    .total)

  # non-linear minimization:
  test=nlm(fntominimize.eps,p=c(0,0),data=cbind(tempx,tempy))

  # Stocking the parameters estimate for the different strains:
  stv=rbind(stv,c(c("SB300","M2702")[i],test$estimate))
}
```

```
283
285 ######## Using the standard curve on in vivo plasmid data:
287 # Gives a mean replication rate from the plasmid dilution p:
    ftheorysimplerev=function(deltad=0,eps=0,p)
289 {
      if (p>=1)
291   {
        r=NA
293   }
      else
295   {
        r=-(log2(p)+deltad)/(1-eps)
297   }
      r
299 }

301 recap=c() # Stores data for each experiment
    # Each line corresponds to one experiments and will contain, in this
        order:
303 # salmonella strain type ("SB300"or"M2702")
    # the inoculum size (10^3, 10^5, 10^7), simplified as 3, 5, 7
305 # the mouse ID, or "all" for measures pooled per salmonella and inoc
    # the mean replication rate
307 # variance on the growth rate
    recap=rbind(recap,cbind(as.character(dataw[dataw$day==1,]$salmonella_
        strain.dilution),
309     inoc.size(dataw[dataw$day==1,]$treatment),
        as.character(dataw[dataw$day==1,]$mouse.ID),
311     rep("replication",sum(dataw$day==1)),
        apply(cbind(dataw[dataw$day==1,]$X..AmpR/100,
313     as.numeric(stv[sapply(dataw[dataw$day==1,]$salmonella_strain.
         dilution,function(x){which(stv[,1]==x)}),2]),
        as.numeric(stv[sapply(dataw[dataw$day==1,]$salmonella_strain.
         dilution,function(x){which(stv[,1]==x)}),3])),1,
315     function(x){ftheorysimplerev(p=x[1],deltad=x[2],eps=x[3])})
    ))
317
    ################## CUTOFF ON THE WITS DATA ####################
319
    temp=unlist(dataw[,10:16]) # 7 columns of 7 WITS data
321 temp=temp[temp>0]

323 for (i in 1:dim(dataw)[1]){
      for (j in 10:16){
325     if (dataw[i,j]<1){dataw[i,j]=0}
      }
327 }

329 ################## LOG-LIKELIHOOD MAXIMIZATION ################

331 ###### Derivative of the log-likelihood
    # wits.lost is a boolean vector being TRUE for the lost WITS
333 derivative.log.likelihood.beta.c=function(beta,initial.wits,wits.lost
        ,r1,r2,q,t,inoctot,finaltot,rmean)
```

```
{
  c=(1/t)*log((beta*inoctot*exp(rmean*t))/finaltot)
  tempdll=((q*(r1-c))/(r1-c*exp(-(r1-c)*t))) + (((r2-c)*(1-q))/(r2-c*
    exp(-(r2-c)*t))) #=1 when c=0
  -sum(initial.wits*wits.lost*tempdll)+sum(initial.wits*(1-wits.lost)
    *tempdll*exp(-beta*initial.wits*tempdll)/(1-exp(-beta*initial.
    wits*tempdll)))
}

##### Function giving beta from the root of the derivative of
# the log-likelihood, all the other parameters being known:
loss.function=function(winoc,w,bounds,r1b,r2b,qb,tb,inoctotb,
    finaltotb,rmeanb)
{
  if (sum(w==0)==0)
  {# no loss
    res=1
  }
  else if (sum(w==0)==length(w))
  {# all lost
    res=0
  }
  else
  {
    res=uniroot(derivative.log.likelihood.beta.c,bounds,
                initial.wits=winoc,
                wits.lost=(w==0),r1=r1b,r2=r2b,q=qb,t=tb,inoctot=
    inoctotb,finaltot=finaltotb,rmean=rmeanb,extendInt="yes")$root

  }
  res
}

######## STOCKING EXPERIMENTAL VARIANCE AND MEAN POP SIZE ########

for (i in 1:dim(dataw)[1])
{
  tempw2=dataw[i,10:16]  # final numbers of WITS for exp. i

  # Put initial numbers of WITS in the inoculum in tempn02:
  if (substr(dataw[i,3],1,1)=="S")
  {
    tempn02=dataw[(dataw$day==0)&(dataw$salmonella_strain.dilution=="
    SB300"),10:16]
    tempn02=(tempn02/sum(tempn02))*(dataw[i,]$WITS.per.inoculum)[1]
    tempn02=unlist(tempn02)
  }

  if (substr(dataw[i,3],1,1)=="M")
  {
    tempn02=dataw[(dataw$day==0)&(dataw$salmonella_strain.dilution=="
    M2702"),10:16]
    tempn02=tempn02/sum(tempn02)*(dataw[i,]$WITS.per.inoculum)[1]
    tempn02=unlist(tempn02)
  }
```

```
  tempw2=as.numeric(tempw2)

  recap=rbind(recap,c(as.character(dataw[i,]$salmonella_strain.
    dilution),
      inoc.size(dataw[i,]$treatment),
      as.character(dataw[i,]$mouse.ID),"variance",
      newnew_variance_norm(tempw2,tempn02)))

  recap=rbind(recap,c(as.character(dataw[i,]$salmonella_strain.
    dilution),
      inoc.size(dataw[i,]$treatment),
      as.character(dataw[i,]$mouse.ID),"mean",
      mean_pop_size_exp(tempw2,tempn02)))
}

################ SIMULATIONS AND RESULTS STOCKING ################

tl = 1
niter = 3000
n.witsl=7
rmaxl = 48*log(2) #(30min=1/48 day, the time unit)

for (strain in c("SB300","M2702"))
{
  for (inocs in c(3,5,7))
  {
    # Getting the initial total number
    inocl=inoc[(substr(inoc$Inoculum,1,5)==strain)&(substr(inoc$
    Inoculum,10,10)==inocs),2] #total initial number of bacteria

    # Selecting from the right inoculum size and strain:
    condi=(dataw$salmonella_strain.dilution==strain)&(sapply(dataw$
    treatment,inoc.size)==inocs)
    tempn02=dataw[(dataw$day==0)&(dataw$salmonella_strain.dilution==
    strain),10:16]
    tempn02=tempn02/sum(tempn02)*(dataw[condi,]$WITS.per.inoculum)[1]
    # Initial numbers of WITS:
    nwitsinocl=as.numeric(unlist(tempn02))

    # Average (3 mice/exp.) of the total final number
    nb.final = mean(dataw[(dataw$day==1)&(dataw$salmonella_strain.
    dilution==strain)&(sapply(dataw$treatment,inoc.size)==inocs),9])

    # Average (3 mice/exp.) of the mean replication rate
    rl=log(2)*mean(as.numeric(recap[((recap[,4]=="replication"))&(
    recap[,1]==strain)&(recap[,2]==inocs),5]),na.rm=TRUE)

    # Final numbers of WITS
    condi=(dataw$salmonella_strain.dilution==strain)&(sapply(dataw$
    treatment,inoc.size)==inocs)&(dataw$day==1)
    tempw3=as.numeric(unlist(dataw[condi,10:16]))

    # Proportion of WITS lost
    p.loss = sum(tempw3==0)/length(tempw3)

    # Parameter space to explore:
```

```
431     ql=0.1*(0:10)
        al = 1+0.5*(0:10)  # alpha > 1
433
        # stocking the results
435     z.var.norm = c() # iterated variance
        z.var.th = c()   # theoretical variance
437     stockbeta=c()    # beta estimate
        stock.c=c()      # c estimate
439
        for (ia in (1:length(al)))
441     {

443       for (iq in (1:length(ql)))
          {
445
            froot <- function(x) {ql[iq]*exp(al[ia]*x)+(1-ql[iq])*exp(x)-
      exp(rmaxl)}
447         rmaxl2 = uniroot(froot,c(0,2*rmaxl),tol=0.000000000001)$root
            rmaxl1 = al[ia]*rmaxl2
449
            bl = loss.function(winoc=rep(nwitsinocl,sum(condi)),
451               w=tempw3,bounds = c(0.0001,5),r1b=rmaxl1,
                  r2b=rmaxl2,qb=ql[iq],tb=tl,inoctotb=inocl,
453               finaltotb=nb.final,rmeanb=rl)
            cl = (1/tl)*log((bl*inocl*exp(rl*tl))/nb.final)
455
            stockbeta=c(stockbeta,bl)
457         stock.c=c(stock.c,cl)

459         Kl = nb.final/(1-cl/rmaxl)

461         temp = c()
            temp.var = c()
463         temp.moy = c()

465         for (iiter in 1:niter)
            {
467           temp = simulation(params=c(inoc=inocl,beta1=bl,
                    beta2=bl,r1=rmaxl1,r2=rmaxl2,
469                 nwitsinoc=nwitsinocl,c=cl,n.wits=n.witsl,
                    K=Kl,q=ql[iq]),tf.days=tl,eps=0.003)
471           temp.var = c(temp.var,temp$final.variance)
              temp.moy = c(temp.moy,mean((temp$WITS.last)/
473                                        nwitsinocl))
            }
475
            z.var.norm = c(z.var.norm,mean(temp.var)/
477                           ((mean(temp.moy,na.rm=TRUE))^2))
            z.var.th = c(z.var.th,newnew_var_norm_th(bl, rmaxl1,
479                 rmaxl2, cl, tl, n.witsl, nwitsinocl,ql[iq]))
          } # on q
481     } # on alpha

483     zt.var.norm = matrix(z.var.norm,length(ql),length(al))
        zt.var.th = matrix(z.var.th,length(ql),length(al))
485     zt.stock.beta = matrix(stockbeta,length(ql),length(al))
```

```
    zt.stock.c = matrix(stock.c,length(ql),length(al))
487
    ##################### PLOT #####################
489
    n.couleurs = 50
491
    ### Color scale for the variance
493 min.local = min(log10(zt.var.norm[is.finite(log10(zt.var.norm))])
    ,log10(zt.var.th))
    max.local = max(log10(zt.var.norm[is.finite(log10(zt.var.norm))])
    ,log10(zt.var.th))
495 breaks.local = min.local + (0:n.couleurs)*(max.local-min.local)/(
    n.couleurs)

497 image(min.local+(0:(n.couleurs-1)-0.5)*(max.local-min.local)/(n.
    couleurs-1),
        1,matrix(min.local+(c(1,1:(n.couleurs-1))-0.5)*(max.local-
    min.local)/(n.couleurs-1),n.couleurs,1),
499         col=heat.colors(n.couleurs),breaks=breaks.local,yaxt="n",
    main="scale variance")
    for (val in c(-2,0,2,4,6,8,10,12))
501 {lines(c(val,val),c(0,2))}

503 ### Simulated variance
    image(ql,al,log10(zt.var.norm),xlab="q",ylab="alpha",main=paste0(
    "Simulated variance_",strain," 10^",inocs),
505         col=heat.colors(n.couleurs),breaks=breaks.local)
    titi=as.numeric(recap[(recap[,4]=="variance")&(recap[,1]==strain)
    &(recap[,2]==inocs),5])
507 titi=titi[!is.na(titi)]
    titi[titi==0]=0.0000000001
509 contour(ql,al, log10(zt.var.norm), col = "black", add = TRUE,
    method = "edge",vfont = c("sans serif", "plain"),levels=log10(
    mean(titi)))
    contour(ql,al, log10(zt.var.norm), col = "black", add = TRUE,
    method = "edge",vfont = c("sans serif", "plain"),levels=log10(
    titi),lty=2)
511

    ### Theoretical variance
513 image(ql,al,log10(zt.var.th),xlab="q",ylab="alpha",main=paste0("
    Theoretical variance_",strain," 10^",inocs),
515         col=heat.colors(n.couleurs),breaks=breaks.local)
    titi=as.numeric(recap[(recap[,4]=="variance")&(recap[,1]==strain)
    &(recap[,2]==inocs),5])
    titi=titi[!is.na(titi)]
517 titi[titi==0]=0.0000000001
    contour(ql,al, log10(zt.var.th), col = "black", add = TRUE,
    method = "edge",vfont = c("sans serif", "plain"),levels=log10(
    mean(titi)))
519 contour(ql,al, log10(zt.var.th), col = "black", add = TRUE,
    method = "edge",vfont = c("sans serif", "plain"),levels=log10(
    titi),lty=2)
  } # on inoculum size
521 } # on the strain
```

figures/code_clean.R

Appendix D

Constant division time instead of constant division rate?

In our models of part I, we have always considered a fixed replication rate for the sake of simplicity. However, fixed replication times are closer to the reality of bacterial replication. In this section, we will study a simple model with one population, identical to the null model defined at the beginning of section 1.2, but with all the bacteria that divide after a fixed replication time τ instead of at a fixed rate r. Death happens at fixed rate c, and the initial size of the population is a random number drawn from a Poisson distribution of mean βN.

D.1 Generating function

We will use the formalism developed in [38]. Let us write Z_i the size of the population at the i^{th} generation (*i.e.* after a time $i\tau$). We make the assumptions that (Z_0, Z_1,...) form a Markov chain, meaning that Z_{n+1} only depends on Z_n, and that the transition probabilities for the chain do not vary with time. We also suppose that each individual is behaving independently from the others, meaning that the number of offspring per individual in the next generation does not depend on the size of the population. With these assumptions, the model we consider is a branching process, also called a Galton-Watson process. Writing p_k the probability for one individual to produce k offspring in the next generation, the probability generating function is defined as:

$$f(s) = \sum_{k=0}^{\infty} p_k s^k$$

where s is a complex variable. In our death-birth process, those p_k are easy to write: either the considered individual is dead before the next generation and hence produces no offspring, either it is not dead and it divides in two new individuals at time τ. Thus only p_0 and p_2 will be non-zero. During the time interval between two replication events, the size of the population $N(t)$ follows an exponential decay of the form $N(t) = Ne^{-ct}$. Thus after the time τ, only a

fraction $e^{-c\tau}$ of the initial population has survived, $e^{-c\tau}$ is thus the probability of survival for one individual during the time τ. Thus $p_0 = 1 - e^{-c\tau}$ and $p_2 = e^{-c\tau}$, and finally:

$$f(s) = 1 - e^{-c\tau} + s^2 e^{-c\tau}$$

As shown in [38], the generating function of Z_n, the size of the population at the n^{th} generation, knowing that $Z_0 = 1$, writes as the function f composed with itself n times $f_n = f \circ f \circ ... \circ f$. Now let us assume that $Z_0 = k$: the generating function of Z_n will be $[f_n]^k$, because Z_n can then be seen as the sum of k groups of individuals of sizes $Z_n^1, ...Z_n^k$, each group coming from one specific of the k initial individuals and resulting from an independent branching process:

$$\begin{aligned}
[f_n(s)]^k &= \left[\sum_{i=0}^{\infty} P(Z_n = i \mid Z_0 = 1)s^i\right]^k \\
&= \sum_{i_1}\sum_{i_2}...\sum_{i_k} P(Z_n^1 = i_1 \mid Z_0 = 1)...P(Z_n^k = i_k \mid Z_0 = 1)s^{i_1}...s^{i_k} \\
&= \sum_{i_1+...+i_k=i} P(Z_n^1 = i_1 \mid Z_0 = 1)...P(Z_n^k = i_k \mid Z_0 = 1)s^i \\
&= \sum_i P(Zn = i \mid Z_0 = k)s^i
\end{aligned}$$

In our case, we want Z_0 to be drawn from a Poissonian distribution of mean βN. Thus the total generating function for the distribution of the size of the population after n generations writes:

$$f_{tot}(s) = \sum_{k=0}^{\infty} P(Z_0 = k)[f_n(s)]^k = \sum_{k=0}^{\infty}[f_n(s)]^k \frac{(\beta N)^k}{k!}e^{-\beta N}$$

$$f_{tot}(s) = \exp[\beta N(f_n(s) - 1)] \tag{D.1}$$

D.2 Mean population size

From the generating function expression D.1, one can get the mean size of the population at the n^{th} generation:

$$\langle Z_n\rangle = \sum_{i=0}^{\infty} iP(Z_n = i) = \left.\frac{\partial f_{tot}}{\partial s}\right|_{s=1} = \left.\frac{\partial f_n}{\partial s}\right|_{s=1} \beta N \exp[\beta N(f_n(1) - 1)]$$

One can show by recurrence that

$$f_n'(s) = \frac{\partial f_n}{\partial s} = \prod_{i=0}^{n-1} f' \circ f_i, \tag{D.2}$$

with f_0 being defined as the identity function and f_i the function f being composed i times with itself. 1 is a fixed point of f, hence $f_i(1) = 1$ for all i, and *in fine*:

$$\langle Z_n\rangle = \beta N[f'(1)]^n = \beta N[2e^{-c\tau}]^n$$

which matches what we expected. To make the final population size correspond to the one we get in the fixed rate model ($\beta N \exp(r-c)t$ after a time $t = n\tau$), the replication rate has to be defined as $r = \ln(2)/\tau$, as it was seen previously already (when calculating r_{mean} from the plasmid dilution in section 2.2.1).

D.3 Fixed point and extinction probability

The extinction probability, *i.e.* the probability that at a certain point all bacteria are dead, is given by:

$$f_{tot}(0) = \sum_{k=0}^{\infty} P(Z_n = k)0^k = P(Z_n = 0)$$

For large n, $f_n(0)$ can be approximated by the first fixed point $s*$ of the function f, given by $f(s*) = s*$ (similarly to a recurring sequence converging to the fixed point of the function which defines it). f admits one fixed point smaller than 1, provided that $2e^{-c\tau} > 1$: $s* = e^{c\tau} - 1$, [1]and thus the probability of extinction is given by:

$$\lim_{n\to+\infty} P(Z_n = 0) = \exp[\beta N(e^{c\tau} - 2)] \tag{D.3}$$

Since we took the limit of a high number of generations n in the fixed replication time model, we should compare it with the large time limit of the extinction probability in the fixed replication rate model (equation 2.1):

$$\lim_{t\to+\infty} P(n = 0, t) = \exp[-\beta N(1 - \frac{c}{r})].$$

Let us consider the survival probability $\mathcal{S}_1 = 1 - s*$ of the lineage of one established bacteria (then the overall survival probability is $1 - \exp(-N\beta\mathcal{S}_1)$, summing over the Poisson distribution). When bacteria divide every τ, $\mathcal{S}_{1,\tau} = 2 - \exp(c\tau)$, while when bacteria divide at a rate r, $\mathcal{S}_{1,r} = 1 - c/r$. We have seen in the previous section that we need to take $\tau = \ln 2/r$ to impose the equivalence in the mean population sizes. Then $\mathcal{S}_{1,\tau} = 2 - 2^{c/r}$. It can be easily shown that for any c/r between 0 and 1, $1 - c/r \leq 2 - 2^{c/r}$, then $S_{1,r} \leq S_{1,\tau}$. Survival is increased when taking fixed division time instead of a fixed division rate.

D.4 Variance and variance on the growth factor

Likewise, we calculate the variance over population size after n generations:

[1] $s*$ is also the probability of extinction for the lineage of one established bacteria (after the Poisson process), since f_n is the generating function for the n^{th} generation knowing that one started from one established bacteria

$$\begin{aligned}\langle var\rangle = \left\langle (Z_n - \langle Z_n\rangle)^2\right\rangle = \langle Z_n^2\rangle - \langle Z_n\rangle^2 &= \sum_{i=0}^{\infty} i^2 P(Z_n = i) - \langle Z_n\rangle^2 \\ &= \sum_{i=0}^{\infty} i(i-1)P(Z_n = i) + \sum_{i=0}^{\infty} iP(Z_n = i) - \langle Z_n\rangle^2 \\ &= \left.\frac{\partial^2 f_{tot}}{\partial s^2}\right|_{s=1} + \langle Z_n\rangle - \langle Z_n\rangle^2\end{aligned}$$

Let us first calculate the second derivative of the total generating function eq. (D.1):

$$f''_{tot}(s) = \frac{\partial^2 f_{tot}}{\partial s^2} = f''_n(s)\beta N \exp[\beta N(f_n(s)-1)] + (f'_n(s))^2(\beta N)^2 \exp[\beta N(f_n(s)-1)]$$

The second derivative of f_n can be deduced from the expression of its first derivative eq. (D.2):

$$f''_n(s) = \left[\prod_{i=0}^{n-1} f' \circ f_i(s)\right]' = \sum_{j=0}^{n-1} [f' \circ f_j]' \prod_{\substack{i=0 \\ i\neq j}}^{n-1} f' \circ f_i$$

Then one can make the following simplification, since $f''(s) = 2e^{-c\tau}$ is a constant:

$$[f' \circ f_i]' = f'_i[f'' \circ f_i] = 2e^{-c\tau} f'_i$$

Remembering that $f_i(1) = 1$ and that $f'_i(1) = [2e^{-c\tau}]^i$ for all i, one finally gets:

$$f''_n(1) = [2e^{-c\tau}] \sum_{j=0}^{n-1} [2e^{-c\tau}]^j [f'(1)]^{n-1} = [2e^{-c\tau}]^n \sum_{j=0}^{n-1} [2e^{-c\tau}]^j = [2e^{-c\tau}]^n \frac{(1 - [2e^{-c\tau}]^n)}{1 - [2e^{-c\tau}]}$$

Replacing in the expression of the total second derivative:

$$f''_{tot}(1) = \beta N [2e^{-c\tau}]^n \frac{(1 - [2e^{-c\tau}]^n)}{1 - [2e^{-c\tau}]} + (\beta N)^2 [2e^{-c\tau}]^{2n}$$

And *in fine* in the variance:

$$\begin{aligned}\langle var\rangle_\tau =& \beta N [2e^{-c\tau}]^n \frac{(1 - [2e^{-c\tau}]^n)}{1 - [2e^{-c\tau}]} + \beta N [2e^{-c\tau}]^n \\ =& \beta^2 N^2 (2e^{-c\tau})^{2n} \frac{1}{\beta N} \left(\frac{1}{2e^{-c\tau} - 1} + (2e^{-c\tau})^{-n} \frac{2e^{-c\tau} - 2}{2e^{-c\tau} - 1}\right).\end{aligned} \tag{D.4}$$

In the fixed replication rate model, we had the following expression for the simple variance:

$$\langle var\rangle_r = \left(\beta N e^{(r-c)t}\right)^2 \frac{1}{\beta N} \left(\frac{2r}{r-c} - \frac{r+c}{r-c} e^{-(r-c)t}\right)$$

The ratio is :

$$\frac{var_r}{var_\tau} = \frac{\frac{2r}{r-c} - \frac{r+c}{r-c}e^{-(r-c)t}}{\frac{1}{2e^{-c\tau}-1} + (2e^{-c\tau})^{-n}\frac{2e^{-c\tau}-2}{2e^{-c\tau}-1}}$$

Then, in the limit of long time (t and n large),

$$\frac{var_r}{var_\tau} \to \frac{2r(2e^{-c\tau}-1)}{r-c} = \frac{2r(2^{1-c/r}-1)}{r-c}$$

In the limit of $c \ll r$, this ratio tends to 2 : the variance is larger in the case of division at a rate r rather than division every τ. The exact ratio is shown on fig. D.1 (replacing t by $n\tau$ and r by $\ln(2)/\tau$).

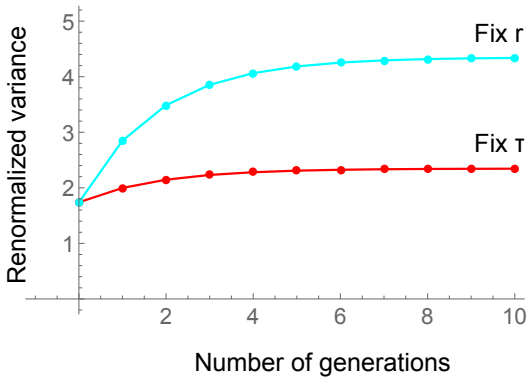


Figure D.1 – Red: variance for the fixed time model; Cyan: variance for the fixed rate model, in function of n (the number of generations, so that $t = n\tau$). Values for the other parameters: $\beta = 0.115$, $c = 0.2r$, $\tau = 1/48$, $r = 48 * \ln(2)$ (time unit is the day).

Then, for the variance over the growth factor, we calculate

$$\left\langle \left(\frac{m_i}{n_i}\right)^2 \right\rangle = \frac{1}{n_i^2}\langle m_i(m_i-1)\rangle + \frac{1}{n_i^2}\langle m_i\rangle = \frac{1}{n_i^2}\left.\frac{\partial^2 f_{tot}}{\partial s^2}\right|_{s=1} + \frac{\beta(2e^{-c\tau})^n}{n_i}$$

$$= \beta(2e^{-c\tau})^n\left(1 + \frac{1-[2e^{-c\tau}]^n}{1-[2e^{-c\tau}]}\right)\frac{1}{h}\sum_{i=1}^{h}\frac{1}{n_i}$$

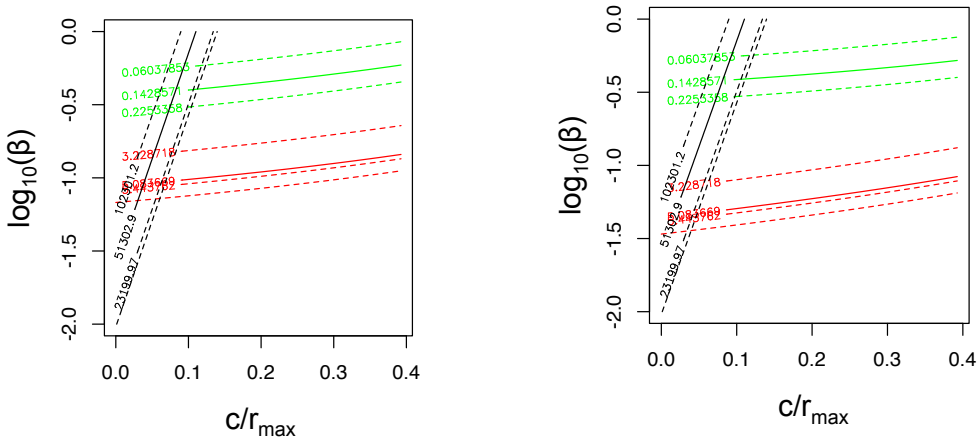
and then with eq. (2.10):

$$\langle var\rangle_\tau = \beta(2e^{-c\tau})^n\left(1 + \frac{1-[2e^{-c\tau}]^n}{1-[2e^{-c\tau}]}\right)\frac{1}{h}\sum_{i=1}^{h}\frac{1}{n_i} \tag{D.5}$$

Note that if we take all $n_i = N$ we recover the simple variance D.4 divided by N^2.

D.5 Comparison with data

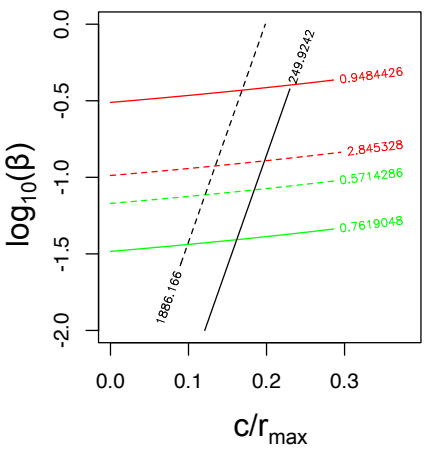
We have seen that the expression of the two observables "WITS loss" and "variance" were changed when we consider a fixed division time instead of a fixed replication rate : both the loss probability and the variance are decreased. We know that bacteria actually divide at constant time rather than at constant rate, thus it is important to check if our hypothesis of constant rate impacts our result considerably or not. It is not trivial to predict the impact on the parameter estimates, since the whole observables landscape will be shifted differently for each observable. Thus we go back to the experimental values contour curves on the observables landscape presented in the one-population model section 2.5.2. The contour curves on the population size are unchanged since we have equivalence between the two models for the population size. The question is to know whether the fixed time model will see the contour for the variance and for the loss get closer, or on the contrary, further apart. After close examination of all the data in this framework, the answer is that it depends on the experiment : in some of them, the fixed time pulls the estimates further apart (as for example in figure D.2), and in some others, it gets them closer together (as for example in figure D.3). Thus, no clear conclusion can be drawn on the impact of the fixed replication rate hypothesis on our study.



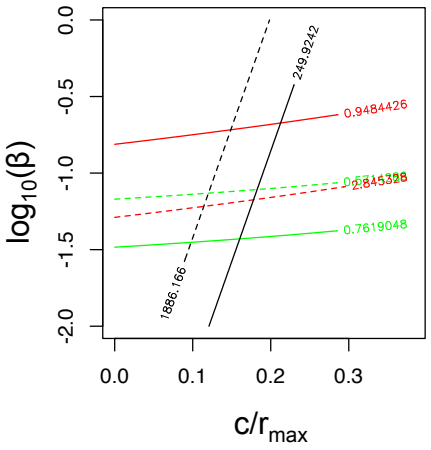
A. Contour lines on the theoretical landscapes in the constant rate model

B. Contour lines on the theoretical landscapes in the fixed division time model

Figure D.2 – Contour lines for the experiment with strain "SB300" starting with 10^3 inoculum. A. is a repetition of figure 2.8B., except that the confidence interval lines (variance of the variance and quantiles for the variance, quantiles of the WITS loss) and the loss as calculated from the log-likelihood have been removed, so as to ease visual comparison (because no equivalent was derived in the framework of the fixed division time model). B. Everything as in A, except that the landscapes for the variance (expression D.5) and the proportion of lost WITS (expression D.3) are calculated in the framework of the fixed division time model. In this case, switching to the fixed division time model pulls the two estimates (from the variance and the loss) further apart.



A. Contour lines on the theoretical landscapes in the constant rate model



B. Contour lines on the theoretical landscapes in the fixed division time model

Figure D.3 – Contour lines for the experiment with strain "SB300" starting with a 10^5 inoculum. Same color code as in figure D.2: Red dashed: final experimental variance in three mice on the theoretical variance landscape (as calculated in the framework of the fixed replication rate for A. and fixed division time for B. Only one appears, because all the WITS were lost in the two other mice and a null variance does not appear on this grid of parameters), and red solid line: the mean value. Green line: contour of the experimental proportion of WITS lost (combining the data on the three mice) on the theoretical proportion of WITS lost landscape (as calculated in the framework of the fixed replication rate for A. and fixed division time for B.). Green doted lines: on the same landscape, contour lines for the proportion of WITS lost +/− its square root. **Black**: on the landscape of the theoretical mean growth factor, the contour for its experimental value in the three mice (dashed lines) and for the average (solid line). In this case, switching to the fixed division time model get the two estimates (from the variance and the loss) closer.

Appendix E

Detailed derivation for the model with force-dependent breaking rate

We present here the details of the derivations for the equations of our force-dependent breaking rate model from section 5.3.5.

A link between bacteria may consist of several sIgA bonds, and the number of bound sIgA may not be exactly the same from one inter-bacteria link to the next, but as sIgA are likely well mixed, many per bacteria and that bacteria are similar to each other, let us assume that link heterogeneity is negligible. The links could break if there was some process degrading the sIgA, but the sIgA are thought to be very stable[66]. Another possible explanation for link breaking is that the antigen get extracted from the bacterial membrane, which may depend exponentially with the force applied on the link[67][63]. If the forces are produced by the bacteria themselves (such as by flagella rotation), there are likely to fluctuate on timescales which are short compared to the time between two bacterial replications, and their distribution is likely to be the same for all links, so it would be appropriate to model their effect as a fixed breaking rate, the same for all the links. Another force is the hydrodynamical force exerted by the flow on the bacterial chain.

The flow in the digestive system is complex and not precisely characterized. Longer bacterial chains may also bend and their shape have complex interactions with the flow. Here, we present the simplest model taking into account the forces exerted by the flow on the link breaking rate. We aim to capture the main plausible effects of the flow when the link breaking rate is force-dependent.

Let us take a linear chain of N bacteria, each of length B. Let us approximate it by a rigid chain with beads linked by straight rods of length B (panel A of figure E.1). Let us assume that the rods are infinitely thin so they do not interact with the flow, and let us neglect the hydrodynamical interaction between the beads, so that they are each subject to the same frictional force for a given fluid velocity, and, given that the typical Reynolds numbers in the digestive tract are relatively low[9], then the viscous force on each bead is proportional to the flow velocity.

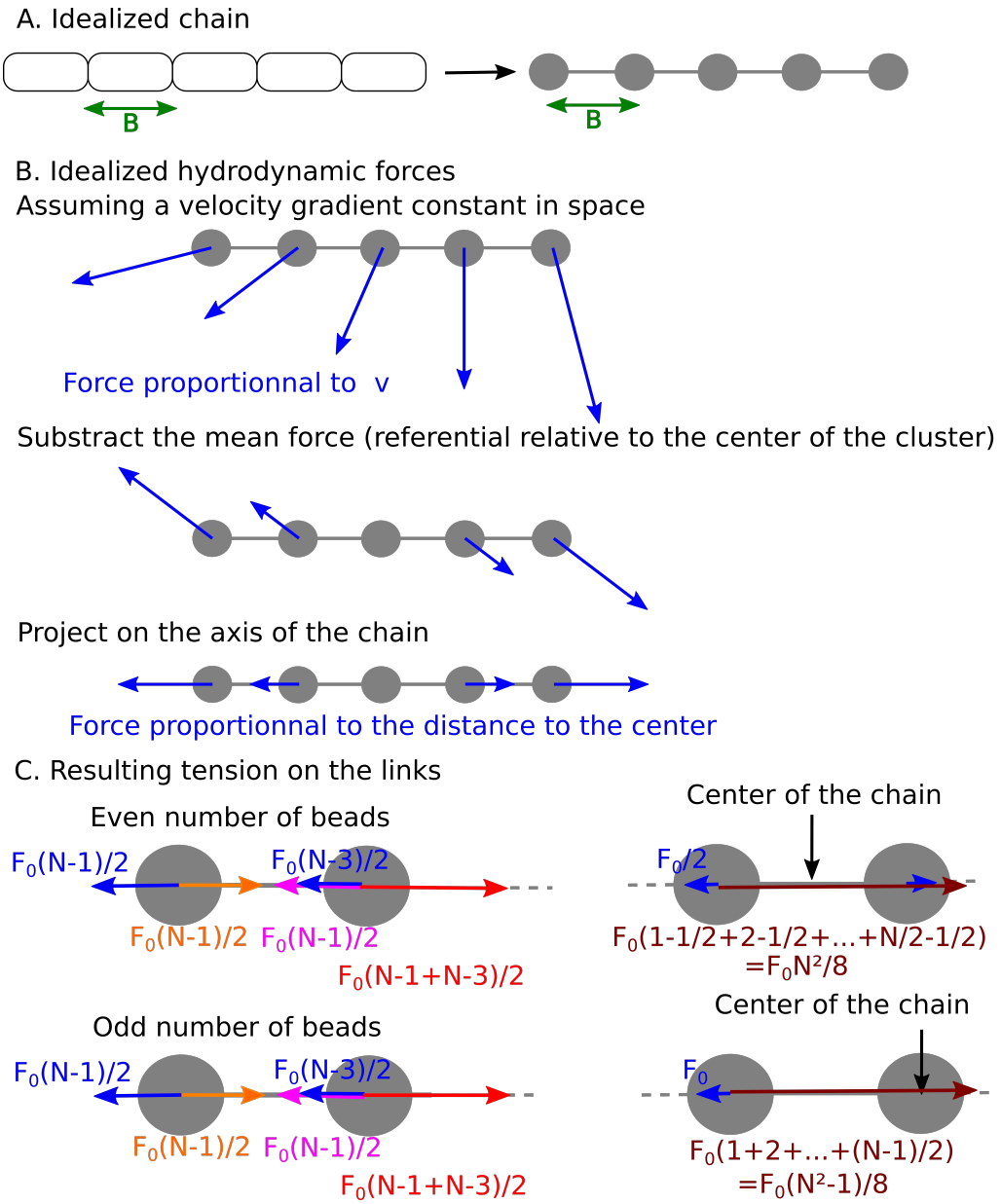


Figure E.1 – Schematic of the forces applied to the chain. **A** We assume a straight chain of beads with no hydrodynamic interactions between them. **B** We subtract the average force to put ourselves in the referential of the center of the chain, as the total force will translate the whole chain and not impact the forces on the links. We focus on the forces parallel to the chain that will impact the tension between the links. **C** Sum of the forces on each bead, for chains with even and odd numbers of beads.

Then, let us assume that the velocity gradient in the fluid is constant around the chain. The rationale for this approximation is that the typical scales of the

flow are of the order of the centimeter / millimeter (for instance in a mouse, the cecum typical size is in the cm range), much larger than typical bacterial chains (the length of one bacterium is about $2\mu m$, so even chains of dozens of bacteria remain small compared to the typical flow scale), thus we take a linear approximation of the velocity field in the vicinity of a bacterial chain.

Then, if we take the sum of the forces on the whole chain, it will be equal to mN multiplied by the acceleration of the center of mass of the chain, with m the mass of each bead. When all the beads move together, there is no force on the links, thus let us take the referential relative to the center of the chain, and subtract the mean force on each bead (panel B of figure E.1). Then, there remains forces perpendicular to the axis of the chains, and forces parallel to the axis of the chain. The forces perpendicular to the axis of the chain will make it rotate, and as they are perpendicular, they have no effect on the tension on the rods. Then, let us consider only the forces parallel to the chain.

In the example portrayed here, the chain is elongated. The reverse could happen, but in this case, the chain would likely buckle, and the force applied on the links would be small. The flow varies considerably in time, due to peristaltic motions[10][9]. There would be moments with no force and little breaking, and moments with larger forces and more breaking. The flow due to peristaltic motions changes on time scales short compared to the typical bacterial division time, thus we will assume that periods of low breaking and high breaking rates will be equivalent to an average effective breaking rate. Then let us consider the case of elongation only, as portrayed here.

Then the force on each bead is equal to F_0 multiplied by the distance to the center divided by B. We assume, following [67][63], that the breaking rate is dependent on the force. Thus, we define α and β such that the breaking rate of a link is $\alpha \exp(\beta F/F_0)$ if a force F is applied to the link. In the limit of small force, the breaking rate will be α, the same for all links, as in the base model. β is some constant caracterizing how much the stability of the link is force-dependent.

We can write the force on each bead (panel C of figure E.1). Then, here, because the chain is rigid and straight, the sum of the forces on each bead has to be zero. The tension on the outermost link will simply be equal to the flow force on the outermost bead, i.e. F_0 multiplied by its distance to the center divided by B, i.e. $(N-1)/2$ (both for chains of odd and even numbers of beads). On the next link, the tension has to compensate for the flow force on the second bead, plus the tension applied by the outermost link. Thus the tension on this link is $F_0((N-1)/2 + (N-1)/2 - 1)$, and so forth (this is analogous to modeling of breaking of polymer chains in elongational flows, as in[65]).

For N even, the force on the j^{th} link starting from the outermost link will be:

$$F_{jth\ link, N\ even} = F_0 \sum_{k=N/2-j+1}^{N/2} (k - 1/2)$$

Using $\sum_{i=1}^{n} i = n(n+1)/2$, it can be rewritten as:

$$F_{jth\ link,N\ even} = F_0 \left(\frac{N(N+2)}{8} - \frac{(N-2j)(N+2-2j)}{8} - j/2 \right),$$

$$F_{jth\ link,N\ even} = F_0 \left(\frac{N^2}{8} - \frac{(N-2j)^2}{8} \right).$$

There are two links j^{th} away from the extremities, for j from 1 to $N/2-1$, and one central link, for which $j = N/2$. This leads to:

$$\frac{dn_i}{dt} = -rin_i - \alpha n_i e^{\beta i^2/8} \left(1 + 2\sum_{k=2}^{i/2} e^{-(k-1)^2\beta/2} \right) + r(i-1)n_{i-1} + 2\alpha n_{i+1} e^{\beta i/2}$$

For N odd, the force on the j^{th} link starting from the outermost link will be:

$$F_{jth\ link,N\ odd} = F_0 \sum_{k=(N-1)/2-j+1}^{(N-1)/2} k.$$

Simiarly to the N even case, we can rewrite:

$$F_{jth\ link,N\ odd} = F_0 \left(\frac{(N-1)(N+1)}{8} - \frac{(N-1-2j)(N+1-2j)}{8} \right),$$

$$F_{jth\ link,N\ odd} = F_0 \left(\frac{N^2}{8} - \frac{(N-2j)^2}{8} \right)$$

Because of the two sides, there are two links j for each chain, for j from 1 to $(N-1)/2$. Then, this leads to the following equation for the evolution in time of the mean number of clusters of odd size i:

$$\frac{dn_i}{dt} = -rin_i - 2\alpha n_i e^{\beta i^2/8} \sum_{k=1}^{(i-1)/2} e^{-(k-1/2)^2\beta/2} + r(i-1)n_{i-1} + 2\alpha n_{i+1} e^{\beta i/2}.$$

Appendix F

Proportion of mixed clusters and probability to transmit at least one mutant

In our model of chapter 6, we take the mean field assumption that the proportion of resistant versus sensitive bacteria at the end of the infection can just be taken as its mean expected value (average over several realization of the intra-host dynamics), consistently with the deterministic ordinary differential equations model. However, fluctuations could be important. For instance, if the infection starts from a small number of bacteria of the same type, a mutant appearing during the first replication will then make an important share of the population at the end of the infection. We also assume that in immune individuals, except if the infection starts with only sensitive bacteria, clusters are only of one type, either all sensitive or all resistant. In this section, we explore how realistic these assumptions are.

First, if the infection starts with a mix of sensitive and resistant bacteria, as loss of bacteria will be rare at the beginning of the infection, and as the population of bacteria grows to large numbers in the infected host, the number of both sensitive and resistant bacteria will be large, so that fluctuations are not expected to play a big role. There will be some mutations, but of very small impact, since a substantial fraction of both types is already present. The case in which mutations and fluctuations are expected to matter is when the infection starts with bacteria of only one type. Let us here discuss the case in which all the infecting bacteria are of one type (it does not matter whether it is resistant or sensitive), with μ the mutation probability to the other type for each daughter cell, and let us look at the probability to transmit at least one mutant bacteria. For the sake of simplicity, we will consider here the case $s = 0$, i.e. both bacterial types have the same fitness.

Let us assume that there are G generations, and that clusters are of size 2^g, with $2^g = N_c$. We here assume that N_c, the cluster size, is equal to N_b, the bottleneck size. We both write them as N. We will consider exponential growth: starting from size N, bacteria divide G times, leading to a final population size

$N2^G$. At each replication, each daughter bacteria has a probability μ of mutating.

In the mean field approximation, let us consider bacteria at the end of the infection. They went through G replications from the start of the infection, so they have a probability μG of being mutant. When transmitting N bacteria (and assuming that N is very small compared to the final population size, and that $N\mu G \ll 1$), the probability to transmit (at least) one mutant is $1-(1-\mu G)^N \simeq N\mu G$ in the naive case; in the immune case, we assume that there is a probability μG to transmit a cluster of mutants only, and only sensitive are transmitted in the other cases. Here we will study the validity of these assumptions. Henceforth, we will consider the case of a very small mutation rate μ so that at most one mutation occurs during the infection of a host.

Naive individual

Consider the lineage of one bacteria: it involves G steps of replication. At step j, this bacteria has 2^j descendants, there is a probability $\mu 2^j$ that one mutation occurs at this step, in which case 2^{G-j} bacteria will carry the mutation in the final population. This will correspond to a proportion $\mathcal{P}_j = 2^{G-j}/(2^G N) = 1/(N2^j)$ in the final population.

Assuming that $2^G \gg 1$, we can neglect the difference between taking a sample with or without replacement, so the probability for a transmission from a naive host to contain at least one mutant will be $1-(1-\mathcal{P}_j)^N$ (probability $1-\mathcal{P}_j$ for one bacteria to be of the initial type, probability $(1-\mathcal{P}_j)^N$ that all bacteria chosen are of the initial type). So, multiplying by the initial N bacteria, and summing over the G replication steps, the probability that a transmission from a naive host includes at least one mutant is:

$$m_{1,naive} = \sum_{j=1}^{G} N\mu 2^j \left(1-\left(1-\frac{1}{N2^j}\right)^N\right) \tag{F.1}$$

In the limit of j large, $(1-\frac{1}{N2^j})^N \simeq 1-1/2^j$, equation (F.1) yields $m_{1,exp,naive} \simeq N\mu G$. Hence, the mean field result is recovered. We know the mean field is an upper bound of the real value: an early mutation will lead to a higher proportion of mutant in an individual, and thus the probability that several mutants are transmitted at the same time will be higher. But, because when we average over all the possible transmissions the mean number of mutants does not change, transmitting several mutants at a time leads to a lower probability of transmitting at least one mutant.

In the limit of N large (but it does not need to be very large), $1-\exp(N\log(1-$

$1/(N2^j))) \geq 1 - \exp(-1/2^j)$. Thus, going back to (F.1):

$$\begin{aligned} m_{1,naive} &\geq \sum_{j=1}^{G} N\mu 2^j \left(1 - \exp(-1/2^j)\right) \\ &\geq \sum_{j=1}^{G} N\mu 2^j \left(1/2^j - 1/(2^{2j+1})\right) \\ &= \sum_{j=1}^{G} N\mu \left(1 - 1/(2^{j+1})\right) \\ &= N\mu \left(G - \frac{1}{2^2} \sum_{j=1}^{G} \frac{1}{2^{j-1}}\right) \\ &= N\mu \left(G - \frac{1}{2}\left(1 - \frac{1}{2^G}\right)\right) \end{aligned}$$

We then have shown that:

$$N\mu G \geq m_{1,exp,naive} \geq N\mu(G - 1/2). \tag{F.2}$$

Immune individual

In the immune case, we assume that one cluster is transmitted. Let us estimate the probability that the cluster transmitted is fully mutant ($m_{N,immune}$) and the probability that the cluster transmitted is mixed ($m_{mixed,immune}$). If a mutation occurs at step j (probability $N\mu 2^j$) between the first step and the $(G-g)^{th}$ step, then there will be 2^{G-g-j} mutant bacteria at the $(G-g)^{th}$ step, yielding 2^{G-g-j} fully mutant clusters at the final G^{th} step. They will then be in proportion $1/(N2^j)$ among the $N2^{G-g}$ clusters. Thus:

$$m_{N,immune} = \sum_{j=1}^{G-g} N\mu 2^j \frac{1}{N2^j} = \mu(G-g). \tag{F.3}$$

If a mutation occurs at step j between the $(G-g+1)$-th step and the G-th step, then it will give one mixed cluster. Thus:

$$m_{mixed,exp,immune} = \sum_{j=G-g+1}^{G} N\mu 2^j \frac{1}{N2^{G-g}} = \sum_{j=1}^{g} \mu 2^j = \mu 2(2^g - 1) = 2\mu(N-1) \tag{F.4}$$

Conclusion

Interestingly, when the host is naive the result is very close to the mean-field case. Indeed, we showed that the probability to transmit at least one mutant is bounded

between $2N\mu(G-1/2)$ and $2N\mu G$ (the mean field result) (see equation (F.2)). The total probability for a transmission from an immune host to include at least one mutant is $m_{tot,immune} = m_{N,immune} + m_{mixed,immune} = \mu(G-g+2(2^g-1)) = \mu(G-\log_2(N)+2(N-1))$ (see equations (F.3) and (F.4)). If $N \gg G$, then it gives $m_{tot,exp,immune} \simeq 2N\mu$, which is $G/2$ times smaller than for the naive case. However in this case most transmissions will be of mixed clusters rather than fully mutant clusters. If $N \ll G$, then $m_{tot,exp,immune} \simeq \mu G$, which is N times smaller than for a naive host, and will be mostly of fully mutant clusters.

In all cases, for naive donor hosts, we will assume that the N transmitted bacteria are of types taken randomly and independently. In particular, when the initial bacteria are all of one type, if the average final proportion of mutants is $\mathcal{P}$ (proportional to the mutation rate μ, thus very small), then the probability of transmitting one mutant bacteria among N will be $\binom{N}{1}\mathcal{P}(1-\mathcal{P})^{N-1} \simeq N\mathcal{P}$, with very little chance of transmitting more than one mutant bacteria.

For immune donor hosts, *if the infection starts with a mixed inoculum*, we will assume that all bacteria in a cluster transmitted to another host are of the same type. Conversely, *if the infection starts with an inoculum of bacteria that are all of the same type*, then other bacterial types are produced only by mutations. With G the number of generations within the host and N the bottleneck size and cluster size, if $G \gg N$, then most clusters will be of one bacterial type only, and the probability to transmit a fully mutant cluster will be m, with negligible amount of mixed clusters transmitted. In the case of $G \ll N$, there will be many more mixed clusters than clusters made of mutant bacteria only, and the proportion of mixed clusters will be of the order of $2N\mu$.

Appendix G

Approximations for the evolution of resistance model

In this section we develop approximations for the equations of the model studied in chapter 6. There are particular limits to consider:

- Either the number of generations within a host G is large enough ($sG \gg 1$) so that infections started with a mixed inoculum end up at proportions close to the mutation selection balance, or the number of generations is small ($sG \ll 1$) so that the final proportions of bacterial types is close to the initial values.

- When starting with a fully sensitive inoculum, in an immune host, if $G \gg N$, then most transmissions containing resistant bacteria will be from clusters containing only resistant bacteria. If $G \ll N$, then most transmissions containing resistant bacteria will be from mixed clusters.

We develop here approximations for 2 combinations of these limits.

G.1 Regime of a few generations within the host: both $sG \ll 1$ and $G \ll N$

Here, when all the infecting bacteria are sensitive, clusters in the immune host will be either all sensitive, or mixed. We neglect the fully resistant clusters. The probability for a cluster to be mixed is, in the limit of no selection, $\mu_1 2(2^g - 1) = \mu_1 2(N-1)$ (see equation F.4) . With $p_0 \simeq \mu_1 G$, then the probability to transmit a mixed cluster is about $2p_0 N/G$. Half the mixed clusters transmitted will contain only one resistant bacteria. A quarter of mixed clusters will contain 2 resistant bacteria. So in most cases, the number of resistant bacteria transmitted is small, so taking the limit of one resistant bacteria transmitted is fair.

G.1.1 First approximation: only states starting from 0, 1 and N resistant bacteria matter

Let us make several approximations:

- Let us assume that $Np_0 \ll 1$. Thus for naive untreated individuals initially infected with sensitive bacteria only, the probability to transmit more than 1 resistant bacteria is negligible. They transmit 0 resistant, N sensitive bacteria to an average of $\lambda(1-p_0)^N \simeq \lambda(1-Np_0)$ individuals, and 1 resistant bacteria and $N-1$ sensitive bacteria to an average of $\lambda N p_0$ individuals.

- Let us assume that the pathways with progressive increase in the proportion of resistant bacteria transmitted are very infrequent, compared to the case in which a resistant bacteria gets into a treated host, which leads directly to a full resistant infection. Then, when an individual starts the infection with one resistant bacteria, let us only compute rigorously cases in which it transmits 0 or N resistant bacteria, and let us assume that in all the other cases, it transmits 1 resistant bacteria. There will be cases in which more than one resistant bacteria are transmitted, but we argue that they would behave similarly enough from the cases in which one 1 resistant bacteria is transmitted.

- Let us assume that reversion mutants are rare, so that when an individual is initially infected with only resistant bacteria, we will consider that only resistant bacteria are transmitted. In practice, there would be cases in which $N-1$ resistant bacteria and 1 sensitive are transmitted, but as most subsequent transmissions will be of resistant bacteria only (because $sG \ll 1$, and because some individuals are treated), then cases starting with $N-1$ resistant bacteria are lumped with cases starting with resistant bacteria only.

- Here, we assume that an immune individual transmits mixed clusters with rate $2\lambda' N p_0/G$, as explained in appendix F.

Then there are only 3 equations:

$$e_N = (1-w)\exp[-\lambda(1-e_N)] + w\exp[-\lambda'(1-e_N)]. \tag{G.1}$$

Remarkably, this equation does not depend neither on e_0 nor on e_1 and thus can be numerically solved independently. It does not depend on neither q nor q'. If $\lambda = \lambda'$, it also does not depend on w.

$$\begin{aligned} e_1 =& (1-w)q\exp[-\lambda(1-e_N)] + wq'\exp[-\lambda'(1-e_N))] \\ &+ (1-w)(1-q)\exp[-\lambda((1-p_1)^N(1-e_0) + (1-(1-p_1)^N)(1-e_1))] \\ &+ w(1-q')\exp[-\lambda'((1-p_1)(1-e_0) + p_1(1-e_N)))] \end{aligned} \tag{G.2}$$

$$\begin{aligned} e_0 =&(1-w)q + wq' + (1-w)(1-q)\exp[-\lambda((1-Np_0)(1-e_0) + Np_0(1-e_1)))] \\ &+ w(1-q')\exp\left[-\lambda'\left(\left(1-\frac{2Np_0}{G}\right)(1-e_0) + \frac{2Np_0}{G}(1-e_1)\right)\right] \end{aligned} \tag{G.3}$$

As p_1 is of the order of $1/N$ (see 6.6), with N large, then $(1-p_1)^N = \exp(N\log(1-p_1)) \simeq \exp(-Np_1)$. We cannot simplify it further, contrary to the term in p_0, because while Np_0 is assumed to be small (because p_0 is of the order of μ_1 which is assumed to be much smaller than $1/N$), we expect Np_1 to be of the order of 1. Although this system still has to be solved numerically, it is much simpler, and it provides a basis for further simplifications.

G.1.2 When extinction is certain in the absence of mutations

Let us go back to equation (G.3). If $\lambda(1-w)(1-q) + \lambda' w(1-q') > 1$, then most spread of the bacteria will come from spread of the sensitive bacteria, which will then have the opportunity to become resistant when present in many individuals. In this regime, clustering could change the timing of apparition of resistance, but will be of little effect on whether the bacterial strain which can evolve to resistance spread. Here, we study the regime in which extinction is certain in the absence of mutations, i.e. $\lambda(1-w)(1-q) + \lambda' w(1-q') < 1$. Then $e_0 = 1-\epsilon$, with ϵ small, expected to be of the order of μ_1. We replace in equation (G.3):

$$\begin{aligned} 1-\epsilon =&(1-w)q + wq' + (1-w)(1-q)\exp(-\lambda((1-Np_0)\epsilon + Np_0(1-e_1))) \\ &w(1-q')\exp(-\lambda'((1-2Np_0/G)\epsilon + (1-e_1)2Np_0/G)). \end{aligned} \tag{G.4}$$

Writing (G.3), we have already assumed that $Np_0 \ll 1$. Thus let us develop the exponential in (G.4). If we keep the first order in ϵ and p_0, then:

$$\epsilon(-1+\lambda(1-w)(1-q)+\lambda' w(1-q')) = -(1-w)(1-q)\lambda Np_0(1-e_1) - w(1-q')\lambda' 2\frac{Np_0(1-e_1)}{G}.$$

$$e_0 \simeq 1 - (1-e_1)Np_0\frac{(1-w)(1-q)\lambda + 2w(1-q')\lambda'/G}{1-\lambda(1-w)(1-q)-\lambda' w(1-q')}.$$

In this regime, e_0 is very close to 1, so (G.2) can be simplified to:

$$\begin{aligned} e_1 \simeq&(1-w)q\exp(-\lambda(1-e_N)) + wq'\exp(-\lambda'(1-e_N)) \\ &+ (1-w)(1-q)\exp(-\lambda(1-e^{-Np_1})(1-e_1)) \\ &+ w(1-q')\exp(-\lambda' p_1(1-e_N)). \end{aligned} \tag{G.5}$$

G.1.3 Ratio of spread of the bacteria in all immune vs. all naive host populations

Remaining in the regime $\lambda(1-w)(1-q) + \lambda' w(1-q') < 1$ (limited spread in the absence of mutations), let us compare the case with either all immune, or all

naive host population.

$$e_{0,naive} = 1 - \frac{(1-q)\lambda N p_0 (1 - e_{1,naive})}{1 - \lambda(1-q)}.$$

$$e_{0,immune} = 1 - \frac{(1-q')\lambda' 2N p_0 (1 - e_{1,immune})}{G(1 - \lambda'(1-q'))}.$$

Using G.5 and G.1,

$$e_{1,naive} \simeq q\exp(-\lambda(1-e_{N,naive})) + (1-q)\exp(-\lambda(1-e^{-Np_1})(1-e_{1,naive})). \quad \text{(G.6)}$$

$$e_{N,naive} = \exp(-\lambda(1 - e_{N,naive})). \quad \text{(G.7)}$$

$$e_{1,immune} = q'\exp(-\lambda'(1 - e_{N,immune})) + (1-q')\exp(-\lambda' p_1(1 - e_{N,immune})). \quad \text{(G.8)}$$

$$e_{N,immune} = \exp(-\lambda'(1 - e_{N,immune})). \quad \text{(G.9)}$$

Let us look at the ratio of survival probability:

$$ratio = \frac{1 - e_{0,naive}}{1 - e_{0,immune}} = \frac{G}{2}\frac{(1-q)\lambda}{(1-q')\lambda'}\frac{(1 - \lambda'(1-q'))}{(1 - \lambda(1-q))}\frac{(1 - e_{1,naive})}{(1 - e_{1,immune})}$$

This ratio does not depend on p_0 any more, i.e. it will not depend on the mutation rate. If we take further $q = q'$, and $\lambda = \lambda'$, the expression simplifies greatly:

$$ratio = \frac{1 - e_{0,naive}}{1 - e_{0,immune}} = \frac{G}{2}\frac{1 - e_{1,naive}}{1 - e_{1,immune}} \quad \text{(G.10)}$$

If $\lambda = \lambda'$, then G.7 and G.9 are the same equations, thus in this regime $e_{N,naive} = e_{N,immune} = e_N$. In this regime we can simplify (G.6) and (G.8):

$$e_{1,naive} \simeq qe_N + (1-q)\exp(-\lambda(1 - e^{-Np_1})(1 - e_{1,naive})).$$

$$e_{1,immune} = qe_N + (1-q)\exp(-\lambda(p_1(1 - e_N))) = qe_N + (1-q)e_N^{p_1}$$

Consequently,

$$e_{1,naive} - e_{1,immune} = (1-q)(\exp(-\lambda(1-e^{-Np_1})(1-e_{1,naive})) - \exp(-\lambda(p_1(1-e_N)))) \quad \text{(G.11)}$$

Going back to (G.6),

$$\begin{aligned} e_{1,naive} =& q\exp(-\lambda(1 - e_N)) + (1-q)\exp(-\lambda(1 - e^{-Np_1})(1 - e_{1,naive})) \\ &< q\exp(-\lambda(1 - e_N)) + 1 - q. \end{aligned}$$

As $\exp(-\lambda(1 - e_N)) = e_N$, then $e_{1,naive} < qe_N + 1 - q$. Then, let us compare $(1 - e^{-Np_1})(1 - e_{1,naive})$ and $p_1(1 - e_N)$. As $e_{1,naive} < qe_N + 1 - q$, then $(1 - e^{-Np_1})(1 - e_{1,naive}) > (1 - e^{-Np_1})q(1 - e_N)$. It is likely that $q > p_1/(1 - e^{-Np_1})$: indeed, p_1 is of the order of $1/N$, thus $p_1/(1 - e^{-Np_1})$ is of the order of $1/(N(1 - e^{-1}))$. N is large in general, and q must be large enough so that extinction

is certain in the absence of mutations, while survival is possible when there are mutations. So, in the most general case, it is the case that $q > p_1/(1-e^{-Np_1})$, and thus $(1-e^{-Np_1})(1-e_{1,naive}) > p_1(1-e_N)$, and $\exp(-\lambda(1-e^{-Np_1})(1-e_{1,naive})) < \exp(-\lambda(p_1(1-e_N)))$. Going back to (G.11), then $e_{1,naive} < e_{1,immune}$. Thus, going back to (G.10), if $q > p_1/(1-e^{-Np_1})$, then:

$$ratio = \frac{1-e_{0,naive}}{1-e_{0,immune}} = \frac{G}{2}\frac{1-e_{1,naive}}{1-e_{1,immune}} > \frac{G}{2}$$

G.2 Limit of a large number of generations, with both $sG \gg 1$ and $G \gg N$

In this case, infections can be sorted in three categories : infections starting with 0 resistant bacteria, infections starting from N resistant bacteria, and infections starting from an intermediate number of resistant bacteria (from 1 to $N-1$, this case will be called *int* for "intermediate"). In the latter case, we assume that the proportion of resistant bacteria will reach the mutation-selection balance (p_{MSB} as discussed in section 6.2.1.3).

When the infection starts with sensitive bacteria only, the average proportion of resistant bacteria at transmission is $p_0 \simeq \mu_1 \frac{1-2^{-sG}}{s\log(2)}$ (see equation (6.5)). Thus when $sG \gg 1$, then $p_0 \simeq p_{MSB} = \frac{\mu_1}{s\log(2)}$, *i.e.* the mutation selection balance is also reached. As a consequence, in the absence of treatment, starting from any number of sensitive bacteria (but at least one) will lead to similar results.

It is also important to compare the quantity of mixed transmissions vs. fully resistant transmission when starting with a fully sensitive infection of an immune host: as we assume that $sg \ll 1$ (*i.e.* the difference in growth rates is negligible over the period of time, or rather the number of generations g it takes to make a cluster), then the proportion of mixed clusters transmitted will be of the order of $2N\mu_1$, the result derived in section F (where we took $s = 0$), equation F.4 . The proportion of fully resistant clusters however will not correspond to the results derived in the same section, because $s > 0$, thus the number of fully sensitive clusters will grow faster than the number of fully resistant clusters. But we know that p_0 tends to the mutation-selection balance rapidly, thus, because $G \gg g$ the proportion of fully resistant clusters transmitted will be of the order of $p_{MSB} = \mu_1/(s\log(2))$. Then if $2N\mu_1$ is larger or of the same order as $\mu_1/(s\log(2))$, *i.e.* if sN is larger or of the same order as 1, then mixed clusters cannot be neglected. If $sN \ll 1$, mixed clusters can be neglected. *We will neglect mixed clusters in all this section.*

When the infection starts with resistant bacteria only, we have shown (6.7) that when $2^{sG}\mu_2 \ll s$, then $p_N \to 1 - 2^{sG}\mu_2/(s\log(2))$. We will remain in this limit.

Another aspect to consider is whether $p_{MSB}N$ is smaller or larger than 1, *i.e.* how $N\mu_1/(s\log(2))$ compares to 1. In the following, we will assume that $p_{MSB}N \ll 1$, so that transmissions from naive hosts at the mutation selection balance will consists most likely of sensitive bacteria only.

G.2.1 Equations

We write the equations in this regime of $G \gg N$ and $sG \gg 1$. Note that in the following, e_i stands for a host starting with a mixed inoculum, whatever be the initial proportion (it will be the same equation for all).

0

$$\begin{aligned}e_N=&(1-w)(1-q)\exp(-\lambda((1-e_i)(1-p_N^N-(1-p_N)^N)+p_N^N(1-e_N)+(1-p_N)^N(1-e_0)))\\&+(1-w)q\exp(-\lambda(1-e_N))\\&+wq'\exp(-\lambda'(1-e_N))+w(1-q')\exp(-\lambda'(1-p_Ne_N-(1-p_N)e_0)).\end{aligned}$$

$$\begin{aligned}e_i=&(1-w)q\exp(-\lambda(1-e_N))+wq'\exp(-\lambda'(1-e_N))\\&+(1-w)(1-q)\exp\big[-\lambda((1-p_{MSB})^N(1-e_0)+(1-(1-p_{MSB})^N-p_{MSB}^N)(1-e_i)\\&\qquad+p_{MSB}^N(1-e_N)\big]\\&+w(1-q')\exp(-\lambda'((1-p_{MSB})(1-e_0)+p_{MSB}(1-e_N)))\end{aligned}$$

$$\begin{aligned}e_0=&(1-w)q+wq'\\&+(1-w)(1-q)\exp(-\lambda((1-p_0)^N(1-e_0)+(1-(1-p_0)^N-p_0^N)(1-e_i)+p_0^N(1-e_N)))\\&+w(1-q')\exp(-\lambda'((1-p_0)(1-e_0)+p_0(1-e_N)))\end{aligned}$$

We can also write the equations for the survival probability instead:

$$\begin{aligned}1-e_N=&(1-w)(1-q)(1-\exp(-\lambda((1-e_i)(1-p_N^N-(1-p_N)^N)+p_N^N(1-e_N)+(1-p_N)^N(1-e_0))))\\&+(1-w)q(1-\exp(-\lambda(1-e_N)))\\&+wq'(1-\exp(-\lambda'(1-e_N)))+w(1-q')(1-\exp(-\lambda'(p_N(1-e_N)+(1-p_N)(1-e_0)))).\end{aligned}$$

(G.12)

$$\begin{aligned}1-e_i=&(1-w)q(1-\exp(-\lambda(1-e_N)))+wq'(1-\exp(-\lambda'(1-e_N)))\\&+(1-w)(1-q)(1-\exp\big[-\lambda((1-p_{MSB})^N(1-e_0)\\&\qquad+(1-(1-p_{MSB})^N-p_{MSB}^N)(1-e_i)+p_{MSB}^N(1-e_N))\big]\\&+w(1-q')(1-\exp(-\lambda'((1-p_{MSB})(1-e_0)+p_{MSB}(1-e_N))))\end{aligned}\qquad\text{(G.13)}$$

$$\begin{aligned}1-e_0=&(1-w)(1-q)(1-\exp(-\lambda((1-p_0)^N(1-e_0)+(1-(1-p_0)^N-p_0^N)(1-e_i)+p_0^N(1-e_N))))\\&+w(1-q')(1-\exp(-\lambda'((1-p_0)(1-e_0)+p_0(1-e_N))))\end{aligned}$$

(G.14)

In all cases, p_{MSB}^N is extremely small, and $1-p_{MSB} \simeq 1$ (but without further assumptions, $(1-p_{MSB})^N \simeq 1$ is not guaranteed). Since $p_0 \simeq p_{MSB}$, the same

can be said about it. Then equations (G.13) and (G.14) can be rewritten as:

$$\begin{aligned}1-e_i =&(1-w)q(1-\exp(-\lambda(1-e_N)))+wq'(1-\exp(-\lambda'(1-e_N)))\\&+(1-w)(1-q)(1-\exp(-\lambda((1-p_{MSB})^N(1-e_0)+(1-(1-p_{MSB})^N)(1-e_i)))\\&+w(1-q')(1-\exp(-\lambda'(1-e_0+p_{MSB}(1-e_N))))\end{aligned} \tag{G.15}$$

$$\begin{aligned}1-e_0 =&(1-w)(1-q)(1-\exp(-\lambda((1-p_0)^N(1-e_0)+(1-(1-p_0)^N)(1-e_i))))\\&+w(1-q')(1-\exp(-\lambda'(1-e_0+p_0(1-e_N))))\end{aligned} \tag{G.16}$$

As for (G.12), it depends on the assumptions. When $2^{sG}\mu_2/(s\log(2)) \ll 1/N$, i.e. $sG < (-\log(N)+\log(s\log(2))-\log(\mu_2))/\log(2)$, then $(1-p_N)$ is small, as well as $1-p_N^N$. For both naive and immune hosts, then most transmissions will be of resistant bacteria only. Thus

$$1-e_N = 1-(1-w)\exp(-\lambda(1-e_N))-w\exp(-\lambda'(1-e_N)) \tag{G.17}$$

G.2.2 Regime of sure extinction in the absence of mutations

Let us take two further simplifications. If we limit ourselves to the case in which extinction is certain in the absence of mutations, $R_{0,WT} = (1-w)(1-q)\lambda + w(1-q')\lambda' < 1$, then $1-e_0$ is very small, of the order of μ_1. Also, we will assume $Np_{MSB} \ll 1$. Then we can simplify the system.

Starting from (G.15):

$$\begin{aligned}1-e_i =&(1-w)q(1-\exp(-\lambda(1-e_N)))+wq'(1-\exp(-\lambda'(1-e_N)))\\&+(1-w)(1-q)\lambda(1-e_0+Np_{MSB}(1-e_i))\\&+w(1-q')\lambda'(1-e_0+p_{MSB}(1-e_N))\end{aligned}$$

As here we assume $1-e_0$ very small, of the order of μ_1,

$$1-e_i \simeq \frac{(1-w)q(1-e^{-\lambda(1-e_N)})+wq'(1-e^{-\lambda'(1-e_N)})+w(1-q')\lambda' p_{MSB}(1-e_N))}{1-(1-w)(1-q)\lambda N p_{MSB}}.$$

And as we assumed that $Np_{MSB} \ll 1$,

$$1-e_i \simeq (1-w)q(1-\exp(-\lambda(1-e_N)))+wq'(1-\exp(-\lambda'(1-e_N)))+w(1-q')\lambda' p_{MSB}(1-e_N)). \tag{G.18}$$

Starting from equation (G.16):

$$1-e_0 = (1-w)(1-q)\lambda(1-e_0+Np_{MSB}(1-e_i))+w(1-q')\lambda'(1-e_0+p_{MSB}(1-e_N)),$$

leading to:

$$1 - e_0 = p_{MSB} \frac{(1-w)(1-q)\lambda N(1-e_i) + w(1-q')\lambda'(1-e_N)}{1-(1-w)(1-q)\lambda + w(1-q')\lambda'}.$$

Then, replacing $1 - e_i$ by expression (G.18),

$$\begin{aligned} 1 - e_0 \simeq & p_{MSB} \left[\frac{(1-w)(1-q)\lambda N((1-w)q(1-e^{-\lambda(1-e_N)}) + wq'(1-e^{-\lambda'(1-e_N)}))}{1-(1-w)(1-q)\lambda + w(1-q')\lambda'} \right. \\ & \left. + \frac{(1+(1-w)(1-q)\lambda N p_{MSB})w(1-q')\lambda'(1-e_N)}{1-(1-w)(1-q)\lambda + w(1-q')\lambda'} \right] \end{aligned}$$

And as we assumed that $Np_{MSB} \ll 1$,

$$\begin{aligned} 1 - e_0 \simeq & p_{MSB} \left[\frac{(1-w)(1-q)\lambda N((1-w)q(1-\exp(-\lambda(1-e_N)))}{1-(1-w)(1-q)\lambda + w(1-q')\lambda'} \right. \\ & \left. + \frac{wq'(1-\exp(-\lambda'(1-e_N)))) + w(1-q')\lambda'(1-e_N)}{1-(1-w)(1-q)\lambda + w(1-q')\lambda'} \right] \end{aligned} \tag{G.19}$$

As for (G.12), it depends on the assumptions.

Here we suppose When $2^{sG}\mu_2/(s\log(2)) \ll 1/N$, i.e. $sG < (-\log(N) + \log(s\log(2)) - \log(\mu_2))/\log(2)$, then $(1-p_N)$ is small, as well as $1-p_N^N$, and thus (G.17) is unchanged. If $q = q'$, (G.18) can be further simplified to:

$$1 - e_i \simeq q(1-e_N) + w(1-q)\lambda' p_{MSB}(1-e_N) \simeq q(1-e_N) \tag{G.20}$$

and (G.19) can be simplified to:

$$1 - e_0 \simeq p_{MSB}((1-w)(1-q)\lambda N q + w(1-q)\lambda')(1-e_N) \tag{G.21}$$

Thus in this regime, and if $q = q'$, e_N can be found numerically solving (G.17), and e_0 and e_i can be straightforwardly obtained using (G.20) and (G.21). The ratio in the emergence probabilities between the case of a fully naive and a fully immune population is:

$$ratio = \frac{1-e_{0,w=0}}{1-e_{0,w=1}} = \frac{\lambda N q(1-e_{N,w=0})}{\lambda'(1-e_{N,w=1})}$$

If we further assume $\lambda = \lambda'$, it is straightforward from equation (G.17) that $e_{N,w=1} = e_{N,w=0}$, and

$$ratio = \frac{1-e_{0,w=0}}{1-e_{0,w=1}} = Nq$$

Thus, in most cases, immunity decreases the probability of emergence by a factor of at least Nq.

Bibliography

[1] Ron Sender, Shai Fuchs, and Ron Milo. Revised estimates for the number of human and bacteria cells in the body. *PLoS biology*, 14(8):e1002533, 2016. [Cited on pages vii, xix, 9, 83, and 125.]

[2] Bärbel Stecher and Wolf-Dietrich Hardt. The role of microbiota in infectious disease. *Trends in microbiology*, 16(3):107–114, 2008. [Cited on pages vii, xix, 9, 83, and 125.]

[3] Geneva, World Health Organization. Global health estimates 2016: Disease burden by cause, age, sex, by country and by region, 2000-2016. `http://www.who.int/healthinfo/global_burden_disease/estimates/en/index1.html`, 2018. [Cited on pages vii and 9.]

[4] Nathan S McClure and Troy Day. A theoretical examination of the relative importance of evolution management and drug development for managing resistance. *Proceedings of the Royal Society of London B: Biological Sciences*, 281(1797):20141861, 2014. [Cited on pages vii, xxiii, 9, and 125.]

[5] Rasika M Harshey. Bacterial motility on a surface: many ways to a common goal. *Annual review of microbiology*, 57:249–73, January 2003. [Cited on pages vii and 9.]

[6] Philippe Thomen, Jerome Robert, Amaury Monmeyran, Anne Florence Bitbol, Carine Douarche, and Nelly Henry. Bacterial biofilm under flow: First a physical struggle to stay, then a matter of breathing. *PLoS ONE*, 12(4):1–24, 2017. [Cited on pages vii and 9.]

[7] Lydia Robert, Jean Ollion, Jerome Robert, Xiaohu Song, Ivan Matic, and Marina Elez. Mutation dynamics and fitness effects followed in single cells. *Science*, 359(6381):1283–1286, 2018. [Cited on pages vii and 10.]

[8] Roberto Rusconi, Melissa Garren, and Roman Stocker. Microfluidics expanding the frontiers of microbial ecology. *Annual review of biophysics*, 43(March):65–91, 2014. [Cited on pages vii and 10.]

[9] RG Lentle and PWM Janssen. Physical characteristics of digesta and their influence on flow and mixing in the mammalian intestine: a review. *Journal of Comparative Physiology B*, 178(6):673–690, 2008. [Cited on pages viii, 10, 66, 121, 185, and 187.]

[10] Corrin Hulls, Roger G Lentle, Clement de Loubens, Patrick WM Janssen, Paul Chambers, and Kevin J Stafford. Spatiotemporal mapping of ex vivo motility in the caecum of the rabbit. *Journal of Comparative Physiology B*, 182(2):287–297, 2012. [Cited on pages viii, 10, 66, 121, and 187.]

[11] Patrick Kaiser, Médéric Diard, Bärbel Stecher, and Wolf-Dietrich Hardt. The streptomycin mouse model for Salmonella diarrhea: functional analysis of the microbiota, the pathogen's virulence factors, and the host's mucosal immune response. *Immunological reviews*, 245(1):56–83, 2012. [Cited on pages viii, xii, 10, and 21.]

[12] Karoline Faust and Jeroen Raes. Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538–550, 2012. [Cited on pages viii and 10.]

[13] Laura Hindersin and Arne Traulsen. Most Undirected Random Graphs Are Amplifiers of Selection for Birth-Death Dynamics, but Suppressors of Selection for Death-Birth Dynamics. *PLoS Computational Biology*, 11(11):1–14, 2015. [Cited on pages viii and 10.]

[14] Michael Sieber, Lucía Pita, Nancy Weiland-Bräuer, Philipp Dirksen, Jun Wang, Benedikt Mortzfeld, Sören Franzenburg, Ruth A. Schmitz, John F. Baines, Sebastian Fraune, Ute Hentschel, Hinrich Schulenburg, Thomas C. G. Bosch, and Arne Traulsen. The neutral metaorganism. *bioRxiv*, 2018. [Cited on pages viii and 10.]

[15] Alan Perelson and Gérard Weisbuch. Immunology for physicists. *Reviews of Modern Physics*, 69(4):1219–1268. [Cited on pages viii and 11.]

[16] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan. Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences*, 107(12):5405–5410, 2010. [Cited on pages viii and 11.]

[17] Bryan G Yipp, Björn Petri, Davide Salina, Craig N Jenne, Brittney N V Scott, Lori D Zbytnuik, Keir Pittman, Muhammad Asaduzzaman, Kaiyu Wu, H Christopher Meijndert, Stephen E Malawista, Anne de Boisfleury Chevance, Kunyan Zhang, John Conly, and Paul Kubes. Infection-induced NETosis is a dynamic process involving neutrophil multitasking in vivo. *Nature Medicine*, 18(9):1386–1393, 2012. [Cited on pages viii and 11.]

[18] Keira Melican, Ruben M Sandoval, Abdul Kader, Lina Josefsson, George a. Tanner, Bruce a. Molitoris, and Agneta Richter-Dahlfors. Uropathogenic escherichia coli P and type 1 fimbriae act in synergy in a living host to facilitate renal colonization leading to nephron obstruction. *PLoS Pathogens*, 7(2):2–13, 2011. [Cited on pages viii and 11.]

[19] Kathrin Moor, Médéric Diard, Mikael E. Sellin, Boas Felmy, Sandra Y. Wotzka, Albulena Toska, Erik Bakkeren, Markus Arnoldini, Florence

Bansept, Alma Dal Co, Tom Völler, Andrea Minola, Blanca Fernandez-Rodriguez, Gloria Agatic, Sonia Barbieri, Luca Piccoli, Costanza Casiraghi, Davide Corti, Antonio Lanzavecchia, Roland R. Regoes, Claude Loverdo, Roman Stocker, Douglas R. Brumley, Wolf-Dietrich Hardt, and Emma Slack. High-avidity IgA protects the intestine by enchaining growing bacteria. *Nature*, 544(7651):498–502, 2017. [Cited on pages x, xi, xix, xx, xxi, xxvii, xxviii, 12, 13, 41, 83, 85, 86, 87, 88, 89, 90, 91, 94, 97, 99, 115, 120, 126, 127, 129, 147, and 148.]

[20] Florence Bansept, Kathrin Moor-Schumann, Mederic Diard, Wolf-Dietrich Hardt, Emma Wetter Slack, and Claude Loverdo. Enchained growth and cluster dislocation: a possible mechanism for microbiota homeostasis. *bioRxiv*, page 298059, 2018. [Cited on pages x, 12, 93, and 131.]

[21] Marco Gherardi, Alberto Amato, Jean-Pierre Bouly, Soizic Cheminant, Maria Immacolata Ferrante, Maurizio Ribera d'Alcalá, Daniele Iudicone, Angela Falciatore, and Marco Cosentino Lagomarsino. Regulation of chain length in two diatoms as a growth-fragmentation process. *Physical Review E*, 94(2):022418, 2016. [Cited on pages xi, xxviii, 13, 110, and 148.]

[22] Yuriy Pichugin, Jorge Peña, Paul B. Rainey, and Arne Traulsen. Fragmentation modes and the evolution of life cycles. *PLoS Computational Biology*, 13(11):1–20, 2017. [Cited on pages xi, xxviii, 13, and 148.]

[23] Jeremie Guedj, Phillip S. Pang, Jill Denning, Maribel Rodriguez-Torres, Eric Lawitz, William Symonds, and Alan S. Perelson. Analysis of hepatitis C viral kinetics during administration of two nucleotide analogues: Sofosbuvir (GS-7977) and GS-0938. *Antiviral Therapy*, 19(2):211–220, 2014. [Cited on pages xi and 17.]

[24] Alan S Perelson. Modelling viral and immune system dynamics. *Nature reviews. Immunology*, 2(1):28–36. [Cited on pages xi and 17.]

[25] Lisa Maier, Médéric Diard, Mikael E Sellin, Elsa-Sarah Chouffane, Kerstin Trautwein-Weidner, Balamurugan Periaswamy, Emma Slack, Tamas Dolowschiak, Bärbel Stecher, Claude Loverdo, Roland R Regoes, and Wolf-Dietrich Hardt. Granulocytes impose a tight bottleneck upon the gut luminal pathogen population during Salmonella typhimurium colitis. *PLoS pathogens*, 10(12):e1004557, 2014. [Cited on pages xi, xiii, 17, 22, 26, and 41.]

[26] Patrick Kaiser, Emma Slack, Andrew J. Grant, Wolf-Dietrich Hardt, and Roland R. Regoes. Lymph Node Colonization Dynamics after Oral Salmonella Typhimurium Infection in Mice. *PLoS Pathogens*, 9(9):e1003532, 2013. [Cited on pages xi, xiii, 17, and 26.]

[27] Bryan Coburn, Guntram a Grassl, and B B Finlay. Salmonella, the host and disease: a brief review. *Immunology and cell biology*, 85(December 2006):112–118, 2007. [Cited on pages xii and 21.]

[28] Scott Sutton. Accuracy of Plate Counts. *Journal of Validation Technology*, Vol. 17(Issue 3):42–46, 2011. [Cited on pages xii and 23.]

[29] Andrew J Grant, Olivier Restif, Trevelyan J McKinley, Mark Sheppard, Duncan J Maskell, and Pietro Mastroeni. Modelling within-host spatiotemporal dynamics of invasive bacterial disease. *PLoS biology*, 6(4):e74. [Cited on pages xiii, 25, and 26.]

[30] Herbert F Helander and Lars Fändriks. Surface area of the digestive tract–revisited. *Scandinavian journal of gastroenterology*, 49(6):681–689, 2014. [Cited on pages xix and 83.]

[31] Christophe Casteleyn, Anamaria Rekecki, A Van der Aa, Paul Simoens, and W Van den Broeck. Surface area assessment of the murine intestinal tract as a prerequisite for oral dose translation from mouse to man. *Laboratory animals*, 44(3):176–183, 2010. [Cited on pages xix and 83.]

[32] RC Williams and RJ Gibbons. Inhibition of bacterial adherence by secretory immunoglobulin a: a mechanism of antigen disposal. *Science*, 177(4050):697–699, 1972. [Cited on pages xix and 83.]

[33] Richard A Strugnell and Odilia LC Wijburg. The role of secretory antibodies in infection immunity. *Nature Reviews Microbiology*, 8(9):656, 2010. [Cited on pages xix and 83.]

[34] Kathrin Moor, Sandra Y. Wotzka, Albulena Toska, Médéric Diard, Siegfried Hapfelmeier, and Emma Slack. Peracetic acid treatment generates potent inactivated oral vaccines from a broad range of culturable bacterial species. *Frontiers in Immunology*, 7(FEB):1–14, 2016. [Cited on pages xx and 85.]

[35] J Carlet and P Le Coz. Rapport du groupe de travail spécial pour la préservation des antibiotiques. Technical report, Ministère des Affaires sociales, de la Santé et des Droits des femmes, 2015. [Cited on pages xxiii and 126.]

[36] Sécurité Sociale. Projet de loi de financement de la sécurité sociale - annexe 1 : Programmes de qualité et d'efficience - programme de qualité et d'efficience ≪ maladie ≫ - sous-indicateur n ° 9-2. Technical report, Sécurité sociale, 2014. [Cited on pages xxiii and 126.]

[37] David G Kendall et al. On the generalized" birth-and-death" process. *The annals of mathematical statistics*, 19(1):1–15, 1948. [Cited on pages xxv, 26, and 133.]

[38] T E Harris. The Theory of Branching Processes, 1963. [Cited on pages xxv, 27, 133, 134, 177, and 178.]

[39] Kenneth Lange. *Applied probability*. Springer Science & Business Media, 2010. [Cited on pages xxv and 133.]

[40] Robert L. Snipes. Anatomy of the cecum of the laboratory mouse and rat. *Anatomy and Embryology*, 162(4):455–474, 1981. [Cited on pages 22 and 66.]

[41] Jonas Cremer, Igor Segota, Chih-yu Yang, Markus Arnoldini, John T Sauls, Zhongge Zhang, Edgar Gutierrez, Alex Groisman, and Terence Hwa. *Effect of flow and peristaltic mixing on bacterial growth in a gut-like channel*, volume 113. National Acad Sciences, 2016. [Cited on pages 23 and 121.]

[42] Jonas Cremer, Markus Arnoldini, and Terence Hwa. Effect of water flow and chemical environment on microbiota growth and composition in the human colon. *Proceedings of the National Academy of Sciences*, 114(25):6438–6443, 2017. [Cited on pages 23, 83, and 121.]

[43] D Gil and J P Bouché. ColE1-type vectors with fully repressible replication. *Gene*, 105(1):17–22, 1991. [Cited on page 23.]

[44] J Paulsson and M Ehrenberg. Noise in a minimal regulatory network: plasmid copy number control. *Quarterly reviews of biophysics*, 34(1):1–59, 2001. [Cited on page 24.]

[45] Artem S. Novozhilov, Georgy P. Karev, and Eugene V. Koonin. Biological applications of the theory of birth-and-death processes. *Briefings in Bioinformatics*, 7(1):70–85, 2006. [Cited on page 26.]

[46] George. Casella and Roger L. Berger. Statistical inference, 2002. [Cited on pages 26 and 30.]

[47] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81(25):2340–2361, 1977. [Cited on pages 26 and 31.]

[48] Yang Cao, Daniel T. Gillespie, and Linda R. Petzold. Adaptive explicit-implicit tau-leaping method with automatic tau selection. *Journal of Chemical Physics*, 126(22), 2007. [Cited on pages 26 and 32.]

[49] J F. Kenney. Mathematics of Statistics, Part II, 1940. [Cited on pages 26 and 33.]

[50] Lokenath Debnath. *Nonlinear Partial Differential Equations*. 2008. [Cited on page 28.]

[51] C. W. Gardiner. *Handbook of stochastic methods for physics, chemistry and the natural sciences*, volume 13 of *Springer Series in Synergetics*. Springer-Verlag, Berlin, third edition, 2004. [Cited on page 30.]

[52] Hanna Tuomisto. An updated consumer's guide to evenness and related indices. *Oikos*, 121(8):1203–1218, 2012. [Cited on page 45.]

[53] Marcus M Mason. A comparison of the maximal growth rates of various bacteria under optimal conditions. *Journal of bacteriology*, 29(2):103, 1935. [Cited on pages 66 and 129.]

[54] Alexander Sturm, Matthias Heinemann, Markus Arnoldini, Arndt Benecke, Martin Ackermann, Matthias Benz, Jasmine Dormann, and Wolf-Dietrich Hardt. The cost of virulence: retarded growth of Salmonella Typhimurium cells expressing type III secretion system 1. *PLoS pathogens*, 7(7):e1002143, jul 2011. [Cited on page 66.]

[55] Mary K. Stewart and Brad T. Cookson. Non-genetic diversity shapes infectious capacity and host resistance. *Trends in Microbiology*, 20(10):461–466, 2012. [Cited on page 66.]

[56] Larry J Dishaw, John P Cannon, Gary W Litman, and William Parker. Immune-directed support of rich microbial communities in the gut has ancient roots. *Developmental & Comparative Immunology*, 47(1):36–51, 2014. [Cited on page 83.]

[57] Gregory P Donaldson, S Melanie Lee, and Sarkis K Mazmanian. Gut biogeography of the bacterial microbiota. *Nature Reviews Microbiology*, 14(1):20, 2016. [Cited on pages 83 and 125.]

[58] Savannah L Logan, Jacob Thomas, Jinyuan Yan, Ryan P Baker, Drew S Shields, Joao B Xavier, Brian K Hammer, and Raghuveer Parthasarathy. The vibrio cholerae type vi secretion system can modulate host intestinal mechanics to displace gut bacterial symbionts. *Proceedings of the National Academy of Sciences*, 115(16):E3779–E3787, 2018. [Cited on pages 83 and 121.]

[59] World Health Organization & Food and Agriculture Organization. *Risk assessments of Salmonella in eggs and broiler chickens*, volume http://www.who.int/foodsafety/publications/micro/salmonella/en/. 2002. [Cited on pages 86 and 127.]

[60] Kathrin Endt, Bärbel Stecher, Samuel Chaffron, Emma Slack, Nicolas Tchitchek, Arndt Benecke, Laurye Van Maele, Jean-Claude Sirard, Andreas J Mueller, Mathias Heikenwalder, et al. The microbiota mediates pathogen clearance from the gut lumen after non-typhoidal salmonella diarrhea. *PLoS pathogens*, 6(9):e1001097, 2010. [Cited on page 94.]

[61] Keita Kudoh, Jun Shimizu, Aki Ishiyama, Masahiro Wada, Toshichika Takita, Yusuke Kanke, and Satoshi Innami. Secretion and excretion of immunoglobulin a to cecum and feces differ with type of indigestible saccharides. *Journal of nutritional science and vitaminology*, 45(2):173–181, 1999. [Cited on page 94.]

[62] D F Tees, O Coenen, and H L Goldsmith. Interaction forces between red cells agglutinated by antibody. IV. Time and force dependence of breakup. *Biophysical Journal*, 65(3):1318–1334, 1993. [Cited on pages 94, 122, and 131.]

[63] Evan A Evans and David A Calderwood. Forces and bond dynamics in cell adhesion. *Science*, 316(5828):1148–1153, 2007. [Cited on pages 94, 112, 131, 185, and 187.]

[64] ED McGrady and Robert M Ziff. "shattering"transition in fragmentation. *Physical review letters*, 58(9):892, 1987. [Cited on pages 95 and 131.]

[65] JA Odell and A Keller. Flow-induced chain fracture of isolated linear macromolecules in solution. *Journal of Polymer Science Part B: Polymer Physics*, 24(9):1889–1916, 1986. [Cited on pages 95, 112, 131, and 187.]

[66] Per Brandtzaeg. Role of secretory antibodies in the defence against infections. *International Journal of Medical Microbiology*, 293(1):3–15, 2003. [Cited on pages 112 and 185.]

[67] E. Evans, D. Berk, and A. Leung. Detachment of agglutinin-bonded red blood cells. I. Forces to rupture molecular-point attachments. *Biophysical Journal*, 59(4):838–848, 1991. [Cited on pages 112, 122, 185, and 187.]

[68] Gunnar Bratbak and Ian Dundas. Bacterial dry matter content and biomass estimations. *Applied and environmental microbiology*, 48(4):755–757, 1984. [Cited on page 116.]

[69] William W Baldwin, Richard Myer, Nicole Powell, Erika Anderson, and Arthur L Koch. Buoyant density of escherichia coli is determined solely by the osmolarity of the culture medium. *Archives of microbiology*, 164(2):155–157, 1995. [Cited on page 116.]

[70] P W M Janssen, R G Lentle, P Asvarujanon, P Chambers, K J Stafford, and Y Hemar. Characterization of flow and mixing regimes within the ileum of the brushtail possum using residence time distribution analysis with simultaneous spatio-temporal mapping. *The Journal of Physiology*, 582(3):1239–1248, 2007. [Cited on page 121.]

[71] Hyun Jung Kim, Dongeun Huh, Geraldine Hamilton, and Donald E. Ingber. Human gut-on-a-chip inhabited by microbial flora that experiences intestinal peristalsis-like motions and flow. *Lab on a Chip*, 12(12):2165, 2012. [Cited on page 121.]

[72] R G Lentle and C Loubens. A review of mixing and propulsion of chyme in the small intestine: fresh insights from new methods. *Journal of Comparative Physiology B*, pages 369–387, 2015. [Cited on page 121.]

[73] Travis J Wiles, Matthew Jemielita, Ryan P Baker, Brandon H Schlomann, Savannah L Logan, Julia Ganz, Ellie Melancon, Judith S Eisen, Karen Guillemin, and Raghuveer Parthasarathy. Host gut motility promotes competitive exclusion within a model intestinal microbiota. *PLoS biology*, 14(7):e1002517, 2016. [Cited on page 121.]

[74] BM Fournier and CA Parkos. The role of neutrophils during intestinal inflammation. *Mucosal immunology*, 5(4):354, 2012. [Cited on page 122.]

[75] Julie Mirpuri, Megan Raetz, Carolyn R Sturge, Cara L Wilhelm, Alicia Benson, Rashmin C Savani, Lora V Hooper, and Felix Yarovinsky. Proteobacteria-specific iga regulates maturation of the intestinal microbiota. *Gut microbes*, 5(1):28–39, 2014. [Cited on page 122.]

[76] Lee K Richman and William R Brown. Immunochemical characterization of igm in human intestinal fluids. *The Journal of Immunology*, 119(4):1515–1519, 1977. [Cited on page 122.]

[77] Ianko D Iankov, Dragomir P Petrov, Ivan V Mladenov, Iana H Haralambieva, and Ivan G Mitov. Lipopolysaccharide-specific but not anti-flagellar immunoglobulin a monoclonal antibodies prevent salmonella enterica serotype enteritidis invasion and replication within hep-2 cell monolayers. *Infection and immunity*, 70(3):1615–1618, 2002. [Cited on page 122.]

[78] Médéric Diard, Victor Garcia, Lisa Maier, Mitja NP Remus-Emsermann, Roland R Regoes, Martin Ackermann, and Wolf-Dietrich Hardt. Stabilization of cooperative virulence by the expression of an avirulent phenotype. *Nature*, 494(7437):353, 2013. [Cited on page 122.]

[79] Carmen Dahms, Nils-Olaf Hübner, Florian Wilke, and Axel Kramer. Minireview: Epidemiology and zoonotic potential of multiresistant bacteria and clostridium difficile in livestock and food. *GMS hygiene and infection control*, 9(3), 2014. [Cited on page 126.]

[80] Timothy F Landers, Bevin Cohen, Thomas E Wittum, and Elaine L Larson. A review of antibiotic use in food animals: perspective, policy, and potential. *Public health reports*, 127(1):4–22, 2012. [Cited on page 126.]

[81] Ewan M Harrison, Francesc Coll, Michelle S Toleman, Beth Blane, Nicholas M Brown, M Estee Török, Julian Parkhill, and Sharon J Peacock. Genomic surveillance reveals low prevalence of livestock-associated methicillin-resistant staphylococcus aureus in the east of england. *Scientific reports*, 7(1):7406, 2017. [Cited on page 126.]

[82] Médéric Diard, Erik Bakkeren, Jeffrey K Cornuault, Kathrin Moor, Annika Hausmann, Mikael E Sellin, Claude Loverdo, Abram Aertsen, Martin Ackermann, Marianne De Paepe, Emma Slack, and Wolf-Dietrich Hardt. Inflammation boosts bacteriophage transfer between salmonella spp. *Science*, 355(6330):1211–1215, 2017. [Cited on page 126.]

[83] B. R. Levin, V. Perrot, and N. Walker. Compensatory mutations, antibiotic resistance and the population genetics of adaptive evolution in bacteria. *Genetics*, 154(3):985–997, Mar 2000. [Cited on page 127.]

[84] D. I. Andersson and D. Hughes. Antibiotic resistance and its cost: is it possible to reverse resistance? *Nat. Rev. Microbiol.*, 8:260–271, 2010. [Cited on page 127.]

[85] P. A. zur Wiesch, R. Kouyos, J. Engelstadter, R. R. Regoes, and S. Bonhoeffer. Population biological principles of drug-resistance evolution in infectious diseases. *Lancet Infect Dis*, 11(3):236–247, Mar 2011. [Cited on page 127.]

[86] J. Moura de Sousa, A. Sousa, C. Bourgard, and I. Gordo. Potential for adaptation overrides cost of resistance. *Future Microbiol*, 10(9):1415–1431, 2015. [Cited on page 127.]

[87] W. Paulander, S. Maisnier-Patin, and D. I. Andersson. Multiple mechanisms to ameliorate the fitness burden of mupirocin resistance in Salmonella typhimurium. *Mol. Microbiol.*, 64(4):1038–1048, May 2007. [Cited on page 127.]

[88] M Park, C Loverdo, S J Schreiber, and J O Lloyd-Smith. Multiple scales of selection influence the evolutionary emergence of novel pathogens. *Philosophical Transactions B*, page 20120333, 2013. [Cited on page 127.]

[89] SJ Schreiber, R Ke, C Loverdo, M Park, P Ahsan, and JO Lloyd-Smith. Crossscale dynamics and the evolutionary emergence of infectious diseases. biorxiv. 2016. [Cited on pages 127 and 133.]

[90] Michael Lynch. Evolution of the mutation rate. *TRENDS in Genetics*, 26(8):345–352, 2010. [Cited on page 127.]

[91] Cameron Myhrvold, Jonathan W Kotula, Wade M Hicks, Nicholas J Conway, and Pamela A Silver. A distributed cell division counter reveals growth dynamics in the gut microbiota. *Nature communications*, 6:10039, 2015. [Cited on page 129.]

[92] Rustom Antia, Roland R Regoes, Jacob C Koella, and Carl T Bergstrom. The role of evolution in the emergence of infectious diseases. *Nature*, 426:658–661, 2003. [Cited on page 133.]

[93] Yoh Iwasa, Franziska Michor, and Martin A Nowak. Evolutionary dynamics of invasion and escape. *Journal of Theoretical Biology*, 226(2):205–214, 2004. [Cited on page 133.]

[94] Jean-Baptiste André and Troy Day. The Effect of Disease Life History on the Evolutionary Emergence of Novel Pathogens. *Proceedings of the Royal Society B: Biological Sciences*, 272:1949–1956, 2005. [Cited on page 133.]

[95] HA Orr and RL Unckless. Population extinction and the genetics of adaptation. *The American Naturalist*, 172(2):160–9, 2008. [Cited on page 133.]

[96] C Loverdo, M Park, S J Schreiber, and J O Lloyd-Smith. Influence of viral replication mechanisms on within-host evolutionary dynamics. *Evolution*, 66:3462–3471, 2012. [Cited on page 143.]

Florence Bansept **5 Décembre 2018**

Sujet : Modélisation biophysique des dynamiques d'une population bactérienne et de la réponse immunitaire dans les intestins

Résumé : La première partie de cette thèse porte sur les dynamiques de colonisation d'une population bactérienne au début d'une infection intestinale. Le but est de déduire des paramètres biologiquement pertinents de données indirectes. Un modèle simple est étudié, et l'on discute de l'observable optimale pour caractériser la variabilité d'une distribution d'étiquettes génétiques. Des arguments biologiques et des incohérences entre des observables expérimentales avec le premier modèle motivent l'étude d'un second, où deux sous-populations se répliquent à des taux différents, mais on ne peut pas conclure clairement sur le jeu de données utilisé. La seconde partie porte sur les mécanismes de la réponse immunitaire. Le principal effecteur du système immunitaire adaptatif dans l'intestin, l'IgA (un anticorps), enchaîne les bactéries-filles en agrégats clonaux lors de la réplication. Nous avons contribué à prouver ce phénomène par un modèle qui prédit la réduction de la diversité bactérienne qui en découle. Au sein de l'hôte, l'interaction entre la croissance et la fragmentation des agrégats a pour conséquence le piégeage préférentiel des bactéries à croissance rapide, ce qui pourrait permettre au système immunitaire de réguler la composition du microbiote. A l'échelle de la population-hôte, et dans le contexte de l'évolution d'une résistance aux antibiotiques, si les bactéries sont transmises sous forme d'amas clonaux, alors la probabilité de transmettre une bactérie résistante est réduite dans une population immunisée. Ainsi, des outils de physique statistique nous permettent d'identifier des mécanismes génériques en biologie.

Mots clés : modélisation biophysique, dynamiques de population, processus stochastique, mécanismes physiques de la réponse immunitaire, infection bactérienne

Subject : Biophysical modeling of bacterial population dynamics and the immune response in the gut

Abstract: The first part of this thesis focuses on the colonization dynamics of a bacterial population in early infection of the gut. The aim is to infer biologically relevant parameters from indirect data. We discuss the optimal observable to characterize the variability in genetic tags distributions. In a first one-population model, biological arguments and inconsistencies between several experimental observables lead to the study of a second model with two-subpopulations replicating at different rates. As expected, this model allows for broader possibilities in observables combination, even though no clear conclusion can be drawn as to a data set on Salmonella in mice. The second part concerns the mechanisms that make the immune response effective. The main effector of the immune system in the gut, IgA (an antibody), enchains daughter bacteria in clonal clusters upon replication. Our model predicting the ensuing reduction of diversity in the bacterial population contributes to evidence this phenomenon, called "enchained growth". Inside the host, the interplay of cluster growth and fragmentation results in preferentially trapping faster-growing and potentially noxious bacteria away from the epithelium, which could be a way for the immune system to regulate the microbiota composition. At the scale of the hosts population, in the context of evolution of antibiotic resistance, if bacteria are transmitted via clonal clusters, the probability to transmit a resistant bacteria is reduced in immune populations. Thus we use statistical physics tools to identify some generic mechanisms in biology.

Keywords : biophysical modeling, population dynamics, stochastic processes, physical mechanisms of the immune response, bacterial infection