Fluctuations in cell lineages and population trees : a thermodynamic perspective

Arthur Genthon

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'ESPCI Paris

Fluctuations in cell lineages and population trees: a thermodynamic perspective

Fluctuations dans les lignées et populations de cellules : un point de vue thermodynamique

Soutenue par
Arthur GENTHON
Le 14 Octobre 2022

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

Spécialité
Physique

Composition du jury :

Erik **Aurell** Prof., KTH Royal Institute of Technology	*Président/Rapporteur*
Philipp **Thomas** Sen. Lect., Imperial College	*Rapporteur*
Marie **Doumic** Ingénieure en Chef des Ponts et Forêts, INRIA & Sorbonne Université	*Examinatrice*
Reinaldo **García-García** Assoc. Prof., University of Navarra	*Examinateur*
Edo **Kussell** Prof., New York University	*Examinateur*
Claude **Loverdo** CR, Sorbonne Université	*Examinatrice*
David **Lacoste** DR, ESPCI	*Directeur de thèse*
Jérémie **Unterberger** MCF, Université de Lorraine	*Codir. de thèse*

ESPCI PARIS | PSL

Abstract

Experiments on growing cells can be carried out either in bulk or in confined geometries that constrain the growth of the colony. How should the population trees be sampled in each setup? Are there statistical biases between them? How to quantify natural selection in these trees? These are the main questions we address in this thesis.

In a first part, we study the statistical bias between the single-lineage and population levels, which has similarities with fluctuation theorems in stochastic thermodynamics. To do so, we develop a theoretical framework based on lineage histories within population trees, and obtain universal constraints that are exploited in two directions.

First, this bias informs on the strength of selection, that quantifies the correlations between the value of a cell trait and the reproductive success of the lineage. This selection results from the variability of lineages in the population, which we analyze using linear response theory. We also extend our framework to allow situations where lineages end before the end of the experiment, due to cell death or dilution. We show how dead lineages should be taken into account in the statistics, and how death impacts the phenotypic variability and therefore the strength of selection.

Second, we show how single-lineage data can be used to infer population-level quantities like the population growth rate, also called Malthus parameter. Focusing on size-regulated populations, we derive steady-state cell size distributions for single-lineage experiments, that can also be used to infer cell cycle parameters such as the single cell elongation rate and the asymmetry of division. In addition, we explore how the lineage-population bias for size statistics is affected by different sources of stochasticity.

In a second independent part, we propose a thermodynamic description of cell growth and division using simple coarse-grained models of cell size control. This question is important to understand how cell colonies are constrained by thermodynamics. Using a decomposition of cell division in two sub-processes: branching (creation of an identical new cell), and resetting (restart of the properties of the two cells), we derive the first and second laws of thermodynamics for a colony of cells, and identify the contribution of each process to the change in average energy and Shannon entropy. This allows us to understand how the distributions of age and size are affected by cell division from an information-theoretic point of view.

Keywords: Cell populations, cell lineages, natural selection, cell size distribution, stochastic thermodynamics, fluctuation theorems.

Résumé

Au cours des dix dernières années, de nouveaux dispositifs microfluidiques ont été conçus pour suivre des lignées de cellules uniques, et des lignées de cellules au sein de populations constantes en géométrie confinée. Ces dispositifs expérimentaux ont permis de surmonter certaines difficultés rencontrées dans les expériences classiques où les populations croissent librement, notamment en rendant possible l'observation de lignées sur de nombreuses générations. Notre travail trouve son origine dans ces expériences : comment utiliser les données de lignées de cellules, maintenant disponibles en grande quantité ? En géométrie confinée, une difficulté supplémentaire intervient, car il faut comprendre comment prendre en compte dans les statistiques les cellules qui sont continuellement expulsées pour maintenir la population constante.

Historiquement, ce nouveau type de données a mené à des avancées importantes dans les domaines du contrôle de la taille des cellules, et dans l'exploration du lien entre l'échelle la cellule unique et celle de la population. En parallèle, de nouvelles technologies ont été développées pour suivre les lignées au sein de populations en croissance, permettant ainsi de reconstruire la généalogie des populations. Ces avancées ont inspiré de nouvelles définitions de la sélection naturelle basées sur les données temporelles des lignées, et en particulier sur le biais entre le niveau de la lignée unique et celui de la population.

Au cours de cette thèse, nous abordons les questions suivantes. Comment relier les niveaux de la cellule unique et de la population ? Comment échantillonner les populations dans chaque configuration expérimentale ? Comment définir et quantifier la sélection naturelle ? Comment utiliser les données de lignées uniques en pratique pour inférer des quantités au niveau de la population ?

Dans les deux premiers chapitres, nous cherchons des résultats universels valables pour n'importe quel arbre branchant représentant une population, indépendamment de sa dynamique. De cette façon, les éléments de réponse que nous apportons sont pertinents à la fois dans le contexte du contrôle de la taille des cellules et dans celui de l'évolution. Nous analysons ces réponses d'un point de vue conceptuel, en montrant par exemple le lien fondamental entre fluctuations et sélection, mais aussi d'un point de vue pratique, en proposant des méthodes d'inférence.

Dans les deux derniers chapitres, nous adoptons un autre point de vue, basé sur des modèles particuliers de contrôle de la taille, et nous posons les questions suivantes. Comment utiliser les données en lignée unique pour inférer les lois de la croissance et de la division cellulaire ? Quelles sont les limites thermodynamiques imposées à la croissance et à la division cellulaire ? La recherche de principes universels et l'approche de modélisation sont complémentaires, car les modèles fournissent des prédictions testables qui peuvent valider ou invalider des hypothèses et fournissent des descriptions quantitatives de systèmes spécifiques.

Chapitre 1: Le biais lignée-population

Dans ce chapitre, nous étudions les biais statistiques entre l'échelle de la cellule unique et celle de la population. Pour cela, nous utilisons et développons le formalisme proposé dans [Nozoe et al., 2017] T. Nozoe et al. (2017). Inferring fitness landscapes and selection on phenotypic states from single-cell genealogical data. *PLoS Genetics* 13.(3), e1006653, qui repose sur deux différents échantillonnages des lignées d'une population. L'échantillonnage rétrospectif (backward) consiste à attribuer un poids égal à chaque lignée, ce qui conduit à une sur-représentation des lignées qui se sont beaucoup divisées et qui ont donc généré beaucoup de descendants. Pour compenser cet effet, le poids de chaque lignée peut être adapté pour tenir compte du nombre de divisions le long de cette lignée, c'est l'échantillonnage chronologique (forward). Dans ce dernier, les deux cellules filles issues d'une même division se voient attribuer le même poids statistique, indépendamment de leur succès reproductif futur. Cette pondération est en fait la même que celle qui émerge naturellement dans les expériences en lignée unique, telle que la mother-machine, où une seule cellule est suivie au moment de la division. Par conséquent, la statistique obtenue en échantillonnant de façon chronologique une population reproduit la statistique de lignée unique, et repésente donc un moyen d'"annuler" la sélection naturelle.

Nous montrons que le biais entre les probabilités chronologique et rétrospective de choisir aléatoirement une lignée avec K divisions a la même forme que les théorèmes de fluctuation en thermodynamique stochastique, qui comparent habituellement les probabilités de courant entre une expérience de référence et une expérience où le protocole expérimental est inversé dans le temps. En nous appuyant sur nos connaissances en thermodynamique stochastique, nous obtenons deux conséquences de ce "théorème de fluctuation" en dynamique des populations. Premièrement, nous dérivons deux inégalités entre le taux de croissance de la population et le nombre moyen de divisions dans chaque échantillonnage. Dans la limite des temps longs, les inégalités se transforment en inégalités entre les temps moyens inter-divisions et le temps de doublement de la population. Ce dernier résultat généralise deux inégalités bien connues pour les modèles en âge, c'est-à-dire les modèles où la division cellulaire est contrôlée par l'âge des cellules, mais dans notre seule l'hypothèse d'un régime stationnaire de croissance exponentielle aux temps longs a été nécessaire, et notre résultat est donc valable pour tous les modèles de contrôle de la taille. Deuxièmement, ce biais peut aussi être utilisé à notre avantage. Nous construisons un estimateur du taux de croissance de la population à partir de données en lignée unique. Cet estimateur repose uniquement sur la statistique du nombre de divisions, et nous testons sa convergence avec des données en mother-machine. Finalement, nous montrons que le théorème de fluctuation, indépendant du modèle dynamique, implique les équations de Powell et Euler-Lotka lorsque l'on considère un modèle en âge.

Ce même biais est au cœur de la définition de la force de sélection proposée dans Nozoe et al., 2017, qui mesure une distance entre les distributions issues des échantillonnages chronologique et rétrospectif. Quantifier la sélection est une étape importante pour comprendre l'évolution d'une population, et depuis les résultats de Fisher dans les années 1930, de nombreuses mesures se sont succédées. Dans les différentes approches dont nous donnons un bref aperçu dans le texte, les mesures de sélection sont reliées à la variabilité de fitness au sein de la population, où fitness prend des sens différents pour chaque

approche. Ces résultats ont une forme similaire aux théorèmes de fluctuation-dissipation où la réponse d'un système a une perturbation est proportionnelle aux fluctuations du système au repos. La force de sélection considérée ici a l'avantage d'être très générale puisqu'applicable à tout arbre de population, et pour cette raison nous cherchons des contraintes universelles sous la forme de relation de fluctuation-dissipation. Nous obtenons des bornes supérieures et inférieures pour la force de sélection, qui impliquent la variance de fitness.

Chapitre 2: Statistiques des populations avec mort

Dans le chapitre précédent, les échantillonnages chronologique et rétrospectif reposent sur la survie de toutes les lignées jusqu'à la fin de l'expérience. Cependant, dans de nombreuses situations ce n'est pas vérifié. Par exemple, les cellules peuvent mourir lors d'expériences en population croissante, pour diverses raisons : fluctuations d'environnement, réaction à des antibiotiques, accumulation de protéines délétères, ... Egalement dans les expériences en population constante, les cellules évacuées par dilution donnent lieu à des lignées qui prennent fin en cours d'expérience. Dans ces cas-là, et plus généralement pour tout arbre branchant impliquant des lignées tronquées (que l'on appelle mortes au sens large quelle que soit la cause), les deux échantillonnages doivent être adaptés.

Dans ce chapitre, nous modifions les échantillonnages chronologique et rétrospectif pour tenir compte des lignées mortes. En prenant une photo de la population à un instant t, seules les cellules survivantes apparaissent, alors dans l'échantillonnage rétrospectif nous attribuons un poids égal à ces cellules, et un poids nul aux lignées mortes avant l'instant t. A l'inverse, en suivant les lignées à partir de l'instant initial et en choisissant une des deux cellules filles avec une probabilité égale à chaque division, toutes les lignées, mortes et vivantes, sont échantillonnées. Par conséquent, les lignées mortes ne sont échantillonnées que dans une des deux procédures et le théorème de fluctuation se généralise aux lignées vivantes. Ce théorème fait apparaître un nouveau terme : la probabilité de survie dans l'échantillonnage chronologique, qui agit comme un facteur de renormalisation. Lorsqu'un certain modèle est considéré pour décrire la dynamique des cellules, nous montrons que cette probabilité de survie est simplement reliée au taux de mort. Les résultats du premier chapitre sont généralisés, en particulier les inégalités entre taux de croissance et nombre de divisions moyen sont vérifiées avec des données en population constante. De même, les équations de Powell et Euler-Lotka pour les modèles en âge sont étendues aux modèles en âge avec corrélations mère-fille et avec mort.

Lorsque la mort affecte les cellules d'un phénotype plutôt qu'un autre, un nouveau biais statistique apparaît : le biais du survivant. Nous obtenons des formules explicites pour exprimer ce biais comme le rapport des probabilités d'observer une trajectoire phénotypique dans des expériences avec et sans mort. La force de sélection résulte donc maintenant à la fois de la sélection présente en absence de mort, en raison de l'avantage reproductif de certains phénotypes, et du biais du survivant. Nous proposons donc une mesure de l'effet de la mort sur la sélection, dont le signe indique si la mort augmente ou diminue la distance entre les distributions chronologique et rétrospective, ou ne la change pas. Avec un

modèle simple à deux états, nous illustrons les différents signes possibles, qui dépendent uniquement des taux de reproduction et les taux de mort de chaque phénotype.

Chapitre 3: Distribution de taille des cellules

La sur-représentation dans la statistique en population (rétrospectif) comparée à la statistique en lignée unique (chronologique) des lignées avec plus de divisions que la moyenne a été étudiée dans les deux premiers chapitres. Si un trait cellulaire, compris au sens large comme une propriété des cellules, est corrélé avec le nombre de divisions le long d'une lignée, alors les valeurs de ce trait associées aux lignées avec beaucoup de divisons seront selectionnées. Par exemple, si la taille des cellules croit de façon déterministe et exponentielle entre les divisions, et que le deux cellules filles héritent chacune exactement de la moitié du volume de la cellule mère à la division, alors la taille à un instant t est une fonction déterministe de la taille initiale et du nombre de divisions avant l'instant t. Les cellules de petites tailles sont donc sur-représentées en population comparée à la statistique lignée unique.

Dans ce chapitre, nous étudions les distributions de taille à l'équilibre pour des cellules régulées en taille, c'est-à-dire où seule la taille contrôle la division. L'objectif est double : (i) obtenir des distributions analytiques pour la statistique de lignées uniques, qui sont comparées à des données expérimentales pour valider ou invalider le modèle et les hypothèses, ainsi que pout inférer les lois de la croissance et la division. (ii) Comparer ces distributions de taille pour les expériences en lignées uniques à celle connues dans la littérature mathématique pour les expériences de population en croissance, et ainsi dériver le biais lignée-population pour la statistique de taille. Nous nous intéressons en particulier à l'influence des différentes sources de bruit sur ce biais.

Lorsque la répartition de volume à la division et la croissance de la cellule unique sont déterministes, nous obtenons des distributions analytiques sous forme de séries. Ces distributions sont en bon accord avec les distributions expérimentales de taille pour E. coli en mother-machine. De cet accord, nous déduisons des estimations de paramètres tels que l'asymétrie de la division, le taux de croissance de la cellule unique ou la force du contrôle de la taille. Quand la répartition de volume est stochastique, nous dérivons le comportement de la distribution dans les limites des grandes et petites tailles. Aux grandes tailles, la distribution ne dépend plus de noyau de division, alors qu'aux petites tailles la distribution ne dépend plus de la force du contrôle sur la taille. Ces comportements asymptotiques sont comparés à ceux connus en population, et l'influence des différents paramètres sur le biais lignée-population est rendu explicite. Ces résultats asymptotiques sont aussi valides pour le mécanisme de adder, où la division est contrôlée par l'incrément de taille depuis la naissance, et non par la valeur absolue de la taille comme pour le sizer. En effet, aux grandes tailles l'incrément de taille et la taille deviennent équivalents, et aux petites tailles la distribution est indépendante du mécanisme de contrôle de la taille. Finalement, lorsque l'on introduit du bruit diffusif sur la croissance de la cellule unique autour de la tendance exponentielle, le comportement aux grandes tailles est obtenu pour la lignée unique et pour la population, et la comparaison entre les deux révèle que ce type de bruit annule le biais lignée-population.

Chapitre 4: Thermodynamique stochastique de la croissance et de la division cellulaires

Comme tout système biologique, les populations de cellules sont régies par les lois de la thermodynamique. La théorie thermodynamique renseigne sur les transformations qui sont interdites, et impose des compromis entre différentes propriétés des processus, comme l'efficacité, la puissance, la dissipation ou la précision. Dans ce chapitre, nous cherchons à identifier les limites d'origine thermodynamique pour la croissance et la division cellulaires. Cette question est délicate car la division cellulaire est un processus absolument irréversible, au sens où le processus inversé dans le temps, la fusion de deux cellules, n'est jamais observé. Or la comparaison entre la probabilité d'un processus et celle du processus inverse est à la base de la définition de la création d'entropie en thermodynamique stochastique. De fait, le formalisme habituel doit être adapté pour traiter le cas des colonies de cellules.

Pour progresser, nous proposons donc une description thermodynamique de la croissance et de la division cellulaires basée sur la décomposition de la division en deux sous-processus simultanés : le branchement (création d'une nouvelle cellule identique), et la réinitialisation (resetting) (modification des propriétés des deux cellules). Notre description repose sur des modèles simples de contrôle de la taille, tels que le sizer et le timer, où la division est contrôlée par la taille et l'âge respectivement; mais aussi le adder où la division est contrôlée par l'incrément de volume depuis la naissance. Pour le sizer par exemple, le branchement correspond à la création d'une nouvelle cellule de taille x identique à la cellule qui se divise, et les deux cellules voient leur tailles instantanément réduites à $x/2$ (pour une division symétrique), c'est le resetting.

Nous dérivons les deux lois de la thermodynamique pour une colonie de cellules, et nous identifions la contribution de chaque sous-processus au changement d'énergie moyenne et d'entropie de Shannon. La seconde loi est un purement un résultat de théorie de l'information, qui permet de quantifier l'impact de la division sur les distributions de taille ou d'âge. En utilisant des hypothèses raisonnables sur le comportement du taux de division, qui sont justifiées par les expériences, les signes de ces contributions sont obtenus ce qui permet une analogie avec les machines thermiques. Pour le sizer comme pour le timer, le resetting apparaît alors comme le processus d'entrée (input/driving) et le branchement comme processus de sortie (output). Cela peut paraître intuitif si on pense que la prolifération des cellules est l'"objectif" d'une population, mais que cela ne peut se faire qu'au prix d'une réduction de la taille des cellules (resetting). A partir de cette analogie, nous définissons l'efficacité de la division cellulaire comme le rapport du taux de production d'entropie dû au branchement avec celui dû au resetting. L'influence des différents paramètres des modèles sur cette efficacité est étudiée. Finalement, nous étendons ce formalisme à des modèles à n variables, ce qui permet de décrire des modèles plus complexes que les mécanismes de contrôle de la taille à une ou deux variables étudiés avant.

Remerciements / Acknowledgments

Je tiens à remercier tous ceux qui m'ont accompagné, soutenu et encouragé pendant ces trois années de thèse, et plus généralement pendant toute ma scolarité.

Tout d'abord merci à vous David. Merci pour votre disponibilité et votre implication, que tous les directeurs de thèse n'ont pas. Je vous suis reconnaissant de m'avoir fait confiance pour ce projet qui était un saut dans l'inconnu pour nous deux, et de m'avoir laissé choisir ma route au gré des découvertes. Vous ne comptez pas vos heures et avez le même enthousiasme qu'aux débuts (j'imagine !), et pour ça je vous souhaite d'encadrer de nouvelles thèses à l'avenir.

Merci à Reinaldo, sans qui ce projet n'aurait pas existé. Par ta patience et l'aide précieuse que tu m'as apportée pendant mon stage de M2, tu as permis de lancer cette thèse sur de bons rails. Ça a été un plaisir de poursuivre cette collaboration pendant la thèse malgré la distance, et j'espère que nos projets en cours prendront forme un jour !

Que seraient trois ans de thèse sans un bon 'partner in crime'? Merci Gabin d'avoir partagé nos moments de joie, de doute et de frustration qui accompagnent la recherche, et toujours avec humour ! Je te souhaite de trouver un poste permanent dans les dix ans ;)

Ce travail de recherche est aussi le résultat d'un ensemble de rencontres, et bien que la pandémie ait fortement limité les interactions 'réelles', certaines personnes ont pris part à cette aventure et je les en remercie maintenant. Merci à Jérémie, mon co-directeur de thèse, pour son intérêt pour mon travail et sa précieuse aide en mathématiques, aux membres de mon comité de suivi Lydia Robert et Ken Sekimoto, et à Luca pour avoir rejoint le projet avec enthousiasme.

Le bon déroulement de la thèse, tant au laboratoire qu'à l'extérieur lors de voyages professionnels, tient à une équipe de personnels administratifs efficace, et pour ça, merci à Elisa Silveira, Fée Sorrentino, Hyo-Jin Cho et David Noël.

Ces trois (en fait quatre) années ont aussi été synonymes d'aventure humaine puisque c'est avec mes compères de la Botulus Triplex que j'ai passé la quasi-totalité de mon temps en dehors du laboratoire. Merci Thomas, Théo et Clément d'avoir fait de notre maison un lieu chaleureux, ouvert, où on ne s'est jamais ennuyé !

Arriver jusqu'à la thèse est aussi un parcours, l'aboutissement d'années d'études qui ont forgé mon caractère et dessiné mes intérêts scientifiques. J'aimerais alors remercier tous les professeurs qui ont entretenu mon goût pour la science depuis tout jeune. En particulier merci à vous, Gérard Rousseau, Laurent Poirier, Marc Pascaud et Christian Devanz.

Merci du fond du cœur à mes parents et à Marie, qui m'ont toujours encouragé, ont toujours cru en moi et ont su créer un environnement dans lequel j'ai pu m'épanouir personnellement et professionnellement. Sans la bienveillance de ses parents, il est bien plus difficile d'arriver au bout de neuf ans d'études.

Finally, I am really grateful to all the members of the jury, and especially to Philipp Thomas and Erik Aurell who accepted to evaluate my work.

Nomenclature

	Cells
x	Size
x_b	Birth size
Δ	Added volume since birth
a	Age
$\mathcal{S}$	General cell trait
s	Value of trait $\mathcal{S}$
τ	Generation time
K	Number of divisions
	Populations
m	Number of daughter cells produced at division
N_0	Initial number of cells
$N(t)$	Number of cells at time t
$n(y,t)$	Number of cells with property y at time t
$\Lambda_p(t)$	Instantaneous population growth rate
Λ_t	Population growth rate
Λ	Steady-state population growth rate, or Malthus parameter
h_t	Fitness landscape
$\Pi_{\mathcal{S}}$	Strength of selection acting on trait $\mathcal{S}$
	Probability distributions
p_{for} / ϕ	Forward probability
p_{back} / ψ / p	Backward probability
f	Distribution of generation times
ρ_{nb}	Distribution at birth
ρ_{d}	Distribution at division
$\langle\cdot\rangle_{\text{p}}$	Average with respect to distribution p
	Rates and coefficients
r	Division rate per unit time
ζ	Division rate per unit size
ν	Single cell growth/elongation rate
Σ	Partitioning kernel
b	Homogeneous partitioning kernel
α	Strength of the control
β	Growth law exponent
	Death-related quantities
$\Gamma_p(t)$	Instantaneous decrease rate of the forward survival probability
Γ_t	Decrease rate of the forward survival probability
Γ	Steady-state decrease rate of the forward survival probability
γ	Death/dilution rate
X°	X in the absence of death
$p^\star_{\text{for}}$	Forward distribution conditioned on survival

List of publications

In preparation

[6] A. Genthon, L. Peliti, T. Nozoe, D. Lacoste,
Sampling lineage trees with death

Submitted

[5] A. Genthon,
Analytical cell size distribution: lineage-population bias and parameter inference,
arXiv:2206.06146 (2022)†

Published

[4] A. Genthon, R. García-García, D. Lacoste,
Branching processes with resetting as a model for cell division,
Journal of Physics A: Mathematical and Theoretical 55, 074001 (2022)

[3] A. Genthon, D. Lacoste,
Universal constraints on selection strength in lineage trees,
Physical Review Research 3, 023187 (2021)

[2] A. Genthon, D. Lacoste,
Fluctuation relations and fitness landscapes of growing cell populations,
Scientific Reports 10, 11889 (2020)

[1] R. García-García, A. Genthon, D. Lacoste,
Linking lineage and population observables in biological branching processes,
Physical Review E 99, 042413 (2019)

†Since the defence of this PhD thesis, this preprint has been published as:
[Genthon, 2022] A. Genthon (2022). Analytical cell size distribution: lineage-population bias and parameter inference. *Journal of The Royal Society Interface* 19.(196), p. 20220405.

Contents

Chapter 1

Introductory chapter

Contents

1 Motivation and outline of the thesis

In the past decade, new microfluidic devices have been designed to follow either single lineages of cells, like the mother machine (Wang et al., 2010), or lineages of cells within finite populations grown in confined geometries (Hashimoto et al., 2016). These experimental setups helped overcoming some difficulties encountered in classical experiments where populations are freely growing in bulk, namely they allowed to follow lineages of cells for many generations, which is difficult in bulk because of the exploding size of the population. At the origin of our work is a question deeply rooted in experiments: how to use the very large amount of lineage data that is now available?

Historically, these new lineage data led to important advances in the field of cell size control (Robert et al., 2014; Amir, 2014). For example, taking advantage of the large and reliable statistics on age and volume obtained in mother-machines, the adder mechanism, that postulates that cells add a certain volume to their birth volume before dividing, has been tested and largely accepted to describe E. coli and other species (Taheri-Araghi et al., 2015). Mother machine data have also been used to probe with unprecedented precision the stochasticity of the cell growth and division processes. Given that the growth of a population of cells is a deterministic process in the long-time limit, with a smooth exponential growth, a new field of research emerged to link the two levels of description (Lin et al., 2017; Jafarpour et al., 2018).

In parallel, information on lineages within growing populations have become available in bulk thanks to time-lapse video microscopy (Stewart et al., 2005). These temporal information allow to reconstruct the population tree with the genealogy of each cell, which led to new perspectives to define natural selection in a population (Leibler et al., 2010; Nozoe et al., 2017). In Nozoe et al., 2017, a strength of selection based on the comparison between two different samplings of the lineages within a population tree is proposed. The backward sampling leads to the classical population statistics with equal weight on the lineages, where natural selection results from the variability between the cell lineages and leads to the over-representation of the fittest phenotypes. On the other hand, the forward sampling cancels selection by weighting daughter cells born from the same division with uniform weight regardless of their future reproductive success, that is the number of offspring they generate at later times.

At the beginning of my PhD we were interested in the questions of cell size control and the biases between the single cell and population levels. It was only later that we discovered the literature on selection and evolution, and understood that the same bias was at the core of the measure of selection proposed in Nozoe et al., 2017. Indeed, the forward sampling is built to reproduce the single lineage statistics obtained in mother-machine, where no selection occurs. We then tried to provide new answers to the question of the use of lineage data, relevant for both the questions of cell size control and evolution. We aimed to make these answers useful at the conceptual level by showing the fundamental interplay between fluctuations and selection, but also at a practical level with examples of inference methods.

In the first half of this thesis, we seek universal results characterizing any population tree, independently of a particular dynamical model. This is a step in the direction of

disentangling signatures of particular size control models and properties of the population tree itself, and toward a universal definition of selection in biological systems. This approach is in line with recent works unveiling universality in biology when analyzed with tools from theoretical physics (Goldenfeld et al., 2011). In particular, statistical physics has been extensively used to understand evolution (Neher et al., 2011; Kussell et al., 2014), and here we study cell colonies through the lens of stochastic thermodynamics (Mustonen et al., 2010; Sughiyama et al., 2015).

Next, we are interested in the fundamental thermodynamic limits on the growth of a cell colony, which, like any bio-physical system, must follow the first and second laws of thermodynamics. This question is challenging since cell division is an absolutely irreversible process, in the sense that the time-reversed counterpart of division, which is merging, is never observed, making the usual formalism of stochastic thermodynamics inadequate here. Even though our work is only a first step in this direction, the motivation behind this project is to provide some elements of comparison between the models of cell size control, based on their relative efficiency and robustness. This question may seem unrelated to those above, however it is not, since the fluctuations at the level of the cell cycle discussed before are also at the heart of the thermodynamic theory.

To sum up, we address the following questions:

- How can we relate the single cell and population levels?
- How should the population trees be sampled for each experimental setup?
- How should dead cells and cells that are evacuated from the setup because of dilution be taken into account in the statistics?
- How to define and quantify selection?
- How can we use single lineage data in practice to infer population-level quantities or the laws of cell growth and division?
- What are the thermodynamic limits imposed on cell growth and division?

This thesis is organized in four chapters, briefly presented below.

In chapter 2, we study the lineage-population bias for general population trees, without any assumption on the dynamics. The lineage-population bias is put in parallel with fluctuation theorems in stochastic thermodynamics, and, inspired by this comparison, we derive two main consequences: (i) universal bounds on the population growth rate (and the population doubling time in steady-state) involving the average values of the number of divisions along lineages, that generalize known results for age models, and (ii) an estimator of the population growth rate based only on single-lineage statistics that can be obtained in mother-machine experiments. The lineage-population bias lies at the heart of the definition of selection, and we seek universal constraints on the strength of selection, involving the variability among the lineages in the fashion of fluctuation-dissipation theorems.

In chapter 3, we extend all the results from chapter 2 to the case where some lineages disappear before the end of the experiment, for example because of cell death and dilution.

In this situation, the way to sample the lineages should be adapted to take into account lineages that disappeared. In particular, the remaining cells are subject to the survivor bias, which affects the phenotypic variability of the population. We try to quantify this new bias and its consequences on the strength of selection, which now results from more complex interactions between reproductive success and survival.

In chapter 4, we focus on size-regulated populations and we seek analytical steady-state size distributions for lineage statistics, that are used in two directions. First, these distributions can be compared to experimental data to test the validity of the model and to infer the parameters of the cell cycle, which is an example of a practical use of single-lineage data. Second, they are compared to population distributions to study the influence of different sources of variability, for example in cell growth or in partitioning of volume, on the lineage-population bias presented in the chapter 2.

Chapter 5 is independent of the other chapters and can be read alone. In this chapter, we explore the thermodynamic constraints on the growth of a colony of cells using simple coarse-grained models of cell size control. Based on a decomposition of cell division into two sub-processes: branching and resetting, we derive the first and second laws of thermodynamics for a colony of cells. We propose a definition for the efficiency of cell division, which relies on the way division modifies the distributions of age and size from an information theoretic point of view.

In the rest of the introductory chapter, we give the main conceptual and technical tools necessary to understand our contributions. Notions from stochastic thermodynamics are introduced in section 2. In section 3, we present in details the different experimental setups mentioned above, and the probability distributions corresponding to the possible samplings of the population trees. The three most popular models of cell size control: the sizer, the timer and the adder, are exposed in section 4, along with and their formulations in terms of partial differential equations. Finally, in section 5, we give a description of the data-sets that we use through the thesis to illustrate our theoretical results.

2 Short introduction to stochastic thermodynamics

In this section we give a short introduction to the fundamental concepts of stochastic thermodynamics that are used in chapters 2 and 5. This introduction is largely based on the book Peliti et al., 2021, Gatien Verley's PhD thesis (Verley, 2012, in french) and the review article Seifert, 2012. In Peliti et al., 2021, stochastic thermodynamics is defined as 'a thermodynamic theory for mesoscopic, non-equilibrium physical systems interacting with equilibrium heat reservoirs'.

The 'mesoscopic' scale is intermediate between microscopic and macroscopic. The microscopic world is ruled by the laws of mechanics, and macroscopic systems are described by the thermodynamic theory, which relies on a few state variables only: temperature, pressure, volume, ... Because of the large number of particles in a macroscopic system, of the order of Avogadro's number $\mathcal{N}_A \sim 10^{23}$, and the central limit theorem, state variables have Gaussian distributions which are very peaked around their mean values. In the thermodynamic limit where the size of the system tends to infinity, state variables are thus taken equal to their average value. In-between, systems can be characterized by a small

ensemble of mesoscopic variables provided equilibrium microscopic degrees of freedom are coarse-grained. Contrary to classical thermodynamics, these variables follow stochastic dynamics because of the random interactions between the system and the equilibrium heat bath. For that reason, a given experimental protocol leads to statistics of random trajectories. Stochastic thermodynamics aims to bridge the gap between the mesoscopic and macroscopic scales by connecting the statistics of trajectories to thermodynamic observables.

In the following, we first explain how quantities such as work, heat and entropy production are defined at the level of single trajectories and of ensembles of trajectories. In a second step, we present fluctuation theorems, which clarify the link between entropy production and the breaking of time-reversal symmetry. Because population balance equations describing the evolution of a colony of cells, given in section 4.1, are continuous partial differential equations, we present the framework of stochastic thermodynamics for continuous-state systems. The stochastic dynamics of such systems are described by two classes of equations: Fokker-Planck and Langevin equations, that we recall in the following.

2.1 Stochastic dynamics

2.1.1 Fokker-Planck equation

Let us consider a system characterized by a single degree of freedom x, like a 1D Brownian particle. The probability density $p(x,t)$ to find the system in state x at time t obeys the following Fokker-Planck equation:

$$\partial_t p(x,t) = -\partial_x j(x,t)\,, \tag{1.1}$$

where we introduced the current

$$j(x,t) = \mu F(x) p(x,t) - D\partial_x \left[p(x,t)\right]\,. \tag{1.2}$$

The first term in the current is convective, with μ the particle mobility and $F(x) = -\partial_x V$ a conservative force deriving from a potential $V(x)$. The second term is diffusive, and accounts for the random interactions with the heat reservoir. To keep this introduction simple, we consider that the diffusion coefficient D does not depend on position x and that there is no non-conservative force.

An important distinction is made between stationarity and equilibrium. A stationary state is reached when $p(x,t)$ becomes time-independent, that is when $\partial_x j(x,t) = 0$, while an equilibrium state requires in addition the absence of currents in the system: $j(x,t) = 0$. The solution to this last condition reads

$$p_{\mathrm{eq}}(x) \propto e^{\frac{-V(x)}{k_B T}}\,, \tag{1.3}$$

which is Boltzmann's distribution, as expected. To derive this solution, we used Einstein's relation

$$D = \mu k_B T\,, \tag{1.4}$$

which links the diffusive coefficient D, the temperature T of the heat bath and the mobility μ of the particle. Although this is an equilibrium relation, it is assumed to be true in stochastic thermodynamics even for non-equilibrium systems, because the heat bath is supposed to always be in equilibrium.

2.1.2 Langevin equation

Alternatively, the same one-dimensional diffusing particle can be described by the following Langevin equation:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \mu F(x) + \sqrt{2D}\xi(t)\,, \tag{1.5}$$

which is a stochastic differential equation because of the presence of the stochastic noise $\xi(t)$. The statistical properties of the noise are prescribed: it is isotropic $\langle\xi(t)\rangle = 0$, and memory-less $\langle\xi(t)\xi(t')\rangle = \delta(t-t')$.

The Brownian velocity $\mathrm{d}x/\mathrm{d}t$ is a ill-defined object since Wiener proved that it is defined only on a set of points of vanishing measure. For that reason, mathematicians prefer the following writing of Langevin equation:

$$\mathrm{d}x = \mu F(x)\mathrm{d}t + \sqrt{2D}\mathrm{d}W\,, \tag{1.6}$$

where W is the Wiener process, whose increments $\mathrm{d}W = \xi(t)\mathrm{d}t$ over a time $\mathrm{d}t$ are Gaussian with zero mean and variance $\mathrm{d}t$. Physicists' noise $\xi(t)$ is then the non-rigorous derivative of Wiener process.

2.2 Information theory

The equivalence between the thermodynamic entropy S_{sys} of a system at equilibrium and the Shannon entropy

$$H[p] = -\int \mathrm{d}x\; p(x)\ln p(x) \tag{1.7}$$

of the equilibrium distribution is a hallmark of classical thermodynamics:

$$S_{\mathrm{sys}} = k_B H[p_{\mathrm{eq}}]\,. \tag{1.8}$$

This result indicates that the thermodynamic entropy can be interpreted as the degree of disorder of the system, and is simply obtained by evaluating $H[p_{\mathrm{eq}}]$ with the equilibrium distribution given by eq. (1.3), where the proportionality constant is $\exp\left[\mathcal{F}/k_BT\right]$ with $\mathcal{F} = \langle V\rangle - TS_{\mathrm{sys}}$ the free energy.

By consistency with equilibrium thermodynamics, the non-equilibrium entropy is defined as

$$S_{\mathrm{sys}}(t) = k_B H[p(x,t)]\,, \tag{1.9}$$

where $p(x,t)$ is in general different from the equilibrium distribution. This non-equilibrium entropy is a measure of the information contained in the distribution $p(x,t)$, and suggests the definition of a stochastic entropy

$$s_{\mathrm{sys}}(t) = -k_B \ln p(x(t),t) \tag{1.10}$$

associated with a single trajectory. The lowercase is used to indicate the dependence on a single trajectory, and the non-equilibrium entropy of the system is then the average of the stochastic entropy over trajectories: $S_{\rm sys}(t) = \langle s_{\rm sys}(t)\rangle$.

2.3 Stochastic thermodynamics

We now manipulate our system via an experimental protocol λ that changes the energy $V(x,\lambda)$, so that an infinitesimal change in energy reads

$$\mathrm{d}V = \partial_x V(x,\lambda)\mathrm{d}x + \partial_\lambda V(x,\lambda)\mathrm{d}\lambda\,. \tag{1.11}$$

We define the infinitesimal work and heat as (Sekimoto, 1998):

$$\delta w = -\partial_\lambda V(x,\lambda)\mathrm{d}\lambda \tag{1.12}$$
$$\delta q = -\partial_x V(x,\lambda)\mathrm{d}x \tag{1.13}$$
$$= F(x)\mathrm{d}x\,, \tag{1.14}$$

which represent the change in energy due to an external manipulation of the potential, and due to the random motion of the particle in a fixed potential, respectively. The stochastic first law of thermodynamics then reads:

$$\mathrm{d}V = -\delta w - \delta q\,, \tag{1.15}$$

where the exchanges are counted positively if they go out of the system. Using these definitions, one can evaluate the work performed and the heat dissipated along a stochastic trajectory $\boldsymbol{x}$ by $w(\boldsymbol{x}) = \int_0^t \delta w$ and $q(\boldsymbol{x}) = \int_0^t \delta q$, but also the rates $\dot{w} = \delta w/\mathrm{d}t$ and $\dot{q} = \delta q/\mathrm{d}t$ at which these energies are exchanged.

These production rates are associated with single realizations of the random process, and can also be defined as the level of ensembles of trajectories. To do so, one can just integrate the above definitions over trajectories, or equivalently start from the Fokker-Planck equation. Indeed, we multiply eq. (1.1) by $V(x,\lambda(t))$ and integrate over x, which leads to:

$$\langle \dot{V}\rangle = -\int \mathrm{d}x\; j(x,t)F(x) + \dot{\lambda}\partial_\lambda\langle V\rangle \tag{1.16}$$
$$= -\langle \dot{q}\rangle - \langle \dot{w}\rangle\,, \tag{1.17}$$

for vanishing currents at the boundaries: $\lim_{x\to-\infty} V(x,\lambda(t))j(x,t) = \lim_{x\to\infty} V(x,\lambda(t))j(x,t) = 0$. Similarly, if we now multiply the Fokker-Planck equation by $k_B \ln p(x,t)$ and integrate over x we obtain:

$$\langle \dot{s}_{\rm sys}\rangle = -k_B\int \mathrm{d}x\;\frac{\mu j(x,t)F(x)}{D} + k_B\int \mathrm{d}x\;\frac{j^2(x,t)}{Dp(x,t)}\,. \tag{1.18}$$

The first term is identified as the entropy exchange rate with the reservoir:

$$\langle \dot{s}_{\rm m}\rangle = k_B\int \mathrm{d}x\;\frac{\mu j(x,t)F(x)}{D} \tag{1.19}$$
$$= \int \mathrm{d}x\;\frac{j(x,t)F(x)}{T} \tag{1.20}$$
$$= \frac{\langle \dot{q}\rangle}{T} \tag{1.21}$$

where we used Einstein's relation (eq. (1.4)) to go from the first to second line. The second term is the total entropy production rate:

$$\langle \dot{s}_{\rm tot} \rangle = k_B \int \mathrm{d}x \; \frac{j^2(x,t)}{Dp(x,t)} \geq 0 \,, \tag{1.22}$$

which is null when the system is at equilibrium, that is when there are no currents, and positive otherwise.

2.4 Fluctuation theorems

In classical thermodynamics, the irreversibility of a transformation is related to the value of the total entropy production: the more 'irreversible' the transformation, the larger the production of entropy. The second law states $\Delta S_{\rm tot} \geq 0$, which becomes an equality only for reversible transformations (when the system stays at equilibrium with the environment at all times). This connection is clarified by fluctuation theorems in stochastic thermodynamics, that put constraints on the distributions of fluctuating quantities, such as entropy production. These theorems rely on the notion of time-reversed, or backward, trajectories and protocols:

$$t^\dagger = t_f - t \tag{1.23}$$

$$x^\dagger(t) = x(t^\dagger) \tag{1.24}$$

$$\lambda^\dagger(t) = \lambda(t^\dagger) \,, \tag{1.25}$$

where t_f is the time at the end of the experiment. It can be shown that the total entropy production $s_{\rm tot}(\boldsymbol{x}) = \Delta s_{\rm sys} + s_{\rm m}(\boldsymbol{x}) = \Delta s_{\rm sys} + q(\boldsymbol{x})/T$ along a single trajectory $\boldsymbol{x}$ is linked to the ratio of the forward to backward path probabilities (Seifert, 2005):

$$s_{\rm tot}(\boldsymbol{x}) = k_B \ln \left[\frac{\mathcal{P}(\boldsymbol{x})}{\mathcal{P}^\dagger(\boldsymbol{x}^\dagger)} \right] \,, \tag{1.26}$$

where $\mathcal{P}^\dagger(\boldsymbol{x}^\dagger)$ is the probability to observe the time-reversed trajectory $\boldsymbol{x}^\dagger$ when starting from the initial condition $x^\dagger(0) = x(t_f)$ and with the backward protocol $\lambda^\dagger$. Thus, $s_{\rm tot}(\boldsymbol{x})$ reflects the degree of dissimilarity between the two path probabilities. Similarly, the entropy $s_{\rm m}(\boldsymbol{x})$ exchanged with the heat bath along the trajectory $\boldsymbol{x}$ is given by a similar formula for the path probabilities conditioned on their initial values x_0 and $x_0^\dagger$:

$$s_{\rm m}(\boldsymbol{x}) = k_B \ln \left[\frac{\mathcal{P}(\boldsymbol{x}|x_0)}{\mathcal{P}^\dagger(\boldsymbol{x}^\dagger|x_0^\dagger)} \right] \,. \tag{1.27}$$

These are examples of detailed fluctuation theorem, i.e. involving a comparison between two probability distributions, and from eq. (1.26) two important consequences are derived. First, the backward probability distribution can be integrated out to obtain an integral fluctuation theorem:

$$\langle e^{-s_{\rm tot}(\boldsymbol{x})/k_B} \rangle_{\mathcal{P}} = 1 \,, \tag{1.28}$$

where we explicitly indicate the probability $\mathcal{P}$ used to compute the average value $\langle\cdot\rangle_{\mathcal{P}}$ since two distributions are involved in eq. (1.26). This was not necessary in the previous section where the averages were non-ambiguously computed with the solution of the Fokker-Planck equation. Second, the convexity inequality $\langle e^x\rangle \geq e^{\langle x\rangle}$ implies that

$$\langle s_{\rm tot}(\boldsymbol{x})\rangle_{\mathcal{P}} \geq 0\,, \tag{1.29}$$

which is the classical second law of thermodynamics. Fluctuation theorems such as eq. (1.26) may be viewed as generalizations of the second law, and eq. (1.28) indicates that, even though the average entropy production is positive or null, the entropy production along certain trajectories must be negative. In the next section, we give two examples of well-known fluctuation theorems and their main use.

2.4.1 The Jarzynski and Crooks relations

Since the ratio of the path probabilities in eq. (1.26) depends only on the functional $s_{\rm tot}(\boldsymbol{x})$ of the path, then the fluctuation theorem can be recast for the marginal probabilities for the value $s_{\rm tot}$ of this functional:

$$p(s_{\rm tot}) = \langle\delta(s_{\rm tot}(\boldsymbol{x}) - s_{\rm tot})\rangle_{\mathcal{P}} \tag{1.30}$$

$$p^\dagger(-s_{\rm tot}) = \langle\delta(s_{\rm tot}(\boldsymbol{x}^\dagger) + s_{\rm tot})\rangle_{\mathcal{P}^\dagger}\,. \tag{1.31}$$

Indeed, when this functional is odd under time-reversal symmetry: $s_{\rm tot}(\boldsymbol{x}^\dagger) = -s_{\rm tot}(\boldsymbol{x})$, we multiply eq. (1.26) by the function $\delta(s_{\rm tot}(\boldsymbol{x}) - s_{\rm tot})$ and then integrate over all trajectories:

$$\langle\delta(s_{\rm tot}(\boldsymbol{x}) - s_{\rm tot})e^{-s_{\rm tot}(\boldsymbol{x})/k_B}\rangle_{\mathcal{P}} = \langle\delta(s_{\rm tot}(\boldsymbol{x}^\dagger) + s_{\rm tot})\rangle_{\mathcal{P}^\dagger}\,, \tag{1.32}$$

which is written in terms of the marginal distributions:

$$\frac{p(s_{\rm tot})}{p^\dagger(-s_{\rm tot})} = e^{s_{\rm tot}/k_B}\,. \tag{1.33}$$

Now, we recall the thermodynamic identity for isothermal processes between two equilibrium states: $s_{\rm tot} = (w - \Delta\mathcal{F})/T = w_{\rm diss}/T$, which expresses that the total entropy production is proportional to the dissipated work, that is the part of the free energy difference that is not converted in useful available work. When replacing $s_{\rm tot}$ in the above relation, we obtain Crooks relation (Crooks, 1999):

$$\frac{p(w)}{p^\dagger(-w)} = e^{(w-\Delta\mathcal{F})/k_BT}\,, \tag{1.34}$$

and the corresponding integral fluctuation theorem, Jarzynski equality (Jarzynski, 1997):

$$\Delta\mathcal{F} = -k_BT\ln\langle e^{-w/k_BT}\rangle_{\mathcal{P}}\,. \tag{1.35}$$

In both relations, even though the initial point must be at equilibrium and $\Delta\mathcal{F}$ is the difference in equilibrium free energy, the system needs not be at equilibrium at final time t_f. Indeed, let us consider the original time interval $[0, t_f]$ during which a work w

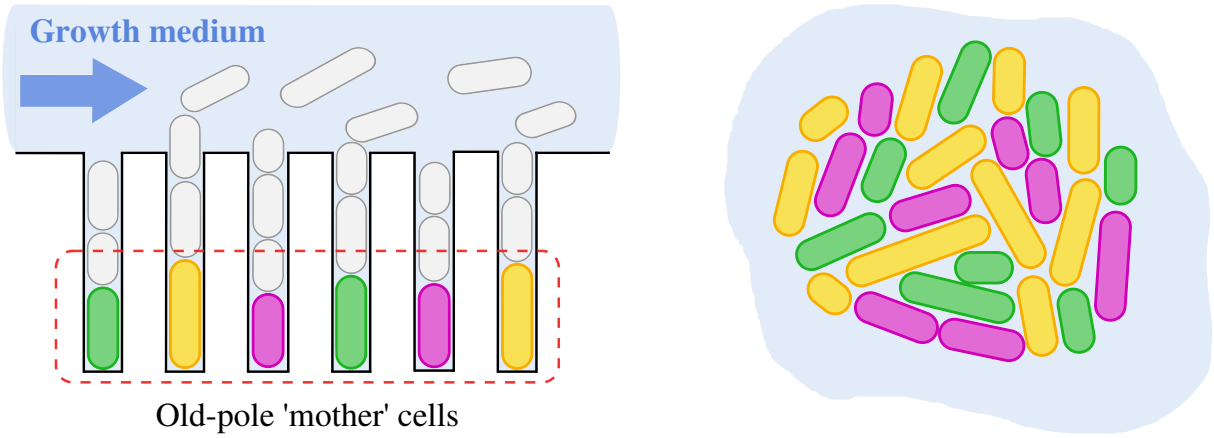


Figure 1.1: Cartoons of the two main experimental setups to study bacterial colonies. Left: mother machine setup with several microchannels. Right: freely-growing population in bulk. In both cases, colors are used to indicate the phenotypic variability among cells.

is performed, and the auxiliary time interval $[0, T]$ with $T > t_f$, for which the system relaxes to equilibrium between t_f and T with a fixed value $\lambda(t_f)$ of the protocol. The distributions of work for the auxiliary process obey Crooks relation (and Jarzynski's), and are equal to the same distributions for the original process since no work is performed in $[t_f, T]$. Therefore, these relations hold irrespective of the speed of the process.

Consequently, these two relations surprisingly allow the inference of equilibrium free energies from non-equilibrium work measurements, either by computing an exponential average over forward trajectories with Jarzynski equality (Liphardt et al., 2002), or by looking at the intersection of the forward and backward work distributions with Crooks relation (Collin et al., 2005), which are both difficult in practice since one needs to sample rare events (Jarzynski, 2006).

3 The different probability distributions in cell experiments

Cell data can be acquired in different experimental configurations:

1. Snapshot: picture of the population at a time t, the past history of the population is not available with this method.

2. Time-lapse: the whole colony is followed in time via video-microscopy, which provides us the population tree.

3. Mother-machine: single lineages confined in microfluidic channels are monitored in time.

The majority of the available cellular data is snapshot data. Indeed, this is the only method accessible for in vivo experiments at the moment, with ongoing progress for in

vivo time-lapse methods. Moreover, mostly snapshot data are available in the field of evolution as well. In evolution, data are represented with a phylogenetic tree, similar to the population tree for bacterial colonies, where the branches represent species or genes, and the splitting of the branches speciations or mutations (Schuh et al., 2009). On the other hand, time-lapse and mother-machine data are accessible in vitro. To compensate for the lack of time-lapse and mother-machine data in vivo and in evolution, it is necessary to link the three types of data, and to develop inference methods from snapshot data only. Here, we only focus on time-lapse and mother-machines data, that is when the entire history of the population/lineage is known, to understand the statistical biases induced by the differences between the two experimental setups. In the next sections, we present these two configurations and the corresponding statistics.

3.1 Lineage distribution

Single lineage experiments are designed to monitor a single cell lineage in time over many generations. The mother machine developed in Wang et al., 2010 and represented on fig. 1.1 is a prototypical example of such experiments. In this setup, cells are confined in microfluidic channels with one closed end and one open end. The cell that remains at the closed end division after division is called the 'mother cell', and measurements are usually made on this cell only. Be careful that throughout the thesis, we use the term 'mother cell' differently, to refer to any dividing cell that gives birth to daughter cells. All the other cells are pushed towards the open end and eventually carried away by the flow of growth medium, which also fills the channels with the necessary nutrients. This setup is mainly used to study rod-shaped bacteria, such as E. coli, growing only in length while keeping their width approximately constant, and the width and height of the channels are engineered to be equal to the cell width so that the lineage grows in one dimension only. Since the cell width depends on the growth medium, the width of the channels should be adapted for each experimental conditions (Taheri-Araghi et al., 2015).

A large number L of channels are observed simultaneously, and we define the *lineage distribution* p_{lin} of any property y, which can be anything at this stage, at time t as:

$$p_{\text{lin}}(y,t) = \frac{1}{L}\sum_{i=1}^{L}\delta(y-y_i) = \frac{n(y,t)}{L}\,, \tag{1.36}$$

where y_i is the value of property y associated with cell i, and $n(y,t)$ is the number of cells with property y at time t.

3.2 How to sample a population tree?

Now we consider a growing population of cells, like the one on fig. 1.1 right. The history of such a colony is represented by a branched tree, starting with N_0 cells at time $t = 0$ and ending with $N(t)$ cells at time t. At each division, m daughter cells are produced, including the mother cell, where m is a non-stochastic quantity. We assume that all lineages survive up to time t, and therefore the final number $N(t)$ of cells corresponds

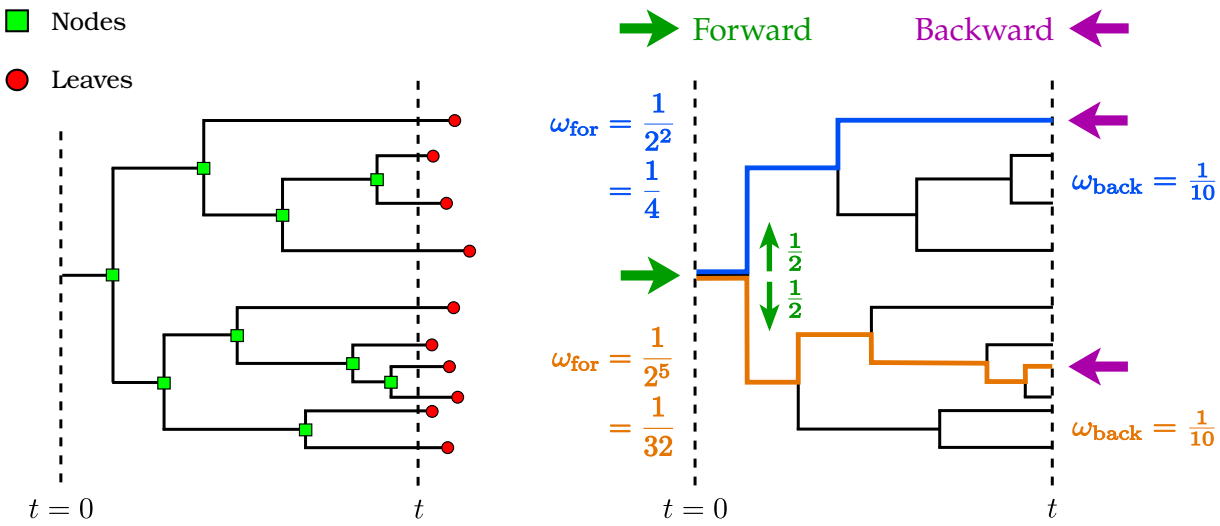


Figure 1.2: Example of a tree with $N_0 = 1$ and $N(t) = 10$ lineages at time t. On the left, divisions that occurred before time t are called nodes, while the future divisions of cells present in the population at time t are called leaves. On the right, two lineages are highlighted, in blue with 2 divisions and in orange with 5 divisions. The forward sampling is represented with the green right arrows: it starts at time $t = 0$ and goes forward in time by choosing one of the two daughters lineages at each division with probability $1/2$. The backward sampling is pictured by the left purple arrows: starting from time t with uniform weight on the 10 lineages it goes backward in time down to time $t = 0$.

to the number of lineages in the tree. An example of such a tree starting with a unique ancestor cell: $N_0 = 1$, and where cells follow binary fission: $m = 2$, is shown on fig. 1.2.

What is the correct way to sample cells in such a population tree?

A division-based approach consists in putting a uniform weight on all divisions in the tree, resulting in the *tree distribution*. There are subtleties on either to count only the $N(t) - N_0$ divisions that happened before final time t, represented by the green squares on fig. 1.2 left and called nodes, or to include the divisions of the cells present at final time t that have not divided yet (Powell, 1956; Lin et al., 2017). These cells, depicted by the red circles, are called leaves. This distribution is limited to variables that are defined at the level of each division: the age at the moment of the division, or the size at birth for example, but cannot be used to sample variables defined at the level of cell lineages: like the number K of divisions along a lineage from initial to final time, a phenotypic trajectory, or even the cells properties that are only defined in a time snapshot, such as age and size. For this reason, we shift the focus from individual division events to individual cell lineages, and recall the definitions of the forward and backward samplings of the lineages proposed in Nozoe et al., 2017.

The most simple way to sample the lineages is to put uniform weight on all of them. Note that this is also the only sampling available for evolutionary data and frequently for in-vivo data, where the history of the lineages is unknown. This sampling is called backward, (or retrospective) because at the end of the experiment one randomly chooses one lineage among the $N(t)$ with a uniform probability and then traces the history of the

lineage backward in time from time t to 0, until reaching the ancestor population. The backward weight associated with a lineage l is defined as

$$\omega_{\text{back}}(l) = N(t)^{-1}\,. \tag{1.37}$$

In a tree, some lineages divide more often than others, which results in an over-representation in the final population of lineages that have divided more than average. Therefore by choosing a lineage with uniform distribution, we are more likely to end up with a lineage with high reproductive success. To balance this effect, the forward (or chronological) sampling puts a weight

$$\omega_{\text{for}}(l) = N_0^{-1} m^{-K(l)}\,. \tag{1.38}$$

on a lineage l with $K(l)$ divisions. This choice of weights is called forward because one starts at time 0 by uniformly choosing one cell among the N_0 initial cells, and goes forward in time up to time t, by choosing one of the m offspring with equal weight $1/m$ at each division. A population with examples of forward and backward weights for different lineages in shown on fig. 1.2 right.

Let us comment on two important properties of the forward sampling. First, it is built to give the same weight to all initial cells (and to all daughter cells born from the same division), regardless of the size of the subpopulations of offspring they generate at later times, which is a measure of their reproductive successes. Therefore, the forward sampling is said to 'cancel selection', understood in the sense of the over-representation of certain lineages because of their reproductive advantage. The second comment is highly connected to the first one: the forward sampling is defined in such a way as to reproduce the lineage statistics introduced in section 3.1. Indeed, no selection occurs in single lineage experiments, where at each division only one of the m daughter cells is followed with probability $1/m$ while the other ones are disregarded. This can be seen more quantitatively at the level of the equations, as later shown in section 4.3. This sampling thus links the scale of tree-structured population and single lineages, and for that reason we will use the terms forward and lineage statistics interchangeably. Similarly, we will use the term *population statistics* and backward statistics equivalently.

We now define the probability to pick a lineage with property y and K divisions as the number $n(y, K, t)$ of lineages with this property times the weight $\omega(K)$ of a lineage with K divisions, in both the forward and backward samplings:

$$\begin{aligned} p_{\text{back}}(y, K, t) &= \omega_{\text{back}}(K) n(y, K, t) && (1.39)\\ &= N(t)^{-1} n(y, K, t) && (1.40)\\ p_{\text{for}}(y, K, t) &= \omega_{\text{for}}(K) n(y, K, t) && (1.41)\\ &= N_0^{-1} m^{-K} n(y, K, t)\,. && (1.42) \end{aligned}$$

By going from the weights ω to the probability distributions p, we shift the domain of definition from the ensemble of lineages, with the normalizations $\sum_{i=1}^{N(t)} \omega_{\text{back}}(l_i) = \sum_{i=1}^{N(t)} \omega_{\text{for}}(l_i) = 1$, to that of the tracked variables: $\sum_{K=0}^{\infty} \int \mathrm{d}y\ p_{\text{back}}(y, K, t) = \sum_{K=0}^{\infty} \int \mathrm{d}y\ p_{\text{back}}(y, K, t) = 1$.

For variables that make sense both in division-based and lineage-based approaches, like the inter-division time, the convergence of the tree distribution to either the forward or backward distributions has been studied in Nakashima et al., 2020.

4 Introduction to cell size control

4.1 Models of cell size control

Although certain results of the first two chapters of this thesis are independent of the underlying dynamics producing the population tree, other features explicitly depend on it, and we illustrate them in the context of cell size control. Moreover, in the last two chapters the cell dynamics is at the center of our attention. For these reasons, we introduce here the different models of cell size control that we use.

The question of how a cell controls its size is a very old one, which despite decades of research is still under intense focus because the old experiments have only provided incomplete answers while a new generation of experiments based on the observation and manipulation of single cells in microfluidic devices is becoming more and more mature. Many models of cell size control have been proposed, and we do not aim to provide an extensive view of their diversity in this introduction, that can be found in the comprehensive reviews Willis et al., 2017; Ho et al., 2018; Jun et al., 2018. In this thesis, we focus on the three most popular models of cell size control: the timer, the sizer and the adder, where division is controlled by age, size and increment of volume respectively. One common aspect of these models is to rely on few macroscopic parameters, thus coarse-graining all the biomolecular machinery. All three models have proven useful, and classifications of species according to their mechanisms of cell size control can be found in Willis et al., 2017; Jun et al., 2018, where some species follow mixed strategies.

Finally, we describe these models using continuous rate models (Robert et al., 2014), based on partial differential equations that describe cell cycles continuously in time, as opposed to discrete stochastic maps (Amir, 2014; Ho et al., 2018) formulated as Markov processes for division times.

4.1.1 The timer

Uncorrelated divisions The simplest model of cell size control is the age model, also called *timer*, or Bellman-Harris process in the mathematical literature (Kimmel et al., 2002). In this model, the probability that a cell of age a divides during a time interval $\mathrm{d}t$ is equal to $r(a)\mathrm{d}t$. After division, each daughter cell starts with age 0, thus there is no memory effect. Age is by definition the time elapsed since birth, so it grows linearly with time: $\mathrm{d}a/\mathrm{d}t = 1$. Note that in the literature on bacterial colonies, age is sometimes defined as replicative age (Wang et al., 2010), that is the number K of consecutive divisions along the lineage under consideration, or as physiological age (Olivier, 2017), which evolves non-linearly in time to describe possible phases with different aging rates. In this thesis, we stick to the usual definition of age.

The Population Balance Equation (PBE) describing the time evolution of the number $n(a,t)$ of cells of age a at time t reads

$$\partial_t n(a,t) = -\partial_a n(a,t) - r(a)n(a,t)\,, \tag{1.43}$$

together with the boundary condition

$$n(a=0,t) = m\int \mathrm{d}a' r(a')n(a',t)\,. \tag{1.44}$$

In eq. (1.43), the first term is convective and accounts for cell aging, and the second term is a loss term coming from the divisions of cells of age a. The boundary condition says that the number of newborn cells is equal to the number of dividing cells of any ages times the number m of daughter cells produced by each division.

Although simple, the timer model has been shown to be unrealistic since it leads to diverging fluctuations in cell size when cells grow exponentially (Trucco et al., 1970; Amir, 2014). However, Robert et al., 2014 suggested that steady size distributions can be reached for the timer model if cells of sizes near zero and infinity grow sub-exponentially.

Correlated divisions Experiments show correlations between the generation time of the mother cell and that of the daughter (Taheri-Araghi et al., 2015). The classical timer model does not account for these correlations, and more complex age-structured models have been proposed. There are at least two ways to model age-structured populations with mother-daughter correlations in inter-division times.

The first approach is to consider that correlations are described in the source term via a kernel $\Sigma(\tau|\tau')$ at division, where τ is the inter-division time (Powell, 1956; Lebowitz et al., 1974; Lin et al., 2020; Levien et al., 2020). Doing so, τ becomes a variable, and the number $n(a,\tau,t)$ of cells of age a at time t that will divide at age τ follows a population balance equation, valid only for $a \leq \tau$, which does not involve a division rate:

$$\partial_t n(a,\tau,t) = -\partial_a n(a,\tau,t) \qquad \text{for } 0 < a \leq \tau \tag{1.45}$$

$$n(a=0,\tau,t) = m\int \mathrm{d}\tau' \Sigma(\tau|\tau')n(\tau',\tau',t)\,. \tag{1.46}$$

The second approach consists in introducing an intermediate variable, let us call it y, on which the division rate $r(a,y)$ depends and which is transmitted at division via kernel $\Sigma(y|y')$. The variable is often thought of as a state or type, meaning it is not evolving during the cell cycle, and using this vocabulary, this timer model with correlations can be called a multitype age process (Sughiyama et al., 2019; Nakashima et al., 2020). For example, in (García-García et al., 2019; Genthon et al., 2020) we proposed the single cell growth rate as a candidate for the variable y. The equation and the boundary condition read

$$\partial_t n(a,y,t) = -\partial_a n(a,y,t) - r(a,y)n(a,y,t) \tag{1.47}$$

$$n(a=0,y,t) = m\int \mathrm{d}a'\mathrm{d}y' r(a',y')\Sigma(y|y')n(a',y',t)\,. \tag{1.48}$$

4.1.2 The sizer

In the sizer model, division is triggered by cell size only, and the probability for a cell of size x to divide during a time interval $\mathrm{d}t$ is equal to $r(x)\mathrm{d}t$. Since rod-shaped bacteria grow along one axis only with constant width (Taheri-Araghi et al., 2015), throughout this thesis we will consider size, volume and length as synonyms. Note that this is simply a convenience, and that the sizer model is a valid model for any cell morphology. At division, the volume x' of the dividing cell is partitioned between the m daughter cells, and the probability for one of them to inherit a volume x is given by the transition kernel $\Sigma(x|x')$, normalized as: $\forall x', \int \mathrm{d}x\ \Sigma(x|x') = 1$. Moreover, the conservation of volume at division between the mother cell and the m daughter cells is imposed through $m \int \mathrm{d}x\ x\Sigma(x|x') = x'$. This family of kernels allows the description of stochastic partitions of volume, and by setting $\Sigma(x|x') = \delta(x - x'/m)$, we recover the deterministic case of equal fission, where each cell inherits a fraction $1/m$ of the mother cell's volume. Within the cell cycles, we describe cell growth with the rate $\mathrm{d}x = \nu(x)\mathrm{d}t$. This general function accounts for the most common growth strategies, such as linear growth when $\nu(x) = \nu$, or exponential growth if $\nu(x) = \nu x$.

The population balance equation for the sizer reads

$$\partial_t n(x,t) = -\partial_x[\nu(x)n(x,t)] - r(x)n(x,t) + m\int \mathrm{d}x'\ \Sigma(x|x')r(x')n(x',t)\,, \tag{1.49}$$

where, unlike the timer, the integral term appears in the main equation.

4.1.3 The adder

In the past decade, more and more organisms have been observed to neither follow the sizer nor the timer mechanisms, but to divide after adding an increment of volume $\Delta_d = x_d - x_b$ between the birth size x_b and the division size x_d, drawn from a probability distribution independent of the birth size.

To model this behavior, we introduce the *adder* mechanism, which relies on two variables $\{x, \Delta\}$ with $\Delta = x - x_b$ the added volume since birth. In this model we impose that the volume evolves as $\mathrm{d}x = \nu(x)\mathrm{d}t$ like for the sizer, and thus $\mathrm{d}\Delta = \nu(x)\mathrm{d}t$, and that the division rate per unit time

$$r(x,\Delta) = \nu(x)\zeta(\Delta) \tag{1.50}$$

is equal to the product of the growth rate $\nu(x)$ and the division rate per unit volume $\zeta(\Delta)$ (Taheri-Araghi et al., 2015). Indeed, the probability that the increment of volume between birth and division takes the value Δ_d knowing that the birth size is x_b is given by the following change of variable

$$p(\Delta_d|x_b)\mathrm{d}\Delta_d = p(\tau|x_b)\mathrm{d}\tau\,, \tag{1.51}$$

where

$$p(\tau|x_b) = r(x(\tau), x(\tau) - x_b)e^{-\int_0^\tau r(x(t),x(t)-x_b)\mathrm{d}t} \tag{1.52}$$

is the probability that the cell divides after a generation time τ knowing it was born with size x_b. Finally, combining eqs. (1.50) to (1.52), we obtain:

$$p(\Delta_d|x_b) = \zeta(\Delta_d)e^{-\int_0^{\Delta_d}\zeta(\Delta)\mathrm{d}\Delta}\,, \tag{1.53}$$

which is independent of the size at birth x_b.

The population balance equation and the boundary condition then read

$$\partial_t n(x,\Delta,t) = -\left(\partial_x + \partial_\Delta\right)\left[\nu(x) n(x,\Delta,t)\right] - \nu(x)\zeta(\Delta) n(x,\Delta,t) \tag{1.54}$$

$$\nu(x) n(x,\Delta=0,t) = m \int \mathrm{d}x' \mathrm{d}\Delta'\ \Sigma(x|x')\nu(x')\zeta(\Delta') n(x',\Delta',t)\,, \tag{1.55}$$

The adder mechanism can also be formulated in terms of other pairs of variable, provided that the increment of volume can be deduced from them. For example, the adder model can be seen a mixed age-size model with variables $\{x, a\}$, because for a known growth function ν, the increment of volume is determined by the values of age and size. For example, in the case of exponential growth with $\nu(x) = \nu x$, then $\Delta = x[1-\exp(-\nu a)]$. Another simple possibility is the couple of size and size at birth $\{x, x_b\}$.

4.2 Division rate and stochasticity

In the following chapters, we will sometimes choose to describe the division rates by power laws. This simplification is justified by the shapes of the division rates inferred from experimental data (see Doumic et al., 2015 for $r(x)$, and the SM of Robert et al., 2014 for $r(a)$). Further possible theoretical justifications for this power law have been discussed in Nieto et al., 2020. We define

$$r(x) = r x^\alpha\,, \tag{1.56}$$

where α is the strength of the size control, and similarly $r(a) = r a^\alpha$ for the timer where α is the strength of the age control.

In the limit where α goes to zero, the division rate becomes independent of the size of the cell, which is then said to be an uncontrolled variable. The resulting dynamics is a simple Poisson process with rate r in this case. On the other hand, in the limit of strong size control $\alpha \to +\infty$, the division rate becomes a step function with value 0 before threshold size 1, and $+\infty$ after. Cells divide deterministically when reaching the size threshold, which can be tuned by the rescaling $r(x) = r(x/x_\dagger)^\alpha$. In this limit, we recover the deterministic versions of the sizer, timer and adder, where cells divide when reaching a certain age or size, or when growing by a constant increment of volume, without variability. The variability in size, age or added volume at division is then a decreasing function of the control strength α.

Before closing this section, we want to warn the reader about alternative definitions of the sizer, timer and adder mechanisms that have been used in recent articles following Nieto et al., 2020. In this work on exponentially growing cells ($\nu(x) = \nu x$), the authors considered what we defined as the sizer model in section 4.1.2, with power-law division rate $r(x) = r x^\alpha$. They argued that this model recovers a 'timer' strategy when $\alpha = 0$, an 'adder' strategy when $\alpha = 1$, and a 'sizer' strategy when $\alpha \to \infty$, because these values yield specific relations between the average increment of volume and the average size at birth, respectively with slopes 1, 0 and -1. Indeed, when $\alpha = 0$ the division rate is constant, size becomes uncontrolled, and the size model falls into the scope of the timer model described by eq. (1.43), with constant division rate $r(a) = r$. Similarly,

when $\alpha = 1$ then $r(x) = rx \propto \nu(x)$, so that the size model fits in with the adder model described by eq. (1.54), with constant division rate per unit volume $\zeta(\Delta) = \zeta = r/\nu$. These represent only simple cases of the timer and adder models, where age and added volume are respectively uncontrolled, and thus Poisson-distributed at division. However, the size model cannot reproduce the timer and adder models as we defined them, with general division rates which are functions of age and added volume respectively.

4.3 Population balance equations at the level of probabilities

The PBE for the different size control models presented in the previous section can be recast at the level of probability distributions using eqs. (1.39) and (1.41). For simplicity, we show how it is done for the sizer, given that the manipulations are identical for the other models.

First, since the backward weight does not depend on the number of divisions, eq. (1.39) can be summed over K to obtain the marginal distribution of x: $p_{\text{back}}(x,t) = N(t)^{-1}n(x,t)$. Now, replacing $n(x,t)$ in eq. (1.49) gives

$$\begin{aligned}\partial_t p_{\text{back}}(x,t) = &- \partial_x[\nu(x)p_{\text{back}}(x,t)] - (r(x) + \Lambda_p(t))\, p_{\text{back}}(x,t) \\ &+ m\int \mathrm{d}x'\ \Sigma(x|x')r(x')p_{\text{back}}(x',t)\,,\end{aligned} \tag{1.57}$$

where

$$\Lambda_p(t) = \frac{1}{N(t)}\frac{\mathrm{d}N}{\mathrm{d}t}\,, \tag{1.58}$$

is the instantaneous population growth rate.

Second, since the forward probability explicitly depends on the number K of division, it is useful to make it appears in the PBE as:

$$\begin{aligned}\partial_t n(x,K,t) = &- \partial_x[\nu(x)n(x,K,t)] - r(x)n(x,K,t) \\ &+ m\int \mathrm{d}x'\ \Sigma(x|x')r(x')n(x',K-1,t) \qquad \text{for } K \geq 1 \end{aligned} \tag{1.59}$$

$$\partial_t n(x,K=0,t) = - \partial_x[\nu(x)n(x,K=0,t)] - r(x)n(x,K=0,t)\,. \tag{1.60}$$

The equation is split into two because the evolution of the number $n(x,K=0,t)$ of cells of size x that have not divided yet has no contribution from larger dividing cells, by definition. The integral term involves the number $n(x',K-1,t)$ of cells of any sizes with $K-1$ divisions before time t that divide into cells of size x at time t, thus increasing by one their number of divisions to K. Now, we use eq. (1.41) to change $n(x,K,t)$ into $p_{\text{for}}(x,K,t)$:

$$\begin{aligned}\partial_t p_{\text{for}}(x,K,t) = &- \partial_x[\nu(x)p_{\text{for}}(x,K,t)] - r(x)p_{\text{for}}(x,K,t) \\ &+ \int \mathrm{d}x'\ \Sigma(x|x')r(x')p_{\text{for}}(x',K-1,t) \qquad \text{for } K \geq 1 \end{aligned} \tag{1.61}$$

$$\partial_t p_{\text{for}}(x,K=0,t) = - \partial_x[\nu(x)p_{\text{for}}(x,K=0,t)] - r(x)p_{\text{for}}(x,K=0,t)\,. \tag{1.62}$$

Summing this equation over K gives the equation for the marginal distribution of size x:

$$\partial_t p_{\text{for}}(x,t) = -\partial_x[\nu(x)p_{\text{for}}(x,t)] - r(x)p_{\text{for}}(x,t) + \int \mathrm{d}x'\ \Sigma(x|x')r(x')p_{\text{for}}(x',t)\,. \tag{1.63}$$

Third, we consider the experimental setup of the mother machine. In this setup, the number of cells $N(t)$ is not evolving with time, and is equal to the number L of microchannels that are monitored. At each division m cells are produced but only one remains at the close end of the channel so that we follow only $m = 1$ cell. The PBE eq. (1.49) is then recast at the level of the lineage distribution by setting $m = 1$ and using eq. (1.36):

$$\partial_t p_{\text{lin}}(x,t) = -\partial_x[\nu(x)p_{\text{lin}}(x,t)] - r(x)p_{\text{lin}}(x,t) + \int \mathrm{d}x'\ \Sigma(x|x')r(x')p_{\text{lin}}(x',t)\,. \tag{1.64}$$

Finally, we showed that the lineage and forward distributions obey the same differential equation.

The difference between the population/backward and the lineage/forward equations is the presence in the former of the instantaneous population growth rate and the number of daughter cells. These two quantities are simply related by:

$$\Lambda_p(t) = (m-1)\int \mathrm{d}x\ r(x)p_{\text{back}}(x,t)\,, \tag{1.65}$$

which follows from the integration of eq. (1.57) over x, using the normalization of p_{back} at any time t: $\int_0^\infty \mathrm{d}x\ p_{\text{back}}(x,t) = 1$ and the no-flux boundary conditions

$$\nu(x)p_{\text{back}}(x,t) \underset{x\to 0}{\to} 0 \tag{1.66}$$

$$\nu(x)p_{\text{back}}(x,t) \underset{x\to +\infty}{\to} 0\,. \tag{1.67}$$

Thus, the instantaneous population growth rate is the backward average of the division rate. We recover that setting $m = 1$ leads to $\Lambda_p(t) = 0$ corresponding to a constant population, as for the lineage distribution.

4.4 Steady-state behavior

In the mathematical literature, the population balance equations eqs. (1.43) and (1.49) for the timer and the sizer are called renewal equation and growth-fragmentation equation, respectively. The conditions for the existence and uniqueness of solutions to these equations have been established (see Doumic et al., 2021 for a review on the subject), and in the long-time limit the population grows exponentially with a rate Λ, called the Malthus parameter. More precisely, it is proven that the solutions to these equations, for y the age or the size, obey

$$\lim_{t\to\infty} n(y,t)e^{-\Lambda t} = p_{\text{back}}(y)\,, \tag{1.68}$$

where $p_{\text{back}}(y)$ is the steady-state backward distribution, and where the Malthus parameter is the asymptotic value of the instantaneous population growth rate:

$$\Lambda = \lim_{t\to\infty} \Lambda_p(t)\,. \tag{1.69}$$

4.5 Analytical results for age models without correlations

Age models without correlations have been long studied because of their analytical simplicity (Kimmel et al., 2002). In particular, three important results were obtained in the 1950s, and we give here a short derivations of them termed with our notations.

The equations for the forward and backward probabilities, together with their boundary conditions, read:

$$\partial_t p_{\text{back}}(a,t) = -\partial_a p_{\text{back}}(a,t) - [r(a) + \Lambda_p(t)]\, p_{\text{back}}(a,t) \tag{1.70}$$

$$p_{\text{back}}(a=0,t) = m \int \mathrm{d}a' r(a') p_{\text{back}}(a',t)\,. \tag{1.71}$$

$$\partial_t p_{\text{for}}(a,t) = -\partial_a p_{\text{for}}(a,t) - r(a) p_{\text{for}}(a,t) \tag{1.72}$$

$$p_{\text{for}}(a=0,t) = \int \mathrm{d}a' r(a') p_{\text{for}}(a',t)\,. \tag{1.73}$$

In the backward statistics, this proportion $p_{\text{back}}(a=0,t)$ of cells of age 0 at time t can also be linked to the population growth rate when combining eq. (1.65) and eq. (1.71):

$$(m-1) p_{\text{back}}(0,t) = m\Lambda_p(t)\,. \tag{1.74}$$

In steady-state, eq. (1.70) and eq. (1.72) can be solved and the time-independent age distributions are given by

$$p_{\text{back}}(a) = p_{\text{back}}(0)\, \exp\left[-\Lambda a - \int_0^a r(a')\mathrm{d}a'\right] \tag{1.75}$$

$$p_{\text{for}}(a) = p_{\text{for}}(0)\, \exp\left[-\int_0^a r(a')\mathrm{d}a'\right]\,. \tag{1.76}$$

We define the distribution $f(\tau,t)$, both forward and backward, of generation times τ as the ratio of the number of cells dividing at age τ at snapshot time t, to the total number of cells dividing in this snapshot, weighted accordingly:

$$f_{\text{back}}(\tau,t) = \frac{r(\tau) p_{\text{back}}(\tau,t)}{\int \mathrm{d}\tau'\; r(\tau') p_{\text{back}}(\tau',t)} \tag{1.77}$$

$$f_{\text{for}}(\tau,t) = \frac{r(\tau) p_{\text{for}}(\tau,t)}{\int \mathrm{d}\tau'\; r(\tau') p_{\text{for}}(\tau',t)} \tag{1.78}$$

Combined with the steady-state age distributions, we obtain the steady-state distributions of generation times:

$$f_{\text{back}}(\tau) = m r(\tau) \exp\left[-\Lambda\tau - \int_0^\tau \mathrm{d}a'\; r(a')\right] \tag{1.79}$$

$$f_{\text{for}}(\tau) = r(\tau) \exp\left[-\int_0^\tau \mathrm{d}a'\; r(a')\right]\,. \tag{1.80}$$

By comparing the two above relations, we obtain the first result, called Powell's relation (Powell, 1956):

$$f_{\text{back}}(\tau) = m f_{\text{for}}(\tau) e^{-\Lambda\tau}\,, \tag{1.81}$$

which shows that the distributions of generation times obtain when following a lineage forward and backward in time are not the same. The backward distribution is indeed biased toward smaller values of generation times, reflecting the over-representation in the population of lineages that divided a lot and along which cell cycles are on average shorter, while no such selection is present in single-lineage experiments.

Second, by integrating eq. (1.81) over τ and using the normalization of f_{back}, we obtain Euler-Lotka equation:

$$1 = m \int_0^\infty \mathrm{d}\tau \; f_{\text{for}}(\tau) e^{-\Lambda\tau} \,. \tag{1.82}$$

This relation is valuable at least in two ways. From a practical point of view, and remembering that the forward and lineage distributions are identical, it can be used to infer the population growth rate from single-lineage measurements if cells are known to be age-controlled with negligible correlations. Equivalently, given a population tree, the validity of the timer without correlation to describe the colony can be tested by comparing the growth rate obtained by eq. (1.82) to the measured population growth rate. On a more conceptual level, Euler-Lotka equation offers insights on the link between single cell variability and population growth: the whole distribution of generation time is shaping the population growth rate, and not only its mean. This idea has been exploited to expand the population growth rate as a function of the moments of $f_{\text{for}}(\tau)$, thus quantifying the correction from the first order result $\Lambda = \ln 2/\langle\tau\rangle_{\text{for}}$ due to the variability in generation times (Lin et al., 2020).

Third, let us now introduce the Kullback-Leibler (KL) divergence between two probability distributions p and q, which is the non-negative and asymmetric information-theoretic distance between them:

$$\mathcal{D}_{\text{KL}}(p||q) = \int \mathrm{d}x \, p(x) \ln \frac{p(x)}{q(x)} \geq 0 \,. \tag{1.83}$$

We now compute the two KL divergences between the forward and backward distributions of generation times:

$$\mathcal{D}_{\text{KL}}(f_{\text{back}}||f_{\text{for}}) = \ln m - \Lambda\langle\tau\rangle_{\text{back}} \tag{1.84}$$

$$\mathcal{D}_{\text{KL}}(f_{\text{for}}||f_{\text{back}}) = -\ln m + \Lambda\langle\tau\rangle_{\text{for}} \,. \tag{1.85}$$

Combining the two inequalities gives an upper and a lower bound for the population growth rate:

$$\frac{\ln m}{\langle\tau\rangle_{\text{for}}} \leq \Lambda \leq \frac{\ln m}{\langle\tau\rangle_{\text{back}}} \,. \tag{1.86}$$

This result is more often presented in terms of the population doubling time $T_d = \ln 2/\Lambda$ (for binary fission: $m = 2$), which is the time necessary to double the number of cells when the population is the regime of exponential growth (Hashimoto et al., 2016). The inequalities then read

$$\langle\tau\rangle_{\text{back}} \leq T_d \leq \langle\tau\rangle_{\text{for}} \,. \tag{1.87}$$

They indicate in particular that measuring the average generation time in a single-lineage experiment over-estimates the population doubling time.

Let us now comment on the mathematical analogy between these three results and relations in stochastic thermodynamics. Powell's relation expresses the exponential bias between two probability distributions, and in that respect is similar to detailed fluctuation theorems like Crooks relation (eq. (1.34)). Once the backward distribution of generation times is integrated, Euler-Lotka equation is analogous to integral fluctuation theorems such as Jarzynski equation (eq. (1.35)), and their uses are similar. In the same way as Jarzynski equation is used to infer equilibrium free energies from non-equilibrium work measurements, Euler-Lotka equation provides the population growth rate from single-lineage data. Finally, the double inequality on the population doubling time, obtained by convexity from Powell's relation, are equivalent to the second law of thermodynamics (eq. (1.29)). The reason why there are two inequalities instead of one is because in stochastic thermodynamics the forward and backward dynamics are related by a time-reversal symmetry, and the observable which is averaged (generally a current) changes sign under time reversal. None of these two properties are true in population dynamics. These three results are only valid in steady state and for age models without correlations, and one goal of chapter 2 is to generalize them for any population tree.

5 Short descriptions of datasets used

In this thesis, we use three sets of experimental data on E. coli to illustrate our theoretical results. We give here a short description of each of them and orders of magnitude on E. coli, that are always useful to keep in mind to understand the challenges faced by experimenters.

E. coli is a rod-shaped bacterium, growing mainly in one direction while maintaining its diameter almost constant around 1 µm. The volume and the length are thus proportional and we use the term 'size' as a catch-all descriptor. The length is typically a few micrometers, depending on the conditions and the stage in the cell cycle. The mean generation time strongly depends on the medium, and varies from ~ 20 min for the optimal temperature 37 °C, up to several hours. E. coli comes in different strains, that represent sub-families of the species characterized by specific properties. Since E. coli bacteria have flagella that give them motility, in the following experiments specific strains were chosen for their poor motility, or genes that encode flagella were knocked-out, so that cells do not escape the experimental setups unwanted.

Mother machine data from Tanouchi et al., 2017 This dataset is used in chapter 2 to test the convergence of a lineage-based estimator for the population growth rate, and in chapter 4 to test theoretical predictions on single lineage cell size distributions. The MC1400 strain is grown under three temperature conditions: 25 °C, 27 °C and 37 °C in the same experimental setup as the original mother machine developed in Wang et al., 2010, and illustrated on fig. 1.1. The dimensions of the channels are 1 µm (w) × 1 µm (h) × 25 µm (l), so that cells roughly fill all the channel in width and height, and are constrained to grow in one direction, along its length. The nutrients are brought by the medium flow at the open end of the mother machine, and the time-scale of diffusion of the nutrients inside the channel (1 s) is much smaller than the

time-scale of the nutrient uptake by E. coli ($2 - 3\,\text{min}$), so that steady-state conditions are ensured for all cells in the channel (Wang et al., 2010). They acquired 279 lineages of 70 consecutive generations: 65 for 25 °C, 54 for 27 °C and 160 for 37 °C, resulting in 4550, 3780 and 11200 cell cycles respectively. The time lapse interval between two measures is 1 minute, so that at least 20 measurement points are obtained per cell cycle. Data are accessible at: https://figshare.com/collections/Data_from_long-term_growth_data_of_Escherichia_coli_at_a_single-cell_level/3493548

Growing population data from Kiviet et al., 2014 We use cell size and age measurements in population to test our bounds on the strength of selection derived in chapter 2. In these experiments, the MG1655 strain of E. coli is cultured on gel pads containing the necessary medium for bacterial growth, and maintained at 37 °C. Initial cells grow in a $2D$-layer for around 9 generations, resulting in approximately 500 cells. They conducted this experiment for 11 different growth media. The dataset was kindly communicated to us by Philippe Nghe.

Constant population data from Hashimoto et al., 2016 We use constant population data from the dynamic cytometer, illustrated on fig. 3.1 right, to test our bounds on population growth rate when cells are diluted before the end of the experiment. In this setup, cells evolve in a channel that is open at both ends. Its dimensions are 3 µm (w) × 1 µm (h) × 30 µm (l), so that cells are constrains on a $2D$ layer where typically 3 rows of cells can coexist. A flow of medium carries away the excess cells at both ends in order to maintain the population constant around $25 \sim 40$ cells, and brings the necessary nutrients. A precise control of the medium inside the channel is allowed by an additional source of nutrients through a cellulose membrane clamped on top of the channel. Data points are taken regularly each $30\,\text{s} \sim 3\,\text{min}$, depending on the condition, such that 50 time points are acquired per cell cycle. Different experiments are realized, for the two strains W3110 and B/r derivatives, in different media, for two temperatures: 30 °C and 37 °C. The dataset was kindly communicated to us by Yuichi Wakamoto.

Bibliography for the introductory chapter

[Amir, 2014] Amir, A. (2014). Cell Size Regulation in Bacteria. *Physical Review Letters* 112.(20), p. 208102.

[Collin et al., 2005] Collin, D., F. Ritort, C. Jarzynski, S. B. Smith, I. Tinoco, and C. Bustamante (2005). Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies. *Nature* 437.(7056), pp. 231–234.

[Crooks, 1999] Crooks, G. E. (1999). Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Physical Review E* 60.(3), pp. 2721–2726.

[Doumic et al., 2021] Doumic, M. and M. Hoffmann (2021). Individual and population approaches for calibrating division rates in population dynamics: Application to the bacterial cell cycle. *arXiv:2108.13155*.

[Doumic et al., 2015] Doumic, M., M. Hoffmann, N. Krell, and L. Robert (2015). Statistical estimation of a growth-fragmentation model observed on a genealogical tree. *Bernoulli* 21.(3), pp. 1760–1799.

[García-García et al., 2019] García-García, R., A. Genthon, and D. Lacoste (2019). Linking lineage and population observables in biological branching processes. *Physical Review E* 99.(4), p. 042413.

[Genthon et al., 2020] Genthon, A. and D. Lacoste (2020). Fluctuation relations and fitness landscapes of growing cell populations. *Scientific Reports* 10.(1), p. 11889.

[Goldenfeld et al., 2011] Goldenfeld, N. and C. Woese (2011). Life is Physics: Evolution as a Collective Phenomenon Far From Equilibrium. *Annual Review of Condensed Matter Physics* 2.(1), pp. 375–399.

[Hashimoto et al., 2016] Hashimoto, M., T. Nozoe, H. Nakaoka, R. Okura, S. Akiyoshi, K. Kaneko, E. Kussell, and Y. Wakamoto (2016). Noise-driven growth rate gain in clonal cellular populations. *Proceedings of the National Academy of Sciences* 113.(12), pp. 3251–3256.

[Ho et al., 2018] Ho, P.-Y., J. Lin, and A. Amir (2018). Modeling Cell Size Regulation: From Single-Cell-Level Statistics to Molecular Mechanisms and Population-Level Effects. *Annual Review of Biophysics* 47.(1), pp. 251–271.

[Jafarpour et al., 2018] Jafarpour, F., C. S. Wright, H. Gudjonson, J. Riebling, E. Dawson, K. Lo, A. Fiebig, S. Crosson, A. R. Dinner, and S. Iyer-Biswas (2018). Bridging the Timescales of Single-Cell and Population Dynamics. *Physical Review X* 8.(2), p. 021007.

[Jarzynski, 1997] Jarzynski, C. (1997). Nonequilibrium Equality for Free Energy Differences. *Physical Review Letters* 78.(14), pp. 2690–2693.

[Jarzynski, 2006] Jarzynski, C. (2006). Rare events and the convergence of exponentially averaged work values. *Physical Review E* 73.(4), p. 046105.

[Jun et al., 2018] Jun, S., F. Si, R. Pugatch, and M. Scott (2018). Fundamental principles in bacterial physiology—history, recent progress, and the future with focus on cell size control: a review. *Reports on Progress in Physics* 81.(5), p. 056601.

[Kimmel et al., 2002] Kimmel, M. and D. E. Axelrod (2002). *Branching processes in biology*. Springer. Interdisciplinary applied mathematics. New York.

[Kiviet et al., 2014] Kiviet, D. J., P. Nghe, N. Walker, S. Boulineau, V. Sunderlikova, and S. J. Tans (2014). Stochasticity of metabolism and growth at the single-cell level. *Nature* 514.(7522), pp. 376–379.

[Kussell et al., 2014] Kussell, E. and M. Vucelja (2014). Non-equilibrium physics and evolution—adaptation, extinction, and ecology: a Key Issues review. *Reports on Progress in Physics* 77.(10), p. 102602.

[Lebowitz et al., 1974] Lebowitz, J. L. and S. I. Rubinow (1974). A theory for the age and generation time distribution of a microbial population. *Journal of Mathematical Biology* 1.(1), pp. 17–36.

[Leibler et al., 2010] Leibler, S. and E. Kussell (2010). Individual histories and selection in heterogeneous populations. *Proceedings of the National Academy of Sciences* 107.(29), pp. 13183–13188.

[Levien et al., 2020] Levien, E., J. Kondev, and A. Amir (2020). The interplay of phenotypic variability and fitness in finite microbial populations. *Journal of The Royal Society Interface* 17.(166), p. 20190827.

[Lin et al., 2017] Lin, J. and A. Amir (2017). The Effects of Stochasticity at the Single-Cell Level and Cell Size Control on the Population Growth. *Cell Systems* 5.(4), 358–367.e4.

[Lin et al., 2020] Lin, J. and A. Amir (2020). From single-cell variability to population growth. *Physical Review E* 101.(1), p. 012401.

[Liphardt et al., 2002] Liphardt, J., S. Dumont, S. B. Smith, I. Tinoco, and C. Bustamante (2002). Equilibrium Information from Nonequilibrium Measurements in an Experimental Test of Jarzynski's Equality. *Science* 296.(5574), pp. 1832–1835.

[Mustonen et al., 2010] Mustonen, V. and M. Lassig (2010). Fitness flux and ubiquity of adaptive evolution. *Proceedings of the National Academy of Sciences* 107.(9), pp. 4248–4253.

[Nakashima et al., 2020] Nakashima, S., Y. Sughiyama, and T. J. Kobayashi (2020). Lineage EM algorithm for inferring latent states from cellular lineage trees. *Bioinformatics* 36.(9), pp. 2829–2838.

[Neher et al., 2011] Neher, R. A. and B. I. Shraiman (2011). Statistical genetics and evolution of quantitative traits. *Reviews of Modern Physics* 83.(4), pp. 1283–1300.

[Nieto et al., 2020] Nieto, C., J. Arias-Castro, C. Sánchez, C. Vargas-García, and J. M. Pedraza (2020). Unification of cell division control strategies through continuous rate models. *Physical Review E* 101.(2), p. 022401.

[Nozoe et al., 2017] Nozoe, T., E. Kussell, and Y. Wakamoto (2017). Inferring fitness landscapes and selection on phenotypic states from single-cell genealogical data. *PLoS Genetics* 13.(3), e1006653.

[Olivier, 2017] Olivier, A. (2017). How does variability in cell aging and growth rates influence the Malthus parameter? *Kinetic and Related Models* 10.(2), pp. 481–512.

[Peliti et al., 2021] Peliti, L. and S. Pigolotti (2021). *Stochastic Thermodynamics: An Introduction*. Princeton University Press. Princeton.

[Powell, 1956] Powell, E. O. (1956). Growth Rate and Generation Time of Bacteria, with Special Reference to Continuous Culture. *Journal of General Microbiology* 15.(3), pp. 492–511.

[Robert et al., 2014] Robert, L., M. Hoffmann, N. Krell, S. Aymerich, J. Robert, and M. Doumic (2014). Division in Escherichia coli is triggered by a size-sensing rather than a timing mechanism. *BMC Biology* 12.(1), p. 17.

[Schuh et al., 2009] Schuh, R. T. and A. V. Z. Brower (2009). *Biological systematics: principles and applications*. 2nd. Ithaca: Comstock Pub. Associates/Cornell University Press.

[Seifert, 2005] Seifert, U. (2005). Entropy Production along a Stochastic Trajectory and an Integral Fluctuation Theorem. *Physical Review Letters* 95.(4), p. 040602.

[Seifert, 2012] Seifert, U. (2012). Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on Progress in Physics* 75.(12), p. 126001.

[Sekimoto, 1998] Sekimoto, K. (1998). Langevin Equation and Thermodynamics. *Progress of Theoretical Physics Supplement* 130, pp. 17–27.

[Stewart et al., 2005] Stewart, E. J., R. Madden, G. Paul, and F. Taddei (2005). Aging and Death in an Organism That Reproduces by Morphologically Symmetric Division. *PLoS Biology* 3.(2), e45.

[Sughiyama et al., 2015] Sughiyama, Y., T. J. Kobayashi, K. Tsumura, and K. Aihara (2015). Pathwise thermodynamic structure in population dynamics. *Physical Review E* 91.(3), p. 032120.

[Sughiyama et al., 2019] Sughiyama, Y., S. Nakashima, and T. J. Kobayashi (2019). Fitness response relation of a multitype age-structured population dynamics. *Physical Review E* 99.(1), p. 012413.

[Taheri-Araghi et al., 2015] Taheri-Araghi, S., S. Bradde, J. T. Sauls, N. S. Hill, P. A. Levin, J. Paulsson, M. Vergassola, and S. Jun (2015). Cell-Size Control and Homeostasis in Bacteria. *Current Biology* 25.(3), pp. 385–391.

[Tanouchi et al., 2017] Tanouchi, Y., A. Pai, H. Park, S. Huang, N. E. Buchler, and L. You (2017). Long-term growth data of Escherichia coli at a single-cell level. *Scientific Data* 4.(1), p. 170036.

[Trucco et al., 1970] Trucco, E. and G. I. Bell (1970). A note on the dispersionless growth law for single cells. *The Bulletin of Mathematical Biophysics* 32.(4), pp. 475–483.

[Verley, 2012] Verley, G. (2012). Fluctuations et réponse des systèmes hors de l'équilibre. PhD thesis.

[Wang et al., 2010] Wang, P., L. Robert, J. Pelletier, W. L. Dang, F. Taddei, A. Wright, and S. Jun (2010). Robust Growth of Escherichia coli. *Current Biology* 20.(12), pp. 1099–1103.

[Willis et al., 2017] Willis, L. and K. C. Huang (2017). Sizing up the bacterial cell cycle. *Nature Reviews Microbiology* 15.(10), pp. 606–620.

Chapter 2

Lineage-population bias and selection[†]

[†]This chapter is based on the articles García-García et al., 2019; Genthon et al., 2020; Genthon et al., 2021 with the authorization of all co-authors; and on some yet unpublished material.

Contents

1 Introduction

Recent advances in single cell experiments, where the growth and divisions of thousand of individual cells can be tracked, have led to the acquisition of an unprecedented amount of single cell data. For instance, time-lapse video-microscopy experiments of growing cell populations provide information on all the lineages in the population tree (Stewart et al., 2005), and experiments carried out with the mother-machine configuration (Wang et al., 2010) allows to monitor single lineages for many generations. The availability of these data has led to many theoretical progresses in different directions, of which we give some examples.

First, while the growth of cell populations is deterministic, single cell data have revealed stochasticity at the single cell level. This variability can arise, among many possibilities, from the stochasticity in the generation times (Sandler et al., 2015), in the partition of volume at division (Campos et al., 2014), or in single cell growth rates (Taheri-Araghi et al., 2015), which are usually linked to stochastic gene expression (Elowitz et al., 2002). If one is able to disentangle the various sources of stochasticity (Barizien et al., 2019), he can predict how they affect macroscopic parameters of the cell population, such as the population growth rate (Olivier, 2017; Thomas, 2017b; Jafarpour et al., 2018; Lin et al., 2020).

Second, as discussed in section 4.5 of chapter 1, Powell revealed a bias between the distributions of generation times at the single-lineage and population levels, for age-controlled populations without mother-daughter correlations and in steady-state (Powell, 1956). This implies in particular a discrepancy between the mean generation time and the population doubling time: populations of Escherichia coli double faster than the mean doubling time of their constituent single cells (Hashimoto et al., 2016). Following the development of the mother-machine, the study of this lineage-population bias is now a very active field of research, and its manifestation for cell size statistics (Thomas, 2017b; Thomas, 2018; Totis et al., 2021) and gene expression (Thomas, 2017a; Thomas, 2019), for example, have been thoroughly investigated. In particular, this bias goes against a naive view of ergodicity, which states that following a single lineage for a long time should lead to the same statistics as that obtained for an ensemble of cells, and efforts have been made to formulate new ergodic principles (Thomas, 2017a; Rochman et al., 2018). In parallel, a pathwise thermodynamic framework was built for population dynamics using large deviation theory (Sughiyama et al., 2015), which was formulated in terms of two key path distributions. In this work and others that followed, the authors compared the lineage-population bias at the level of path probabilities to fluctuation theorems in stochastic thermodynamics (see section 2 of chapter 1), which map typical behaviors in one ensemble (here the population level) to atypical behaviors in another one (here the single lineage level).

Third, the temporal information available for the lineages within a growing population inspired new developments in the field of evolution. By tracking phenotypes on cell lineages, one can measure selection more accurately than using classical population growth rate measurements (Leibler et al., 2010), and optimal lineage principles have been established to infer the population growth rate (Wakamoto et al., 2012) or selective forces

(Lambert et al., 2015) from lineage statistics. In Nozoe et al., 2017, the authors proposed a measure of selection which can be defined for any branching tree, independently of its dynamics. This measure, called strength of selection, only relies on the forward and backward samplings of the lineages detailed in section 3.2 of chapter 1, and quantifies the distance between the distributions of phenotypic traits of interest when in single lineages or in populations.

Many of the results cited in the second point are model-dependent, but they suggest the existence of more universal lineage-population biases for any branching tree. The framework proposed in Nozoe et al., 2017 is an important step in that direction. In this chapter, we thus build on their framework to relate the single cell and population levels for universal population trees, independently of the model.

In section 2, we show that the comparison between the forward and backward samplings of the lineages of a population tree is similar to fluctuation theorems in stochastic thermodynamics, and thus leads to two important consequences. First, we derive general bounds on the population growth rate involving the forward and backward averages of the number of divisions. In the long time limit, these bounds turn into inequalities for the mean generation times and the population doubling time, that generalize Powell's inequality known for age models. Second, we build an estimator of the population growth rate based on mother-machine data only, and test its convergence with experimental data on E. coli. The estimator is only a function of the distribution for the number of divisions, and thus connects single cell stochasticity and population growth.

In section 3, we derive universal constraints on the strength of selection, interpreted with linear-response theory. We first obtain an inequality for the change in the average value of a function of a cell trait between the forward and backward statistics, which involves the variability of this function and an information-theoretic distance between the two distributions. When applied to the fitness landscape itself, this inequality bounds the strength of selection. In addition, we also derive a set of lower bounds for the strength of selection. These results help understanding the link between selection and variability in fitness in the same way as the link between fluctuations and response is rationalized by fluctuation-dissipation theorems in physical systems.

2 Fluctuation theorem and consequences

2.1 Fluctuation theorem

The forward weight of a lineage l given in eq. (1.38) depends only on its reproductive success, measured by $m^{K(l)}$, which is the size of a population where all lineages would be equivalent to lineage l and have the same number $K(l)$ of divisions. However, it is unaffected by the reproductive performance of the other lineages in the population tree, captured by the number $N(t)$ of cells in the population. The two scales are not independent since the reproductive success of the population can be expressed as a forward

average of that of the lineages:

$$\frac{N(t)}{N_0} = \sum_{i=1}^{N(t)} m^{K(l_i)}\ \omega_{\text{for}}(l_i) = \langle m^K \rangle_{\text{for}}\,. \tag{2.1}$$

On the contrary, the backward weight put on a lineage, given in eq. (1.37), depends on the number of cells at time t, but it is unaffected by the reproductive performance of the lineage considered. Therefore, the ratio of the two weights for a particular lineage informs on the bias between the reproductive performance of that lineage with respect to the colony:

$$\frac{\omega_{\text{back}}(l)}{\omega_{\text{for}}(l)} = \frac{m^{K(l)}}{\langle m^K \rangle_{\text{for}}}\,. \tag{2.2}$$

An equivalent relation can also be obtained in terms of the backward sampling:

$$\frac{\omega_{\text{back}}(l)}{\omega_{\text{for}}(l)} = \frac{\langle m^{-K} \rangle_{\text{back}}}{m^{-K(l)}}\,, \tag{2.3}$$

where we used

$$\frac{N_0}{N(t)} = \sum_{i=1}^{N(t)} m^{-K(l_i)}\ \omega_{\text{back}}(l_i) = \langle m^{-K} \rangle_{\text{back}}\,. \tag{2.4}$$

We now recast the forward/backward bias as a relation for the distributions of joint property y and number of divisions K defined by eq. (1.39) and eq. (1.41):

$$p_{\text{back}}(y, K, t) = p_{\text{for}}(y, K, t)\ e^{K \ln m - t\Lambda_t}\,, \tag{2.5}$$

where we defined the population growth rate

$$\Lambda_t = \frac{1}{t} \ln\left(\frac{N(t)}{N_0}\right)\,. \tag{2.6}$$

The population growth rate is linked to the instantaneous population growth rate $\Lambda_p(t)$ defined in eq. (1.58) by

$$\Lambda_t = \frac{1}{t}\int_0^t \mathrm{d}t'\ \Lambda_p(t')\,. \tag{2.7}$$

Importantly, the bias between the two distributions expressed by eq. (2.5) is only dependent on K and not on y, therefore we give two useful versions of this relation.

The most fundamental version is obtained when no property y is tracked:

$$p_{\text{back}}(K, t) = p_{\text{for}}(K, t)\ e^{K \ln m - t\Lambda_t}\,, \tag{2.8}$$

which indicates that lineages that divided more than average, in the sense of $K \ln m > t\Lambda_t$, are exponentially over-represented in the population as compared to single-lineage experiments. We emphasize that this relation only relies on the forward and backward samplings of the lineages within the population tree, and is thus independent of the dynamics of the system. Moreover, it is valid at any time t, and does not requires a steady-state assumption.

A second version can be written at the level of path probabilities. Let us introduce a vector $s = \{s_i\}$ of variables, possibly of high dimension, to describe the dynamical state of the system. For models of cell size control, the variables s_i can typically be the size and age of the cell, or the concentration of a key protein. A path is then fully characterized by the values $s(t)$ of these variables in time, the number of divisions K along the path, and the generation times of each cycle $\{\tau_k\}_{k=1}^{K}$. We call the collection of these values over time a path $\boldsymbol{\chi}$, and the bold font is used throughout this thesis to indicate trajectories. The probabilities of such a path are linked by

$$p_{\text{back}}(\boldsymbol{\chi}, t) = p_{\text{for}}(\boldsymbol{\chi}, t)\ e^{K[\boldsymbol{\chi}] \ln m - t\Lambda_t}\,. \tag{2.9}$$

In appendix A, we provide a third and complementary version of the fluctuation theorem, using an operator-based framework.

The two versions eqs. (2.8) and (2.9) of the forward/backward bias are akin to detailed fluctuation theorems from stochastic thermodynamics, discussed in section 2.4 of chapter 1. Indeed, eq. (2.8) has a form similar to Crooks relation (eq. (1.34)), where the number of divisions K plays the role of the work w, and the population growth rate Λ_t is the analog of the free energy difference $\Delta\mathcal{F}$. Similarly, eq. (2.9) resembles fluctuation theorems at the level of path probabilities like eq. (1.26), where $t\Lambda_t - K[\boldsymbol{\chi}] \ln m$ is analog to the entropy production $s_{\text{tot}}(\boldsymbol{x})$. Two differences between fluctuation relations in stochastic thermodynamics and in population dynamics must be emphasized. First, thermodynamic fluctuation theorems describe non-autonomous systems which are driven out of equilibrium by the application of a time-dependent protocol, whereas the relations for cell growth derived here concern autonomous systems, in the absence of any external protocol. Second, in stochastic thermodynamics the backward and forward dynamics are linked by time-reversal symmetry, which is not the case for the two samplings of the lineages that are independent of the dynamics on the tree.

These fluctuation theorems for population dynamics have in fact been known in different forms before. First, similar results for simple lineage dynamics were derived in the mathematical literature (see Baake et al., 2007 and Bansaye et al., 2011 for example), and referred to as *many-to-one formulas*. More recently, the Japanese group led by Tetsuya Kobayashi worked on these relations in a series of articles (Kobayashi et al., 2015; Sughiyama et al., 2015; Kobayashi et al., 2017; Sughiyama et al., 2017; Sughiyama et al., 2019), and explored the rich underlying thermodynamic and information-theoretic structure of population dynamics. In this series of works, the authors compared the forward/backward bias to fluctuation theorems on path probabilities in stochastic thermodynamics, and investigated the questions of sensing, feedback and response when the population is in a fluctuating environment, through the lens of information theory.

The connection between free energy and population growth rate has also been observed before, in Baake et al., 2007 and the Japanese articles, using the formalism of large deviations (Touchette, 2009). In this formalism, the rescaled number of divisions $k = K/t$ is said to follow a large deviation principle if the limit $-\lim_{t\to\infty} \ln p_{\text{for}}(k, t)/t$ exists, and is called the rate function $I(k)$. This states that averages of functions of k over $p_{\text{for}}(k, t)$ are dominated in the long time limit by the typical value of k which minimizes the rate function. A similar rate function is also defined for the backward distribution, and the

fluctuation relation can thus be expressed at the level of the rate functions (see next section and Sughiyama et al., 2019). The Gartner-Ellis theorem links the rate function $I(k)$ to the population growth rate via the following variational principle (Levien et al., 2020):

$$\Lambda = \sup_k [k \ln m - I(k)] \,, \tag{2.10}$$

which is obtained by a saddle point approximation. This relation indicates a competition between lineages that divided a lot and those which minimize the rate function $I(k)$. Similar optimal lineage principles have been studied for age-controlled lineages in Wakamoto et al., 2012 for example. By analogy with the variational principle known in statistical physics between the densities of energy, free energy and entropy, the population growth rate and the rate function are identified as the free energy and the entropy respectively.

Importantly, in all these works, the forward/backward bias was obtained for specific lineage dynamics with restrictive assumptions. Only in Nozoe et al., 2017 eq. (2.5) has been derived for general branching trees regardless of the dynamics. However, in this last work, the connection with thermodynamics has not been made explicit.

In addition to these studies, some direct consequences of the detailed fluctuation theorem eq. (2.8) can still be explored, and lead to meaningful messages in the context of population dynamics. In section 2.4 of chapter 1, we recalled how a detailed fluctuation theorem can be turned into an integral fluctuation theorem, and used to infer equilibrium free energies from non-equilibrium work measurements; and how it implies the inequality of the second law of thermodynamics, which expresses a constraint on the trajectories that are allowed or not. Similarly, in the next sections we derive (i) inequalities of the type of the second law generalizing the bounds on the population doubling time known for age models (eq. (1.87)), and (ii) an integral fluctuation theorem that allows the inference of the population growth rate from single-lineage measurements. Importantly, these consequences are very general since eq. (2.8) is independent of the dynamics.

2.2 Bounds on the population doubling time

The inequalities (1.87) between the population doubling time and the average values of the generation time in the forward and backward dynamics, detailed in section 4.5 of chapter 1, are fundamental properties of age models without correlations. We show in this section that the fluctuation theorem can be used to derive universal bounds on the population growth rate, and on the population doubling time when it is defined, independently of the control mechanism.

Using eq. (2.8), we compute the two Kullback-Leibler divergences between the forward and backward distributions for the number of divisions:

$$\mathcal{D}_{\rm KL}(p_{\rm back}||p_{\rm for}) = \langle K \rangle_{\rm back} \ln m - t\Lambda_t \geq 0 \tag{2.11}$$

$$\mathcal{D}_{\rm KL}(p_{\rm for}||p_{\rm back}) = t\Lambda_t - \langle K \rangle_{\rm for} \ln m \quad \geq 0 \,. \tag{2.12}$$

When combined, the two inequalities give

$$\frac{\langle K \rangle_{\rm for} \ln m}{t} \leq \Lambda_t \leq \frac{\langle K \rangle_{\rm back} \ln m}{t} \,, \tag{2.13}$$

which represent constraints on the population growth rate equivalent to the second law of thermodynamics, which classically follows from the fluctuation relations. Note that because the forward and backward samplings are not related by a timer-reversal symmetry as previously mentioned, we obtain two inequalities instead of one for the second law.

For cells following binary fission ($m = 2$), the population doubling time is defined in steady state as $T_d = \ln 2/\Lambda$, where $\Lambda = \lim_{t\to\infty} \Lambda_t$ is the steady-state population growth rate in the regime of exponential growth. For branching trees following an exponential growth in the long time limit, the inequalities for Λ_t are then reshaped as inequalities for T_d:

$$\lim_{t\to\infty} \frac{t}{\langle K\rangle_{\text{back}}} \leq T_d \leq \lim_{t\to\infty} \frac{t}{\langle K\rangle_{\text{for}}} \,. \tag{2.14}$$

Note that eq. (2.14) is a priori different from eq. (1.87) for age models without correlations, since the upper and lower bounds are a priori different from the backward and forward average generation times: $\langle\tau\rangle = \lim_{t\to+\infty} \langle t/K\rangle$. They are equal in particular for age models without correlations where successive generation times are independent, because of the fundamental theorem for renewal processes. More generally, by Jensen's inequality we have $t/\langle K\rangle \leq \langle t/K\rangle$ so that $\lim_{t\to+\infty} t/\langle K\rangle \leq \langle\tau\rangle$ and the right hand side of eq. (1.87) is generalized for any model which ensures a steady-state exponential growth:

$$T_d \leq \langle\tau\rangle_{\text{for}} \,. \tag{2.15}$$

Therefore, the fact that populations double faster than the mean doubling time of their constituent single cells is not a signature of uncorrelated age models but a consequence of the branching structure of the population tree. To define the population doubling time we needed one additional assumption compared to eq. (2.13) which is true independently of the dynamics: the existence of a phase of exponential growth in the long time limit. Thus, eq. (2.15) is valid for any model of cell size control (sizer, timer, adder, ...), irrespective of the precise form of the division rate and of the partition at division, given that this model ensures an exponential growth regime in the long time limit (which has been proven for classical models of cell size control (Doumic et al., 2021)).

We now show a stronger result when the probability $p(k,t)$ for the rescaled number of divisions $k = K/t$ satisfies a large deviation principle:

$$-\lim_{t\to\infty} \frac{1}{t} \ln p(k,t) = I(k) \,, \tag{2.16}$$

where $I(k)$ is the rate function. This function is positive and there is a single point $k^* = \langle k\rangle$ such that $I(k^*) = 0$. Note that we did not precise which distribution follows this principle, since we get from eq. (2.8) that if either the forward or backward distribution follows a large deviation principle, it is also true for the second distribution, with the following relation on the rate functions:

$$I_{\text{back}}(k) = I_{\text{for}}(k) + \Lambda - k \ln m \,. \tag{2.17}$$

In the context of semi-Markov processes, which are processes depending on both the previous state and the time elapsed since the previous jump (like for example multitype

age models presented in section 4.1.1 of chapter 1), large deviation principles are well established (Maes et al., 2009), the rate functions have been computed explicitly (Maes et al., 2009; Sughiyama et al., 2018), and similar fluctuation theorems at the level of rate functions has been previously obtained. For example, in Sughiyama et al., 2019 the authors derived $I_{\text{back}}(j) = I_{\text{for}}(j) + \Lambda - k \ln m$ for $j(x, x', \tau')$ the empirical density of transitions from state x' to sate x at age τ', where the bias is the same as in eq. (2.17). Beyond simple Markov-like cases, it remains open to determine under which dynamical conditions the large deviation principle eq. (2.16) is satisfied.

Let us now re-write:

$$\langle \tau \rangle = \lim_{t\to\infty} \left\langle \frac{t}{K(t)} \right\rangle \tag{2.18}$$

$$= \lim_{t\to\infty} \left\langle \int_0^\infty \mathrm{d}s \; e^{-sK(t)/t} \right\rangle \tag{2.19}$$

$$= \lim_{t\to\infty} \int_0^\infty \mathrm{d}s \int_0^\infty \mathrm{d}k \; p(k) e^{-sk} \, . \tag{2.20}$$

Then, we express $p(k)$ with the rate function, and use a saddle-point approximation of the integral in the limit $t \to \infty$, which leads to:

$$\langle \tau \rangle = \lim_{t\to\infty} \int_0^\infty \mathrm{d}s \int_0^\infty \mathrm{d}k \; e^{-sk - tI(k)} \tag{2.21}$$

$$= \int_0^\infty \mathrm{d}s \; e^{-s\langle k \rangle} \tag{2.22}$$

$$= \langle k \rangle^{-1} \tag{2.23}$$

$$= \lim_{t\to\infty} \frac{t}{\langle K(t) \rangle} \, . \tag{2.24}$$

Finally, the bounds in eq. (2.14) tend to $\langle \tau \rangle_{\text{back}}$ and $\langle \tau \rangle_{\text{for}}$, and we thus recover, for any model ensuring a large deviation principle for the rescaled number of divisions, the inequalities known for uncorrelated age models. Since large deviation principles have been proven for multitype age models, then inequalities eq. (1.87) on mean generation times hold in particular for age models with correlations.

2.3 Inference of the population growth rate from single-lineage measurements

One of the main applications of integral fluctuation theorems like Jarzynski equality concerns the thermodynamic inference of equilibrium free energies from non-equilibrium fluctuations of work. In the same spirit we show here that single lineage data can be used to infer the growth rate of the corresponding freely-growing population. The detailed fluctuation theorem eq. (2.8) can be integrated over the number of divisions K, leading to:

$$\Lambda_t = \frac{1}{t} \ln \langle m^K \rangle_{\text{for}} \, . \tag{2.25}$$

Note that it is also obtained by combining eq. (2.1) and eq. (2.6). Using the equivalence between the statistics obtained in single lineage experiments and the forward procedure,

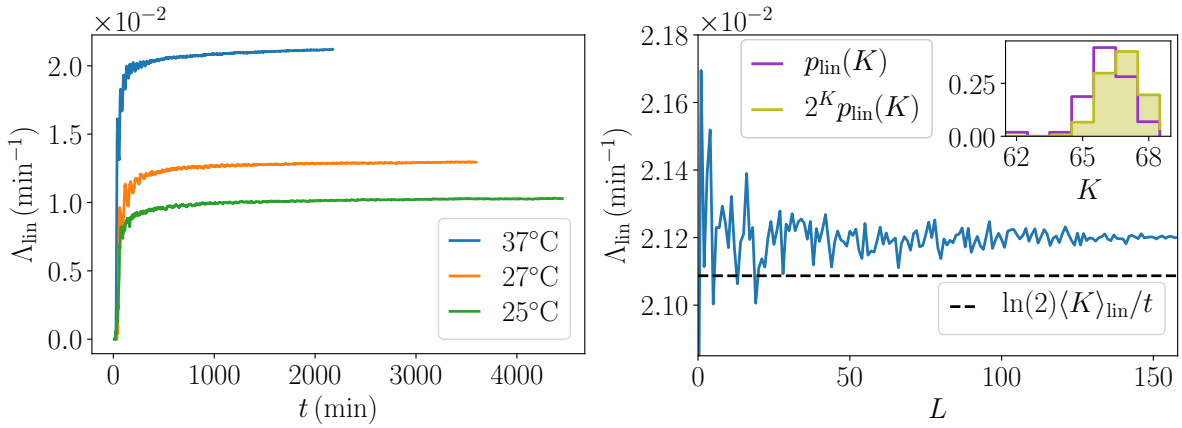


Figure 2.1: Estimator Λ_{lin} of the population growth rate based on eq. (2.26), (left) as function of the the length t of the lineages and (right) as function of the number L of lineages used in the estimation. On the left plot, the curves for the three temperatures converge to a constant value. On the right plot, only the curve for 37 °C is shown and the horizontal dashed line represents the quantity $\ln(2)\langle K\rangle_{\text{lin}}/t$, which is smaller than the limit value of Λ_{lin}, as expected from the second law-like inequality, namely eq. (2.13). In the inset, the purple histogram is the distribution of the number of divisions, while the green filled histogram is the histogram deduced from it by weighting it by a factor 2^K and normalizing. All the 160 lineages were used to plot these histograms.

we built the estimator Λ_{lin} of the population growth rate from the lineage distribution of numbers of divisions:

$$\Lambda_{\text{lin}} = \frac{1}{t} \ln \left[\frac{1}{L} \sum_{i=1}^{L} m^{K_i} \right] , \tag{2.26}$$

where L is the number of independent single lineages.

The population growth rate is often understood as the fitness of the population, and is a central quantity in the understanding of the evolution and natural selection in the population. However, following a large number of cells in time-lapse is challenging, making the measure of the population growth rate difficult. The specific advantage of this estimator is that it only requires single lineage statistics obtained from mother machine experiments, for which cells can be followed for much longer.

Let us now show how this can be done in practice. We use the data from Tanouchi et al., 2017, described in section 5 of chapter 1, where the growth of many independent lineages of E. coli have been recorded over 70 generations in a mother machine at three different temperatures (25 °C, 27 °C, and 37 °C). For each temperature condition, we study the convergence of the estimator of the population growth rate Λ_{lin} as a function of the length t of the lineages for a fixed number of independent lineages L, and as a function of the number of independent lineages for a fixed observation time. Of course, we can only study the convergence of the estimator and not its accuracy, since the true value of the population growth rate cannot be measured independently from the evolution of the population in the mother machine setup.

First, for each temperature, we take into account all the lineages available and truncate

them at an arbitrary time t smaller than the length of the shortest lineage of the set. On these portions of lineages of length t, we compute Λ_{lin} versus the time t as shown in fig. 2.1 left. We see that the estimator Λ_{lin} starts from zero, increases and quickly converges towards a limiting value. The limits we found agree with those from the independent analysis carried out in Levien et al., 2020 on the same data, where the difference in the monotony of the estimator Λ_{lin} with time has been clarified in Levien et al., 2021.

Second, we truncate all the lineages at a fixed time equal to the length of the shortest lineage of the set, and compute Λ_{lin} versus the number L of lineages considered for the estimation, which have been randomly selected from the ensemble of available lineages. As shown in fig. 2.1 right for the experiment at 37 °C (curves for the other temperatures look similar), the convergence is also good in that case. The figure also confirms that the value of the population growth rate deduced from the estimator Λ_{lin} is larger than $\langle K\rangle_{\text{lin}} \ln(2)/t$, as predicted by the right inequality of eq. (2.13).

It is well understood in the context of stochastic thermodynamics that the exponential average of the estimator is dominated by rare values, which are not accessible or not well sampled (Jarzynski, 2006). Indeed, the typical values are the most commonly obtained and for which the work distribution is peaked, whereas the dominating values are the ones for which the work distribution shifted by the exponential bias is peaked, and can be far from typical values. Therefore the inference can require a great number of experiments to correctly sample the region of dominating values. To understand why this problem does not arise here, we show in inset of fig. 2.1 right the distribution $p_{\text{lin}}(K)$ of the number of divisions together with the same distribution weighted by the factor 2^K and normalized. Here, we observe that both distributions have a narrow support and are close to each other: the weighted distribution is peaked at $K = 67$ while $p_{\text{lin}}(K)$ is peaked at $K = 66$, therefore typical and dominating values are very close, which explains why the estimator is good. Of course, a direct consequence of this narrow support is precisely that the estimator Λ_{lin} is very close to the naive estimator $\ln 2\langle K\rangle_{\text{lin}}/t$, thus decreasing its interest. Indeed, looking at the y-axis scale, we see that the difference between Λ_{lin} and $\ln 2\langle K\rangle_{\text{lin}}/t$ is below experimental precision. Applying this inference method to data with more variability in the distribution of the number of divisions, coming for example from noisier cell cycles, would be more useful but at the same time it would require a larger number of independent lineages to converge.

The idea of a lineage-based estimator for the population growth rate has been explored in parallel in Levien et al., 2020 and Pigolotti, 2021 using the framework of large deviation theory mentioned in sections 2.1 and 2.2. In Levien et al., 2020, assuming a large deviation principle and using a Gaussian approximation for the distribution $p_{\text{lin}}(k)$, where $k = K/t$ is the rescaled number of divisions, the authors studied the accuracy of the estimator and pointed out that the errors caused by the finite number L of lineages and by the finite duration t of the experiment, despite being canceled separately in the limits $L \to \infty$ and $t \to \infty$ respectively, could not be canceled simultaneously. This suggests that the convergence of this estimator can be an issue, and that a trade-off between long lineages and numerous lineages has to be found. They showed that the error between the estimator and the true population growth rate is minimized for intermediate lineage length t and large numbers of lineages L.

2.4 Powell's relation for age models

In this section, we show that the fluctuation theorem for path probabilities eq. (2.9) can be used to derive in a simple way Powell's relation for age models in exponential growth, without having to solve for the age or generation time distributions. In age models, a cell trajectory $\boldsymbol{\chi}$ is only characterized by its number K of divisions, the generation times τ_k for each cell cycle, and the final age a of the cell. A fundamental property of age models without correlations is that consecutive cell cycles are independent and the path probability can thus be factorized as a product of generation time distributions $f(\tau)$:

$$p_{\text{for}}(a, \{\tau_k\}, K, t) = g_{\text{for}}(a) \prod_{k=1}^{K} f_{\text{for}}(\tau_k)\ \delta\left(t - a - \sum_{k=1}^{K} \tau_k\right) , \tag{2.27}$$

where $g_{\text{for}}(a)$ is the probability for the cell not to divide from age 0 at time $t-a$ to age a at final time t.

We define the corresponding distributions $f_{\text{back}}(\tau)$ and $g_{\text{back}}(a)$ in the backward dynamics:

$$p_{\text{back}}(a, \{\tau_k\}, K, t) = g_{\text{back}}(a) \prod_{k=1}^{K} f_{\text{back}}(\tau_k)\ \delta\left(t - a - \sum_{k=1}^{K} \tau_k\right) . \tag{2.28}$$

Using the fluctuation relation at the path level, namely eq. (2.9), and making the choice $\forall k \in [1, K],\ \tau_k = \tau = (t-a)/K$, we obtain

$$f_{\text{back}}(\tau) = m f_{\text{for}}(\tau) e^{-\tau \Lambda_t} \left[e^{-a\Lambda} \frac{g_{\text{for}}(a)}{g_{\text{back}}(a)} \right]^{1/K} . \tag{2.29}$$

In a steady state, when $t \to \infty$ and $K \to \infty$, then $\left[e^{-a\Lambda} g_{\text{for}}(a)/g_{\text{back}}(a)\right]^{1/K} \to 1$, so Powell's relation (eq. (1.81)) is recovered:

$$f_{\text{back}}(\tau) = m f_{\text{for}}(\tau) e^{-\tau \Lambda_t} . \tag{2.30}$$

This relation can be seen as a fluctuation theorem at the scale of the single cell cycle, while eq. (2.8) is a fluctuation theorem at the scale of the entire lineage, consisting in K cell cycles.

Although we did not need the expressions of the forward and backward distributions of generation times to derive Powell's relation, they can be obtained from the analytical path probability derived in appendix B:

$$p_{\text{for}}(a, \{\tau_k\}, t, K) = e^{-\int_0^a \mathrm{d}a'\ r(a')} \prod_{k=1}^{K} r(\tau_k) e^{-\int_0^{\tau_k} \mathrm{d}a'\ r(a')}\ \delta\left(t - a - \sum_{k=1}^{K} \tau_k\right) . \tag{2.31}$$

The above solution can be compared with the definition eq. (2.27) to obtain the expressions of the probability not to divide up to age a and of the generation time distribution, in the forward dynamics:

$$g_{\text{for}}(a) = e^{-\int_0^a \mathrm{d}a'\ r(a')} \tag{2.32}$$

$$f_{\text{for}}(\tau) = r(\tau) e^{-\int_0^\tau \mathrm{d}a\ r(a)} . \tag{2.33}$$

The second line shows the consistency between the generation time distribution defined along lineage histories, and that defined with the snapshot age distribution eq. (1.78).

Extensions of Powell's relation to age models with mother-daughter correlations and death are presented in section 3 of chapter 3.

3 Quantifying selection

3.1 On the definitions of fitness and selection

Quantifying the strength of selection in populations is an essential step in any description of evolution. However, the very notion of selection is not agreed upon and many measures of selection have been proposed through the years to overcome successive conceptual difficulties. In this introduction we give an overview of some attempts to define selection and highlight the conceptual link they have with fluctuation-dissipation theorems.

The first quantitative definition of selection is the so-called fundamental theorem of natural selection developed in Fisher, 1930. Two versions of the theorem exist: in discrete time, where each time step is a generation, and in continuous time. In both cases, the fitness associated with a phenotype s is defined as the reproductive success of individuals carrying it. In discrete time, this means the number of offspring of one individual in one generation, and in continuous time it is the division rate, which in this context of quantitative genetics is called the reproduction rate instead. Without lack of generality, in the following we focus on the continuous formulation. The population Fisher studied is very simple: individuals are characterized by their phenotypes s, associated with a fitness $r(s)$, which is not time-dependent, and they can never switch phenotype:

$$\partial_t n(s,t) = r(s) n(s,t)\,. \tag{2.34}$$

This equation is recast at the backward probability level, and since forward probabilities never appear in this introduction, we note p the backward probability without ambiguity:

$$\partial_t p(s,t) = \left[r(s) - \langle r\rangle\right] p(s,t)\,, \tag{2.35}$$

where we used $\Lambda_p(t) = \langle r\rangle$ (eq. (1.65)) with $m = 2$. In this context, the population growth rate is defined as the population fitness, and is thus equal to the average individual fitness value in the population. Fisher's theorem is simply obtained by multiplying this equation by $r(s)$ and integrating over s:

$$\frac{\mathrm{d}\Lambda_p}{\mathrm{d}t} = \mathrm{Var}(r)\,, \tag{2.36}$$

and states that the rate of increase of the population fitness is equal to the variance of fitness in the population. Fisher thus proposed the fitness variance a measure of selection: it is positive when the population is heterogeneous, and when the fittest individuals eventually invade the population, the population growth rate stabilizes at the value of the largest individual fitness.

While insightful, this relation is limited. First of all, it suffers from dynamical insufficiency. Fisher's theorem was intended to predict the evolution of the population fitness

from the knowledge of the fitness variance at an initial time; however, to compute $\mathrm{d}\Lambda_p/\mathrm{d}t$ at later times, you also need the variance at later times, whose evolution is given by

$$\frac{\mathrm{dVar}(r)}{\mathrm{d}t} = \langle (r - \langle r \rangle)^3 \rangle \,, \tag{2.37}$$

and so on. Indeed, eq. (2.35) gives rise to infinitely many moment equations. Although in some cases higher moments can be suppressed (Neher et al., 2011), in general you need to know all moments, that is the full distribution, to predict the population fitness at later times. This issue has been tackled in some works, like Smerlak et al., 2017, in which the authors explore the possibility to predict the long run dynamics from the knowledge of the high fitness tail of the initial fitness distribution only, characterized by a single number.

Second and more important, the theorem is only valid for the very simple model defined by eq. (2.34), that does not capture very important phenomena in biology such as mutations, genetic drift, environment fluctuations, ... This criticism has triggered many developments, some of which we present now.

Price generalized Fisher's theorem for a general function $g(s,t)$ of trait s and time (Price, 1972). After multiplication of eq. (2.35) by $g(s,t)$ and integration, Price's equation is obtained:

$$\frac{\mathrm{d}\langle g \rangle}{\mathrm{d}t} = \mathrm{Cov}(g,r) + \langle \partial_t g \rangle \,. \tag{2.38}$$

The time evolution of the average value of an arbitrary trait is split into two contributions: the covariance term, akin to Fisher's variance, describes the effect of natural selection, and the second accounts for the variations in $\langle g \rangle$ that are due to anything but selection. More precisely, the last term is called 'environment change term' and reflects a dynamical effect when trait g is time-dependent. Then any phenomenon changing the values of the trait in time is included in this term. When applied to $g(s,t) = r(s)$, Price's equation gives back Fisher's theorem. This result suffers from the same dynamic insufficiency as Fisher's theorem, and its very purpose is not to provide a quantitative prediction for the evolution of $\langle g \rangle$ but rather to propose a general definition of selection effect and environment effect (Gardner, 2020). By allowing the trait to be time-dependent, Price provided some partial answers to the lack of generality of Fisher's theorem, but his derivation is still based on the simple model eq. (2.34).

When phenotype switching is allowed for example, the dynamical equation reads:

$$\partial_t n(s,t) = r(s)n(s,t) + \int \mathrm{d}s' \; T(s,s')n(s',t) \,, \tag{2.39}$$

where $T(s,s')$ is the rate of switching from s' to s. Price's equation can of course be generalized for this model and an extra term appears in eq. (2.38) due to switching (Leibler et al., 2010).

In Leibler et al., 2010, the authors proposed to shift the perspective from individuals to individuals' histories, and to define the historical fitness associated with phenotypic trajectory $\boldsymbol{s}$:

$$H_t(\boldsymbol{s}) = \frac{1}{t}\int_0^t r(s(t'))\mathrm{d}t' \,, \tag{2.40}$$

which takes advantage of the large amount of single cell data to define selection more accurately while maintaining the intuitive understanding of Fisher and Price's results. Note that we included a factor $1/t$ in the definition of H_t which is not present in the original paper, to make it consistent with the population growth rate and the fitness landscape presented below. The authors proposed the following measure of selection: when all replications rates are multiplied by $\beta \geq 1$, which amplifies the selective differences, the historical mean fitness $\langle H_t \rangle$ is changed by $\partial_\beta \langle H_t \rangle$. This quantity gauges the population response to a change in selective differences, and provides a measure of selection different from the increase of population growth rate in time proposed by Fisher. In the context of eq. (2.39), which accounts for independent replications and mutations, this measure is related to the variance in historical fitness:

$$\partial_\beta \langle H_t \rangle = t \mathrm{Var}(H_t)\,. \tag{2.41}$$

similarly to Fisher's theorem. Of course, one cannot easily increase all the replication rates by a factor β in an experiment, but when evaluated at $\beta = 1$, the variance in the right hand side is the variance in the original experiment. The authors argued that this variance in historical fitness provides a better measure of selection than the variance in reproductive rate proposed by Fisher, because the latter captures both individual responses and selection effects at the level of the population, which are difficult to disentangle. For example, if phenotype switching is much faster than replication, the variance in replication rates is largely due to the changes in phenotype, and does not reflect a selection effect. Again, eq. (2.41) was derived only in the context of eq. (2.39), and we show in appendix C.3 how this result is modified when applied to models of cell size control. In this context, $\partial_\beta \langle H_t \rangle$ is no longer equal to $\mathrm{Var}(H_t)$ and is no longer positive-definite, and may not be a suitable measure of selection.

An alternative method to define selection focuses on population trajectories of the frequency distribution $p(s,t)$ instead of individual trajectories, and introduces the notion of fitness flux to characterize the adaptation of a population by taking inspiration from stochastic thermodynamics (Mustonen et al., 2009; Mustonen et al., 2010). In a different direction, an optimal lineage principle can be used to infer the population growth rate (Wakamoto et al., 2012) or selective forces (Lambert et al., 2015) from lineage statistics. All these methods contribute to bridging the gap between single-cell experiments at the population level and molecular mechanisms.

Let us now comment on the link between all these results and the fluctuation-dissipation theorem. This theorem, and the field of linear response theory to which it belongs, states that near equilibrium the response of a system following a small perturbation is proportional to the fluctuations of the system in equilibrium (Kubo, 1966). The first and most simple example is Einstein's relation $D = \mu k_B T$ (Einstein, 1905), obtained in equilibrium by balancing the diffusion current with the particles flow due to an external driving force. The mobility μ, which is the inverse of the drag coefficient and which represents viscous dissipation following the application of a driving force, is proportional to fluctuations in Brownian velocity captured by the diffusion coefficient D. These ideas have important implications in biology explored for example in Sato et al., 2003, where the change in the average value of a variable following the small change in an exter-

nal parameter is proven in simple cases proportional to its variance in the unperturbed ensemble.

All the results mentioned above on selection are also expressed in the form fluctuation-response relations. In these cases, the responses of the population to the differences in reproductive rates, captured either by $\mathrm{d}\Lambda_p/\mathrm{d}t$ or $\partial_\beta\langle H_t\rangle$, are related to the fluctuations in the population in the form of the variance of fitness. In Leibler et al., 2010, this view is even at the core of the definition of $\partial_\beta\langle H_t\rangle$, which is similar to slightly change the temperature of a system in thermal equilibrium and see how the system responds.

A generalization of the notion of historical fitness was proposed in Nozoe et al., 2017, based on the forward and backward lineage phenotypic histories. This framework gave rise to a new measure of the strength of selection, which overcomes the difficulty of the previous measures: it is model-independent, and can be evaluated for any population tree, regardless of the dynamics. The authors gave a first hint at a linear-response theory linking the strength of selection to the variance in fitness landscape, in the case where the latter follows a Gaussian distribution. In the next sections, we use this framework and seek more general linear-response relations for this universal measure of selection, beyond the Gaussian case.

3.2 Fitness landscape

3.2.1 Definition and properties

We now introduce a general cell trait $\mathcal{S}$, understood in the broad sense of any cell property. It can be quantitative, either defined in a snapshot like cell size, or defined as a time-integrated quantity like the average single cell elongation rate along the cell lineage, for example. In this case the trait can take different *values*, that we note s. The trait can also be non-quantitative and represent a state, like for instance the belonging to the group of wild-type cells or mutants. In that case, the trait comes in different *versions*, or *phenotypes*, that we also note s. For simplicity, in the following we will not distinguish between these cases, and call traits all these properties. In both cases, trait trajectories are indicated with the bold symbols $\boldsymbol{\mathcal{S}}$ and $\boldsymbol{s}$.

In Nozoe et al., 2017, the authors suggested that one way to define the fitness associated with the value s of trait $\mathcal{S}$ could be to compare the chronological and retrospective marginal probabilities of that trait, and they defined the fitness landscape:

$$h_t(s) = \Lambda_t + \frac{1}{t}\ln\left[\frac{p_{\mathrm{back}}(s,t)}{p_{\mathrm{for}}(s,t)}\right], \tag{2.42}$$

where the marginal probabilities are obtained from the joint distributions of s and K:

$$\begin{aligned} p_{\mathrm{back}}(s,t) &= \sum_{K=0}^{\infty} p_{\mathrm{back}}(s,K,t) \\ &= N(t)^{-1} n(s,t)\,, \end{aligned} \tag{2.43}$$

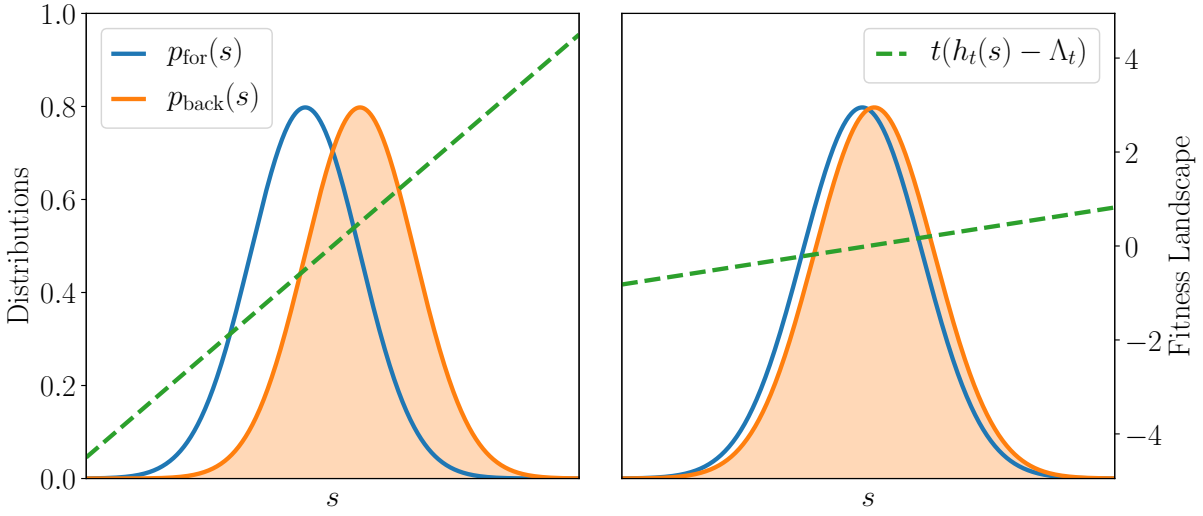


Figure 2.2: Illustrations of the fitness landscape for strong and weak selection. On the left plot, the backward and forward distributions of trait $\mathcal{S}$ are significantly different and the fitness landscape is then much steeper than on the right plot where it is almost flat because the two distributions nearly coincide. On both plots, the fitness landscapes are linear because the two distributions are Gaussian with the same variance.

and

$$p_{\text{for}}(s,t) = \sum_{K=0}^{\infty} p_{\text{for}}(s,K,t) \\ = N_0^{-1} \sum_{K=0}^{\infty} m^{-K} n(s,K,t)\,. \tag{2.44}$$

Note that in the classical framework of evolutionary dynamics, the notion of fitness landscape finds its origin in Wright's seminal work (Wright, 1932), and is defined as a mapping between the values or versions of a phenotype or a genotype, with their associated fitnesses (Peliti, 1997).

The definition eq. (2.42) can be reshaped as a fluctuation theorem

$$p_{\text{back}}(s,t) = p_{\text{for}}(s,t) \exp\left[t(h_t(s) - \Lambda_t)\right]\,, \tag{2.45}$$

which suggests that the fitness landscape $h_t(s)$ plays a role similar to that of an effective division rate, depending on the trait s. This expression indicates that cells with values s of trait $\mathcal{S}$ are over-represented in the backward sampling as compared to the forward one if the corresponding fitness landscape $h_t(s)$ is larger than the population growth rate, and vice versa. This is illustrated on fig. 2.2 for two situations where the forward and backward distributions are Gaussian with the same standard deviation, so that the fitness landscape is linear. On the left, there is a significant difference between the two distributions, with large values s more represented in the backward distribution, so that the fitness landscape has a large positive slope. On the contrary, on the right plot, the two distributions almost coincide, resulting in a near-flat fitness landscape.

We warn the reader right away about the subtle differences between this fitness landscape and the other notions of fitness mentioned in the introduction. The growth rate of the subpopulation carrying value s of trait $\mathcal{S}$ is defined as

$$\Lambda_t(s) = \frac{1}{t} \ln \left[\frac{n(s,t)}{n(s,0)} \right] = \Lambda_t + \frac{1}{t} \ln \left[\frac{p_{\text{back}}(s,t)}{p_{\text{back}}(s,0)} \right] . \tag{2.46}$$

Therefore the fitness landscape and the subpopulation growth rate differ by

$$h_t(s) - \Lambda_t(s) = \frac{1}{t} \ln \left[\frac{p_{\text{back}}(s,0)}{p_{\text{for}}(s,t)} \right] = \frac{1}{t} \ln \left[\frac{p_{\text{for}}(s,0)}{p_{\text{for}}(s,t)} \right] , \tag{2.47}$$

where the last equality follows from the fact that, at $t = 0$, cells have not divided yet and so the forward and backward samplings of the population are identical. The two notions of fitness match when the forward distribution of trait $\mathcal{S}$ is constant, which is for example the case in Fisher's very simple example where there is no mutation. Otherwise, the sign of $h_t(s) - \Lambda_t(s)$ indicates the evolution of the frequency of the value s of trait $\mathcal{S}$ as time goes by, due to every phenomenon but selection. This measure is complementary to $h_t(s) - \Lambda_t$ which quantifies the separate effect of selection. Therefore, $h_t(s_1) > h_t(s_2)$ means that the trait value s_1 benefits more from selection than s_2, in the sense that its frequency is increased by a greater amount when going from single lineage statistics to population statistics, but not necessarily that s_1 has a greater reproductive success than s_2. As a consequence, cells carrying the value s_1 could still be less represented in the population than those carrying trait value s_2.

The fitness landscape is also in general different from the historical fitness, although it is built to recover the latter in simple situations. For models of independent reproductions and mutations, the fitness landscape and the historical fitness are identical (Nozoe et al., 2017, SM), and we give an alternative proof of this equality in appendix C.1. Nonetheless, the fitness landscape is easier to evaluate experimentally: one has to compare the backward and forward probabilities of a path, that can be followed by fluorescence for example, while evaluating the historical fitness requires an estimation of the reproductive rate. On the other hand, we show in appendix C.2 that for models of cell size, the two notions of fitness are different, but still linked in a non-trivial way.

To gain further insight, we rewrite the definition of $h_t(s)$ in a slightly different way using (Nozoe et al., 2017, SM)

$$\begin{aligned} p_{\text{back}}(s,t) &= \sum_K p_{\text{back}}(s,K,t) \\ &= e^{-t\Lambda_t} \sum_K 2^K p_{\text{for}}(s,K,t) \\ &= e^{-t\Lambda_t} p_{\text{for}}(s,t) \sum_K 2^K R_{\text{for}}(K,t|s), \end{aligned} \tag{2.48}$$

where we have used the fluctuation theorem eq. (2.5) with $y = s$ and where we introduced the probability of the number of divisions conditioned on trait s at the forward level, $R_{\text{for}}(K,t|s)$. The fitness landscape then reads

$$h_t(s) = \frac{1}{t} \ln \left[\sum_K 2^K R_{\text{for}}(K,t|s) \right] . \tag{2.49}$$

In this form, it appears clearly that the fitness landscape measures the degree of correlation between the final value of a cell trait and the number of divisions along the lineage. Since lineages that divided more than average are over-represented in the population as compared to the single-lineage statistics, if the number of divisions is positively correlated with certain values of the trait, then their frequencies are also increased.

Two limit cases are worth mentioning. First, when the trait $\mathcal{S}$ and the number $\mathcal{K}$ of divisions are uncorrelated, then $R_{\text{for}}(K,t|s) = p_{\text{for}}(K,t)$ and eq. (2.49) reads $h_t(s) = \ln\left[\sum_K 2^K p_{\text{for}}(K,t)\right]/t$, which is equal to Λ_t according to eq. (2.25). In this case the fitness landscape is flat. On the other hand, if $\mathcal{S}$ and $\mathcal{K}$ are fully correlated, in the sense that the number K of divisions is determined by a function $K(s)$ of trait value s at final time, then $R_{\text{for}}(K,t|s) = \delta(K - K(s))$ and the fitness landscape is equal to the lineage fitness $h_t(s) = \hat{h}_t(K(s)) = K(s)\ln 2/t$, which is how we call the fitness landscape when considering the number of divisions as the trait $\mathcal{S}$. A simple example of that is the complete correlation of size and number of division in the absence of noise. Indeed, for deterministic and symmetrical partition of volume ($\Sigma(x|x') = \delta(x - x'/2)$) and exponential growth ($\nu(x) = \nu x$), for a given initial size x_0, the accessible sizes after a time t are given by: $x(t) \in \{x_0 \exp[\nu t]/2^K, K \in \mathbb{N}\}$ independently of the division times and of the size control mechanism. Therefore, K is determined by the function $K(x, x_0) = \ln[x_0 \exp(\nu t)/x]/\ln 2$, and

$$\begin{aligned} h_t(x, x_0) &= K(x, x_0)\ln 2/t \\ &= \nu + \frac{1}{t}\ln\left(\frac{x_0}{x}\right). \end{aligned} \tag{2.50}$$

The fact that this fitness landscape is a decreasing function of size x is coherent with the over-representation of cells that divided a lot, since these cells are mechanically smaller due to the numerous divisions. Reporting this result in eq. (2.48), we obtain a fluctuation relation for the size

$$p_{\text{back}}(x, x_0, t) = e^{(\nu - \Lambda_t)t}\frac{x_0}{x}p_{\text{for}}(x, x_0, t)\,. \tag{2.51}$$

When introducing noises either on partitioning or growth, the correlations between size and divisions can be blurred and the fitness landscape can become non-trivial. The study of the lineage-population bias for the cell size distribution is the topic of chapter 4.

3.2.2 A digression: detection of mother-daughter correlations

Although the main use of the fitness landscape is probably the inference of selection from lineage data, we explore in this short section the possibility to use it to detect mother-daughter correlations. If the duration of all cell cycles in the tree are recorded, then mother-daughter correlations in generation times are straight-forwardly analyzed by computing the covariances between consecutive generation times (Taheri-Araghi et al., 2015). However, this method can be expensive, and if only sparse information is known, like the number of divisions along the lineages and the steady-state age distributions but not the actual division times, then correlations can be revealed by looking at age distributions.

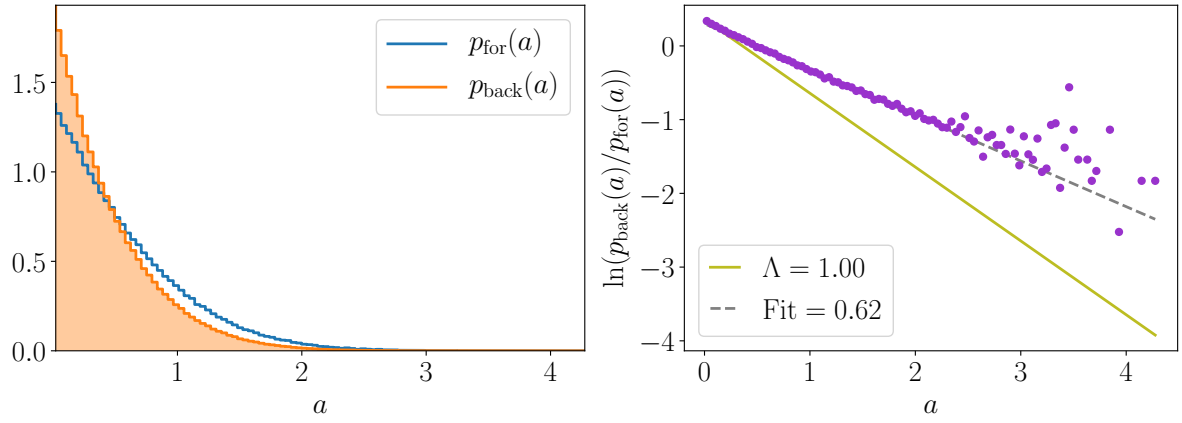


Figure 2.3: Simulation of a cell colony following a sizer mechanism, with $N_0 = 1$, $r(x) = \nu(x) = x$, $\Sigma(x|x') = b(x/x')/x'$ with $b(x)$ uniform in $[0.35, 0.65]$ and 0 otherwise, ending with more than 1.6×10^6 cells at final time $t = 15$. Left: forward and backward age distributions shown with blue empty histogram and orange filled histogram respectively. Right: $\ln[p_{\text{back}}(a)/p_{\text{for}}(a)]$ is linear with age a but with a slope ~ 0.62 different from the measured steady-state population growth rate $\Lambda = 1$, indicating the presence of mother-daughter correlations in generation times.

Indeed, in the timer model without correlations, the backward and forward age distributions are analytical and given by eqs. (1.75) and (1.76). Thus, computing

$$\ln\left[\frac{p_{\text{back}}(a)}{p_{\text{for}}(a)}\right] = -\Lambda a + \ln\left[\frac{p_{\text{back}}(0)}{p_{\text{for}}(0)}\right] \tag{2.52}$$

shows a linear dependence in age with a slope equal to the steady-state population growth rate. From this result, we deduce that if $\ln[p_{\text{back}}(a)/p_{\text{for}}(a)]$ is not linear with a slope equal to Λ, then the population is incompatible with the classical timer description, which indicates the presence of correlations in generation times. The opposite is not true, as eq. (2.52) could hold for models with specific correlations.

We illustrate this by simulating the evolution of a cell colony following a sizer mechanism. We show on fig. 2.3 left the steady state forward and backward age distributions, and on fig. 2.3 right the logarithm of the ratio of the two probabilities. We see that $\ln[p_{\text{back}}(a)/p_{\text{for}}(a)]$ is also linear in that case, but with a slope significantly different from the measured population growth rate, which indicates non-ambiguously that these age statistics are incompatible with the assumption of independent generation times.

This discussion is an opportunity to mention a property of the fitness landscape. From its definition eq. (2.42), we see that if steady-state distributions are reached for $p_{\text{back}}(s)$ and $p_{\text{for}}(s)$, then in the long time limit the fitness landscape becomes equal to the population growth rate. This is why we focused on the quantity $\ln[p_{\text{back}}(a)/p_{\text{for}}(a)] = t(h_t(a) - \Lambda)$ instead of $h_t(a)$, because it remains a non-flat function of age in the long time limit and still encodes information about the forward/backward bias.

3.3 Strength of selection

The strength of selection $\Pi_{\mathcal{S}}$ acting on trait $\mathcal{S}$ has been defined in Nozoe et al., 2017 as the Jeffrey's divergence between the forward and backward distributions of that trait, which is a non-negative and symmetric information-theoretic distance between the two distributions:

$$\Pi_{\mathcal{S}} = \frac{1}{t}\mathcal{J}(p_{\text{back}}(s,t)|p_{\text{for}}(s,t)) \tag{2.53}$$

$$= \frac{1}{t}\int \mathrm{d}s \ [p_{\text{back}}(s,t) - p_{\text{for}}(s,t)] \ln\left(\frac{p_{\text{back}}(s,t)}{p_{\text{for}}(s,t)}\right) . \tag{2.54}$$

This distance can be written in a different form:

$$\Pi_{\mathcal{S}} = \langle h_t \rangle_{\text{back}} - \langle h_t \rangle_{\text{for}} , \tag{2.55}$$

showing that the strength of selection is the change in mean fitness landscape between the ensembles with and without selection, or equivalently the response of the system to the presence of selection. Please be aware that this strength of selection should not be confused with the coefficient of selection, usually defined as the relative difference in fitness associated with two values of a phenotypic trait (Mustonen et al., 2010). Note also that this strength of selection is defined for a given trait $\mathcal{S}$ and reflects the degree of correlation between this trait and reproductive success, and thus the selection of certain values of this trait. This is a difference with some measures presented in section 3.1 such as the one proposed by Fisher, which quantify the selection of the fittest individuals.

Nozoe et al. proved that when the fitness landscape $h_t(s)$ is a bijection and is normally-distributed, the strength of selection is proportional to the variance in fitness landscape:

$$\Pi_{\mathcal{S}} = t\mathrm{Var}(h_t) , \tag{2.56}$$

where the variance can evaluated either with the forward or backward distribution: $\mathrm{Var}(h_t) = \mathrm{Var}_{\text{back}}(h_t) = \mathrm{Var}_{\text{for}}(h_t)$. This result provides a link between response and variability, understood within the framework of linear response theory as detailed in section 3.1. We provide an alternative proof of this linear relation in appendix D.1.

However, the Gaussian case only covers a small portion of realistic cases, and fitness landscapes can exhibit strong deviations from Gaussian distributions. For example, we show on fig. 2.4 experimental examples of fitness landscape distributions for size (left) and age (right) that are non Gaussian, using population data from Kiviet et al., 2014 described in section 5 of chapter 1.

In the following, we derive universal relations going beyond the Gaussian assumption, and obtain a set of upper and lower bounds for the strength of selection, in terms of both the forward and backward variances in fitness landscape. To do so, let us first derive in the next section a general inequality constraining the difference in average value for an observable between two probability distributions.

3.3.1 General fluctuation-response inequality

We consider a general system described by a reference probability distribution $p_{\text{a}}(s,t)$, where s is the value taken by a state variable $\mathcal{S}$. By perturbing the system, we change

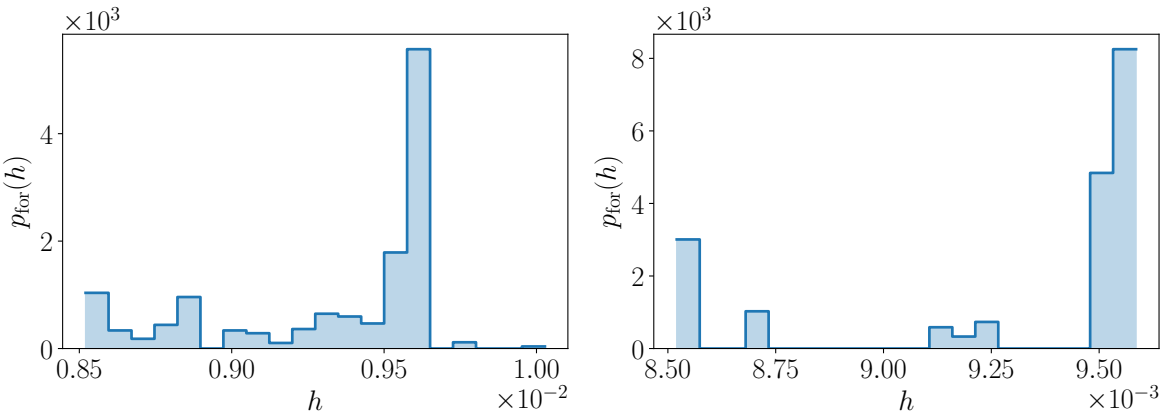


Figure 2.4: Forward distributions of fitness landscape for size (left plot) and age (right plot) for one experiment from Kiviet et al., 2014. For both variable, the distribution is far from a Gaussian.

the distribution of the variable $\mathcal{S}$ from $p_{\mathrm{a}}(s,t)$ to $p_{\mathrm{b}}(s,t)$. We consider an observable depending on the variable $\mathcal{S}$, through a function $g_t(s)$, and ask the question of how the mean value of this observable is modified when the system is perturbed.

Assuming that $p_{\mathrm{a}}(s,t)$ and $p_{\mathrm{b}}(s,t)$ have the same support, we can define the ratio

$$q_t(s) = \frac{p_{\mathrm{b}}(s,t)}{p_{\mathrm{a}}(s,t)}\,. \tag{2.57}$$

Let us now compute the covariance between $g_t(s)$ and $q_t(s)$ with respect to $p_{\mathrm{a}}(s,t)$:

$$\begin{aligned}\mathrm{Cov}_{\mathrm{a}}(g_t,q_t) &= \langle g_t q_t\rangle_{\mathrm{a}} - \langle g_t\rangle_{\mathrm{a}}\langle q_t\rangle_{\mathrm{a}} \\ &= \langle g_t\rangle_{\mathrm{b}} - \langle g_t\rangle_{\mathrm{a}}\,,\end{aligned} \tag{2.58}$$

where we used $\langle q_t\rangle_{\mathrm{a}} = 1$, due to the normalization of p_{b}, and $\langle q_t g\rangle_{\mathrm{a}} = \langle g_t\rangle_{\mathrm{b}}$. Following the method used in Dinis et al., 2020 to derive mean-variance trade-off bounds in horse race gambling, we use the Cauchy-Schwarz inequality for the covariance:

$$\mathrm{Cov}_{\mathrm{a}}(g_t,q_t)^2 \leq \sigma_{\mathrm{a}}^2(g_t)\sigma_{\mathrm{a}}^2(q_t)\,, \tag{2.59}$$

with σ_{a}^2 the variance with respect to $p_{\mathrm{a}}(s,t)$. Finally, by combining eqs. (2.58) and (2.59), we obtain a general bound for the difference in average values:

$$|\langle g_t\rangle_{\mathrm{b}} - \langle g_t\rangle_{\mathrm{a}}| \leq \sigma_{\mathrm{a}}(g_t)\sigma_{\mathrm{a}}(q_t)\,. \tag{2.60}$$

The inequality can be understood as an out-of-equilibrium generalization of the fluctuation-dissipation theorem, because it involves a comparison between a reference unperturbed dynamics and a perturbed dynamics. The difference between the unperturbed and the perturbed averages of the function $g_t(s)$ is bounded by the unperturbed fluctuations of this function, measured by $\sigma_{\mathrm{a}}(g_t)$, times $\sigma_{\mathrm{a}}(q_t)$ which can be seen as an

information-theoretic distance between the two probability distributions. Indeed, since $\langle q_t \rangle_{\mathrm{a}} = 1$ by construction, the variance of q_t is given by

$$\sigma_{\mathrm{a}}^2(q_t) = \int \mathrm{d}s \, \left(\frac{p_{\mathrm{b}}(s)}{p_{\mathrm{a}}(s)} - 1 \right)^2 p_{\mathrm{a}}(s) \,, \tag{2.61}$$

which is none other than $\chi^2(p_{\mathrm{b}}; p_{\mathrm{a}})$: the χ^2 divergence from p_{a} to p_{b}. Therefore, the eq. (2.60) we derived is a particular form of Chapman-Robbins bound:

$$\sigma_{\mathrm{a}}^2(g_t) \geq \sup_{p_{\mathrm{b}}} \frac{\left(\langle g_t \rangle_{\mathrm{b}} - \langle g_t \rangle_{\mathrm{a}} \right)^2}{\chi^2(p_{\mathrm{b}}; p_{\mathrm{a}})} \,, \tag{2.62}$$

which we did not know at that time. The perspective is reversed in the latter: while our inequality bounds the change in the mean value of a function between two ensembles, Chapman-Robbins bound is a lower bound on the variance of this function, which is typically a function of a parameter to estimate.

To derive eq. (2.60), we adopted the point of view of the unperturbed statistics $p_{\mathrm{a}}(s,t)$ as reference, but a similar bound can be obtained in terms of standard deviations with respect to the perturbed dynamics $p_{\mathrm{b}}(s,t)$. We consider the covariance between $g_t(s)$ and $q_t^{-1}(s)$, with respect to $p_{\mathrm{b}}(s,t)$:

$$\mathrm{Cov}_{\mathrm{b}}(g_t, q_t^{-1}) = \langle g_t \rangle_{\mathrm{a}} - \langle g_t \rangle_{\mathrm{b}} \,. \tag{2.63}$$

Following the same steps, and using the Cauchy-Schwarz inequality for this covariance we finally obtain

$$|\langle g_t \rangle_{\mathrm{b}} - \langle g_t \rangle_{\mathrm{a}}| \leq \sigma_{\mathrm{b}}(g_t) \sigma_{\mathrm{b}}(q_t^{-1}) \,, \tag{2.64}$$

where the term $\sigma_{\mathrm{b}}(q_t^{-1})$ is similarly interpreted as an information-theoretic distance measure between the two distributions, and equal to the χ^2 divergence from p_{b} to p_{a}

Thus, combining eqs. (2.60) and (2.64), the change in mean value of the function g_t is bounded by

$$|\langle g_t \rangle_{\mathrm{b}} - \langle g_t \rangle_{\mathrm{a}}| \leq \min \left(\sigma_{\mathrm{a}}(g_t) \sigma_{\mathrm{a}}(q_t), \sigma_{\mathrm{b}}(g_t) \sigma_{\mathrm{b}}(q_t^{-1}) \right) \,. \tag{2.65}$$

A similar bound for $|\langle g_t \rangle_{\mathrm{b}} - \langle g_t \rangle_{\mathrm{a}}|$ was derived in Dechant et al., 2020, using Jensen's inequality. Their bound (eq. 5 or 11 in their text) also involves a measure of the distance between the two probability distributions (Kullback-Leibler divergence) and the standard deviation of the observable considered in the unperturbed dynamics. When applied to path probabilities for observables that are odd under time reversal symmetry, their result recovers Thermodynamic Uncertainty Relations. These relations take the form of inequalities, which generalize fluctuation-response relations far from equilibrium and which capture important trade-offs for thermodynamic and non-thermodynamic systems as recently reviewed in Horowitz et al., 2020. We carry out a numerical comparison between the two bounds in appendix E which shows that the relative performance of the two bounds depends on the shape of the perturbed and unperturbed distributions. In any case, our bound is easy to evaluate since it does not require an optimization over a free parameter, as it is the case in Dechant et al., 2020 (see eq. (2.147)).

3.3.2 Fluctuation-response inequality for the strength of selection

The results derived in the previous section for general distributions a and b are now used to obtain constraints on the strength of selection. Indeed, by setting the unperturbed distribution a to be the forward distribution of a phenotypic trait $\mathcal{S}$ and the perturbed distribution b to be the backward distribution of this trait (which is allowed since the forward and backward distributions have the same support), the difference $\langle g_t \rangle_{\text{back}} - \langle g_t \rangle_{\text{for}}$ is the change of mean value for function g_t between an ensemble without selection (forward) and with selection (backward).

An important application of the above results is when the arbitrary function $g_t(s)$ is the fitness landscape $h_t(s)$ itself. In this case, eqs. (2.58) and (2.63) read

$$\Pi_{\mathcal{S}} = \text{Cov}_{\text{for}}(h_t, e^{th_t})\, e^{-t\Lambda_t} \tag{2.66}$$

$$= \text{Cov}_{\text{back}}(h_t, e^{-th_t})\, e^{t\Lambda_t}\,, \tag{2.67}$$

which generalize the linear relation between the strength of selection and the variance of the fitness landscape, valid in the Gaussian case (eq. (2.56)). To make explicit the role of the variability of the fitness landscape in the fashion of fluctuation-response relations, we write eq. (2.65) in this context:

$$\Pi_{\mathcal{S}} \leq \min\left(\sigma_{\text{for}}(h_t)\sigma_{\text{for}}(q_t), \sigma_{\text{back}}(h_t)\sigma_{\text{back}}(q_t^{-1})\right)\,, \tag{2.68}$$

where the absolute values in the left hand side can be removed because the strength of selection is defined positive, as deduced from eq. (2.53). We finally obtained a universal upper bound for the strength of selection acting on trait $\mathcal{S}$, which involves the χ^2 distances $\sigma_{\text{for}}(q_t)$ and $\sigma_{\text{back}}(q_t^{-1})$ between the backward and forward statistics, and the variances of the fitness landscape in both ensembles, which are in general different from each other.

Let us make some important comments on this result.

First, in this context, the ratio $q_t(s)$ and the fitness landscape $h_t(s)$ are linked by the simple relation $q_t(s) = \exp\left[t\left(h_t(s) - \Lambda_t\right)\right]$. Consequently, even though $\sigma(h_t)$ is interpreted as the variability of fitness in the unperturbed dynamics and $\sigma(q_t)$ as the distance between the two ensembles, their roles can be exchanged since $\sigma(h_t) = \sigma(\ln q_t)/t$ is also a valid measure of the distance between the forward and backward distributions.

Second, our perturbed and unperturbed distributions, namely backward and forward, are particular in that they are computed from the same population tree, and cannot be evaluated independently in different experiments. This is different from usual systems where the statistics in the presence and absence of a perturbation are measured in two separate experiments. Thus, the philosophy behind usual linear-response relations, that is the use of the variability of a reference system to predict its response to a perturbation without having to actually perturb it, does not apply here.

Third, we can actually take advantage of the equivalence between the single lineage distribution and the forward distribution to infer the response of the colony from mother-machine data only, where no selection is present. We build a lineage-based estimator of the fitness landscape, using its definition eq. (2.49) relying only on the forward distribution, similarly to what we did for the population growth rate in section 2.3:

$$h_{\text{lin}}(s) = \frac{1}{t}\ln\left[\frac{\sum_{i=1}^{L} 2^{K_i}\delta(s_i - s)}{\sum_{i=1}^{L}\delta(s_i - s)}\right]\,. \tag{2.69}$$

In the same way, $q_t(s)$ can be estimated from single-lineage data, meaning that the distance between the two arbitrarily far ensembles is available from single lineage experiments, unlike the distance in most far-from-equilibrium linear-response relations which is not in general available from the unperturbed dynamics. Note that these estimators could suffer from the same convergence difficulties in the limit $t \to \infty$ as Λ_{lin}. However, if we are interested in finite time fitness landscapes and selection strengths, this limit does not appear. Of course, in this case the bound is not really useful, since these estimators can be used to estimate directly the strength of selection rather than the bound. Indeed, the fitness landscape is estimated with eq. (2.69), the population growth rate with eq. (2.26) and by combining these two quantities we obtain the backward distribution and finally the strength of selection.

From these observations, we conclude that the appeal of eq. (2.68) is conceptual rather than practical. It provides a universal link between the variability in fitness landscape and the strength of selection in the form of a fluctuation-response relation.

3.3.3 Linear response equalities

The linear-response inequality eq. (2.65) becomes a linear-response equality when Cauchy-Schwarz inequality is saturated, that is when g_t and q_t (or q_t^{-1}) are linearly dependent. This is true in the small variability limit $\sigma(th_t) \to 0$ (which is equivalent to $\sigma(q_t) \to 0$) where the two probability distributions approach each other, and when function g_t is the fitness landscape h_t.

We show in appendix D.2 that the left hand side of eq. (2.65) reads

$$\langle g_t \rangle_{\text{back}} - \langle g_t \rangle_{\text{for}} \underset{\sigma(th_t)\to 0}{\sim} t \ \text{Cov}(h_t, g_t)\,, \tag{2.70}$$

where the covariance term between the general variable g_t and the fitness landscape h_t is reminiscent of the covariance term in Price's equation eq. (2.38). Note however, that in our result there is no 'environment term' like in Price's equation, because the difference $\langle g_t \rangle_{\text{back}} - \langle g_t \rangle_{\text{for}}$ is defined to capture the effect of selection only. For general functions g_t, eq. (2.65) is not saturated since g_t and h_t are not linearly dependent in general. However, when $g_t = h_t$ then

$$\Pi_{\mathcal{S}} \underset{\sigma(th_t)\to 0}{\sim} t \ \text{Var}(h_t)\,, \tag{2.71}$$

which saturates the bound. In both eq. (2.70) and eq. (2.71), the variance and the covariance can be equivalently taken over the forward or backward sampling.

The limit of small variability can also be written $t \ll \sigma(h_t)^{-1}$ which defines a characteristic timescale of the system. In practice, this limit can be reached either for short times or in the case of a strong control mechanism on the divisions, leading the lineages to stay synchronized even after a finite time. It is also possible to regard this limit as a regime of weak selection (Neher et al., 2011), since the strength of selection is small precisely because of eq. (2.68).

3.3.4 Enhanced lower bound for the strength of selection

To complement the upper bound on the strength of selection given by eq. (2.68), we now derive a non-trivial lower bound, which presents an interest to quantify the minimal effect of selection on a particular trait.

We know that the strength of selection is positive because Jeffrey's divergence is positive, which comes from Jensen's inequality. We can therefore improve this trivial lower bound on the strength of selection by using a sharpened version of Jensen's inequality recently derived in Liao et al., 2019. In practice, the strength of selection is decomposed as a sum of two Kullback-Leibler (KL) divergences: $\mathcal{J}(p_{\rm back}|p_{\rm for}) = \mathcal{D}_{\rm KL}(p_{\rm back}||p_{\rm for}) + \mathcal{D}_{\rm KL}(p_{\rm for}||p_{\rm back})$, and the new version of Jensen's inequality is used to enhance the trivial lower bound of each KL divergence.

We define the convex functions $\varphi_{\rm for}(x) = e^{tx}$, $\varphi_{\rm back}(x) = e^{-tx}$ and the function

$$\Psi(\varphi, x, y) = \frac{\varphi(x) - \varphi(y)}{(x-y)^2} - \frac{\varphi'(y)}{x-y}\,, \tag{2.72}$$

where φ' stands for the derivative of φ. The sharpened version of Jensen's inequality reads

$$\langle e^{th_t}\rangle_{\rm for} - e^{t\langle h_t\rangle_{\rm for}} \geq \sigma^2_{\rm for}(h_t) \inf_h \Psi(\varphi_{\rm for}, h, \langle h_t\rangle_{\rm for})\,. \tag{2.73}$$

We then divide this expression by $\exp(t\Lambda_t)$:

$$\langle e^{t(h_t-\Lambda_t)}\rangle_{\rm for} - e^{t(\langle h_t\rangle_{\rm for}-\Lambda_t)} \geq \frac{\sigma^2_{\rm for}(h_t)}{\exp(t\Lambda_t)} \inf_h \Psi(\varphi_{\rm for}, h, \langle h_t\rangle_{\rm for})\,, \tag{2.74}$$

where the first term is 1 because of the normalization of the probability distribution $p_{\rm back}$ in eq. (2.45), so that

$$\Lambda_t - \langle h_t\rangle_{\rm for} \geq -\frac{1}{t}\ln\left(1 - \frac{\sigma^2_{\rm for}(h_t)}{\exp(t\Lambda_t)} \inf_h \Psi(\varphi_{\rm for}, h, \langle h_t\rangle_{\rm for})\right)\,. \tag{2.75}$$

Similarly, we find

$$\langle h_t\rangle_{\rm back} - \Lambda_t \geq -\frac{1}{t}\ln\left(1 - \frac{\sigma^2_{\rm back}(h_t)}{\exp(-t\Lambda_t)} \inf_h \Psi(\varphi_{\rm back}, h, \langle h_t\rangle_{\rm back})\right)\,. \tag{2.76}$$

Liao et al. proved that when $\varphi'(x)$ is a convex (resp. concave) function, then $\Psi(\varphi, x, y)$ is an increasing (resp. decreasing) function of x, and thus the infimum of function $\Psi(\varphi, x, y)$ on x is reached for $x = x_{\rm min}$ (resp. $x = x_{\rm max}$). Because of the convexity of $\varphi'_{\rm for}(x) = t\exp[tx]$, the minimum of Ψ is reached when evaluating Ψ at the minimum value $h_{\rm min}$ of the fitness landscape $h_t(s)$. Similarly, because of the concavity of $\varphi'_{\rm back}(x) = -t\exp[-tx]$, the minimum of Ψ is reached when evaluating Ψ at the maximum value $h_{\rm max}$. Finally, we use the relation $-\ln(1-x) \geq x$ valid for any real number $x \leq 1$ and we combine the two inequalities to obtain the improved lower bound on the strength of selection:

$$\Pi_{\mathcal{S}} \;\geq\; \frac{1}{t}\left[\frac{\sigma^2_{\rm for}(h_t)}{\exp(t\Lambda_t)}\,\Psi(\varphi_{\rm for}, h_{\rm min}, \langle h_t\rangle_{\rm for}) + \frac{\sigma^2_{\rm back}(h_t)}{\exp(-t\Lambda_t)}\,\Psi(\varphi_{\rm back}, h_{\rm max}, \langle h_t\rangle_{\rm back})\right]. \tag{2.77}$$

Note that the right hand side of eq. (2.77) is positive due to the convexity of φ_{for} and φ_{back}, and therefore does represent an improvement with respect to the trivial bound which is zero.

This lower bound depends on the forward and backward variances of the fitness landscape, similarly to the upper bound eq. (2.68), and is therefore a kind of linear-response inequality as well. In addition, it also depends on the average and extreme values of the distributions of fitness. When the fitness landscape is a monotonic function of the value of the trait, which is the case for cell age and size for example, these extreme values are given by the extreme values of the trait itself.

Liao et al. also proposed another lower bound, looser but simpler than the one involving the function Ψ. Indeed, one can replace $\inf_h \Psi(\varphi_{\text{for}}, h, \langle h_t \rangle_{\text{for}})$ by $\inf_h \varphi''_{\text{for}}(h)/2$ in eq. (2.73). Moreover $\inf_h \varphi''_{\text{for}}(h)/2 = \varphi''_{\text{for}}(h_{\min})/2 = t^2 \exp(t h_{\min})$ since $\varphi''_{\text{for}}(h)$ is an increasing function of h. The same goes for the other inequality, and combining the two leads to

$$\Pi_{\mathcal{S}} \;\geq\; \frac{t}{2}\Big[\sigma^2_{\text{for}}(h_t) e^{t(h_{\min}-\Lambda_t)} + \sigma^2_{\text{back}}(h_t) e^{t(\Lambda_t - h_{\max})}\Big]. \tag{2.78}$$

3.3.5 Illustrations of the linear response relations

We now illustrate our results and test the tightness of the different bounds for growing cell populations, using both simulations and time-lapse video-microscopy experimental data from Kiviet et al., 2014.

First, we illustrate eq. (2.60) for the number of divisions $\mathcal{K}$, and for the linear function $g_t(K) = K$, so that the inequality bounds $\langle K \rangle_{\text{back}} - \langle K \rangle_{\text{for}}$. We simulate lineage trees starting from one cell, and following a sizer mechanism. Each simulation of such a tree yields a single point on fig. 2.5, which shows the ratio of $\sigma_{\text{for}}(K)\sigma_{\text{for}}(q_t)$ to $\langle K \rangle_{\text{back}} - \langle K \rangle_{\text{for}}$ versus the population growth rate Λ_t. Two sets of points are presented, which only differ in the final time of the simulation. As expected from eq. (2.60), all points in both sets are above 1. When the duration of the simulation is small ($t = 3$), the final population is small, around $N \sim 20$, therefore for a given tree the lineages do not have time to differentiate significantly and the variability in the number of divisions among the lineages is small. In that case, simulations points are approaching the horizontal dashed line at $y = 1$ corresponding to the saturation of the inequality. The final population N fluctuates significantly from one simulation to the next, because the simulation time is short and all simulations start with a single cell with random initial size. As a result, the dispersion of values of Λ_t is large. Now, when doubling the duration of the simulation, the cloud of scattered points is considerably reduced in both directions. The horizontal dispersion reduces because as t increases, the state of the system at the final time becomes less and less affected by the initial condition. On the vertical axis, there is a gap between the lower part of the scatter plot and the horizontal line at $y = 1$ due to the increase of heterogeneity in the number of divisions in the lineages with the simulation time.

Second, we test the upper and lower bounds on the strength of selection acting on cell size and age using data from Kiviet et al., 2014. We show on fig. 2.6 the upper bounds $\mathcal{U}$ given by eq. (2.68) and the lower bounds $\mathcal{L}$ given by eq. (2.77), normalized by

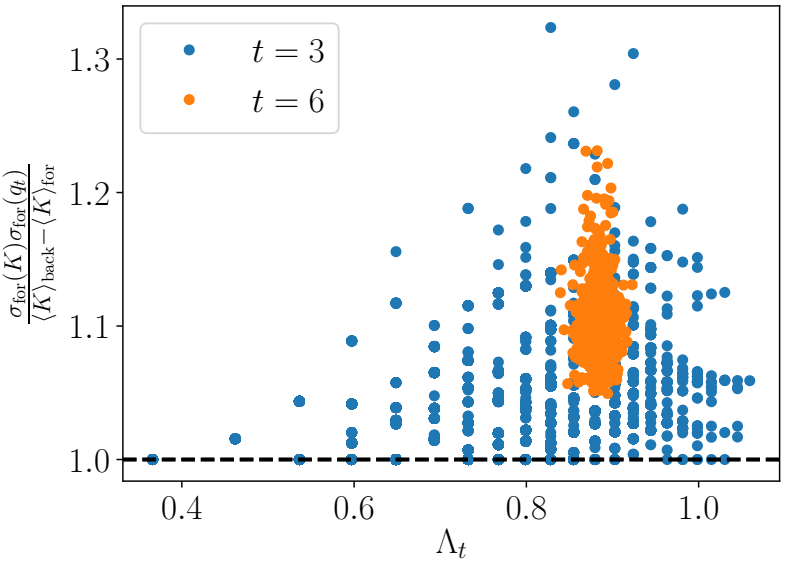


Figure 2.5: Points of $\sigma_{\text{for}}(K)\sigma_{\text{for}}(q_t)/(\langle K\rangle_{\text{back}} - \langle K\rangle_{\text{for}})$ against Λ_t for many tree simulations using a size-controlled model. Each dot corresponds to a single tree, the two sets of data have the same parameters except for the final times of the simulation, which are $t = 3$ (blue) and $t = 6$ (orange). The black horizontal dashed line at $y = 1$ represents the point where the inequality of eq. (2.68) is saturated.

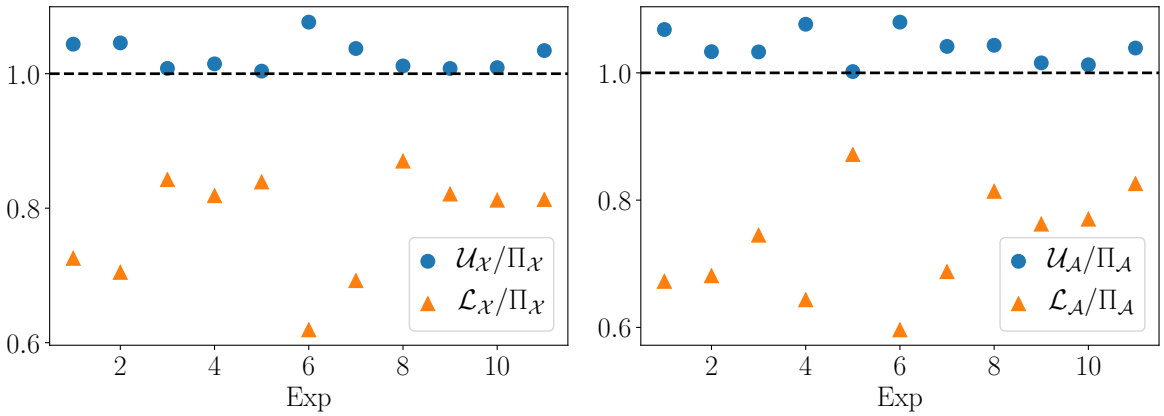


Figure 2.6: Upper bounds $\mathcal{U}$ (blue dots) and lower bounds $\mathcal{L}$ (orange triangles) for the strength of selection acting on size $\Pi_\mathcal{X}$ (left plot) and age $\Pi_\mathcal{A}$, normalized by the latter. The x-axis represents the 11 colonies in different growth conditions from Kiviet et al., 2014, in no particular order.

the strength of selection Π, for size $\mathcal{X}$ on the left and age $\mathcal{A}$ on the right. The x-axis labels in no particular order the 11 colonies which have grown in different conditions. As expected, points representing the upper bound and those representing the lower bound are respectively above and below the horizontal dashed line at $y = 1$. Experiments for which the normalized upper bound approaches 1 indicate that there is small variability in terms of number of divisions among the lineages. We see that the upper bound is typically tight and gives a good approximation for the strength of selection. The lower bound is less tight, but is nevertheless around 70% of the value of the strength of selection, which is a significant improvement compared to the trivial lower bound which is zero.

4 Conclusion

We have studied the relation between two different samplings of lineages in a general branching tree, proposed in Nozoe et al., 2017: the backward sampling presents a statistical bias with respect to the forward sampling, an observation which is important to relate experiments carried out at the population level with the ones carried out at the single lineage level. This bias has been exploited in two directions: to extend relations between single cell stochasticity and population growth beyond age models, and to characterize natural selection and the role of fluctuations for the latter.

This statistical bias can be rationalized by a set of fluctuation relations, which relate the probability distributions in the two ensembles and which are similar to fluctuation relations known in stochastic thermodynamics. This analogy leads to an efficient method to infer the population growth rate from an analysis of single lineages, as we demonstrated by the analysis of the mother machine data from Tanouchi et al., 2017. A second important consequence of the fluctuation relations in the context of population dynamics is the set of inequalities for the mean numbers of divisions and the population growth rate. In the phase of exponential growth in the long time limit, these inequalities generalize for any model of cell size control the inequalities between mean generation times and population doubling time known for age models.

The second axis is the study of the strength of selection, and its interpretation as a response to the lineages variability. The general idea of comparing the response of a system in the presence of a perturbation to its fluctuations in the absence of the perturbation lies at the heart of the fluctuation-dissipation theorem, which has a long history in physics (Kubo, 1966), with some applications to evolution (Kaneko et al., 2018). Remarkably, the present framework with forward (unperturbed) and backward (perturbed) dynamics can be conveniently applied to population dynamics without having to perform additional experiments, since both probabilities can be calculated with the same lineage tree. We derived a set of inequalities for the average of an arbitrary function of a trait and for its fitness landscape, valid beyond the Gaussian assumption, and which constrain the strength of selection, even in the presence of time-dependent selection pressures. These inequalities may also be interpreted as trade-offs between the strength of selection and the similarity between lineages (inverse of the variability).

We point out again that these results are universal in the sense that they only rely on the branching structure of the population tree and are independent of the dynamics of the

tree, that is the ensemble of rules governing the division of the branches. In the context of cell populations, this means for instance that our results are valid for any mechanism of cell size control, in presence of any source of noise and possible mutations, and regardless of the nature of the cell. Although we illustrated our results with cell populations, we hope they could be insightful in other contexts as well. In ecology for example, such trees are used to represent phylogeny (Schuh et al., 2009), where lineages represent species or versions of genes, and divisions represent speciations or mutations. This could also open new perspectives to address a number of important problems like the differentiation of stem cells (Tak et al., 2021) or virus evolution.

As mentioned in the introductory chapter, the use of this framework is limited to situations where the genealogy of the tree is accessible, which may not be the case for various settings, as for in vivo experiments or in the context of ecology. A next logical step would then be trying to elucidate the links between the state of a population at observation time and its past history. A second difficulty lies in one assumption of the framework, namely that all lineages survive up to final time, which prevents us from studying datasets where cells die or are diluted. This obstacle is precisely addressed in chapter 3.

Finally, we hope that our work contributes to clarifying the connection between single lineage and population statistics and to understanding the fundamental constraints which cell growth and division must obey. To further test predictions which involve a comparison between these two levels, it would be useful to perform experimental studies in bulk and in single-lineage setups with the same strains and conditions.

5 Appendices

A Fluctuation theorem at the level of operators

In this appendix, we present an operator-based framework which provides an alternate route to eq. (2.8) and eq. (2.25). Unlike the main derivation of these results, the present approach relies on the population balance equation. For simplicity, we illustrate this method for the sizer, given that other size control mechanisms like the timer and the adder can be treated along the same lines and lead to similar results.

Let us first recall the population balance equation for the backward probability with explicit dependence on the number K of divisions:

$$\begin{aligned}\partial_t p_{\text{back}}(x,K,t) = &-\partial_x[\nu(x)p_{\text{back}}(x,K,t)] - (r(x)+\Lambda_p(t))\,p_{\text{back}}(x,K,t)\\ &+ m\int \mathrm{d}x'\ \Sigma(x|x')r(x')p_{\text{back}}(x',K-1,t)\,.\end{aligned} \tag{2.79}$$

We define the backward generating function $G_{\text{back}}(x,\lambda,t)$ as

$$G_{\text{back}}(x,\lambda,t) = \sum_{K=0}^{\infty} e^{-\lambda K} p_{\text{back}}(x,K,t)\,. \tag{2.80}$$

We then multiply eq. (2.79) by $e^{-\lambda K}$ and sum over K to obtain

$$\partial_t G_{\text{back}}(x,\lambda,t) = \mathcal{L}_{\text{back}}(\lambda)\ G_{\text{back}}(x,\lambda,t)\,, \tag{2.81}$$

where linear operator $\mathcal{L}_{\text{back}}(\lambda)$, acting on $G_{\text{back}}(x,\lambda,t)$, is defined on a test function f as

$$\mathcal{L}_{\text{back}}(\lambda)f = -\partial_x[\nu(x)f] - (r(x)+\Lambda_p(t))f + me^{-\lambda}\int \mathrm{d}x' r(x')\Sigma(x|x')f\,. \tag{2.82}$$

Although this operator depends on the size x, we choose not to write this dependence explicitly to ease the reading. By the same method, we obtain the operator at the forward level:

$$\mathcal{L}_{\text{for}}(\lambda)f = -\partial_x[\nu(x)f] - r(x)f + e^{-\lambda}\int \mathrm{d}x' r(x')\Sigma(x|x')f\,. \tag{2.83}$$

By direct comparison, we obtain the fluctuation relation at the level of operators

$$\mathcal{L}_{\text{back}}(\lambda) = \mathcal{L}_{\text{for}}(\lambda - \ln m) - \Lambda_p(t)\mathbb{1}\,. \tag{2.84}$$

This equality between two operators implies relations between their eigenvalues and eigenvectors as well. Let us call $g_{\text{back}}(x,\lambda)$ (resp. $g_{\text{for}}(x,\lambda)$) an eigenvalue of the operator $\mathcal{L}_{\text{back}}(\lambda)$ (resp. $\mathcal{L}_{\text{for}}(\lambda)$), and $V^g_{\text{back}}(x,\lambda)$ (resp. $V^g_{\text{for}}(x,\lambda)$) the associated eigenvector. Then eq. (2.84) gives

$$g_{\text{back}}(x,\lambda) = g_{\text{for}}(x,\lambda - \ln m) - \Lambda_p(t)\,, \tag{2.85}$$

$$V^g_{\text{back}}(x,\lambda) = V^g_{\text{for}}(x,\lambda - \ln m)\,. \tag{2.86}$$

When solving eq. (2.81), the long-time behavior of $G_{\text{back}}(x,\lambda,t)$ is controlled by the largest eigenvalue of $\mathcal{L}_{\text{back}}(\lambda)$, which we call $\mu_{\text{back}}(x,\lambda)$, and reads

$$G_{\text{back}}(x,\lambda,t) \underset{t\to\infty}{\sim} C^{\mu}_{\text{back}}(x,\lambda)V^{\mu}_{\text{back}}(x,\lambda)\,e^{\mu_{\text{back}}(x,\lambda)t}\,, \tag{2.87}$$

where $C^{\mu}_{\text{back}}(x,\lambda)$ is the constant coefficient of the eigenvector $V^{\mu}_{\text{back}}(x,\lambda)$ associated with the largest eigenvalue in the decomposition of the initial condition $V_{\text{back}}(x,\lambda,t=0)$ on the set of the eigenvectors of $\mathcal{L}_{\text{back}}(\lambda)$. We investigate the particular case $\lambda = 0$. On the one hand, using the definition eq. (2.80) we obtain $G_{\text{back}}(x,0,t) = p_{\text{back}}(x,t)$ and thus the normalization of the probability gives

$$\int \mathrm{d}x\; G_{\text{back}}(x,0,t) = 1\,. \tag{2.88}$$

On the other hand, integrating the long time behavior eq. (2.87) over x, we get

$$\int \mathrm{d}x\; G_{\text{back}}(x,0,t) \underset{t\to\infty}{\sim} \int \mathrm{d}x\; C^{\mu}_{\text{back}}(x,0) V^{\mu}_{\text{back}}(x,0)\, e^{\mu_{\text{back}}(x,0)t}\,. \tag{2.89}$$

The only solution to satisfy both conditions eqs. (2.88) and (2.89) is

$$\forall x,\; \mu_{\text{back}}(x,0) = 0 \tag{2.90}$$

$$\int \mathrm{d}x\; C^{\mu}_{\text{back}}(x,0) V^{\mu}_{\text{back}}(x,0) = 1\,. \tag{2.91}$$

Since the largest backward eigenvalue is independent of the size x for $\lambda = 0$, we define the size-independent backward eigenvalue $\hat{\mu}_{\text{back}}(\lambda = 0) = 0$. Reporting $\hat{\mu}_{\text{back}}(\lambda = 0)$ in eq. (2.85), the right hand side of the equation has to be independent of the size x as well, so we define the size-dependent forward eigenvalue: $\hat{\mu}_{\text{for}}(\lambda = -\ln m) = \mu_{\text{for}}(x, \lambda = -\ln m)$ for any x. Finally, eq. (2.85) gives

$$\hat{\mu}_{\text{for}}(\lambda = -\ln m) = \Lambda\,, \tag{2.92}$$

where Λ is the steady-state population growth rate.

Let us now propose a second derivation of the link between the population growth rate and the forward statistics for the number of divisions eq. (2.25). On the one hand, using the definition of the generation function we get

$$\int \mathrm{d}x\; G_{\text{for}}(x, -\ln m, t) = \sum_K m^K p_{\text{for}}(K,t)\,. \tag{2.93}$$

On the other hand,

$$V_{\text{for}}(x, -\ln m, t) \underset{t\to\infty}{\sim} C^{\mu}_{\text{for}}(x, -\ln m) V^{\mu}_{\text{for}}(x, -\ln m)\, e^{\hat{\mu}_{\text{for}}(-\ln m)t}\,. \tag{2.94}$$

Using eqs. (2.86) and (2.91) and the fact that the initial condition is the same for both backward and forward samplings: $G_{\text{for}}(x,\lambda,t=0) = G_{\text{back}}(x,\lambda,t=0)$ for any x and λ, we prove that

$$\int \mathrm{d}x\; C^{\mu}_{\text{for}}(x, -\ln m) V^{\mu}_{\text{for}}(x, -\ln m) = 1\,. \tag{2.95}$$

Finally, integrating eq. (2.94) over x and combining eqs. (2.92), (2.93) and (2.95), we recover

$$\Lambda = \lim_{t\to\infty} \frac{1}{t} \ln \sum_K m^K p_{\text{for}}(K,t)\,. \tag{2.96}$$

B Path integral solution to the uncorrelated age model

In this appendix, we provide a path-integral solution to the age model without correlation (eq. (1.43)) using the method of characteristics. We re-parameterize $n(a,t,K)$ as $\hat{n}(z,K) = n(a(z),t(z),K)$ with $\mathrm{d}a/\mathrm{d}z = 1$ and $\mathrm{d}t/\mathrm{d}z = 1$, so that the equation on $\hat{n}$ reads:

$$\frac{\mathrm{d}\hat{n}(z,K)}{\mathrm{d}z} = -r(a(z))\hat{n}(z,K)\,. \tag{2.97}$$

The solution to this equation is given by

$$\hat{n}(z,K) = \hat{n}(0,K)e^{-\int_0^z \mathrm{d}z'\ r(a(z'))}\,. \tag{2.98}$$

We now choose the parameterization $a(z) = z$ and $t(z) = z + t(0)$ so that $n(a(0),t(0),K) = n(0,t-a,K) = m\int_0^\infty \mathrm{d}\tau_K\ r(\tau_K)n(\tau_K,t-a,K-1)$, and

$$n(a,t,K) = me^{-\int_0^a \mathrm{d}a'\ r(a')}\int_0^\infty \mathrm{d}\tau_K\ r(\tau_K)n(\tau_K,t-a,K-1)\,. \tag{2.99}$$

This relation can be iterated by expressing $n(\tau_K,t-a,K-1)$ as a function of $n(\tau_{K-1},t-a-\tau_K,K-2)$ and so on until reaching cells with no divisions, given by the boundary term:

$$n(\tau_1,t,K=0) = \delta(\tau_1 - t)e^{-\int_0^{\tau_1}\mathrm{d}a'\ r(a')}N_0\,. \tag{2.100}$$

The final solution reads

$$n(a,t,K) = m^K N_0 e^{-\int_0^a \mathrm{d}a'\ r(a')}\prod_{k=1}^K\int_0^\infty \mathrm{d}\tau_k\ r(\tau_k)e^{-\int_0^{\tau_k}\mathrm{d}a'\ r(a')}\,\delta\left(t-a-\sum_{k=1}^K\tau_k\right)\,, \tag{2.101}$$

and thus the number of lineages following a certain path is given by

$$n(a,\{\tau_k\},t,K) = m^K N_0 e^{-\int_0^a \mathrm{d}a'\ r(a')}\prod_{k=1}^K r(\tau_k)e^{-\int_0^{\tau_k}\mathrm{d}a'\ r(a')}\,\delta\left(t-a-\sum_{k=1}^K\tau_k\right)\,. \tag{2.102}$$

The forward path probability is then obtained by dividing the above formula by $m^K N_0$:

$$p_{\mathrm{for}}(a,\{\tau_k\},t,K) = e^{-\int_0^a \mathrm{d}a'\ r(a')}\prod_{k=1}^K r(\tau_k)e^{-\int_0^{\tau_k}\mathrm{d}a'\ r(a')}\,\delta\left(t-a-\sum_{k=1}^K\tau_k\right)\,, \tag{2.103}$$

where $f_{\mathrm{for}}(\tau) = r(\tau)\exp\left[-\int_0^\tau \mathrm{d}a'\ r(a')\right]$ is the forward distribution of generation times.

C Comments on historical fitness

C.1 Link between historical fitness and fitness landscape for models of independent mutations and divisions

In Nozoe et al., 2017 (SM), the authors proved that when considering independent divisions and mutations, described by eq. (2.39), the fitness landscape and the historical fitness of a trajectory $\boldsymbol{s}$ are equal:

$$h_t(\boldsymbol{s}) = H_t(\boldsymbol{s})\,. \tag{2.104}$$

In this appendix, we provide an alternative proof of eq. (2.104) in the case where trait $\mathcal{S}$ takes only discrete values $s \in \{s_i\}_{i=1}^n$. In this case, the probability inside the sum of eq. (2.49) is given by the heterogeneous Poisson distribution:

$$p_{\text{for}}(K, t|\boldsymbol{s}) = \frac{(\sum_{i=1}^n r(s_i)t_i)^K}{K!} e^{-\sum_{i=1}^n r(s_i)t_i}\,, \tag{2.105}$$

where t_i is the time spent in state s_i for the trajectory $\boldsymbol{s}$, which is called occupation time, such that $\sum_{i=1}^n t_i = t$.

This result comes from the fact that the division rate only depends on the current sate of the cell, and that divisions do not affect the trajectory, so that the number of divisions on different portions of the trajectory are independent:

$$p_{\text{for}}(K_1, ..., K_n, t|\boldsymbol{s}) = \prod_{i=1}^n p_{\text{for}}(K_i, t|\boldsymbol{s})\,, \tag{2.106}$$

where K_i is the number of divisions that happened during the duration t_i when the cell was in state s_i (even if this duration is discontinuous). In each state, the division rate is constant, so that each term in the product is a Poisson distribution:

$$p_{\text{for}}(K_i, t|\boldsymbol{s}) = \frac{(r(s_i)t_i)^{K_i}}{K_i!} e^{-r(s_i)t_i}\,. \tag{2.107}$$

Combining these results, we obtain

$$\begin{aligned} p_{\text{for}}(K, t|\boldsymbol{s}) &= \sum_{K_1+...+K_n=K} \prod_{i=1}^n \frac{(r(s_i)t_i)^{K_i}}{K_i!} e^{-r(s_i)t_i} \\ &= e^{-\sum_{i=1}^n r(s_i)t_i} \sum_{K_1+...+K_n=K} \prod_{i=1}^n \frac{(r(s_i)t_i)^{K_i}}{K_i!}\,, \end{aligned} \tag{2.108}$$

where we recognize the multinomial development:

$$\sum_{K_1+...+K_n=K} \prod_{i=1}^n \frac{(r(s_i)t_i)^{K_i}}{K_i!} = \frac{(\sum_{i=1}^n r(s_i)t_i)^K}{K!}\,, \tag{2.109}$$

which proves eq. (2.105).

Finally, plugging eq. (2.105) inside the fitness landscape eq. (2.49) leads to:

$$h_t(\boldsymbol{s}) = \sum_{i=1}^n r(s_i)\frac{t_i}{t} \tag{2.110}$$

$$= \frac{1}{t}\int_0^t \mathrm{d}t'\; r(s(t')) \tag{2.111}$$

$$= H_t(\boldsymbol{s}) \tag{2.112}$$

C.2 Link between historical fitness and fitness landscape for models of cell size control

Not all traits of interest follow an equation like eq. (2.39). For example, in classical models of cell size control, age and size evolve continuously during cell cycles, and their values are reset at division. The important difference between these two families of models is that in eq. (2.39) mutations and divisions are uncoupled, while for models of cell size control the sudden change in the values of age and size occur only at division and occur for all divisions. Because of this, we show in this appendix that the fitness landscape and historical fitness are in general different, but still linked by an exponential average relation. For simplicity, we illustrate our point with the sizer model, but the argument can be made more general.

Using a path integral approach, similar to that of appendix B, one can obtain a solution to the sizer population balance equation eq. (1.49) for a trajectory $\boldsymbol{x}$ of cell size x, at the level of the forward and backward probabilities (García-García et al., 2019):

$$p_{\text{back}}(\boldsymbol{x}) = 2^K \exp\left[-t\Lambda_t - \int_0^t \mathrm{d}t' r(x(t'))\right] \prod_{k=1}^{K} r(x(t_k^-))\Sigma(x(t_k^+)|x(t_k^-))p_{\text{back}}(x(0)) \quad (2.113)$$

$$p_{\text{for}}(\boldsymbol{x}) = \exp\left[-\int_0^t \mathrm{d}t' r(x(t'))\right] \prod_{k=1}^{K} r(x(t_k^-))\Sigma(x(t_k^+)|x(t_k^-))p_{\text{for}}(x(0))\,, \quad (2.114)$$

where t_k^- and t_k^+ indicate the times just before and just after the k-th division.

If we consider a forward dynamics where the division rate is doubled as compared to the backward dynamics, we obtain a particular lineage-population bias involving the historical fitness:

$$p_{\text{for}}^{2r}(\boldsymbol{x}) = p_{\text{back}}^{r}(\boldsymbol{x}) \exp\left[t(\Lambda_t - H_t(\boldsymbol{x}))\right]\,. \quad (2.115)$$

By comparing this relation with the definition of the fitness landscape eq. (2.45), we conclude that the fitness landscape $h_t(\boldsymbol{x})$ and the historical fitness $H_t(\boldsymbol{x})$ are in general different since the forward probability with modified division rate $2r$ is in general different from that with original division rate r. However, integrating eqs. (2.45) and (2.115) over $\boldsymbol{x}$ and using the normalization of the forward distributions, we obtain

$$\langle e^{-H_t}\rangle_{\text{back}} = \langle e^{-h_t}\rangle_{\text{back}}\,. \quad (2.116)$$

C.3 Variance of historical fitness as a measure of selection for models of cell size control?

In section 3.1, we mentioned the measure of selection proposed in Leibler et al., 2010 for models of independent mutations and divisions described by eq. (2.39), namely

$$\partial_\beta \langle H_t\rangle = \int \mathcal{D}\boldsymbol{x}\; H_t(\boldsymbol{x})\partial_\beta p_{\text{back}}^{\beta r}(\boldsymbol{x}) = t\text{Var}(H_t) \geq 0\,, \quad (2.117)$$

where β is a multiplicative factor for all the division rates. The variance in the right hand side is computed with the backward distribution with modified rates βr; but for simplicity we omit the subscript for all variances and averages in the following, which are

all computed with respect to $p^{\beta r}_{\text{back}}(\boldsymbol{x})$. The derivation of this result relies on the fact that the number of cells following trajectory $\boldsymbol{x}$ can be written as

$$n(\boldsymbol{x}, \beta) = p_{\text{for}}(\boldsymbol{x}) e^{t\beta H_t(\boldsymbol{x})}\,, \tag{2.118}$$

where $p_{\text{for}}(\boldsymbol{x})$ only depends on the mutation rates, and is independent of the division rates and thus of β.

Since the forward distribution in eq. (2.115) explicitly depend on division rate r, we expect a quantitative difference when computing $\partial_\beta \langle H_t \rangle$ for models of cell size control. Indeed, $p^{\beta r}_{\text{back}}(\boldsymbol{x})$ is given by eq. (2.115) with r replaced by βr, so that:

$$\partial_\beta p^{\beta r}_{\text{back}}(\boldsymbol{x}) = N(\beta)^{-1} \exp\left[t\beta H_t(\boldsymbol{x})\right] \partial_\beta p^{2\beta r}_{\text{for}}(\boldsymbol{x}) + t H_t(\boldsymbol{x}) p^{\beta r}_{\text{back}}(\boldsymbol{x}) - p^{\beta r}_{\text{back}}(\boldsymbol{x}) N(\beta)^{-1} \partial_\beta N(\beta)\,. \tag{2.119}$$

Using eq. (2.114), with division rate $2\beta r$, we compute:

$$\partial_\beta p^{2\beta r}_{\text{for}}(\boldsymbol{x}) = K(\boldsymbol{x}) \beta^{-1} p^{2\beta r}_{\text{for}}(\boldsymbol{x}) - 2t H_t(\boldsymbol{x}) p^{2\beta r}_{\text{for}}(\boldsymbol{x})\,. \tag{2.120}$$

Finally, $N(\beta)$ is given by the integration of eq. (2.115) over all paths $\boldsymbol{x}$, so:

$$\begin{aligned} N(\beta)^{-1} \partial_\beta N(\beta) =& N(\beta)^{-1} \int \mathcal{D}\boldsymbol{x}\ \partial_\beta p^{2\beta r}_{\text{for}}(\boldsymbol{x}) \exp\left[\beta t H_t(\boldsymbol{x})\right] \\ &+ N^{-1}(\beta) \int \mathcal{D}\boldsymbol{x}\ p^{2\beta r}_{\text{for}}(\boldsymbol{x}) t H_t(\boldsymbol{x}) \exp\left[\beta t H_t(\boldsymbol{x})\right] && (2.121) \\ =& \beta^{-1} \langle K \rangle - 2t \langle H_t \rangle + t \langle H_t \rangle && (2.122) \\ =& \beta^{-1} \langle K \rangle - t \langle H_t \rangle\,. && (2.123) \end{aligned}$$

Combining the above results, we find

$$\partial_\beta p^{\beta r}_{\text{back}}(\boldsymbol{x}) = p_{\text{back}}(\boldsymbol{x}) \left[t \left(\langle H_t \rangle - H_t(\boldsymbol{x}) \right) + \beta^{-1} \left(K(\boldsymbol{x}) - \langle K \rangle \right) \right]\,, \tag{2.124}$$

so that finally

$$\partial_\beta \langle H_t \rangle = \beta^{-1} \text{Cov}(K, H_t) - t \text{Var}(H_t)\,. \tag{2.125}$$

The change in mean historical fitness now involves both the variability in historical fitness, with a minus sign, and the correlations between number of divisions and historical fitness. This quantity has no clear sign, and it is easy to find examples where it is negative. Let us consider that the division rate $r(x)$ is an increasing function of x, and that division is symmetrical: $\Sigma(x|x') = \delta(x - x'/2)$. Histories with large fitness are histories where the cell is large on average along the trajectory, thus with few divisions. The covariance between number of divisions and historical fitness is therefore negative, and so is the left hand side of eq. (2.125). This is a consequence of the dependence of the forward probability on reproductive rates: increasing reproductive rates favors histories with large fitness through the term $\exp(t\beta H_t(\boldsymbol{x}))$, but this makes these histories less probable via the term $p^{2\beta r}_{\text{for}}(\boldsymbol{x})$. Finally, $\text{Var}(H_t)$ and $\partial_\beta \langle H_t \rangle$ may no longer be adequate measures of selection, while then strength of selection Π proposed in Nozoe et al., 2017 is defined and positive for any branching tree, including for models of cell size control.

D Linear response equality for the strength of selection

D.1 Gaussian case

We show in this section that a linear relation between the strength of selection and the variance of the fitness landscape holds in the case where the fitness landscape is normally distributed. To do so, let us first derive a first result: when integrating eq. (2.45) over s and using the normalization of either p_{back} or p_{for}, the population growth rate is expressed as

$$e^{t\Lambda_t} = \langle e^{th_t} \rangle_{\text{for}} \tag{2.126}$$

$$e^{-t\Lambda_t} = \langle e^{-th_t} \rangle_{\text{back}} \,. \tag{2.127}$$

We now assume that $h_t(s)$ can be accounted for by a continuous probability distribution even when trait $\mathcal{S}$ is discrete. We set a Gaussian forward distribution with mean $\langle h_t \rangle_{\text{for}}$ and variance $\sigma_{\text{for}}(h_t)^2$ for the fitness landscape $h_t(s)$, then $\exp(th_t(s))$ follows a log-normal distribution of mean

$$\langle e^{th_t} \rangle_{\text{for}} = e^{t\langle h_t \rangle_{\text{for}} + (t\sigma_{\text{for}}(h_t))^2/2} \,. \tag{2.128}$$

This relation shows that for a given forward average fitness landscape, the growth rate is positively affected by the variability between the lineages.

The backward average of the fitness landscape is given by the forward average of a biased fitness landscape:

$$\langle h_t \rangle_{\text{back}} = e^{-t\Lambda_t} \int h_t(s) e^{th_t(s)} p_{\text{for}}(s,t) \mathrm{d}s \tag{2.129}$$

We make the hypothesis that the fitness landscape is a bijective function of the trait value and use the conservation of the probability: $p_{\text{for}}(s,t)\mathrm{d}s = p_{\text{for}}(h)\mathrm{d}h$, leading to

$$\begin{aligned} \langle h_t \rangle_{\text{back}} &= \frac{e^{-t\Lambda_t}}{\sqrt{2\pi\sigma_{\text{for}}(h_t)^2}} \int h e^{th} e^{-(h-\langle h_t \rangle_{\text{for}})^2/(2\sigma_{\text{for}}(h_t)^2)} \mathrm{d}h \\ &= e^{-t\Lambda_t} \left(\langle h_t \rangle_{\text{for}} + t\sigma_{\text{for}}(h_t)^2 \right) e^{t\langle h_t \rangle_{\text{for}} + (t\sigma_{\text{for}}(h_t))^2/2} \,. \end{aligned} \tag{2.130}$$

Finally, combining eqs. (2.126), (2.128) and (2.130), we obtain

$$\langle h_t \rangle_{\text{back}} = \langle h_t \rangle_{\text{for}} + t\sigma_{\text{for}}(h_t)^2 \,, \tag{2.131}$$

and thus

$$\Pi_{\mathcal{S}} = t\sigma_{\text{for}}(h_t)^2 \,. \tag{2.132}$$

Moreover, combining eqs. (2.128) and (2.132) we deduce that $\langle h_t \rangle_{\text{for}}$ and $\langle h_t \rangle_{\text{back}}$ are symmetrical around Λ_t: $\langle h_t \rangle_{\text{back}} - \Lambda_t = \Lambda_t - \langle h_t \rangle_{\text{for}} = t\ \sigma^2_{\text{for}}(h_t)/2$. In other words, in this particular case, the Kullback-Leibler divergence is symmetrical: $\mathcal{D}_{\text{KL}}(p_{\text{for}}||p_{\text{back}}) = \mathcal{D}_{\text{KL}}(p_{\text{back}}||p_{\text{for}})$.

When $h_t(s)$ follows a Gaussian distribution in the forward statistics, it also follows a Gaussian distribution with mean $\langle h_t \rangle_{\text{back}}$ and standard deviation $\sigma_{\text{back}}(h_t)$ in the backward

statistics, because the bias of the fluctuation relation between p_{back} and p_{for} is exponential in h_t. Then $\exp[-th_t(s)]$ follows a log-normal distribution of mean

$$\langle e^{-th_t}\rangle_{\text{back}} = e^{-t\langle h_t\rangle_{\text{back}}+(t\sigma_{\text{back}}(h_t))^2/2}\,. \tag{2.133}$$

We now use eqs. (2.127) and (2.131) to replace the backward average:

$$e^{t\Lambda_t} = e^{t\left(\langle h_t\rangle_{\text{for}}+t\sigma_{\text{for}}(h_t)^2\right)-(t\sigma_{\text{back}}(h_t))^2/2}\,. \tag{2.134}$$

By comparing eqs. (2.128) and (2.134), it follows that $\sigma_{\text{back}}(h_t) = \sigma_{\text{for}}(h_t)$. Finally, the standard deviation in eq. (2.132) can be taken indifferently with respect to both statistics:

$$\Pi_{\mathcal{S}} = t\text{Var}(h_t)\,. \tag{2.135}$$

D.2 Small variability limit

In this appendix, we study the two sides of the fluctuation-response inequality on an arbitrary function g_t of a phenotypic trait $\mathcal{S}$ (eq. (2.65)) in the limit where the forward and backward distributions approach each other, and show that they are mathematically equivalent in this limit in the case where the function g_t is the fitness landscape h_t. The difference between the two distributions is captured by the distance term $\sigma(q_t)$, or equivalently $\sigma(\ln q_t) = \sigma(th_t)$, where the standard deviations can be taken either in the backward or forward statistics.

We first use this limit in the forward statistics: $\sigma_{\text{for}}(th_t) \to 0$. From eq. (2.45), we get that:

$$\langle g_t\rangle_{\text{back}} = e^{t(\langle h_t\rangle_{\text{for}}-\Lambda_t)}\int g_t(s)e^{t(h_t(s)-\langle h_t\rangle_{\text{for}})}p_{\text{for}}(s,t)\mathrm{d}s\,. \tag{2.136}$$

Since $\sigma_{\text{for}}(th_t)$ is the characteristic distance of th_t to its mean, the small variability limit implies that $t(h_t(s)-\langle h_t\rangle_{\text{for}})$ is small, and therefore a first order expansion of the exponential function gives

$$\langle g_t\rangle_{\text{back}} \underset{\sigma(th_t)\to 0}{\sim} e^{t(\langle h_t\rangle_{\text{for}}-\Lambda_t)}\left[\langle g_t\rangle_{\text{for}} + t\text{Cov}_{\text{for}}(h_t,g_t)\right]\,, \tag{2.137}$$

where the covariance is a first-order correction to $\langle g_t\rangle_{\text{for}}$ in $\sigma_{\text{for}}(th_t)$. Now we compute the prefactor $\exp[-t\Lambda_t]$, starting with eq. (2.126), and a second-order expansion of the exponential since the first order vanishes:

$$\begin{aligned} e^{t\Lambda_t} &= \int e^{th_t(s)}p_{\text{for}}(s,t)\mathrm{d}s \\ &\underset{\sigma(th_t)\to 0}{\sim} e^{t\langle h_t\rangle_{\text{for}}}\int\left[1+t\left(h_t(s)-\langle h_t\rangle_{\text{for}}\right)+\frac{t^2}{2}\left(h_t(s)-\langle h_t\rangle_{\text{for}}\right)^2\right]p_{\text{for}}(s,t)\mathrm{d}s \\ &\sim e^{t\langle h_t\rangle_{\text{for}}}\left[1+\frac{(t\sigma_{\text{for}}(h_t))^2}{2}\right]\,, \end{aligned} \tag{2.138}$$

which is a second-order correction to $\exp[t\langle h_t\rangle_{\text{for}}]$ in $\sigma_{\text{for}}(th_t)$. Combining eqs. (2.137) and (2.138) we find at first order

$$\langle g_t\rangle_{\text{back}} - \langle g_t\rangle_{\text{for}} \underset{\sigma(th_t)\to 0}{\sim} t\ \text{Cov}_{\text{for}}(h_t,g_t)\,. \tag{2.139}$$

Doing the same calculation from the backward point of view: $\sigma_{\text{back}}(th_t) \to 0$, we obtain

$$\begin{aligned}\langle g_t \rangle_{\text{for}} &= e^{t(\Lambda_t - \langle h_t \rangle_{\text{back}})} \int g_t(s) e^{t(\langle h_t \rangle_{\text{back}} - h_t(s))} p_{\text{back}}(s,t) \mathrm{d}s \\ &\underset{\sigma(th_t)\to 0}{\sim} e^{t(\Lambda_t - \langle h_t \rangle_{\text{back}})} \left[\langle g_t \rangle_{\text{back}} - t \mathrm{Cov}_{\text{back}}(h_t, g_t) \right] . \end{aligned} \tag{2.140}$$

The prefactor is computed with the same expansion starting from eq. (2.127):

$$\begin{aligned} e^{-t\Lambda_t} &= \int e^{-th_t(s)} p_{\text{back}}(s,t) \mathrm{d}s \\ &\underset{\sigma(th_t)\to 0}{\sim} e^{-t\langle h_t \rangle_{\text{back}}} \int \left[1 + t \left(\langle h_t \rangle_{\text{back}} - h_t(s) \right) + \frac{t^2}{2} \left(\langle h_t \rangle_{\text{back}} - h_t(s) \right)^2 \right] p_{\text{back}}(s,t) \mathrm{d}s \\ &\sim e^{-t\langle h_t \rangle_{\text{back}}} \left[1 + \frac{(t\sigma_{\text{back}}(h_t))^2}{2} \right] . \end{aligned} \tag{2.141}$$

Combining eqs. (2.140) and (2.141), we find at first order:

$$\langle g_t \rangle_{\text{back}} - \langle g_t \rangle_{\text{for}} \underset{\sigma(th_t)\to 0}{\sim} t \ \mathrm{Cov}_{\text{back}}(h_t, g_t) . \tag{2.142}$$

When comparing eqs. (2.139) and (2.142), we conclude that the covariance can be taken equivalently in the forward or backward statistics.

Let us now turn to the right hand side of inequality eq. (2.65). Using the same kind of Taylor expansion for $q_t = \exp[t(h_t(s) - \Lambda_t)]$, it is straight-forward to show that

$$\sigma_{\text{for}}(g_t)\sigma_{\text{for}}(q_t) \underset{\sigma(th_t)\to 0}{\sim} t\sigma_{\text{for}}(h_t)\sigma_{\text{for}}(g_t) \tag{2.143}$$

$$\sigma_{\text{back}}(g_t)\sigma_{\text{back}}(q_t^{-1}) \underset{\sigma(th_t)\to 0}{\sim} t\sigma_{\text{back}}(h_t)\sigma_{\text{back}}(g_t) . \tag{2.144}$$

Therefore, the inequality eq. (2.65) does not get necessarily saturated in this limit. However, in the particular case where $g_t(s)$ is the fitness landscape $h_t(s)$, then eq. (2.139) reads

$$\Pi_{\mathcal{S}} \underset{\sigma(th_t)\to 0}{\sim} t \ \mathrm{Var}(h_t) , \tag{2.145}$$

and thus the inequality eq. (2.68) is saturated in this limit.

E Upper bounds numerical comparison

In this section we compare numerically the upper bound $\mathcal{U}_{\text{GL}}$ obtained in eq. (2.65):

$$\mathcal{U}_{\text{GL}} = \min \left(\sigma_{\text{a}}(g_t)\sigma_{\text{a}}(q_t), \sigma_{\text{b}}(g_t)\sigma_{\text{b}}(q_t^{-1}) \right) , \tag{2.146}$$

to the upper bound $\mathcal{U}_{\text{DS}}$ derived by Dechant and Sasa (Eq. 5 in Dechant et al., 2020):

$$\mathcal{U}_{\text{DS}} = \inf_{\gamma > 0} \left(K^a_{g_t}(\gamma\sigma) - \gamma\sigma \langle g_t \rangle_{\text{a}} + \mathcal{D}_{\text{KL}}(p_{\text{b}} || p_{\text{a}}) \right) , \tag{2.147}$$

where $\sigma = \text{sign}(\langle g_t \rangle_{\text{b}} - \langle g_t \rangle_{\text{a}})$ and $K^a_{g_t}(\gamma) = \ln \langle \exp(\gamma g_t) \rangle_{\text{a}}$ the cumulant-generating function of g_t. Both quantities $\mathcal{U}_{\text{GL}}$ and $\mathcal{U}_{\text{DS}}$ bound the difference $|\langle g_t \rangle_{\text{b}} - \langle g_t \rangle_{\text{a}}|$ between the

average values of a function g_t of a variable $\mathcal{S}$ with respect to probability distributions p_a and p_b.

To compare them, we took beta distributions for p_a and p_b, having the same support $[0, 1]$ so that both bounds are defined. Beta-distributed variables admit a probability density function (pdf) of the form $f(s, \alpha, \beta) = s^{\alpha-1}(1-s)^{\beta-1}/B(\alpha, \beta)$, where $B(\alpha, \beta)$ is a normalization constant, and their mean is given by $\langle s \rangle = \alpha/(\alpha + \beta)$. We fix the pdf in the ensemble b to $p_b(s) = f(s, 3, 3)$, whose bell shape is reminiscent of a Gaussian distribution, on a finite interval. The pdf in the ensemble a is taken as $p_a(s) = f(s, \alpha, \beta)$, where α and β are varied in $[2, 4]$. We choose the simple function $g_t(s) = s$.

Results are shown on fig. 2.7. The first row of figures shows the differences between the upper bounds and the actual difference $|\langle s \rangle_\mathrm{b} - \langle s \rangle_\mathrm{a}|$, for our bound $\mathcal{U}_\mathrm{GL}$ (fig. 2.7a), and for Dechant-Sasa's bound $\mathcal{U}_\mathrm{DS}$ (fig. 2.7b). As expected, all points on these two plots are positive. We plot on fig. 2.7c the real difference $\langle s \rangle_\mathrm{b} - \langle s \rangle_\mathrm{a}$, which is in complete agreement with the theory: $\langle s \rangle_\mathrm{b} - \langle s \rangle_\mathrm{a} = 1/2 - \alpha/(\alpha + \beta)$. Finally, fig. 2.7d shows a comparison between our bound and Dechant-Sasa's bound: blue regions represent sets of parameters (α, β) where our bound is numerically tighter, and the opposite is true in red regions. We note that, if the blue region is larger than the red region, on the other hand the advantage of one bound over the other $|\mathcal{U}_\mathrm{GL} - \mathcal{U}_\mathrm{DS}|$, is generally larger in the red region. Therefore, the answer to the question 'Which bound is tighter?' depends on the actual distributions $p_a(s)$ and $p_b(s)$. However, we note that our bound is easier to compute since it does not require the optimization over an external parameter, which is the case for $\mathcal{U}_\mathrm{DS}$ in Dechant et al., 2020 (parameter γ in eq. (2.147)). Note that this optimization can be bypassed by choosing a specific value γ in eq. (2.147), but then the corresponding bound is less tight than the version with the infimum.

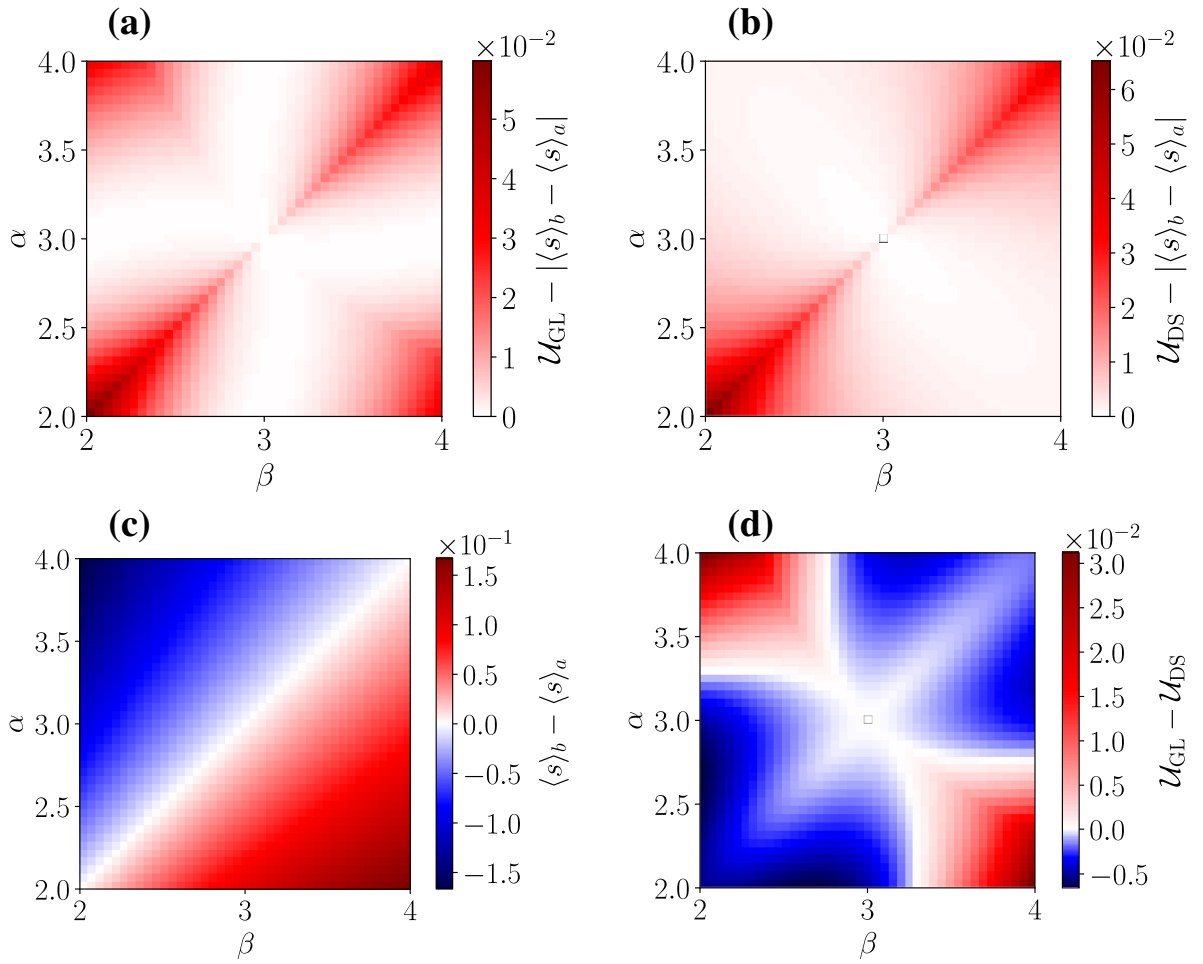


Figure 2.7: Comparison between the upper bounds $\mathcal{U}_{\mathrm{GL}}$ (eq. (2.146)) and $\mathcal{U}_{\mathrm{DS}}$ derived in Dechant et al., 2020 (eq. (2.147)), for beta distributions $p_a(s) = f(s, \alpha, \beta)$ and $p_b(s) = f(s, 3, 3)$. Parameters α and β are varied between 2 and 4. First row: difference between the upper bounds and $|\langle s\rangle_{\mathrm{b}} - \langle s\rangle_{\mathrm{a}}|$, **(a)** for our bound, and **(b)** for Dechant-Sasa's bound, showing that all points are indeed above 0. **(c)**: exact difference $\langle s\rangle_{\mathrm{b}} - \langle s\rangle_{\mathrm{a}}$, in agreement with the theoretical value $\langle s\rangle_{\mathrm{b}} - \langle s\rangle_{\mathrm{a}} = 1/2 - \alpha/(\alpha + \beta)$. **(d)**: comparison between $\mathcal{U}_{\mathrm{GL}}$ and $\mathcal{U}_{\mathrm{DS}}$, blue regions indicate where our bound is tighter, i.e. smaller, and red regions indicate where Dechant-Sasa's bound is tighter. For all four plots, the grid is 41×41, and numerical values are rounded to 10^{-15} to avoid python floats precision errors.

Bibliography for Chapter 2

[Baake et al., 2007] Baake, E. and H.-O. Georgii (2007). Mutation, selection, and ancestry in branching models: a variational approach. *Journal of Mathematical Biology* 54.(2), pp. 257–303.

[Bansaye et al., 2011] Bansaye, V., J.-F. Delmas, L. Marsalle, and V. C. Tran (2011). Limit theorems for Markov processes indexed by continuous time Galton–Watson trees. *The Annals of Applied Probability* 21.(6), pp. 2263–2314.

[Barizien et al., 2019] Barizien, A., M. S. Suryateja Jammalamadaka, G. Amselem, and C. N. Baroud (2019). Growing from a few cells: combined effects of initial stochasticity and cell-to-cell variability. *Journal of The Royal Society Interface* 16.(153), p. 20180935.

[Campos et al., 2014] Campos, M., I. V. Surovtsev, S. Kato, A. Paintdakhi, B. Beltran, S. E. Ebmeier, and C. Jacobs-Wagner (2014). A Constant Size Extension Drives Bacterial Cell Size Homeostasis. *Cell* 159.(6), pp. 1433–1446.

[Dechant et al., 2020] Dechant, A. and S.-i. Sasa (2020). Fluctuation–response inequality out of equilibrium. *Proceedings of the National Academy of Sciences* 117.(12), pp. 6430–6436.

[Dinis et al., 2020] Dinis, L., J. Unterberger, and D. Lacoste (2020). Phase transitions in optimal betting strategies. *EPL (Europhysics Letters)* 131.(6), p. 60005.

[Doumic et al., 2021] Doumic, M. and M. Hoffmann (2021). Individual and population approaches for calibrating division rates in population dynamics: Application to the bacterial cell cycle. *arXiv:2108.13155*.

[Einstein, 1905] Einstein, A. (1905). Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik* 17.(4), pp. 549–560.

[Elowitz et al., 2002] Elowitz, M. B., A. J. Levine, E. D. Siggia, and P. S. Swain (2002). Stochastic Gene Expression in a Single Cell. *Science* 297.(5584), pp. 1183–1186.

[Fisher, 1930] Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford, UK: Clarendon Press.

[García-García et al., 2019] García-García, R., A. Genthon, and D. Lacoste (2019). Linking lineage and population observables in biological branching processes. *Physical Review E* 99.(4), p. 042413.

[Gardner, 2020] Gardner, A. (2020). Price's equation made clear. *Philosophical Transactions of the Royal Society B: Biological Sciences* 375.(1797), p. 20190361.

[Genthon et al., 2020] Genthon, A. and D. Lacoste (2020). Fluctuation relations and fitness landscapes of growing cell populations. *Scientific Reports* 10.(1), p. 11889.

[Genthon et al., 2021] Genthon, A. and D. Lacoste (2021). Universal constraints on selection strength in lineage trees. *Physical Review Research* 3.(2), p. 023187.

[Hashimoto et al., 2016] Hashimoto, M., T. Nozoe, H. Nakaoka, R. Okura, S. Akiyoshi, K. Kaneko, E. Kussell, and Y. Wakamoto (2016). Noise-driven growth rate gain in clonal cellular populations. *Proceedings of the National Academy of Sciences* 113.(12), pp. 3251–3256.

[Horowitz et al., 2020] Horowitz, J. M. and T. R. Gingrich (2020). Thermodynamic uncertainty relations constrain non-equilibrium fluctuations. *Nature Physics* 16.(1), pp. 15–20.

[Jafarpour et al., 2018] Jafarpour, F., C. S. Wright, H. Gudjonson, J. Riebling, E. Dawson, K. Lo, A. Fiebig, S. Crosson, A. R. Dinner, and S. Iyer-Biswas (2018). Bridging the Timescales of Single-Cell and Population Dynamics. *Physical Review X* 8.(2), p. 021007.

[Jarzynski, 2006] Jarzynski, C. (2006). Rare events and the convergence of exponentially averaged work values. *Physical Review E* 73.(4), p. 046105.

[Kaneko et al., 2018] Kaneko, K. and C. Furusawa (2018). Macroscopic Theory for Evolving Biological Systems Akin to Thermodynamics. *Annual Review of Biophysics* 47.(1), pp. 273–290.

[Kiviet et al., 2014] Kiviet, D. J., P. Nghe, N. Walker, S. Boulineau, V. Sunderlikova, and S. J. Tans (2014). Stochasticity of metabolism and growth at the single-cell level. *Nature* 514.(7522), pp. 376–379.

[Kobayashi et al., 2015] Kobayashi, T. J. and Y. Sughiyama (2015). Fluctuation Relations of Fitness and Information in Population Dynamics. *Physical Review Letters* 115.(23), p. 238102.

[Kobayashi et al., 2017] Kobayashi, T. J. and Y. Sughiyama (2017). Stochastic and information-thermodynamic structures of population dynamics in a fluctuating environment. *Physical Review E* 96.(1), p. 012402.

[Kubo, 1966] Kubo, R. (1966). The fluctuation-dissipation theorem. *Reports on Progress in Physics* 29.(1), pp. 255–284.

[Lambert et al., 2015] Lambert, G. and E. Kussell (2015). Quantifying Selective Pressures Driving Bacterial Evolution Using Lineage Analysis. *Physical Review X* 5.(1), p. 011016.

[Leibler et al., 2010] Leibler, S. and E. Kussell (2010). Individual histories and selection in heterogeneous populations. *Proceedings of the National Academy of Sciences* 107.(29), pp. 13183–13188.

[Levien et al., 2020] Levien, E., T. GrandPre, and A. Amir (2020). Large Deviation Principle Linking Lineage Statistics to Fitness in Microbial Populations. *Physical Review Letters* 125.(4), p. 048102.

[Levien et al., 2021] Levien, E., T. GrandPre, and A. Amir (2021). Erratum: Large Deviation Principle Linking Lineage Statistics to Fitness in Microbial Populations [Phys. Rev. Lett. **125** , 048102 (2020)]. *Physical Review Letters* 126.(7), p. 079901.

[Liao et al., 2019] Liao, J. G. and A. Berg (2019). Sharpening Jensen's Inequality. *The American Statistician* 73.(3), pp. 278–281.

[Lin et al., 2020] Lin, J. and A. Amir (2020). From single-cell variability to population growth. *Physical Review E* 101.(1), p. 012401.

[Maes et al., 2009] Maes, C., K. Netočný, and B. Wynants (2009). Dynamical fluctuations for semi-Markov processes. *Journal of Physics A: Mathematical and Theoretical* 42.(36), p. 365002.

[Mustonen et al., 2010] Mustonen, V. and M. Lassig (2010). Fitness flux and ubiquity of adaptive evolution. *Proceedings of the National Academy of Sciences* 107.(9), pp. 4248–4253.

[Mustonen et al., 2009] Mustonen, V. and M. Lässig (2009). From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends in Genetics* 25.(3), pp. 111–119.

[Neher et al., 2011] Neher, R. A. and B. I. Shraiman (2011). Statistical genetics and evolution of quantitative traits. *Reviews of Modern Physics* 83.(4), pp. 1283–1300.

[Nozoe et al., 2017] Nozoe, T., E. Kussell, and Y. Wakamoto (2017). Inferring fitness landscapes and selection on phenotypic states from single-cell genealogical data. *PLoS Genetics* 13.(3), e1006653.

[Olivier, 2017] Olivier, A. (2017). How does variability in cell aging and growth rates influence the Malthus parameter? *Kinetic and Related Models* 10.(2), pp. 481–512.

[Peliti, 1997] Peliti, L. (1997). Introduction to the statistical theory of Darwinian evolution. *arXiv:cond-mat/9712027*.

[Pigolotti, 2021] Pigolotti, S. (2021). Generalized Euler-Lotka equation for correlated cell divisions. *Physical Review E* 103.(6), p. L060402.

[Powell, 1956] Powell, E. O. (1956). Growth Rate and Generation Time of Bacteria, with Special Reference to Continuous Culture. *Journal of General Microbiology* 15.(3), pp. 492–511.

[Price, 1972] Price, G. R. (1972). Fisher's 'fundamental theorem' made clear. *Annals of Human Genetics* 36.(2), pp. 129–140.

[Rochman et al., 2018] Rochman, N. D., D. M. Popescu, and S. X. Sun (2018). Ergodicity, hidden bias and the growth rate gain. *Physical Biology* 15.(3), p. 036006.

[Sandler et al., 2015] Sandler, O., S. P. Mizrahi, N. Weiss, O. Agam, I. Simon, and N. Q. Balaban (2015). Lineage correlations of single cell division time as a probe of cell-cycle dynamics. *Nature* 519.(7544), pp. 468–471.

[Sato et al., 2003] Sato, K., Y. Ito, T. Yomo, and K. Kaneko (2003). On the relation between fluctuation and response in biological systems. *Proceedings of the National Academy of Sciences* 100.(24), pp. 14086–14090.

[Schuh et al., 2009] Schuh, R. T. and A. V. Z. Brower (2009). *Biological systematics: principles and applications*. 2nd. Ithaca: Comstock Pub. Associates/Cornell University Press.

[Smerlak et al., 2017] Smerlak, M. and A. Youssef (2017). Limiting fitness distributions in evolutionary dynamics. *Journal of Theoretical Biology* 416, pp. 68–80.

[Stewart et al., 2005] Stewart, E. J., R. Madden, G. Paul, and F. Taddei (2005). Aging and Death in an Organism That Reproduces by Morphologically Symmetric Division. *PLoS Biology* 3.(2), e45.

[Sughiyama et al., 2018] Sughiyama, Y. and T. J. Kobayashi (2018). The explicit form of the rate function for semi-Markov processes and its contractions. *Journal of Physics A: Mathematical and Theoretical* 51.(12), p. 125001.

[Sughiyama et al., 2017] Sughiyama, Y. and T. J. Kobayashi (2017). Steady-state thermodynamics for population growth in fluctuating environments. *Physical Review E* 95.(1), p. 012131.

[Sughiyama et al., 2015] Sughiyama, Y., T. J. Kobayashi, K. Tsumura, and K. Aihara (2015). Pathwise thermodynamic structure in population dynamics. *Physical Review E* 91.(3), p. 032120.

[Sughiyama et al., 2019] Sughiyama, Y., S. Nakashima, and T. J. Kobayashi (2019). Fitness response relation of a multitype age-structured population dynamics. *Physical Review E* 99.(1), p. 012413.

[Taheri-Araghi et al., 2015] Taheri-Araghi, S., S. Bradde, J. T. Sauls, N. S. Hill, P. A. Levin, J. Paulsson, M. Vergassola, and S. Jun (2015). Cell-Size Control and Homeostasis in Bacteria. *Current Biology* 25.(3), pp. 385–391.

[Tak et al., 2021] Tak, T., G. Prevedello, G. Simon, N. Paillon, C. Benlabiod, C. Marty, I. Plo, K. R. Duffy, and L. Perié (2021). HSPCs display within-family homogeneity in differentiation and proliferation despite population heterogeneity. *eLife* 10, e60624.

[Tanouchi et al., 2017] Tanouchi, Y., A. Pai, H. Park, S. Huang, N. E. Buchler, and L. You (2017). Long-term growth data of Escherichia coli at a single-cell level. *Scientific Data* 4.(1), p. 170036.

[Thomas, 2017a] Thomas, P. (2017a). Making sense of snapshot data: ergodic principle for clonal cell populations. *Journal of The Royal Society Interface* 14.(136), p. 20170467.

[Thomas, 2017b] Thomas, P. (2017b). Single-cell histories in growing populations: relating physiological variability to population growth. *BiorXiv*.

[Thomas, 2018] Thomas, P. (2018). Analysis of Cell Size Homeostasis at the Single-Cell and Population Level. *Frontiers in Physics* 6, p. 64.

[Thomas, 2019] Thomas, P. (2019). Intrinsic and extrinsic noise of gene expression in lineage trees. *Scientific Reports* 9.(1), p. 474.

[Totis et al., 2021] Totis, N., C. Nieto, A. Kuper, C. Vargas-Garcia, A. Singh, and S. Waldherr (2021). A Population-Based Approach to Study the Effects of Growth and Division Rates on the Dynamics of Cell Size Statistics. *IEEE Control Systems Letters* 5.(2), pp. 725–730.

[Touchette, 2009] Touchette, H. (2009). The large deviation approach to statistical mechanics. *Physics Reports* 478.(1), pp. 1–69.

[Wakamoto et al., 2012] Wakamoto, Y., A. Y. Grosberg, and E. Kussell (2012). Optimal lineage principle for age-structured populations. *Evolution* 66.(1), pp. 115–134.

[Wang et al., 2010] Wang, P., L. Robert, J. Pelletier, W. L. Dang, F. Taddei, A. Wright, and S. Jun (2010). Robust Growth of Escherichia coli. *Current Biology* 20.(12), pp. 1099–1103.

[Wright, 1932] Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the Sixth International Congress on Genetics*. Vol. 1, pp. 356–366.

Chapter 3

Sampling lineage trees with death[†]

[†]This chapter is based on the article in preparation Genthon et al., 2022.

Contents

1 Introduction

In the previous chapter, we made the important assumption that all lineages survive up to final time. In many situations however, lineages can end before the end of the experiment for various reasons. For example, in experiments on growing populations, cells can die as a reaction to antibiotics (Wakamoto et al., 2013), when placed in nutrient-poor environments (Schink et al., 2019), or because of the accumulation of damage proteins (Wang et al., 2010). In confined geometries, populations are maintained constant inside a chamber, and cells are flushed away by dilution in order to balance divisions (Hashimoto et al., 2016; Koldaeva et al., 2022). These two situations are represented on fig. 3.1. On the left, a freely-growing population in bulk is represented with dead cells pictured with black and white hatching. On the right, a population is maintained constant in a microfluidic channel called the dynamics cytometer (Hashimoto et al., 2016). This setup is open at both ends, and cells are continuously carried away from the chamber by a flow of growth medium that also brings the necessary nutrients. Note that we also included the possibility for cells to die inside the chamber so that lineage ending has two possible origins in this case.

The issue in all these situations is how early-ending lineages should be taken into account in the analysis. Indeed, the large amount of cytometer data exploited to analyze the relations between single cell stochasticity and population growth, and to infer the mechanism of cell size control, are meaningful only if the role of diluted lineages is well established. Otherwise, unwanted biases between surviving cells in the chamber and diluted cells could distort the conclusions. Another question naturally arises: what can we say about growing populations from measurements made in finite population setups? This question is reminiscent of the link between mother machine data and population growth rate explored in section 2.3 of chapter 2.

In this short chapter, we extend the results of the previous chapter to the cases mentioned above. Once again, the formalism we develop here applies to any branching tree independently of the dynamics. Therefore, we use 'death' as a catch-all term for any termination of lineages before final time regardless of its cause, examples of which have been given above. After adapting the forward and backward samplings to population trees with lineages that stop before final time, we show that most of our results hold in a modified form. We discuss the difference between the selection bias which reflects the difference in distribution between single lineage and population experiments, and the survivor bias that arises if surviving lineages have different features from lineages that end before. These two biases can be entangled, thus we propose a measure of the effect of death on the strength of selection.

2 Extension of the formalism

2.1 The backward and forward samplings in presence of death

We now consider a branching tree starting with N_0 cells at time $t = 0$ and ending with $N(t)$ living cells at time t, where lineages can either survive up to time t or die before,

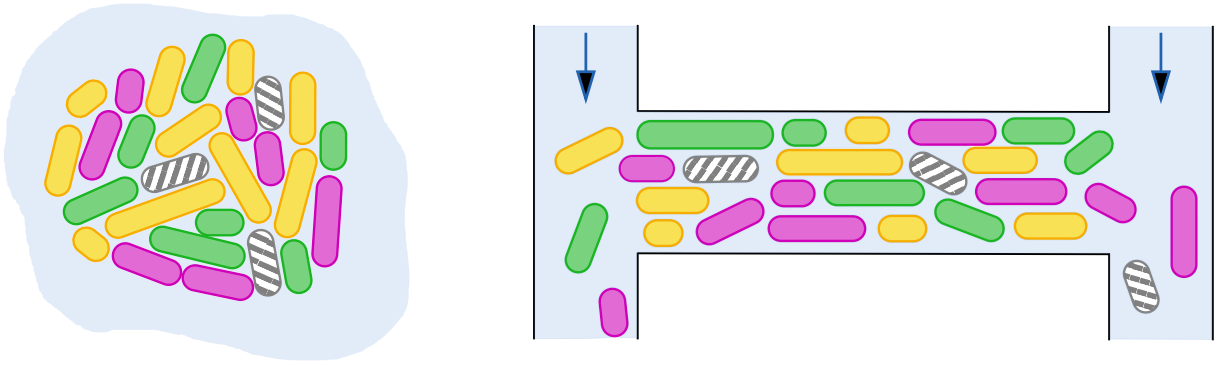
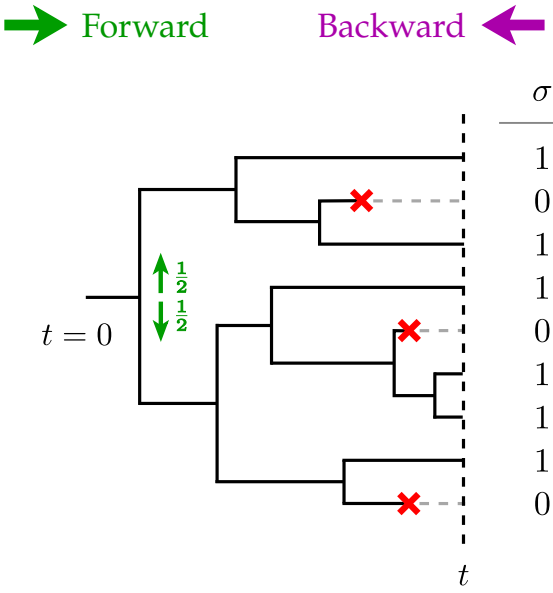


Figure 3.1: Cartoons of the two main experimental setups where lineages can end before the end of the experiment. On the left, free growth in bulk with dead cells represented with black and white hatching. On the right, the dynamics cytometer from Hashimoto et al., 2016 where cells grow in a chamber open at both ends, and are evacuated by the flow of growth medium in order to maintain the population constant inside the chamber. Some cells are also represented with hatching in this setup to indicate the possibility that cells die inside the chamber.



σ	K	ω_{back}	ω_{for}
1	2	1/6	1/4
0	3	0	1/8
1	3	1/6	1/8
1	3	1/6	1/8
0	4	0	1/16
1	5	1/6	1/32
1	5	1/6	1/32
1	3	1/6	1/8
0	3	0	1/8

Figure 3.2: Example of branching tree starting with $N_0 = 1$ lineage and ending with $n(\sigma = 0, t) = 3$ dead lineages indicated by a red crosses, and $N(t) = n(\sigma = 1, t) = 6$ alive lineages at time t. In the array, the survival status σ takes value 1 for surviving lineages and 0 for dead lineages, and the other columns indicate the number K of divisions along the lineage and the backward and forward weights computed with eqs. (3.1) and (3.2).

as represented on fig. 3.2. The survival status of the lineages at time t is indicated by a Boolean variable σ, taking value 1 for alive lineages and 0 for dead lineages, irrespective of the time of death.

The forward and backward samplings of the lineages, introduced in Nozoe et al., 2017 and presented in section 3.2 of chapter 1, are affected by the presence of death in the following way. When taking a snapshot of the population at time t, only living lineages appear, and in the backward sampling we sample them uniformly with weights

$$\omega_{\text{back}}(l) = N(t)^{-1}\delta(\sigma(l) - 1)\,. \tag{3.1}$$

On the other hand, starting from $t = 0$ and following the lineages up to time t by choosing with uniform probability $1/m$ one of the m daughter cell born at each division, both dead or living lineages are sampled with the forward weights

$$\omega_{\text{for}}(l) = N_0^{-1} m^{-K(l)}\,, \tag{3.2}$$

where $K(l)$ is the number of divisions along lineage l up to time t. We give a simple example of how these weights are computed in practice on fig. 3.2. A major difference with the deathless case immediately appears: some lineages are sampled in the forward manner but not in the backward manner.

Let us now recast these weights at the level of the probabilities to pick a lineage with K divisions. We call $\mathcal{L}(t)$ the ensemble of all lineages at time t, both dead and alive, and we define the number of lineages with K division and with survival status σ, and the probability to pick one, as:

$$n(K, \sigma, t) = \sum_{l \in \mathcal{L}(t)} \delta(K - K(l))\delta(\sigma - \sigma(l)) \tag{3.3}$$

$$p(K, \sigma, t) = \sum_{l \in \mathcal{L}(t)} \delta(K - K(l))\delta(\sigma - \sigma(l))\omega(l)\,. \tag{3.4}$$

The numbers of cells alive in the population at time t is thus given by $N(t) = n(\sigma = 1, t) = \sum_K n(K, \sigma = 1, t)$. The forward and backward probabilities finally read

$$p_{\text{for}}(K, \sigma, t) = N_0^{-1} m^{-K} n(K, \sigma, t) \tag{3.5}$$

$$p_{\text{back}}(K, \sigma = 0, t) = 0 \tag{3.6}$$

$$p_{\text{back}}(K, \sigma = 1, t) \equiv p_{\text{back}}(K, t) = N(t)^{-1} n(K, \sigma = 1, t)\,. \tag{3.7}$$

2.2 Fluctuation relation and consequences

A fluctuation relation similar to eq. (2.8) can be obtained for surviving lineages, that are the ones with a non-zero weight in both samplings. To do so, let us introduce some notations. The forward probability of survival

$$p_{\text{for}}(\sigma = 1, t) = \sum_K p_{\text{for}}(K, \sigma = 1, t) = N_0^{-1} \sum_K m^{-K} n(K, \sigma = 1, t) \tag{3.8}$$

is a central quantity in this problem. Importantly, it should not be confused with the ratio $n(\sigma = 1, t)/|\mathcal{L}(t)|$ of living lineages amongst all lineages, dead and alive. In particular,

the latter can increase or decrease with time, depending on the prevalence between death and division events, whereas $p_{\text{for}}(\sigma = 1, t)$ is a strictly decreasing function of the number of death events, and is unaffected by divisions, therefore it tends to 0 as $t \to \infty$. We now define the decrease rate of the forward probability of survival

$$\Gamma_t = \frac{1}{t} \ln p_{\text{for}}(\sigma = 1, t) \,. \tag{3.9}$$

By comparing the backward and forward probabilities $p_{\text{back}}(K, t)$ and $p_{\text{for}}(K, \sigma = 1, t)$ to pick a living lineage with K divisions, we obtain the following fluctuation theorem:

$$p_{\text{back}}(K, t) = p^{\star}_{\text{for}}(K, t)\ e^{K \ln m - t(\Lambda_t - \Gamma_t)} \,, \tag{3.10}$$

where we defined the conditional probability $p_{\text{for}}(K, t|\sigma = 1) = p_{\text{for}}(K, \sigma = 1, t)/p_{\text{for}}(\sigma = 1, t)$ and introduced the notation shorthand

$$p^{\star}_{\text{for}}(\cdot, t) = p_{\text{for}}(\cdot, t|\sigma = 1) \tag{3.11}$$

for the forward distribution conditioned on survival. Note that death enters eq. (3.10) both via the term $p_{\text{for}}(\sigma = 1, t)$ and by lowering the population growth rate Λ_t, which is no longer necessarily positive.

Similarly to the previous chapter, such a detailed fluctuation theorem can be used to derive an integral fluctuation theorem and two inequalities. Integrating the forward probability gives:

$$\langle e^{t\Lambda_t - K \ln m} \rangle_{\text{back}} = 1 - p_{\text{for}}(\sigma = 0, t) \,. \tag{3.12}$$

This result is analogous to the generalization of Jarzynski's equality for absolutely irreversible processes obtained in Murashita et al., 2014, namely $\langle \exp(-\Delta s_{\text{tot}}) \rangle = 1 - \lambda$. In thermodynamics, these processes are characterized by trajectories that are allowed only in one direction (either forward or backward in time) and have a zero probability to happen in the other direction, and λ is the total statistical weight of the unidirectional transitions. Similarly here, dead lineages have a positive weight in the forward sampling and a null one in the backward sampling. In this analogy, $K \ln m - t\Lambda_t$ plays the role of the total entropy production, and $p_{\text{for}}(\sigma = 0, t)$ is the probability λ of the singular part, that is the set of 'irreversible lineages'.

Inequalities of the type of the second law are obtained using the positivity of the two following Kullback-Leibler divergences:

$$\mathcal{D}_{\text{KL}}(p_{\text{back}} || p^{\star}_{\text{for}}) = \langle K \rangle_{\text{back}} \ln m - t\,(\Lambda_t - \Gamma_t) \geq 0 \tag{3.13}$$

$$\mathcal{D}_{\text{KL}}(p^{\star}_{\text{for}} || p_{\text{back}}) = -\langle K \rangle_{\text{for}^\star} \ln m + t\,(\Lambda_t - \Gamma_t) \geq 0 \,. \tag{3.14}$$

Since the number K of divisions is positively defined, eq. (3.14) implies that $\Lambda_t - \Gamma_t$ is a positive quantity. Combined, the two inequalities read

$$\langle K \rangle_{\text{for}^\star} \ln m \leq t\,(\Lambda_t - \Gamma_t) \leq \langle K \rangle_{\text{back}} \ln m \,, \tag{3.15}$$

which generalizes eq. (2.13).

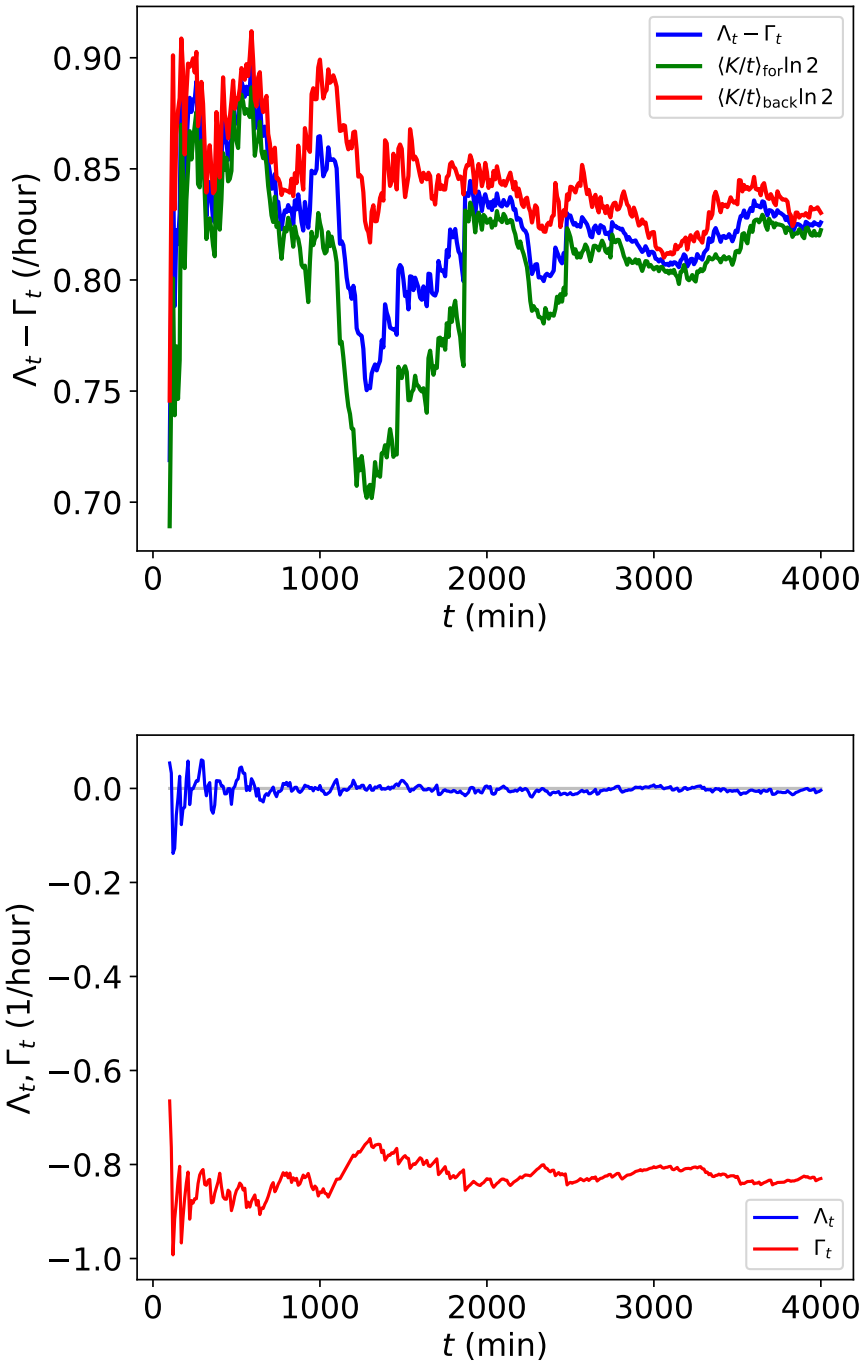


Figure 3.3: Analysis of data for a single constant population with dilution from Hashimoto et al., 2016. The top plot shows the test of eq. (3.15), where the bounds on $\Lambda_t - \Gamma_t$ are tighter as time increases. The bottom plot shows the evolution of Λ_t and Γ_t with time separately, with a clear convergence of Λ_t toward 0, which is expected for constant populations, and a slower convergence of Γ_t toward a steady value, not fully achieved within the reach of the experiment. The data were analyzed and the plots kindly communicated to us by Takashi Nozoe.

We tested these inequalities on data from Hashimoto et al., 2016, presented in section 5 of chapter 1, and the results are shown on fig. 3.3 (top). The values of $\langle K \rangle_{\text{for}^\star}$, $\langle K \rangle_{\text{back}}$, Λ_t and Γ_t are computed for a single population, maintained approximately constant with $20 \sim 40$ cells at any time. We see that the relative discrepancies between the curves are reducing with time from $t \sim 1300$ min, meaning that the bounds on $\Lambda_t - \Gamma_t$ are getting tighter. This is due to the fact that, because of dilution, all cells in the cytometer are likely to have a common ancestor which is close in the past, thus leading to a small variability in the numbers of divisions amongst lineages. Steady-state is not fully reached in this experiment, as shown on the bottom plot where the separate evolution of Λ_t and Γ_t are displayed. As expected for constant populations, Λ_t tends to 0 quickly, while the convergence of Γ_t is slower.

2.3 Link with population dynamics

In the previous chapter, we insisted on the fact that the formalism is independent of the dynamics and can be used to study any branching tree. In particular, the population growth rate $\Lambda_t = \ln\left(N(t)/N_0\right)/t$ and the instantaneous population growth rate $\Lambda_p(t) = \partial_t N/N$ are only functions of the number $N(t)$ of lineages. If we now investigate a particular dynamics, these quantities become related to the division rate of the model by eq. (1.65). Similarly, the decrease rate Γ_t is defined for any tree by simply counting the living lineages weighted by their forward weights, but when considering a particular dynamics it can be linked to the death rate of the model, as shown below.

The precise variables of the model do not matter here, so we consider the general case of mixed age-size controlled populations, where the division rate r depends on both age a and size x. This covers the most popular models of cell size control presented in section 4 of chapter 1: namely the sizer, where r depends only on the size; the timer, for which r depends only on the age; and the adder where r depends on the increment of volume since birth which can be deduced from the knowledge of age and size. We now add a death rate γ_t, where the subscript indicates a possible time dependence. This death rate is a function of an aging factor z, which is typically a number of damage proteins (Maisonneuve et al., 2008). We model the dynamics of this factor with the accumulation rate λ during the cell cycle, and the transmission kernel $\Sigma(x, z|x', z', a')$ at division. In all generality, the death rate also depends on the age and the size, and the growth rate ν and the division rate r are also function of z, to account for possible aging effects (Lindner et al., 2008). For simplicity we define the vector $y = \{x, a, z\}$ of variables.

The population balance equation for the backward probability is given by

$$\begin{aligned} \partial_t p_{\text{back}}(y,t) =& -\partial_x\left[\nu(y)p_{\text{back}}(y,t)\right] - \partial_z\left[\lambda(y)p_{\text{back}}(y,t)\right] - \partial_a\left[p_{\text{back}}(y,t)\right] \\ & - \left[r(y) + \gamma_t(y) + \Lambda_p(t)\right] p_{\text{back}}(y,t) \end{aligned} \tag{3.16}$$

$$p_{\text{back}}(x, a=0, z, t) = m \int \mathrm{d}y'\ \Sigma(x, z|y') r(y') p_{\text{back}}(y', t)\,. \tag{3.17}$$

Direct integration of eq. (3.16) over y leads to

$$\Lambda_p(t) = \int \mathrm{d}y\ \left[(m-1)r(y) - \gamma_t(y)\right] p_{\text{back}}(y,t)\,. \tag{3.18}$$

In presence of death, the population growth rate is the net difference between the backward averaged division and death rates.

Similarly, the equation for the forward probability reads

$$\begin{aligned}\partial_t p_{\text{for}}(y,\sigma=1,t) = &-\partial_x\left[\nu(y)p_{\text{for}}(y,\sigma=1,t)\right]-\partial_z\left[\lambda(y)p_{\text{for}}(x,z,\sigma=1,t)\right]\\ &-\partial_a\left[p_{\text{for}}(y,\sigma=1,t)\right]-\left[r(y)+\gamma_t(y)\right]p_{\text{for}}(y,\sigma=1,t)\end{aligned} \tag{3.19}$$

$$p_{\text{for}}(x,a=0,z,\sigma=1,t) = \int \mathrm{d}y'\ \Sigma(x,z|y')r(y')p_{\text{for}}(y',\sigma=1,t)\,. \tag{3.20}$$

Integrating this equation over y gives

$$\partial_t p_{\text{for}}(\sigma=1,t) = -p_{\text{for}}(\sigma=1,t)\int \mathrm{d}y\ \gamma_t(y)p_{\text{for}}(y,t|\sigma=1)\,. \tag{3.21}$$

We then define the instantaneous forward death rate as

$$\Gamma_p(t) = \frac{\partial_t p_{\text{for}}(\sigma=1)}{p_{\text{for}}(\sigma=1)} = -\int \mathrm{d}y\ \gamma_t(y)p_{\text{for}}(y,t|\sigma=1)\,. \tag{3.22}$$

Note the similarity in construction between $\Lambda_p(t)$ and $\Gamma_p(t)$, and between their time-averaged versions Λ_t and Γ_t:

$$\Lambda_t = \frac{1}{t}\int_0^t \mathrm{d}t'\ \Lambda_p(t') \tag{3.23}$$

$$\Gamma_t = \frac{1}{t}\int_0^t \mathrm{d}t'\ \Gamma_p(t')\,. \tag{3.24}$$

Finally we showed that Λ_t results from the competition between division and death, while Γ_t depends only on the death rate γ_t. Note that the mixed age-size model we considered was only an example, and that eqs. (3.18) and (3.22) hold for any vector y of variables obeying a population balance equation.

2.4 The case of uniform dilution

In the context of confining geometries such as the dynamics cytometer, the general 'death' rate we introduced in the previous section is called the dilution rate. It is almost always postulated for these constant population experiments that the dilution rate is uniform: $\gamma_t(y) \equiv \gamma_t$ (Powell, 1956; Hashimoto et al., 2016; Levien et al., 2020), and balances division:

$$\gamma_t = (m-1)\int \mathrm{d}y\ r(y)p_{\text{back}}(y,t)\,. \tag{3.25}$$

This states that for each cell division, a random cell is chosen with uniform probability and removed from the population, which is known as the Moran process (Moran, 1958). This is of course a simple modeling of the dilution that really occurs, and we discuss the implications of this hypothesis in section 4.4.

In general, the terms $\langle\gamma_t\rangle_{\text{back}}$ in $\Lambda_p(t)$ (eq. (3.18)) and $\langle\gamma_t\rangle_{\text{for}^\star}$ in $\Gamma_p(t)$ (eq. (3.22)) do not cancel. However, they are equal for uniform dilution rates, so that the bias in the fluctuation theorem eq. (3.10) reduces to $\Lambda_t - \Gamma_t = (m-1)\int_0^t \mathrm{d}t'\ \langle r\rangle_{\text{back}}$. This is the same bias as the one when there is no dilution, and as we shall see in the following, when dilution (or death) is uniform, the results become identical to their versions without dilution. In this case, the rate $\Gamma_p(t)$ is simply equal to the dilution rate $\Gamma_p(t) = -\gamma_t = -(m-1)\langle r\rangle_{\text{back}}$, and understood as the opposite of the population growth rate.

3 Generalized Powell's relation

Age models without mother-daughter correlations are characterized by a series of results on the distribution of generation times detailed in section 4.5 of chapter 1, namely Powell's relation and Euler-Lotka equation, that give important relations between the variability in generation times and the population growth rate. Since the original article Powell, 1956, these results have been extended in several directions, in particular to include mother-daughter correlations and for populations with cell death or dilution.

Even though Powell himself considered in Powell, 1956 and in later works Powell, 1964; Powell, 1969 the case of 'continuous cultures', which are populations maintained constant with uniform dilution, and the case of correlations between mother and daughter, the next major advances have been made in Lebowitz et al., 1974. In this work, the authors considered populations with an age-dependent death rate $\gamma(a)$, which includes Powell's continuous cultures when $\gamma(a) \equiv \gamma$ is a constant, and with mother-daughter correlations mediated by a transition kernel at division $\Sigma(\tau|\tau')$, as presented in section 4.1.1 of chapter 1. They obtained a generalization of Euler-Lotka equation involving the distribution of generation times for newborn cells, and, for uniform dilution only, a series of relations between this newborn distribution and the forward and backward distributions. Recently, the Markovian assumption behind the kernel $\Sigma(\tau|\tau')$ has been released in Pigolotti, 2021. Note that the case of continuous cultures is still under focus today, with an emphasis on the competition between multiple species (Levien et al., 2020).

With the development of mother-machines, relations comparing forward and backward distributions received a renewed interest, in particular in the recent works Sughiyama et al., 2019; Nakashima et al., 2020. In these two articles, the authors described multitype age models, as presented in section 4.1.1 of chapter 1, where the correlations are accounted for by cell types y, an age-and-type-dependent death rate $\gamma(a, y)$ and mother-daughter correlations in types via kernel $\Sigma(y|y')$. The advantages of this choice of model are that it may be closer to the biological mechanism causing correlations, and that it allows them to derive a generalized Powell's relation for each type y, which involves a type-dependent correction $Z(y)$ to the classical Powell's relation. Note that their results are particular in that they compare the forward distribution of generation times in the absence of death to the backward distribution in the presence of death, and for that reason they explicitly involve the death rate.

We aim here to complete this set of results for age models with mother-daughter correlations and age-dependent death rate. The case of Lebowitz's model with kernel $\Sigma(\tau|\tau')$ is treated in appendix A, where we obtain new relations between the forward and backward distributions for non-uniform death rates. In the following, we focus on multitype age models with kernel $\Sigma(y|y')$, and we offer an alternative derivation of the result from Sughiyama et al., 2019. In addition to their result, (i) we obtain an explicit and simple expression for the constant $Z(y)$, (ii) we show that when comparing the forward and backward distributions both in the presence of death the bias is simplified as it does not involve the death rate anymore, and (iii) we derive an inequality on average generation times conditioned on type.

Before going into details, note that for models with age-dependent death rate but

without correlations, the derivation of Powell's relation starting from the fluctuation theorem, presented in section 2.4 of chapter 2, holds. The key point is that the probability of a trajectory with K divisions can be factorized as a product of K probabilities for single cycles because there is no memory between mother and daughter, and this remains true even when adding an age-dependent death rate. The resulting Powell's relation and Euler-Lotka equation are then identical to those without death, except for Λ being replaced by $\Lambda - \Gamma$ and for the forward distribution $f_{\text{for}}(\tau)$ being replaced by the same distribution $f^\star_{\text{for}}(\tau)$ conditioned on survival.

3.1 Age distributions

Let us recap the equations for the forward and backward probabilities together with their boundary conditions:

$$\partial_t p_{\text{back}}(a,y,t) = -\partial_a p_{\text{back}}(a,y,t) - \left[r(a,y) + \gamma(a,y) + \Lambda_p(t)\right] p_{\text{back}}(a,y,t) \tag{3.26}$$

$$p_{\text{back}}(a=0,y,t) = m \int \mathrm{d}a'\mathrm{d}y' r(a',y') \Sigma(y|y') p_{\text{back}}(a',y',t) \tag{3.27}$$

$$\partial_t p^\star_{\text{for}}(a,y,t) = -\partial_a p^\star_{\text{for}}(a,y,t) - \left[r(a,y) + \gamma(a,y) + \Gamma_p(t)\right] p^\star_{\text{for}}(a,y,t) \tag{3.28}$$

$$p^\star_{\text{for}}(a=0,y,t) = \int \mathrm{d}a'\mathrm{d}y' r(a',y') \Sigma(y|y') p^\star_{\text{for}}(a',y',t)\,, \tag{3.29}$$

where the death rate γ depends on both age a and type y.

The steady-state solutions to these equations read

$$p_{\text{back}}(a,y) = p_{\text{back}}(0,y) \exp\left[-\Lambda a - \int_0^a \mathrm{d}a' \;\; \left(r(a',y) + \gamma(a',y)\right)\right] \tag{3.30}$$

$$p^\star_{\text{for}}(a,y) = p^\star_{\text{for}}(0,y) \exp\left[-\Gamma a - \int_0^a \mathrm{d}a' \;\; \left(r(a',y) + \gamma(a',y)\right)\right]\,. \tag{3.31}$$

3.2 Powell's relation with joint probabilities

We define the steady-state joint distribution $f(\tau,y)$ of generation time τ and state y as the ratio of the number of cells dividing at age τ while in state y at a given snapshot time, to the total number of cells dividing in this snapshot, weighted in both the forward and backward samplings:

$$f_{\text{back}}(\tau,y) = \frac{r(\tau,y) p_{\text{back}}(\tau,y)}{\int \mathrm{d}\tau'\mathrm{d}y' \; r(\tau',y') p_{\text{back}}(\tau',y')} \tag{3.32}$$

$$f^\star_{\text{for}}(\tau,y) = \frac{r(\tau,y) p^\star_{\text{for}}(\tau,y)}{\int \mathrm{d}\tau'\mathrm{d}y' \; r(\tau',y') p^\star_{\text{for}}(\tau',y')}\,. \tag{3.33}$$

These distributions are independent of the time of the snapshot in steady-state. Integrating the boundary conditions eqs. (3.27) and (3.29) over y and using the normalization of

kernel Σ, we get that the denominators of eqs. (3.32) and (3.33) are given by:

$$\int \mathrm{d}y\ p_{\mathrm{back}}(0,y) = m \int \mathrm{d}\tau \mathrm{d}y' r(\tau,y') p_{\mathrm{back}}(\tau,y') \tag{3.34}$$

$$\int \mathrm{d}y\ p^{\star}_{\mathrm{for}}(0,y) = \int \mathrm{d}\tau \mathrm{d}y' r(\tau,y') p^{\star}_{\mathrm{for}}(\tau,y')\,. \tag{3.35}$$

Then, we identify the distribution of types y for newborn cells:

$$\rho^{\mathrm{nb}}(y) = \frac{p(0,y)}{\int \mathrm{d}y\ p(0,y)}\,, \tag{3.36}$$

for both the forward and backward probabilities.

Finally, combining the results above we obtain

$$f_{\mathrm{back}}(\tau,y) = m\rho^{\mathrm{nb}}_{\mathrm{back}}(y) r(\tau,y) \exp\left[-\Lambda\tau - \int_0^\tau \mathrm{d}a\ \ (r(a,y)+\gamma(a,y))\right] \tag{3.37}$$

$$f^{\star}_{\mathrm{for}}(\tau,y) = \rho^{\mathrm{nb}}_{\mathrm{for}}(y) r(\tau,y) \exp\left[-\Gamma\tau - \int_0^\tau \mathrm{d}a\ \ (r(a,y)+\gamma(a,y))\right]\,, \tag{3.38}$$

and the generalized Powell's equation reads

$$f_{\mathrm{back}}(\tau,y) = m\frac{\rho^{\mathrm{nb}}_{\mathrm{back}}(y)}{\rho^{\mathrm{nb}}_{\mathrm{for}}(y)} f^{\star}_{\mathrm{for}}(\tau,y) e^{-(\Lambda-\Gamma)\tau}\,. \tag{3.39}$$

In the absence of mother-daughter correlations, that is when $\Sigma(y|y') \equiv \hat{\Sigma}(y)$, the newborn distributions are unbiased: $\rho^{\mathrm{nb}}_{\mathrm{back}} = \rho^{\mathrm{nb}}_{\mathrm{for}} = \hat{\Sigma}$, which is a direct consequence of the boundary conditions. Therefore, the fraction in eq. (3.39) cancel, and eq. (3.39) can be integrated over states y to recover Powell's relation in presence of death but without correlations: $f_{\mathrm{back}}(\tau) = m f^{\star}_{\mathrm{for}}(\tau) e^{-(\Lambda-\Gamma)\tau}$.

3.3 Powell's relation with conditional probabilities

It can be useful to recast this result for the distributions $f(\tau|y)$ of generation time conditioned on state y, defined as:

$$f(\tau|y) = \frac{r(\tau,y)p(\tau,y)}{\int \mathrm{d}\tau'\ r(\tau',y)p(\tau',y)} \tag{3.40}$$

$$= f(\tau,y)\ \frac{\int \mathrm{d}\tau'\mathrm{d}y'\ r(\tau',y')p(\tau',y')}{\int \mathrm{d}\tau'\ r(\tau',y)p(\tau',y)} \tag{3.41}$$

$$= \frac{f(\tau,y)}{\rho^{\mathrm{d}}(y)}\,, \tag{3.42}$$

where we identified the distribution of states at division:

$$\rho^{\mathrm{d}}(y) = \frac{\int \mathrm{d}\tau'\ r(\tau',y)p(\tau',y)}{\int \mathrm{d}\tau'\mathrm{d}y'\ r(\tau',y')p(\tau',y')}\,. \tag{3.43}$$

The distributions of states at birth and at division are related by a simple relation, both in the forward and backward statistics:

$$\rho^{\rm nb}(y) = \int {\rm d}y' \; \Sigma(y|y')\rho^{\rm d}(y') \,, \tag{3.44}$$

which is a normalized version of the boundary conditions eqs. (3.27) and (3.29).

The conditioned distributions are thus given by:

$$f_{\rm back}(\tau|y) = m\frac{\rho^{\rm nb}_{\rm back}(y)}{\rho^{\rm d}_{\rm back}(y)} r(\tau,y) \exp\left[-\Lambda\tau - \int_0^\tau {\rm d}a \;\; (r(a,y)+\gamma(a,y))\right] \tag{3.45}$$

$$f^\star_{\rm for}(\tau|y) = \frac{\rho^{\rm nb}_{\rm for}(y)}{\rho^{\rm d}_{\rm for}(y)} r(\tau,y) \exp\left[-\Gamma\tau - \int_0^\tau {\rm d}a \;\; (r(a,y)+\gamma(a,y))\right] . \tag{3.46}$$

Finally, Powell's relation on conditional distributions reads:

$$f_{\rm back}(\tau|y) = \frac{m f^\star_{\rm for}(\tau|y) e^{-(\Lambda-\Gamma)\tau}}{Y(y)} \,, \tag{3.47}$$

with the state-dependent normalization constant

$$Y(y) = \frac{\rho^{\rm nb}_{\rm for}(y)}{\rho^{\rm nb}_{\rm back}(y)} \frac{\rho^{\rm d}_{\rm back}(y)}{\rho^{\rm d}_{\rm for}(y)} \,. \tag{3.48}$$

Let us show how the result from Sughiyama et al., 2019, namely

$$f_{\rm back}(\tau|y) = \frac{m f^\circ_{\rm for}(\tau|y) e^{-\Lambda\tau - \int_0^\tau {\rm d}a \;\; \gamma(a,y)}}{Z(y)} \,, \tag{3.49}$$

where $f^\circ_{\rm for}(\tau|y)$ is the forward distribution in the absence of death, is recovered from eq. (3.47). In the absence of death ($\gamma = \Gamma = 0$), the normalization of $f^\circ_{\rm for}$ in eq. (3.46) reads $\int_0^\infty {\rm d}\tau \; f^\circ_{\rm for}(\tau|y) = \rho^{{\rm nb},\circ}_{\rm for}(y)/\rho^{{\rm d},\circ}_{\rm for}(y) \times \int_0^\infty {\rm d}\tau \; r(\tau,y) \exp\left[-\int_0^\tau {\rm d}a \; r(a,y)\right] = \rho^{{\rm nb},\circ}_{\rm for}(y)/\rho^{{\rm d},\circ}_{\rm for}(y) = 1$. Therefore, we obtain

$$f^\star_{\rm for}(\tau|y) = \frac{\rho^{\rm nb}_{\rm for}(y)}{\rho^{\rm d}_{\rm for}(y)} f^\circ_{\rm for}(\tau|y) \exp\left[-\Gamma\tau - \int_0^\tau {\rm d}a \; \gamma(a,y)\right] . \tag{3.50}$$

When plugging eq. (3.50) into eq. (3.47), we obtain eq. (3.49), with $Z(y) = Y(y)\rho^{\rm d}_{\rm for}(y)/\rho^{\rm nb}_{\rm for}(y) = \rho^{\rm d}_{\rm back}(y)/\rho^{\rm nb}_{\rm back}(y)$.

The advantages of our formulation are that the two distributions can be evaluated on the same population tree, and that knowledge of the shape of the death rate $\gamma(a,y)$ is not required since Γ can be computed simply by counting surviving lineages. Moreover we obtain explicit expressions for the constants $Y(y)$ and $Z(y)$, in terms of the distributions of types y at birth and at division.

3.4 Euler-Lotka equations

We saw in section 4.5 of chapter 1 that Euler-Lotka equation can be derived by simply integrating Powell's relation over generation time τ. Therefore, generalized Euler-Lotka equations in presence of age-and-type-dependent death and mother-daughter correlations can be derived from eqs. (3.39) and (3.47):

$$1 = m \int \mathrm{d}y \, \frac{\rho^{\mathrm{nb}}_{\mathrm{back}}(y)}{\rho^{\mathrm{nb}}_{\mathrm{for}}(y)} \int \mathrm{d}\tau \; f^{\star}_{\mathrm{for}}(\tau, y) e^{-(\Lambda-\Gamma)\tau} \tag{3.51}$$

$$1 = mY(y)^{-1} \int \mathrm{d}\tau \; f^{\star}_{\mathrm{for}}(\tau|y) e^{-(\Lambda-\Gamma)\tau} \, . \tag{3.52}$$

These relations link the shifted population growth rate $\Lambda - \Gamma$ to the variabilities of the different distributions involved: forward distributions of generation times, either joint with or conditioned on type y, and the forward and backward distributions of state at birth and at division.

3.5 Inequality on average generation times

Before closing this section, we show that a weaker form of the inequalities on mean generation times eq. (1.87) known in uncorrelated age models can be derived in presence of correlations and death. Using the positivity of the two Kullback-Leibler divergences between $f_{\mathrm{back}}(\tau|y)$ and $f^{\star}_{\mathrm{for}}(\tau|y)$, we obtain from eq. (3.47):

$$\ln m - \ln Y(y) - (\Lambda - \Gamma) \int \mathrm{d}\tau \; \tau f_{\mathrm{back}}(\tau|y) \geq 0 \tag{3.53}$$

$$-\ln m + \ln Y(y) + (\Lambda - \Gamma) \int \mathrm{d}\tau \; \tau f^{\star}_{\mathrm{for}}(\tau|y) \geq 0 \, , \tag{3.54}$$

which leads for any type y to:

$$\langle \tau|y \rangle_{\mathrm{for}^{\star}} \geq \langle \tau|y \rangle_{\mathrm{back}} \, , \tag{3.55}$$

where we defined $\langle \tau|y \rangle = \int \mathrm{d}\tau \; \tau f(\tau|y)$.

4 Quantifying selection in population trees with death

One of the main applications of the formalism with the forward and backward samplings of the lineages is the definition of a model-independent measure of selection, as exposed in section 3 of chapter 2. This selection involves the ratio between the forward and the backward frequencies of a phenotypic trait, where a bias between the two distributions results, in the absence of death, from the correlations between the value of the trait and the number of divisions along the lineage. When adding death, the frequency of a value s in the population is due to both the correlations between this trait and the divisions, and the correlations between this trait and survival.

4.1 The survivor bias

In this section we show how the forward and backward probabilities are modified when death is non-uniform. For that, let us consider a general model with an unspecified vector χ of cell properties, possibly of high dimension. We imagine two experiments: one in which the population is not subject to death, indicated by a superscript $^\circ$, and the other one where cells die with a rate $\gamma_t(\chi)$. Moreover, we make the important assumptions that the rate at which χ evolves inside cell cycles, the division rate, and the partition of χ at division are the same in the two experiments, and that there is no extra term for one of the two experiments, so that the two population balance equations differ only by the death term proportional to $\gamma_t(\chi)$.

The expected numbers of lineages following the path $\boldsymbol{\chi}$ are then linked by

$$n(\boldsymbol{\chi}) = n^\circ(\boldsymbol{\chi}) \exp\left[-\int_0^t \mathrm{d}t' \gamma_t(\chi(t'))\right] . \tag{3.56}$$

where the exponential term is called the survival probability for trajectory $\boldsymbol{\chi}$:

$$p_{\mathrm{surv}}(\boldsymbol{\chi}) = \exp\left[-\int_0^t \mathrm{d}t' \gamma_{t'}(\chi(t'))\right] . \tag{3.57}$$

We divide eq. (3.56) by $N_0^{-1} m^{-K[\boldsymbol{\chi}]}$ (for simplicity we consider that $N_0 = N_0^\circ$ here), and we obtain

$$p_{\mathrm{for}}(\boldsymbol{\chi}, \sigma = 1) = p^\circ_{\mathrm{for}}(\boldsymbol{\chi}) p_{\mathrm{surv}}(\boldsymbol{\chi}) , \tag{3.58}$$

We then condition the probability in the left hand side on survival: $p_{\mathrm{for}}(\boldsymbol{\chi}, \sigma = 1) = p^\star_{\mathrm{for}}(\boldsymbol{\chi}) p_{\mathrm{for}}(\sigma = 1)$. The term $p_{\mathrm{for}}(\sigma = 1)$ could be expressed in terms of the death rate by combining eqs. (3.9), (3.22) and (3.24) (with $y = \boldsymbol{\chi}$ here), but instead we write it as a normalizing factor $p_{\mathrm{for}}(\sigma = 1) = \langle p_{\mathrm{surv}}(\boldsymbol{\chi}) \rangle_{\mathrm{for}^\circ}$:

$$p^\star_{\mathrm{for}}(\boldsymbol{\chi}) = p^\circ_{\mathrm{for}}(\boldsymbol{\chi}) \, \frac{p_{\mathrm{surv}}(\boldsymbol{\chi})}{\langle p_{\mathrm{surv}}(\boldsymbol{\chi}) \rangle_{\mathrm{for}^\circ}} , \tag{3.59}$$

A similar bias can be obtained at the level of backward probabilities. First, integrating eq. (3.56) over every paths leads to

$$N_t = N_t^\circ \, \langle p_{\mathrm{surv}}(\boldsymbol{\chi}) \rangle_{\mathrm{back}^\circ} . \tag{3.60}$$

Then when dividing eq. (3.56) by N_t we obtain

$$p_{\mathrm{back}}(\boldsymbol{\chi}) = p^\circ_{\mathrm{back}}(\boldsymbol{\chi}) \, \frac{p_{\mathrm{surv}}(\boldsymbol{\chi})}{\langle p_{\mathrm{surv}}(\boldsymbol{\chi}) \rangle_{\mathrm{back}^\circ}} . \tag{3.61}$$

Equations (3.59) and (3.61) express survivor biases, which involve the comparison between the survival probability of a given path and its average value. Of course, if death is uniform these two biases are canceled.

Note that, when combining the version of eq. (3.10) at the level of path probabilities with eq. (3.59), we recover the fluctuation theorem between the backward probability with death and the forward probability without death derived in Sughiyama et al., 2019. This theorem is only defined at the level of trajectories, and is model dependent, while eq. (3.10) is more general.

4.2 Effect of death on fitness and selection

The measures of fitness and selection discussed in chapter 2 can be used for populations where some lineages end prematurely, we must simply adapt them to compare the forward and backward distributions of surviving lineages only. The fitness landscape is then defined as

$$h_t(s) = \Lambda_t - \Gamma_t + \frac{1}{t} \ln \left[\frac{p_{\text{back}}(s,t)}{p^\star_{\text{for}}(s,t)} \right] . \tag{3.62}$$

A similar function comparing the backward distribution to the forward distribution, not conditioned on survival, could also be considered. However, such a definition would face at least two problems: first, one needs a proper definition of the forward probability of a cell state s at time t for lineages that died before time t, and second, the support of the two distributions could be different, leading to a diverging fitness landscape when $p_{\text{for}}(s,t) \neq 0$ and $p_{\text{back}}(s,t) = 0$. Moreover, the interpretation of this function in terms of selection would be less clear. For these reasons, we stick to the definition given by eq. (3.62).

Like in chapter 2, this definition can be combined with the fluctuation relation for the number of divisions eq. (3.10) and turned into

$$h_t(s) = \frac{1}{t} \ln \left[\sum_K m^K p_{\text{for}}(K,t|s,\sigma=1) \right] . \tag{3.63}$$

When $p_{\text{for}}(K,t|s,\sigma=1) = p_{\text{for}}(K,t|\sigma=1)$ for any s, then the fitness landscape is flat and equal to the population growth rate. This condition is called the conditional independence of K and s knowing $\sigma = 1$. Be careful that the conditional independence does not imply, and is not implied by, the regular independence between K and s. This means in particular that the trait s can be correlated to both K and σ, and still have a flat landscape. Similarly, the strength of selection is now defined as the Jeffrey's divergence between the backward distribution and forward distribution conditioned on survival:

$$\Pi_{\mathcal{S}} = \frac{1}{t} \int \mathrm{d}s \ \left(p_{\text{back}}(s,t) - p^\star_{\text{for}}(s,t) \right) \ln \left(\frac{p_{\text{back}}(s,t)}{p^\star_{\text{for}}(s,t)} \right) . \tag{3.64}$$

Even though showing that these measures of fitness and selection can be generalized to surviving lineages is interesting in itself, in this section we focus on the quantitative effect of death on the strength of selection. Indeed, the strength of selection $\Pi_{\mathcal{S}}$ results from the combination of the intrinsic selection effect, present in the absence of death and due to the variability in lineage reproductive successes and to the correlations between phenotype and reproductive success; and of the survival biases that impact the forward and backward distributions. For this reason, we propose the following measure for the effect of death on selection:

$$\Delta\Pi_{\mathcal{S}} = \Pi_{\mathcal{S}} - \Pi^\circ_{\mathcal{S}} . \tag{3.65}$$

The sign of $\Delta\Pi_{\mathcal{S}}$ indicates if death increases or decreases the distance between the forward and backward distributions, when compared to the deathless case.

When focusing on phenotypic trajectories $\boldsymbol{\mathcal{S}}$, and using eqs. (3.59) and (3.61), $\Delta\Pi_{\boldsymbol{\mathcal{S}}}$ can be made more explicit, as shown in appendix B:

$$\Delta\Pi_{\boldsymbol{\mathcal{S}}} = \frac{\mathrm{Cov}_{\mathrm{back}^\circ}\left(h_t^\circ, p_{\mathrm{surv}}\right)}{\langle p_{\mathrm{surv}}\rangle_{\mathrm{back}^\circ}} - \frac{\mathrm{Cov}_{\mathrm{for}^\circ}\left(h_t^\circ, p_{\mathrm{surv}}\right)}{\langle p_{\mathrm{surv}}\rangle_{\mathrm{for}^\circ}} \,. \tag{3.66}$$

The covariances between the fitness landscape in the absence of death $h_t^\circ(\boldsymbol{s})$, representing the intrinsic selection effect, and the probability of survival up to time t: $p_{\mathrm{surv}}(\boldsymbol{s})$; express that only the correlations between reproductive success and survival are determinant. This also indicates that one cannot disentangle the two effects into two separate terms. There are two situations where death does not change the strength of selection, i.e. $\Delta\Pi_{\boldsymbol{\mathcal{S}}} = 0$: (i) when both covariances are null, and (ii) when they are non-zero but cancel. (i) They are both null when there are no correlations between selection and survival, which is the case in particular when either (a) the survival probability or (b) the fitness landscape are constant functions. Case (a) has been discussed previously: the survival probability is uniform when the death rate is uniform, and in this case there is no survivor bias, and thus no effect on the strength of selection. Case (b) may seem less intuitive: if the fitness landscape in the absence of death h_t° is constant, then even with a phenotype-dependent death rate leading to strong survivor biases, death has no effect on selection. Since in this case $\Pi_{\boldsymbol{\mathcal{S}}}^\circ = 0$, this implies that $\Pi_{\boldsymbol{\mathcal{S}}} = 0$ as well, in other words, death cannot create selection if it is not present beforehand. (ii) Correlations between survival and fitness with the forward and backward distributions are non zero but identical if death impacts the forward and backward distributions 'equivalently', in such a way that the distance between them (in the sense of the Jeffrey's divergence) remains the same as compared to the case without death. The simple example of a two-state model is provided in the next section, where we show under which conditions $\Delta\Pi_{\boldsymbol{\mathcal{S}}}$ is positive, negative or null.

Finally, a linear response inequality, akin to the constraints on the strength of selection we derived in section 3.3.2 of chapter 2, can be derived using Cauchy-Schwarz inequality:

$$|\Delta\Pi_{\boldsymbol{\mathcal{S}}}| \leq \frac{\sigma_{\mathrm{for}^\circ}(h_t^\circ)\sigma_{\mathrm{for}^\circ}\left(p_{\mathrm{surv}}\right)}{\langle p_{\mathrm{surv}}\rangle_{\mathrm{back}^\circ}} + \frac{\sigma_{\mathrm{back}^\circ}(h_t^\circ)\sigma_{\mathrm{back}^\circ}\left(p_{\mathrm{surv}}\right)}{\langle p_{\mathrm{surv}}\rangle_{\mathrm{for}^\circ}} \,. \tag{3.67}$$

Unlike in eq. (3.66), here selection and survival are decoupled and the bound involves products of the variabilities in fitness and in survival probability.

4.3 Illustrative example

Let us illustrate the possible values of $\Delta\Pi_{\boldsymbol{\mathcal{S}}}$ with a simple two state model: cells come in two phenotypes a and b, with division rates r_a and r_b and death rates γ_a and γ_b. In order to avoid extinction, we suppose that $r_a > \gamma_a$ and $r_b > \gamma_b$, and we start with a large even number N_0 of initial cells. For simplicity, we consider that in this initial population, phenotypes are equally represented: $N_0(a) = N_0(b) = N_0/2$. Moreover, there are no mutations, that is individuals cannot switch to the other phenotype between divisions, and at division two cells of the same phenotype as the mother are produced.

We show in appendix C that the strength of selection is given by:

$$\Pi_{\boldsymbol{\mathcal{S}}} = \frac{r_b e^{t(r_b-\gamma_b)} + r_a e^{t(r_a-\gamma_a)}}{e^{t(r_b-\gamma_b)} + e^{t(r_a-\gamma_a)}} - \frac{r_b e^{-\gamma_b t} + r_a e^{-\gamma_a t}}{e^{-\gamma_b t} + e^{-\gamma_a t}} \,. \tag{3.68}$$

In the absence of death ($\gamma_a = \gamma_b = 0$), the long time limit of the first fraction is dominated by the phenotype that divides faster. Without loss of generality we consider that $r_a > r_b$, so that the strength of selection is given by:

$$\Pi^{\circ}_{\boldsymbol{\mathcal{S}}} \underset{t\to\infty}{\to} \frac{r_a - r_b}{2} = \frac{\Delta r}{2}\,. \tag{3.69}$$

Note that in this case, $\Pi^{\circ}_{\boldsymbol{\mathcal{S}}}$ is proportional to the selection coefficient in population genetics, which is defined as the difference in reproduction rates $\Delta r = r_a - r_b$.

With death, the asymptotic behavior of the first fraction is controlled by the phenotype that have the largest net reproductive rate, resulting from the balance between division and death, while the second fraction is controlled by the phenotype that dies the slowest. We still suppose that $r_a > r_b$, then the possible outcomes for $\Delta\Pi_{\boldsymbol{\mathcal{S}}}$ are:

$$\Delta\Pi_{\boldsymbol{\mathcal{S}}} \underset{t\to\infty}{\to} \begin{cases} \Delta r/2 & \text{if } \gamma_a > \gamma_b \text{ and } r_a - \gamma_a > r_b - \gamma_b \\ -\Delta r/2 & \text{if } \gamma_a > \gamma_b \text{ and } r_b - \gamma_b > r_a - \gamma_a \\ -\Delta r/2 & \text{if } \gamma_b > \gamma_a \\ 0 & \text{if } \gamma_a = \gamma_b \\ 0 & \text{if } r_a - \gamma_a = r_b - \gamma_b\,. \end{cases} \tag{3.70}$$

This simple two-state model provides cases where death increases selection, decreases selection, or has no effect, corresponding respectively to $\Delta\Pi_{\boldsymbol{\mathcal{S}}} > 0$, $\Delta\Pi_{\boldsymbol{\mathcal{S}}} < 0$ and $\Delta\Pi_{\boldsymbol{\mathcal{S}}} = 0$.

Death has no effect on selection in two situations. First, when death is uniform: $\gamma_a = \gamma_b$ (then each covariance in eq. (3.66) is null), and second, when $r_a - \gamma_a = r_b - \gamma_b$ (which implies that $\gamma_a > \gamma_b$) meaning that death 'shifts' the forward and backward distributions equally, so that their distance remains constant. The last situation corresponds to the equality between the two terms in the right hand side of eq. (3.66), each representing the correlations between death and reproductive success in one sampling. Indeed, it is easy to show:

$$\frac{\mathrm{Cov}^{\circ}_{\mathrm{back}}\left(h^{\circ}_t, p_{\mathrm{surv}}\right)}{\langle p_{\mathrm{surv}}\rangle_{\mathrm{back}^{\circ}}} = \frac{\mathrm{Cov}^{\circ}_{\mathrm{for}}\left(h^{\circ}_t, p_{\mathrm{surv}}\right)}{\langle p_{\mathrm{surv}}\rangle_{\mathrm{for}^{\circ}}} \underset{t\to\infty}{\to} -\frac{\Delta r}{2} \tag{3.71}$$

when $r_a - \gamma_a = r_b - \gamma_b$.

We now associate the trait values $s = 1$ ans $s = 0$ to the phenotypes a and b respectively. Then, to analyze the other cases, we show in appendix C how the distributions themselves are expressed in the long-time limit:

$$p_{\mathrm{back}}(s) \underset{t\to\infty}{\to} \begin{cases} \delta(s) & \text{if } r_b - \gamma_b > r_a - \gamma_a \\ \delta(1-s) & \text{if } r_a - \gamma_a > r_b - \gamma_b \\ \left(\delta(s) + \delta(1-s)\right)/2 & \text{if } r_a - \gamma_a = r_b - \gamma_b \end{cases} \tag{3.72}$$

$$p^{\star}_{\mathrm{for}}(s) \underset{t\to\infty}{\to} \begin{cases} \delta(s) & \text{if } \gamma_a > \gamma_b \\ \delta(1-s) & \text{if } \gamma_b > \gamma_a \\ \left(\delta(s) + \delta(1-s)\right)/2 & \text{if } \gamma_a = \gamma_b \end{cases} \tag{3.73}$$

In the absence of death, $p^{\circ}_{\mathrm{back}}(s) \underset{t\to\infty}{\to} \delta(1-s)$ because cells of type a divide faster and then represent an increasing fraction of the population, and $p^{\circ}_{\mathrm{for}}(s)$ is equal to $1/2$ for

each phenotype for any time t because each subpopulation starts with an initial forward weight 1/2. When introducing death, the backward distribution is controlled by the net reproduction rate which is the difference between the division and death rates, so that both phenotypes can invade the population in the long time limit. When a cell dies, the corresponding subpopulation permanently loses a fraction of its initial forward weight 1/2, therefore the phenotype that dies the slowest is increasingly represented in the forward sampling, up to having a weight 1.

There is only one case where selection is increased by death: when cells that divide faster also die faster, while keeping a larger division-death balance. In that way, phenotype a remains over-represented in the population like in the absence of death, but because it dies faster, its forward weight tends to 0, against 1/2 without death. As a consequence, the distance between the two samplings increases.

On the other hand, if the balance between divisions and death favors phenotype b while maintaining a smaller death rate for b, then phenotype b invades the two statistics and thus there is no difference between forward and backward. Similarly, when phenotype b that divides slower also dies faster, it remains under-represented in the population, and also becomes under-represented in the forward sampling, so that forward and backward distributions are identical. In these two cases, the strength of selection in the presence of death is null, and so the distance between the two samplings is decreased by death.

4.4 A digression: inference of the bulk growth rate from cytometer measurements

Bulk population experiments are challenging to carry out, and steady-state exponential growth is particularly difficult to reach. On the other hand, the dynamics cytometer offers lineages with many generations. A natural question is then: how to use these finite population measurements to infer the steady-state population growth rate? If the hypotheses of section 4.1 are valid for this kind of experimental setup, then we can answer this question using eq. (3.56). The assumptions involve in particular that the dilution rate γ_t can be expressed as a function of a vector χ of variables that would behave similarly in a free growth experiment. In particular, the confined geometry of the setup should have no consequence on the rate at which cells grow. When this is true, we multiply eq. (3.56) by the inverse $p_{\text{surv}}(\boldsymbol{\chi})^{-1}$ of the survival probability and integrate over all trajectories to obtain

$$N_t^\circ = N_t \, \langle p_{\text{surv}}(\boldsymbol{\chi})^{-1} \rangle_{\text{back}} \,, \tag{3.74}$$

which is an alternative form of eq. (3.60). The population growth rate is then given by

$$\Lambda_t^\circ = \Lambda_t + \frac{1}{t} \ln\left(\frac{N_0}{N_0^\circ}\right) + \frac{1}{t} \ln\langle p_{\text{surv}}(\boldsymbol{\chi})^{-1} \rangle_{\text{back}} \,, \tag{3.75}$$

where N_0 and N_0° are the initial numbers of cells in each setup. In confined geometries the number of cells allowed in the chamber is bounded by the chamber capacity $N_{\max}$, so that the population growth rate is vanishing in the long time limit: $\Lambda_t \leq \ln(N_{\max}/N_0)/t \to 0$ when $t \to \infty$. Moreover, the initial numbers of cells N_0 and N_0° are fixed, so the second

term in the right hand side also vanishes in the long time limit. Finally, the long time population growth rate in free growth is given by

$$\Lambda^\circ = \lim_{t\to\infty} \frac{1}{t} \ln\langle p_{\mathrm{surv}}(\boldsymbol{\chi})^{-1}\rangle_{\mathrm{back}} \,. \tag{3.76}$$

The dilution-less population growth rate is related to the backward average of the survival probability inverse. Indeed, paths with a probability to disappear which is null have a weight one in the average: we know that all lineages following these paths are observed so no bias is required. On the other hand, a path with a low probability of survival is highly weighted to compensate the loss of the many lineages following the same path which are not observed because of dilution.

In practice, inferring the population growth rate from cytometer measurements is challenging for two reasons. First, the backward average in eq. (3.76) is computed with the empirical backward distribution:

$$\Lambda^\circ = \lim_{t\to\infty} \frac{1}{t} \ln \left[\frac{1}{n(\sigma=1)} \sum_{i=1}^{n(\sigma=1)} \exp\left[\int_0^t \mathrm{d}t' \gamma_{t'}(\chi_i(t')) \right] \right] . \tag{3.77}$$

For this empirical distribution to accurately reproduce the backward distribution, that is the solution of the partial differential equation, a large number of living lineages is required. However the number of cells in these setups is typically small, of the order of $20 \sim 40$, thus rare lineages are not be observed in general, although they are the lineages with the largest weights in the average. Second, one need to identify the relevant variables χ which control the dilution rate.

Both these difficulties are resolved when supposing that dilution is uniform, that is when γ_t does not depend on $\boldsymbol{\chi}$. In this case eq. (3.76) reduces to

$$\Lambda^\circ = \lim_{t\to\infty} \frac{1}{t} \int_0^t \mathrm{d}t' \gamma_{t'} \,. \tag{3.78}$$

The integral term in the right hand side is simply evaluated by counting the number of cells that are evacuated by the medium. This is how the population growth rate is estimated for example in Hashimoto et al., 2016 (actually they count the number of cells born in a time interval, which is equivalent to the number of diluted cells in a constant-population setup). The question of the validity of the uniform dilution in the dynamics cytometer is still to be clarified, and work should be done to investigate the potential variables χ controlling dilution.

5 Conclusion

The development of experiments on finite populations in confined geometry created a new kind of cellular data, which takes the form of sparse population trees where most lineages end before the end of the experiment because cells are evacuated from the chamber in order to balance divisions. Given the importance of these experiments which allow a precise control of the conditions everywhere inside the chamber and where single cells can

be tracked with accuracy, it is important to understand how this kind of population tree should be sampled, and how selection should be characterized.

In this chapter, we showed how to adapt the forward and backward samplings of the lineages proposed in Nozoe et al., 2017. Unlike the simple case treated in chapter 2 where all lineages are taken into account in each sampling, here dead lineages are sampled only in the forward manner, given that they do not appear in the population at final time. This difference leads to modified versions of the results from chapter 2, where the fluctuation theorem on the number of divisions is recast for living lineages only, and involves a normalizing term Γ_t accounting for the forward weight of the surviving lineages (or dead lineages, the sum of the two being 1). The consequences of the fluctuation theorem such as the inequalities between population growth rate and average numbers of divisions, and Powell's relation for age models are also generalized by simply replacing Λ_t by $\Lambda_t - \Gamma_t$.

Even though our inspiration came from datasets obtained in constant population experiments, the results in this chapter are more general, and apply to any branching tree where some lineages do not survive up to the end. In particular, the population can remain constant but also increase or decrease, depending on the sign of $\Lambda_t - \Gamma_t$. Moreover, the cause of the death is not specified so that the death rate can be a function of any cell trait. This is a progress compared to the majority of articles on the subject where dilution was considered uniform and equal to the population growth rate. Consequently, this framework could be useful in other situations, like for in vitro population experiments on antibiotic resistance (Lambert et al., 2015), in which case death means biological death of the cell, for in vivo time-lapse experiments, or in evolutionary context where species have become extinct (Stadler, 2013).

Note, however, that this framework suffers from the same limitation as in the case without death, namely the genealogy is needed to build the forward statistics. When this information is not accessible only the backward sampling can be performed, and work needs to be done to define selection and infer the past history of the population from snapshot data only. Moreover, relations involving a comparison between the two samplings are valid only for surviving lineages, and thus do not take advantage of the large amount of data from dead lineages. These data can be exploited with the tree sampling (Levien et al., 2020), and linking the tree and forward/backward samplings would be a logical next step.

The strength of selection, as defined in the previous chapter, is a measure of the distance between the forward and backward distributions for a particular cell trait. In the presence of death, these two distributions and thus the strength of selection can be perturbed by a survivor bias, which reflects the variability in the survival probabilities. The overall selection then results from the correlations between the intrinsic selection in the absence of death and the survivor bias. We propose a general measure of the effect of death on selection whose sign indicates if death tends to increase or reduce the difference between the forward and backward statistics.

Finally, the survivor bias implies that the surviving population may not be representative of the whole population, because some phenotypes may be more likely to die than others. As a consequence, the growth rate of the surviving population might be differ-

ent from the growth rate of the whole population. In the context of finite populations where death means dilution, this suggests the following effect: if cells that are diluted and cells staying inside the chambers do not have the same distribution of division rates, then one cannot infer the population growth rate by simply counting the number of cells born/evacuated. Even though the formula we proposed cannot be used in practice until the variables that control the dilution rate are identified, which could be difficult to investigate, we hope that it will encourage people to test the validity of the uniform dilution hypothesis. The analysis carried out in Hashimoto et al., 2016 (figure S3) suggests that the distribution of generation times is independent of the position inside the chamber, in particular at the ends of the channel where cells are about to be expelled; which supports the hypothesis of uniform dilution. However, this analysis relied on the assumption of the timer mechanism, and results could be different for other species following other mechanisms of cell size control.

6 Appendices

A Powell's relation and Euler-Lotka equation for age models with correlations and age-dependent death rate

In section 4.1.1 of chapter 1, we presented two different ways to model mother-daughter correlations in the generation times for age-controlled populations. The first one relies on an intermediate variable, representing a cell type, which is transmitted at division with correlations, and on which the division and death rates depend. The generalization of Powell's relation for this model in presence of an age-and-type-dependent death rate is treated in section 3 of the main text. In this appendix, we present the equivalent derivation for the second model, where generation time τ is considered as a variable and inherited at division via kernel $\Sigma(\tau|\tau')$, normalized as $\forall\tau',\ \int \mathrm{d}\tau\Sigma(\tau|\tau') = 1$. The population balance equation reads (Lebowitz et al., 1974):

$$\partial_t n(a,\tau,t) = -\partial_a n(a,\tau,t) - \gamma(a,\tau)n(a,\tau,t) \qquad \text{for } 0 < a \leq \tau \tag{3.79}$$

$$n(a=0,\tau,t) = m\int \mathrm{d}\tau'\Sigma(\tau|\tau')n(\tau',\tau',t)\,. \tag{3.80}$$

As pointed out in section 4.1.1 of chapter 1, there is no division rate in this model; however there is a death rate $\gamma(a,\tau)$. This implies that even if the newborn cell is 'programmed' via the kernel Σ, at the moment of the division of its mother, to divide after a time τ, it will actually survive until reaching age τ with probability $\exp\left[-\int_0^\tau \mathrm{d}a\ \gamma(a,\tau)\right]$. Note that we allow the death rate to be a function of the age τ at which the cell is programmed to divide, unlike the model proposed in Lebowitz et al., 1974 where it is only a function of the age a of the cell.

This population balance equation and its boundary condition are recast at the levels of the backward probability and forward probability conditioned on survival:

$$\partial_t p_{\text{back}}(a,\tau,t) = -\partial_a p_{\text{back}}(a,\tau,t) - \left[\gamma(a,\tau) + \Lambda_p(t)\right] p_{\text{back}}(a,\tau,t) \quad \text{for } 0 < a \leq \tau \tag{3.81}$$

$$p_{\text{back}}(a=0,\tau,t) = m\int \mathrm{d}\tau'\Sigma(\tau|\tau')p_{\text{back}}(\tau',\tau',t) \tag{3.82}$$

$$\partial_t p^\star_{\text{for}}(a,\tau,t) = -\partial_a p^\star_{\text{for}}(a,\tau,t) - \left[\gamma(a,\tau) + \Gamma_p(t)\right] p^\star_{\text{for}}(a,\tau,t) \qquad \text{for } 0 < a \leq \tau \tag{3.83}$$

$$p^\star_{\text{for}}(a=0,\tau,t) = \int \mathrm{d}\tau'\Sigma(\tau|\tau')p^\star_{\text{for}}(\tau',\tau',t)\,, \tag{3.84}$$

and the steady state solutions to these two equations read

$$p_{\text{back}}(a,\tau) = \begin{cases} p_{\text{back}}(0,\tau)\exp\left[-\Lambda a - \int_0^a \mathrm{d}a'\ \gamma(a',\tau)\right] & \text{for } 0 \leq a \leq \tau \\ 0 & \text{for } a > \tau \end{cases} \tag{3.85}$$

$$p^\star_{\text{for}}(a,\tau) = \begin{cases} p^\star_{\text{for}}(0,\tau)\exp\left[-\Gamma a - \int_0^a \mathrm{d}a'\ \gamma(a',\tau)\right] & \text{for } 0 \leq a \leq \tau \\ 0 & \text{for } a > \tau\,. \end{cases} \tag{3.86}$$

Similarly to the case treated in the main text, we define the steady-state distribution of generation times as the ratio of the number of cells dividing at age τ to the total number

of cells dividing at the same instant, weighted in the two samplings:

$$f(\tau) = \frac{p(\tau,\tau)}{\int \mathrm{d}\tau'\ p(\tau',\tau')}\,. \tag{3.87}$$

We also define the newborn distribution of generation times:

$$\rho^{\mathrm{nb}}(\tau) = \frac{p(0,\tau)}{\int \mathrm{d}\tau'\ p(0,\tau')}\,, \tag{3.88}$$

which represents the proportion of newborn cells that are programmed to divide after a time τ among all the newborn cells, in both samplings. Combining the above definitions, the distributions of generation time are given by:

$$f_{\mathrm{back}}(\tau) = m\rho^{\mathrm{nb}}_{\mathrm{back}}(\tau)\exp\left[-\Lambda\tau - \int_0^\tau \mathrm{d}a\ \gamma(a,\tau)\right] \tag{3.89}$$

$$f^\star_{\mathrm{for}}(\tau) = \rho^{\mathrm{nb}}_{\mathrm{for}}(\tau)\exp\left[-\Gamma\tau - \int_0^\tau \mathrm{d}a\ \gamma(a,\tau)\right]\,, \tag{3.90}$$

and the generalized version of Powell's relation reads

$$f_{\mathrm{back}}(\tau) = m\frac{\rho^{\mathrm{nb}}_{\mathrm{back}}(\tau)}{\rho^{\mathrm{nb}}_{\mathrm{for}}(\tau)} f^\star_{\mathrm{for}}(\tau)e^{-(\Lambda-\Gamma)\tau}\,. \tag{3.91}$$

Once again, this bias does not depend explicitly on the death rate γ but instead on the decrease rate of the forward probability of survival Γ.

When integrating this relation over τ, we obtain a generalization of Euler-Lotka equation. Since four probability distributions appear in this version of Powell's relation, each of them can be isolated and integrated out which leads to four different Euler-Lotka equations. They all link the modified population growth rate $\Lambda - \Gamma$ to the combined variabilities in generation time of the three remaining distributions. A different set of Euler-Lotka equations can be obtained by plugging the solutions eqs. (3.85) and (3.86) in the boundary conditions eqs. (3.82) and (3.84) and integrating over τ:

$$1 = m\int_0^\infty \mathrm{d}\tau\ \exp\left[-\Lambda\tau - \int_0^\tau \mathrm{d}a\ \gamma(a,\tau)\right]\rho^{\mathrm{nb}}_{\mathrm{back}}(\tau) \tag{3.92}$$

$$1 = \int_0^\infty \mathrm{d}\tau\ \exp\left[-\Gamma\tau - \int_0^\tau \mathrm{d}a\ \gamma(a,\tau)\right]\rho^{\mathrm{nb}}_{\mathrm{for}}(\tau)\,, \tag{3.93}$$

where we recover with eq. (3.92) a result originally derived in Lebowitz et al., 1974. These expressions have the advantage of being simpler because they involve only one probability distribution: the newborn distribution of generation times, but they involve the death rate γ explicitly.

B Measure of the effect of death on the strength of selection

In this appendix, we prove the covariance formula eq. (3.66). We start from the definition of the strength of selection in the presence of death:

$$\Pi_{\boldsymbol{s}} = \frac{1}{t}\int \mathcal{D}\boldsymbol{s}\ \left(p_{\mathrm{back}}(\boldsymbol{s},t) - p^\star_{\mathrm{for}}(\boldsymbol{s},t)\right)\ln\left(\frac{p_{\mathrm{back}}(\boldsymbol{s},t)}{p^\star_{\mathrm{for}}(\boldsymbol{s},t)}\right)\,, \tag{3.94}$$

where the distributions inside the logarithm can be expressed using eqs. (3.59) and (3.61). Doing so, the survival probability cancel in the numerator and the denominator:

$$\Pi_{\boldsymbol{\mathcal{S}}} = \frac{1}{t}\int \mathcal{D}\boldsymbol{s}\ \left(p_{\text{back}}(\boldsymbol{s},t) - p^{\star}_{\text{for}}(\boldsymbol{s},t)\right)\ln\left(\frac{p^{\circ}_{\text{back}}(\boldsymbol{s},t)}{p^{\circ}_{\text{for}}(\boldsymbol{s},t)}\right) \tag{3.95}$$

$$= \int \mathcal{D}\boldsymbol{s}\ h^{\circ}_t(\boldsymbol{s})\left(p_{\text{back}}(\boldsymbol{s},t) - p^{\star}_{\text{for}}(\boldsymbol{s},t)\right)\ . \tag{3.96}$$

Then, we compute $\Delta\Pi_{\boldsymbol{\mathcal{S}}}$ by subtracting $\Pi^{\circ}_{\boldsymbol{\mathcal{S}}} = \int \mathcal{D}\boldsymbol{s}\ h^{\circ}_t(\boldsymbol{s})\left(p^{\circ}_{\text{back}}(\boldsymbol{s},t) - p^{\circ}_{\text{for}}(\boldsymbol{s},t)\right)$ from the above relation:

$$\Delta\Pi_{\boldsymbol{\mathcal{S}}} = \int \mathcal{D}\boldsymbol{s}\ h^{\circ}_t(\boldsymbol{s})\left[\left(p_{\text{back}}(\boldsymbol{s},t) - p^{\circ}_{\text{back}}(\boldsymbol{s},t)\right) - \left(p^{\star}_{\text{for}}(\boldsymbol{s},t) - p^{\circ}_{\text{for}}(\boldsymbol{s},t)\right)\right] \tag{3.97}$$

$$= \frac{1}{\langle p_{\text{surv}}(\boldsymbol{\chi})\rangle_{\text{back}^{\circ}}}\int \mathcal{D}\boldsymbol{s}\ h^{\circ}_t(\boldsymbol{s})p^{\circ}_{\text{back}}(\boldsymbol{s},t)\left[p_{\text{surv}}(\boldsymbol{\chi}) - \langle p_{\text{surv}}(\boldsymbol{\chi})\rangle_{\text{back}^{\circ}}\right]$$

$$- \frac{1}{\langle p_{\text{surv}}(\boldsymbol{\chi})\rangle_{\text{for}^{\circ}}}\int \mathcal{D}\boldsymbol{s}\ h^{\circ}_t(\boldsymbol{s})p^{\circ}_{\text{for}}(\boldsymbol{s},t)\left[p_{\text{surv}}(\boldsymbol{\chi}) - \langle p_{\text{surv}}(\boldsymbol{\chi})\rangle_{\text{for}^{\circ}}\right] \tag{3.98}$$

$$= \frac{\text{Cov}_{\text{back}^{\circ}}\left(h^{\circ}_t, p_{\text{surv}}\right)}{\langle p_{\text{surv}}\rangle_{\text{back}^{\circ}}} - \frac{\text{Cov}_{\text{for}^{\circ}}\left(h^{\circ}_t, p_{\text{surv}}\right)}{\langle p_{\text{surv}}\rangle_{\text{for}^{\circ}}}\ . \tag{3.99}$$

C Simple two-state example

In this appendix, we give the detailed analysis of the two-state model introduced in section 4.3. The number of cells in the subpopulation a evolves as $n(a,t) = N_0\exp\left[t(r_a-\gamma_a)\right]/2$ and similarly for the subpopulation b. The total number of cells is given by the sum of these two subpopulations, so that the backward probability reads

$$p_{\text{back}}(s,t) = \frac{e^{-t\Lambda_t}}{2}\left[e^{t(r_a-\gamma_a)}\delta(1-s) + e^{t(r_b-\gamma_b)}\delta(s)\right] \tag{3.100}$$

$$= \frac{e^{t(r_a-\gamma_a)}\delta(1-s) + e^{t(r_b-\gamma_b)}\delta(s)}{e^{t(r_a-\gamma_a)} + e^{t(r_b-\gamma_b)}}\ . \tag{3.101}$$

In the dynamics without death, since the initial distribution of phenotypes is even and since there is no phenotype switching, then for any time t:

$$p^{\circ}_{\text{for}}(s,t) = \frac{\delta(s) + \delta(1-s)}{2}\ . \tag{3.102}$$

To obtain the forward probability with death, we use transformation eq. (3.59):

$$p^{\star}_{\text{for}}(s,t) = \frac{e^{-t\Gamma_t}}{2}\left(\delta(s)e^{-t\gamma_b} + \delta(1-s)e^{-t\gamma_a}\right) \tag{3.103}$$

$$= \frac{\delta(s)e^{-t\gamma_b} + \delta(1-s)e^{-t\gamma_a}}{e^{-t\gamma_b} + e^{-t\gamma_a}}\ . \tag{3.104}$$

The fitness landscape $h_t(s)$, defined by eq. (3.62), is then obtained by computing the ratio of the above distributions:

$$h_t(s) = r_b\delta(s) + r_a\delta(1-s)\ . \tag{3.105}$$

We recover that in this situation the fitness landscape is equal to the historical fitness.

Finally, the strength of selection is given by:

$$\Pi_{\mathcal{S}} = \langle h_t \rangle_{\text{back}} - \langle h_t \rangle_{\text{for}^\star} \tag{3.106}$$

$$= \frac{r_b e^{t(r_b-\gamma_b)} + r_a e^{t(r_a-\gamma_a)}}{e^{t(r_b-\gamma_b)} + e^{t(r_a-\gamma_a)}} - \frac{r_b e^{-\gamma_b t} + r_a e^{-\gamma_a t}}{e^{-\gamma_b t} + e^{-\gamma_a t}} \,. \tag{3.107}$$

The behaviors of $p_{\text{back}}(s,t)$, $p^\star_{\text{for}}(s,t)$ and $\Pi_{\mathcal{S}}$ in the long time limit are directly obtained from the above equations.

Bibliography for Chapter 3

[Genthon et al., 2022] Genthon, A., L. Peliti, T. Nozoe, and D. Lacoste (2022). "Sampling lineage trees with death".

[Hashimoto et al., 2016] Hashimoto, M., T. Nozoe, H. Nakaoka, R. Okura, S. Akiyoshi, K. Kaneko, E. Kussell, and Y. Wakamoto (2016). Noise-driven growth rate gain in clonal cellular populations. *Proceedings of the National Academy of Sciences* 113.(12), pp. 3251–3256.

[Koldaeva et al., 2022] Koldaeva, A., H.-F. Tsai, A. Q. Shen, and S. Pigolotti (2022). Population genetics in microchannels. *Proceedings of the National Academy of Sciences* 119.(12), e2120821119.

[Lambert et al., 2015] Lambert, G. and E. Kussell (2015). Quantifying Selective Pressures Driving Bacterial Evolution Using Lineage Analysis. *Physical Review X* 5.(1), p. 011016.

[Lebowitz et al., 1974] Lebowitz, J. L. and S. I. Rubinow (1974). A theory for the age and generation time distribution of a microbial population. *Journal of Mathematical Biology* 1.(1), pp. 17–36.

[Levien et al., 2020] Levien, E., J. Kondev, and A. Amir (2020). The interplay of phenotypic variability and fitness in finite microbial populations. *Journal of The Royal Society Interface* 17.(166), p. 20190827.

[Lindner et al., 2008] Lindner, A. B., R. Madden, A. Demarez, E. J. Stewart, and F. Taddei (2008). Asymmetric segregation of protein aggregates is associated with cellular aging and rejuvenation. *Proceedings of the National Academy of Sciences* 105.(8), pp. 3076–3081.

[Maisonneuve et al., 2008] Maisonneuve, E., B. Ezraty, and S. Dukan (2008). Protein Aggregates: an Aging Factor Involved in Cell Death. *Journal of Bacteriology* 190.(18), pp. 6070–6075.

[Moran, 1958] Moran, P. A. P. (1958). Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society* 54.(1), pp. 60–71.

[Murashita et al., 2014] Murashita, Y., K. Funo, and M. Ueda (2014). Nonequilibrium equalities in absolutely irreversible processes. *Physical Review E* 90.(4), p. 042110.

[Nakashima et al., 2020] Nakashima, S., Y. Sughiyama, and T. J. Kobayashi (2020). Lineage EM algorithm for inferring latent states from cellular lineage trees. *Bioinformatics* 36.(9), pp. 2829–2838.

[Nozoe et al., 2017] Nozoe, T., E. Kussell, and Y. Wakamoto (2017). Inferring fitness landscapes and selection on phenotypic states from single-cell genealogical data. *PLoS Genetics* 13.(3), e1006653.

[Pigolotti, 2021] Pigolotti, S. (2021). Generalized Euler-Lotka equation for correlated cell divisions. *Physical Review E* 103.(6), p. L060402.

[Powell, 1956] Powell, E. O. (1956). Growth Rate and Generation Time of Bacteria, with Special Reference to Continuous Culture. *Journal of General Microbiology* 15.(3), pp. 492–511.

[Powell, 1964] Powell, E. O. (1964). A Note on Koch & Schaechter's Hypothesis about Growth and Fission of Bacteria. *Journal of General Microbiology* 37.(2), pp. 231–249.

[Powell, 1969] Powell, E. O. (1969). Generation Times of Bacteria: Real and Artificial Distributions. *Journal of General Microbiology* 58.(1), pp. 141–144.

[Schink et al., 2019] Schink, S. J., E. Biselli, C. Ammar, and U. Gerland (2019). Death Rate of E. coli during Starvation Is Set by Maintenance Cost and Biomass Recycling. *Cell Systems* 9.(1), 64–73.e3.

[Stadler, 2013] Stadler, T. (2013). Recovering speciation and extinction dynamics based on phylogenies. *Journal of Evolutionary Biology* 26.(6), pp. 1203–1219.

[Sughiyama et al., 2019] Sughiyama, Y., S. Nakashima, and T. J. Kobayashi (2019). Fitness response relation of a multitype age-structured population dynamics. *Physical Review E* 99.(1), p. 012413.

[Wakamoto et al., 2013] Wakamoto, Y., N. Dhar, R. Chait, K. Schneider, F. Signorino-Gelo, S. Leibler, and J. D. McKinney (2013). Dynamic Persistence of Antibiotic-Stressed Mycobacteria. *Science* 339.(6115), pp. 91–95.

[Wang et al., 2010] Wang, P., L. Robert, J. Pelletier, W. L. Dang, F. Taddei, A. Wright, and S. Jun (2010). Robust Growth of Escherichia coli. *Current Biology* 20.(12), pp. 1099–1103.

Chapter 4

The cell size distribution[†]

[†]This chapter is based on the preprint Genthon, 2022a. Since the defence of this PhD thesis, this preprint has been published as: [Genthon, 2022b] A. Genthon (2022b). Analytical cell size distribution: lineage-population bias and parameter inference. *Journal of The Royal Society Interface* 19.(196), p. 20220405.

Contents

1 Introduction

In chapter 2 we presented a framework based on the forward and backward samplings of the lineages within a population tree, which allowed us to derive general relations between the statistics obtained in population and mother machine setups, and to formulate universal constraints on the strength of selection acting on cell traits correlated with divisions. Using the notion of fitness landscape, we derived eq. (2.51), a very simple bias for the cell size distributions, valid when growth is deterministic and exponential and when volume partitioning is deterministic and symmetric. This relation involves a factor x, the size, that biases the population distribution towards smaller cells, which is intuitive since when there is no noise, more divisions mechanically leads to smaller sizes. However, when introducing noises either in the partitioning of volume at division or in single-cell growth, the formalism of fitness landscape can become difficult to use. An alternative route to investigate this bias and the importance of these sources of stochasticity, that we take in this chapter, is to solve independently the population and lineage equations.

Understanding cell size statistics can be interesting because it offers many insights on the laws of growth and division, and because the size framework also applies to molecular-level quantities. These could be a number of proteins or mRNA, which also grow within the cell cycle and are split between the daughter cells at division. Lineage-population biases for cell size have been recently derived at the level of the first moments of the size distribution (Totis et al., 2021), and of the distribution of size at birth (Thomas, 2017; Thomas, 2018) in some particular cases. However, general understanding of the bias at the level of distributions, illustrated in fig. 4.1, is lacking. In the mathematical literature, the existence of a solution to the growth-fragmentation equation modeling the time evolution of the population cell-size distribution and its convergence have been largely analyzed (Michel, 2006; Doumic Jauffret et al., 2010; Balagué et al., 2013). Moreover, under the assumptions of deterministic partitioning of volume amongst the daughter cells, exponential growth and power-law division rate, the growth-fragmentation equation reduces to a pantograph equation, whose solution is analytical (Hall et al., 1990; Zaidi et al., 2021). Surprisingly, these analyses have not been applied to lineage statistics.

In chapter 2, the fluctuations in the number of divisions were taken advantage of in order to infer the population growth rate form mother machine data. Another possible use of such data has recently been explored by Jia et al., who derived steady-state size distributions for lineage data and used them to infer single-cell parameters describing the laws of cell growth and division, both for bacteria in exponential growth (Jia et al., 2021) and yeasts (Jia et al., 2022). Along the same line, in this chapter we derive analytical cell-size lineage distributions, different from those obtained in Jia et al., 2021 by allowing growth laws that are more general than the exponential growth, and use them to fit mother machine data and thus: (i) check the validity of the model and the assumptions to describe the data, and (ii) infer cell cycle parameters.

This chapter is organized as follows: after introducing the hypotheses of the size-control model in section 2, we derive in section 3 exact steady-state lineage cell-size distributions, in the case of deterministic volume partitioning, both symmetric and asymmetric. The lineage-population bias is then obtained by comparison with population

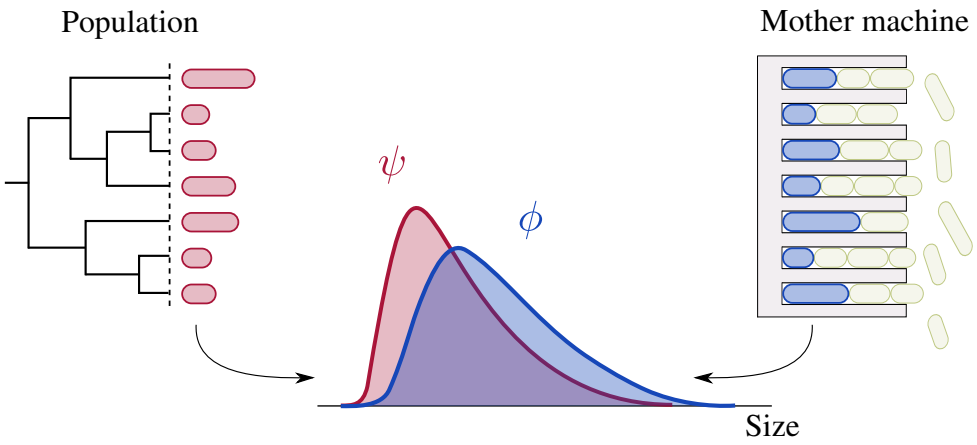


Figure 4.1: Snapshot cell-size distributions for two different experimental setups. The population distribution ψ (red) is computed by uniformly sampling cell sizes in a freely growing population in a batch culture, and the lineage distribution ϕ (blue) is obtained by uniformly sampling the sizes of the constant number of mother cells (in blue) at the bottoms of each micro-channel in a mother machine device.

distributions from the mathematical literature. We also show that these lineage distributions can account for experimental data on E. coli, and illustrate how they can be used for parameters inference. When introducing stochasticity in volume partitioning, we seek in section 4 large and small size asymptotic lineage distributions, and show that only a subpart of the model parameters control these tails, and thus the lineage-population in these limits. In addition, we show that these asymptotic behaviors hold when considering the more realistic adder model. In section 5, we introduce in-cycle noise around exponential growth, and show how it affects the large-size behavior and the lineage-population bias. Finally, we comment in section 6 on the effect of maintaining the population constant on the size distributions in simple cases.

2 Preliminaries

2.1 Model and definitions

We consider size-regulated populations, for which the number $n(x,t)$ of cells of size x at time t follows the population balance equation eq. (1.49). We restrict ourselves to the class of homogeneous kernels Σ, which only depends on the ratio of volume between the daughter and mother cells:

$$\Sigma(x|y) = \frac{1}{y} b\left(\frac{x}{y}\right) , \tag{4.1}$$

and we call $b(x)$ the partition kernel. The partition kernel is normalized as $\int_0^1 \mathrm{d}x\, b(x) = 1$, and the conservation of volume at division imposes that $m \int_0^1 \mathrm{d}x\, xb(x) = 1$. Moreover, we forbid births of cells of size 0 by setting $b(0) = b(1) = 0$. The kernel b is very general, and accounts for *deterministic symmetric partition* observed in bacteria and fission yeast:

$b(x) = \delta(x-1/m)$; *deterministic asymmetric partition* characterizing for example budding yeast: $b(x) = \sum_{i=1}^{m} \delta(x - 1/\omega_i)/m$ with $\omega_i > 1$ and $\sum_{i=1}^{m} 1/\omega_i = 1$; and *stochastic partition* which can be modeled for example as a Beta distribution for size or as a Binomial distribution for protein segregation.

For better readability, in this chapter we call ϕ and ψ the forward and backward probability distributions:

$$\phi(x,t) = p_{\text{for}}(x,t) \tag{4.2}$$
$$\psi(x,t) = p_{\text{back}}(x,t)\,, \tag{4.3}$$

and recall the population balance equations at the probability level with these new notations:

$$\partial_t \phi(x,t) = -\partial_x[\nu(x)\phi(x,t)] - r(x)\phi(x,t) + \int \frac{\mathrm{d}x'}{x'} b(x/x')r(x')\phi(x',t) \tag{4.4}$$
$$\partial_t \psi(x,t) = -\partial_x[\nu(x)\psi(x,t)] - [r(x) + \Lambda_p(t)]\,\psi(x,t) + m\int \frac{\mathrm{d}x'}{x'} b(x/x')r(x')\psi(x',t)\,. \tag{4.5}$$

In section 4.2 of chapter 1, we argued that a power law was a good approximation for the division rate, as indicated by the inference from experimental data. Moreover, a power law for the growth rate includes the most common growth strategies, which we call *linear growth* for $\beta = 0$, and *exponential growth* characterizing most bacteria (Taheri-Araghi et al., 2015) for $\beta = 1$. Since in sections 4 and 5 we will investigate the behavior of the tails of the size distributions in the small and large size limits, we make the power-law assumptions in these limits:

$$r(x) \sim r_0 x^{\alpha_0} \quad \text{as } x \to 0 \tag{4.6}$$
$$r(x) \sim r_\infty x^{\alpha_\infty} \quad \text{as } x \to \infty \tag{4.7}$$
$$\nu(x) \sim \nu_0 x^{\beta_0} \quad \text{as } x \to 0 \tag{4.8}$$
$$\nu(x) \sim \nu_\infty x^{\beta_\infty} \quad \text{as } x \to \infty \tag{4.9}$$
$$b(x) \sim b_0 x^{\kappa_0} \quad \text{as } x \to 0\,. \tag{4.10}$$

The last line accounts for a broad class of kernels defined on $[0,1]$, including the Beta distribution commonly used for volume partitioning (Jia et al., 2021). These different exponents are not independent: the population grows exponentially with a rate $\Lambda = \lim_{t\to\infty} \Lambda_p(t)$ in the long-time limit and reaches steady-state size distributions only if certain conditions are met (Balagué et al., 2013), among which are:

$$\alpha_0 - \beta_0 + 1 > 0 \tag{4.11}$$
$$\alpha_\infty - \beta_\infty + 1 > 0 \tag{4.12}$$
$$\kappa_0 - \beta_0 > 0\,. \tag{4.13}$$

The first two lines reflect the necessary balance between growth and division: eq. (4.11) ensures that there is more growth than division for small cells, to avoid the creation of

cells of vanishing sizes; and eq. (4.12) guarantees that there is more division than growth for large cells, to prevent the survival of cells of diverging sizes. Additionally, eq. (4.13) imposes that there is enough growth to counterbalance the birth of cells with vanishing volumes. In sections 4 and 5, we suppose that these conditions are fulfilled.

Finally, in the next sections we use moments of order k of the distributions b, ψ and ϕ, which are the Mellin transforms of these distributions (up to a constant 1 in the exponent):

$$L_k = \int_0^1 x^k b(x)\mathrm{d}x \tag{4.14}$$

$$M_k = \int_0^\infty x^k \psi(x)\mathrm{d}x \tag{4.15}$$

$$N_k = \int_0^\infty x^k \phi(x)\mathrm{d}x\,. \tag{4.16}$$

2.2 The special case of exponential growth

Cells that grow exponentially with a rate $\nu(x) = \nu x$ (for all x) are characterized by two important properties.

First, for any partitioning kernel b, the steady-state population growth rate matches the single cell growth rate (Hall et al., 1990). This follows from the integration of eq. (4.5) after multiplication by x, and using the mass conservation property of kernel b:

$$\partial_t \langle x\rangle_{\text{back}} = [\nu - \Lambda_p(t)]\,\langle x\rangle_{\text{back}}\,, \tag{4.17}$$

provided that the no-flux boundary conditions $x^2 p_{\text{back}}(x,t) \to 0$ when $x \to 0$ and $x \to +\infty$ are met. Therefore, in steady-state the left hand side is null and

$$\Lambda_p(t) \underset{t\to\infty}{\to} \Lambda = \nu\,. \tag{4.18}$$

Second, the lineage-population bias is analytical for exponentially-growing cells in steady-state. Let us multiply the population equation eq. (4.5) by x, and recast it for the function $q(x) = x\psi(x)$ in steady-state:

$$0 = -\nu x\partial_x q(x) - [r(x)+\nu]\,q(x) + \int \frac{dx'}{x'}\frac{mx}{x'} b(x/x')r(x')q(x')\,. \tag{4.19}$$

We identify the derivative of a product $-\nu x\partial_x q(x) - \nu q(x) = -\nu\partial_x[xq(x)]$:

$$0 = -\nu\partial_x[xq(x)] - r(x)q(x) + \int \frac{dx'}{x'}\frac{mx}{x'} b(x/x')r(x')q(x')\,. \tag{4.20}$$

This equation is eq. (4.4), obeyed by the lineage distribution ϕ with modified partition kernel $\hat{b}(x) = mxb(x)$, which we note $\phi^{mxb(x)}(x)$. Therefore $q(x)^{b(x)}$ is proportional to $\phi^{mxb(x)}$:

$$\phi^{mxb(x)}(x) = Kx\psi^{b(x)}(x)\,, \tag{4.21}$$

with $K = \left(\int_0^\infty \mathrm{d}x\; x\psi^{b(x)}(x)\right)^{-1}$ a normalization constant. Importantly, $\hat{b}$ is a proper kernel, which is normalized as a consequence of the conservation of volume of the original kernel b: $\int_0^1 \mathrm{d}x\;\hat{b}(x) = m\int_0^1 \mathrm{d}x\; xb(x) = 1$.

In the case of deterministic symmetric partitioning, the modified partition kernel $\hat{b}(x) = mxb(x) = mx\delta(x - 1/m) = \delta(x - 1/m) = b(x)$ in the single lineage dynamics is equal to the partition kernel $b(x)$ in the population dynamics and we recover a result from Doumic et al., 2021:

$$\phi(x) = Kx\psi(x)\,. \tag{4.22}$$

Although the above relation and eq. (2.51) have been obtained with different approaches, and although eq. (4.22) is a steady-state statement while eq. (2.51) is a time-dependent relation, in both cases the biases involve a factor x, which comes from the correlations between size and number of divisions. It is clear that if no such correlations were present, the over-representation of lineages with high-reproductive success in a population would not affect the size distribution, so that the lineage and population size distributions would be identical.

From eq. (4.22), it is straightforward to show that the average size is larger in lineages than in populations:

$$\langle x\rangle_{\text{for}} = \langle x\rangle_{\text{bak}} + \frac{\text{Var}_{\text{back}}(x)}{\langle x\rangle_{\text{back}}} \geq \langle x\rangle_{\text{back}}\,. \tag{4.23}$$

3 Exact lineage distributions for deterministic partitioning

Exact population solutions to eq. (4.5) have been obtained in the particular case of exponential growth ($\nu(x) = \nu x$), for deterministic symmetric (Hall et al., 1990) and asymmetric (Zaidi et al., 2021) partitioning, and power law division rates ($r(x) = x^\alpha$). The same method can be adapted to derive the lineage distribution, for which the hypothesis of exponential growth can even be relaxed and replaced by a more general power law growth rate: $\nu(x) = \nu x^\beta$. Note that in order to obtain an analytical solution for all sizes, we need to impose power law division and growth rates with respective exponents α and β for all sizes, with $\alpha - \beta + 1 > 0$. This hypothesis will be relaxed in the next sections concerning the distribution tails, and replaced by eqs. (4.6) to (4.10).

For symmetric partitioning between the m daughter cells, we show in appendix A.1 that the solution reads

$$\phi(x) = \frac{C}{x^\beta}\sum_{k=0}^{\infty} c_k \exp\left[-m^{k(\alpha-\beta+1)}\frac{r}{\nu}\,\frac{x^{\alpha-\beta+1}}{\alpha-\beta+1}\right]\,, \tag{4.24}$$

where the coefficients c_k are given in appendix A.1 and C is a normalization constant. For exponential growth, we find that this lineage distribution is related to the population distribution $\psi(x)$ obtained in Hall et al., 1990 by $\phi(x) = x\psi(x)$, which was expected in the light of section 2.2.

It is worth mentioning that this distribution takes a very simple form in the limit of strong control $\alpha \to +\infty$, where cells divide deterministically when reaching size 1. In this limit, $c_0 = 1$, c_1 tends to -1 and all other c_k tend to 0, such that the lineage size distribution reduces to $\phi(x) = Cx^{-\beta}$ for $x \in [1/m, 1]$ and 0 otherwise. This result is analogous to the one for populations: $\psi(x) \propto x^{-2}$ for $x \in [1/2, 1]$ and 0 otherwise, obtained

for binary fission and exponential growth (Hall et al., 1990; Thomas, 2018). Note also that in this limit, there is no randomness in the dynamics of cell growth and division, and thus the steady-state size distribution is simply the solution to the flux-balance equation $\partial_x[\nu(x)\phi(x,t)] = 0$.

In the case of asymmetric partitioning, for simplicity we choose to focus on binary fission ($m = 2$). The volume of the dividing cell is split unequally between the daughters: one inherits a fraction $1/\omega_1$ of the mother size and the other daughter a fraction $1/\omega_2$, with $\omega_1 > \omega_2 > 1$ and $1/\omega_1 + 1/\omega_2 = 1$. The choice of the protocol to track one of the two daughters is of major importance (Jia et al., 2021). If one chooses to always track the smallest of the two daughters, then the partition kernel is given by: $b(x) = \delta(x - 1/\omega_1)$. This is equivalent to the partition kernel for symmetric partitioning between m daughter cells where m is replaced by ω_1, and the size distribution is therefore given by eq. (4.24), when replacing m by ω_1. On the other hand, in the random tracking protocol, each cell is tracked at division with probability 1/2, so that the partition kernel is given by $b(x) = \delta(x - 1/\omega_1)/2 + \delta(x - 1/\omega_2)/2$. In this case, we show in appendix A.2 that the size distribution reads

$$\phi(x) = \frac{C}{x^\beta} \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} c_{k,l} \exp\left[-\omega_1^{k(\alpha-\beta+1)} \omega_2^{l(\alpha-\beta+1)} \frac{r}{\nu} \frac{x^{\alpha-\beta+1}}{\alpha-\beta+1} \right], \tag{4.25}$$

where the coefficients $c_{k,l}$ are given in appendix A.2 and C is a normalization constant. For exponential growth, we find that this lineage distribution is related to the population distribution $\psi(x)$ obtained in Zaidi et al., 2021 by $\phi^{2xb(x)}(x) = x\psi^{b(x)}(x)$, which we expected in the light of section 2.2.

3.1 Shapes of the theoretical solutions

We numerically investigate the influence of the parameters of the model on the analytical steady-state size distributions eq. (4.24) and eq. (4.25), and show the results on fig. 4.2. The first row corresponds to symmetric partitioning and the second one to asymmetric partitioning. On the top left plot, as the strength of the size control α is increased the distribution gets narrower, and in the limit of large control, division becomes deterministic and $\phi(x) = Cx^{-\beta}$ for $x \in [1/2, 1]$. On the top right plot, the growth rate power β is varied. For $\beta = 0$, ϕ presents a flat maximum and a fast decline for large size. As β increases, the maximum becomes more peaked and the decrease at large size is slowed, which follows from the fact that increasing the growth rate allows cells to reach larger sizes. On the bottom left plot, we vary the volume ratio $1/\omega_1$ of the smallest daughter cell that we follow at each division. The smaller the daughter we follow, the wider the curve on the left hand side. Finally, the bottom right plot corresponds to the random tracking protocol, where $1/\omega_1$ is the ratio of the smallest of the two daughters cells. As the asymmetry is increased, the curve becomes bimodal, intuitively corresponding to the two subpopulations produced by the smaller and larger daughters at each division.

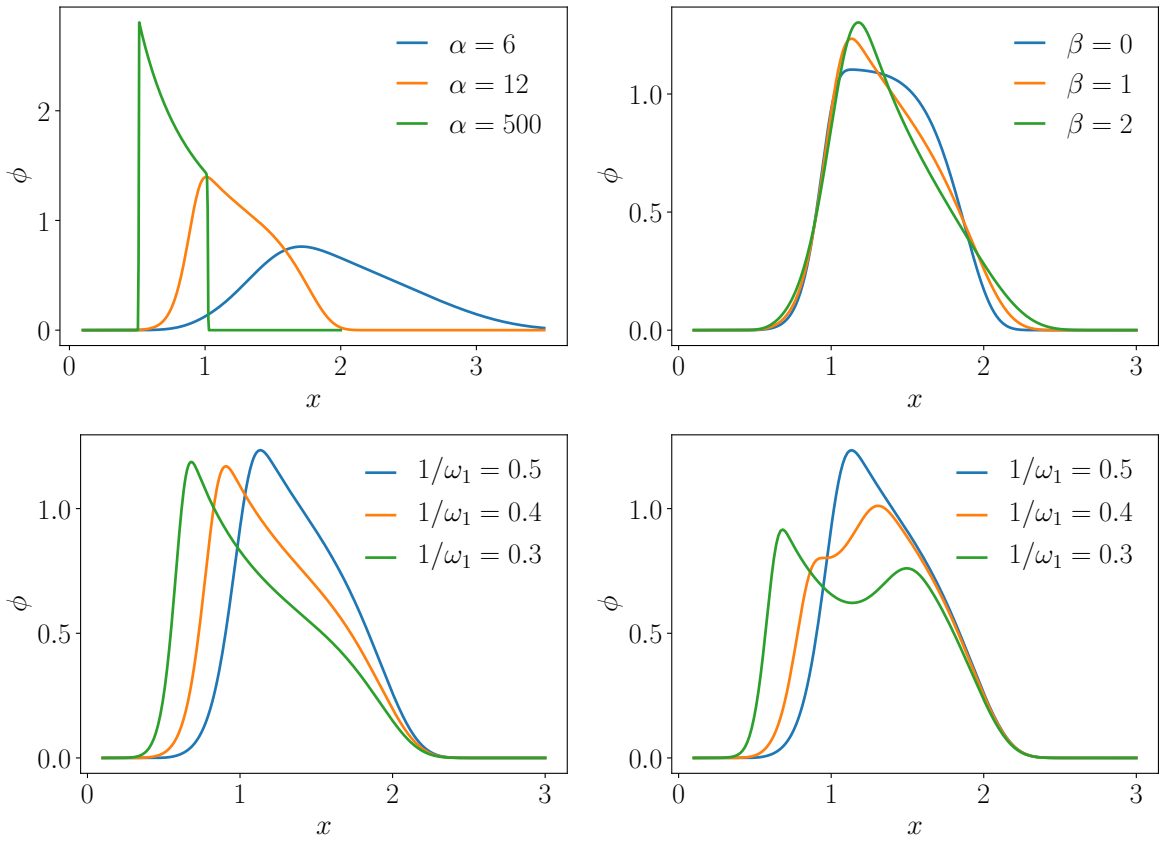


Figure 4.2: Theoretical lineage distributions for binary fission $m = 2$, for symmetric partitioning on the first row, and asymmetric partitioning on the second row. On the first row, the distribution is computed with eq. (4.24), with $\beta = 1$ and α varying on the left, and $\alpha = 10$ and β varying on the right. For asymmetric partitioning, the left plot was generated with the smallest daughter tracking protocol (eq. (4.24) with $m = \omega_1$), and the right plot with the random tracking protocol (eq. (4.25)). For both plots, we fixed $\alpha = 10$ and $\beta = 1$, and varied the asymmetry ω_1. For all four plots we fixed $r/\nu = 0.01$.

3.2 Test on experimental data: parameters inference

In experimental systems, the partitioning is stochastic rather than deterministic, however for E. coli data obtained in mother machine (Tanouchi et al., 2017), the coefficient of variation of the volume ratio at division was found to be smaller than 10% (Jia et al., 2021). This encourages us to test the validity of our theoretical distributions. We use data from Tanouchi et al., 2017, described in section 5 of chapter 1, where the size of many independent cell lineages of E. coli has been recorded every minute over 70 generations at three different temperatures (25 °C, 27 °C, and 37 °C), precisely 65 lineages for 25 °C, 54 for 27 °C, and 160 for 37 °C. We fit the experimental distributions for the three temperatures using the three models: symmetric partitioning, asymmetric partitioning with smaller/larger cell tracking and random tracking. For each temperature, the best fit is shown on fig. 4.3, and the fitting parameters are given in table 4.1.

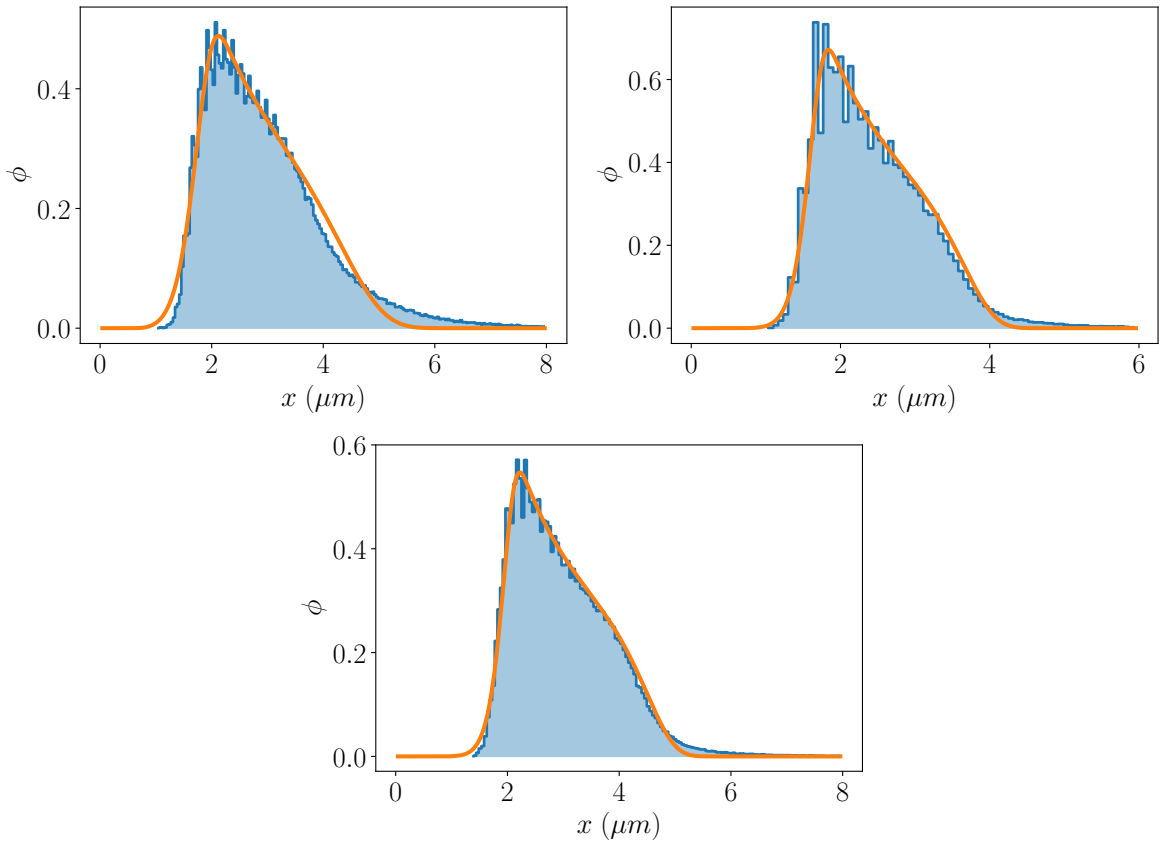


Figure 4.3: Experimental lineage distributions (blue histograms) for E. coli data from Tanouchi et al., 2017 in three temperature conditions: 25 °C (left), 27 °C (right) and 37 °C (bottom). The best fits (orange curves) are computed with eq. (4.24) or eq. (4.25), and the fitting parameters are given in table 4.1.

First of all, we observe that our model is in very good agreement with experiments at 27 °C and 37 °C, and that the fit at 25 °C is average but fails to capture the decay of the right tail. In particular, our results reproduce the three-stages discussed in Jia et al., 2021: fast increase for small cells, slow decay for medium-sized cells, and fast decay for large cells. In the following we analyze the values of the parameters for the condition 27 °C and 37 °C in particular, given than they provide the best fits to experimental data. Surprisingly, for all temperatures the best fit is given by asymmetric partitioning with smallest-daughter tracking protocol, where the daughter cell which is followed inherits a fraction $1/\omega_1 = 0.43$ of the mother volume. This value is really close to the value $0.44 - 0.45$ obtained by direct analysis of the sizes at birth and division along the lineages (Jia et al., 2021). The strength of the control α tends to increase with temperature, in qualitative agreement with what was found in Jia et al., 2021, and the ratio r/ν tends to decrease with temperature. Note that we cannot disentangle the values of r and ν only from the steady-state profile. Finally, the power β in the growth rate is equal to 1.26 for 27 °C at 1.23 for 37 °C, which suggests that in these conditions E. coli grows slightly faster than exponential for large sizes ($x > 1$), and slightly slower than exponential for small

	25 °C	27 °C	37 °C
Tracking protocol	Smallest daughter	Smallest daughter	Smallest daughter
α	7.89	11.49	11.92
β	1.02	1.26	1.23
$1/\omega_1$	0.40	0.43	0.43
r/ν	4.5×10^{-5}	3.5×10^{-6}	1.8×10^{-7}

Table 4.1: Parameters of the best fits to E. coli data from Tanouchi et al., 2017 shown on fig. 4.3. In all cases, the best fit was given by the tracking protocol where partitioning is asymmetric and the smallest cell is always followed, given by eq. (4.24) with $m = \omega_1$.

sizes ($x < 1$). This may be linked to the super-exponential growth observed for E. coli in Kar et al., 2021, where the exponential growth rate ν increases during the cell cycle.

To conclude, in spite of the stochasticity in partitioning present in experimental systems, our model for deterministic partitioning gives a very good description of the data for two temperature conditions, and a correct fit for the last temperature. It captures the complexity of the distributions, and the inferred parameters show the same trends as the ones obtained from the model proposed in Jia et al., 2021, based on a N-step description of the cell cycle. In contrast to this work, our approach allows growth laws that are more general than exponential, and the dependency of the distributions eq. (4.24) and eq. (4.25) on size is more explicit in our model. Also, the present model is simpler in that it involves only one step in the cell cycle. Even though our result produces a fit for the condition 25 °C that is less precise than the one obtained with the N-step model, it performs much better than the N-step model when N is fixed to 1, suggesting that models with N steps may not be minimal.

4 Asymptotic behavior for general partitioning kernel

In this section, we seek large-size and small-size asymptotic solutions to eq. (4.4) for general kernels b. We shall see that the tails of the lineage distribution only depend on the behaviors of the division rate, growth rate and partitioning kernel at large and small sizes, like what happens for the population distribution (Balagué et al., 2013). Therefore, the following results apply to cells obeying more complex growth laws than in the previous section. For example, fission yeasts have been observed to follow piecewise growing patterns (Horváth et al., 2013; Pesti et al., 2021), either bi-linear or bi-exponential during the elongation phase. The bacterium Corynebacterium glutamicum exhibits asymptotically linear growth (Messelink et al., 2021), unlike most bacteria. In these examples, the growth phases are dictated by the age of the cell, but in size-controlled populations, large (resp. small) cells are on average old (resp. young) cells so that the

growth rate at large (resp. small) age is also the growth rate at large (resp. small) size. Moreover, E. coli has also be shown to deviate from exponential growth for large and small sizes (Robert et al., 2014).

In the following, we then suppose that the different rates follow power laws in the small and large size limits, given by eqs. (4.6) to (4.10), and that conditions eqs. (4.11) to (4.13) ensuring the existence of a solution are fulfilled.

4.1 Large size limit

For stochastic partitioning, the large-size population distribution has been derived in Balagué et al., 2013:

$$\psi(x) \underset{x\to\infty}{\sim} \nu(x)^{-1} \exp\left[-\int^x \mathrm{d}y\ \frac{\Lambda + r(y)}{\nu(y)}\right] \tag{4.26}$$

$$\underset{x\to\infty}{\sim} x^{-\beta_\infty} \exp\left[-\frac{r_\infty}{\nu_\infty}\frac{x^{\alpha_\infty-\beta_\infty+1}}{\alpha_\infty-\beta_\infty+1} - \frac{\Lambda}{\nu_\infty}\int^x y^{-\beta_\infty}\mathrm{d}y\right]. \tag{4.27}$$

This result can be understood intuitively as follows: if the distribution is decreasing fast enough in the large-size limit, we can neglect the integral term corresponding to the divisions of larger cells in eq. (4.5), then the resulting equation is exactly solvable and the solution is eq. (4.26), as shown by Friedlander et al., 2008.

We prove in appendix B.1 and appendix C that the lineage distribution reads

$$\phi(x) \underset{x\to\infty}{\sim} x^{-\beta_\infty} \exp\left[-\frac{r_\infty}{\nu_\infty}\ \frac{x^{\alpha_\infty-\beta_\infty+1}}{\alpha_\infty-\beta_\infty+1}\right], \tag{4.28}$$

which is the population distribution eq. (4.26) when setting $\Lambda = 0$. The behavior for large sizes is thus independent of the partition kernel b, which implies in particular that it coincides with the large-size behavior for symmetric partitioning obtained by keeping only the first term the series of eq. (4.24). In order to test this expression, we numerically solve the PBE using a finite difference method with an implicit scheme. Results are shown on fig. 4.4 left for three different values of the strength of size control $\alpha = 2$, 3 and 5. In all three cases, the large-size behavior is in very good agreement with the theory.

For cells growing exponentially in the large-size limit, comparing eqs. (4.27) and (4.28) leads to the following lineage-population bias:

$$x^{\Lambda/\nu_\infty}\psi(x) \underset{x\to\infty}{\sim} \phi(x) \qquad \text{if } \beta_\infty = 1\,. \tag{4.29}$$

If cells grow exponentially with rate $\nu \equiv \nu_\infty$ for all sizes and not only for large-sizes, then we showed in section 2.2 that the population growth rate matches the single cell growth rate $\Lambda = \nu$, so that the lineage-population bias eq. (4.29) for an arbitrary kernel in the large-size limit is the same as the bias for deterministic symmetric partitioning derived in section 2.2. This is not surprising since in this limit the behavior does not depend on the kernel.

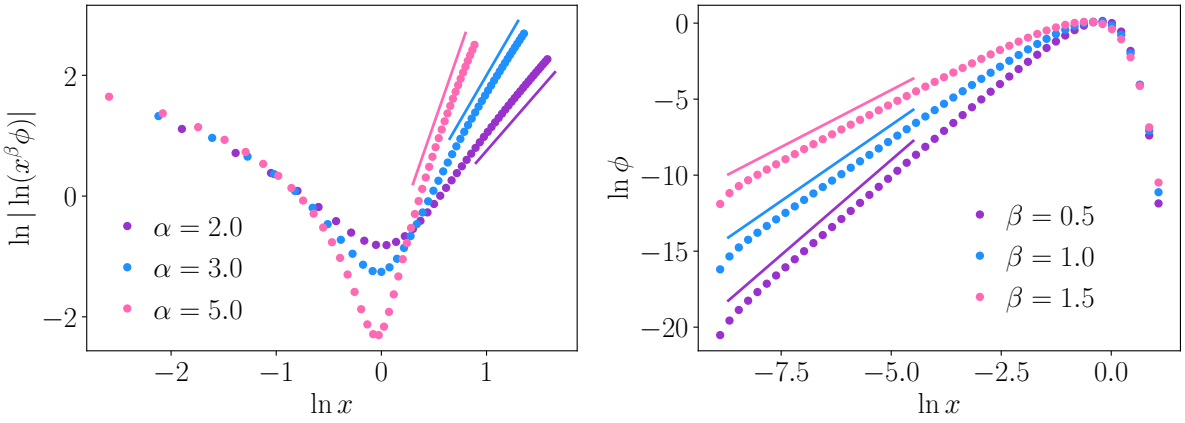


Figure 4.4: Large and small sizes asymptotic behaviors of the lineage distribution ϕ. For both plots, we chose $r(x) = x^\alpha$, $\nu(x) = x^\beta$, and $b(x) = x^2(1-x)^2/B(3,3)$ for all x, where $B(x,y)$ is the Beta function. Left: large size limit given by eq. (4.28) for $\beta = 1$, and for three different values of the strength control α. The slopes $\alpha - \beta + 1 = \alpha$ of the solid lines are, from left to right: 5, 3 and 2. Right: small size limit given by eq. (4.34) for $\alpha = 5$ and for three different values of β. The slopes $\kappa_0 + 1 - \beta = 3 - \beta$ of the solid lines are, from top to bottom: 1.5, 2 and 2.5.

For non-exponential growth in the large-size limit, the lineage-population bias is given by:

$$\frac{\psi}{\phi} \underset{x\to\infty}{\sim} \exp\left[-\frac{\Lambda}{\nu_\infty}\frac{x^{1-\beta_\infty}}{1-\beta_\infty}\right] \qquad \text{if } \beta_\infty \neq 1\,, \tag{4.30}$$

For any value $\beta_\infty \neq 1$, the right hand side of eq. (4.30) is a decreasing function of x, showing that large cells are under-represented in the population statistics as compared to the lineage statistics, similarly to what happens for exponential growth.

4.2 Small size limit

Balagué et al. showed that the population distribution was given in the small-size limit by (Balagué et al., 2013):

$$\psi(x) \underset{x\to 0}{\sim} \begin{cases} x^{\kappa_0+1-\beta_0} & \text{if } \beta_0 < 1 \\ x^{\kappa_0} & \text{if } \beta_0 \geq 1\,. \end{cases} \tag{4.31}$$

In order to understand intuitively the case splitting into two regimes, we give here an argument adapted from the fragmentation theory (Cheng et al., 1988), which is also easily generalizable to the lineage case. We multiply eq. (4.5) by x^k and integrate over x:

$$r_0(mL_k - 1)M_{k+\alpha_0} = \Lambda M_k - \nu_0 k M_{k+\beta_0-1}\,. \tag{4.32}$$

From the power-law behavior of b near 0, we get that not all moments L_k exists: there is a critical $k_c < 0$ under which L_k diverges. First we consider the case $\alpha_0 > 0$. When letting

$k \to k_c^+$, the left hand side of eq. (4.32) diverges, and so must the moment M of lowest order in the right hand side. When $\beta_0 - 1 \geq 0$, M_k is the moment of lowest order and must diverge, while $M_{k+\beta_0-1}$ and $M_{k+\alpha_0}$ converge, so that $M_k \propto L_k$ where the proportionality constant is positive, and thus ψ coincide with b: $\psi(x) \underset{x\to 0}{\sim} x^{\kappa_0}$. On the other hand, when $\beta_0 - 1 < 0$, the moment of lowest order is $M_{k+\beta_0-1}$ and thus $M_{k+\beta_0-1} \propto L_k$, where the proportionality constant is positive because $k_c < 0$. In that case, $\psi(x) \underset{x\to 0}{\sim} x^{\kappa_0+1-\beta_0}$. When $\alpha_0 = 0$, the stability condition reads $\beta_0 - 1 < 0$, so that we are in the second case.

The lineage equation on moments is obtained by following the same steps:

$$r_0(L_k - 1)N_{k+\alpha_0} = -\nu_0 k N_{k+\beta_0-1}\,. \tag{4.33}$$

The fundamental difference with the population case is the absence of terms in N_k. As a consequence, we obtain $N_{k+\beta_0-1} \propto L_k$ regardless of the value of β_0, so that:

$$\phi(x) \underset{x\to 0}{\sim} x^{\kappa_0+1-\beta_0}\,. \tag{4.34}$$

This analytical prediction is in perfect agreement with numerical resolutions of the PBE using a finite difference method with an implicit scheme, shown on fig. 4.4 right, for three different values of $\beta = 0.5$, 1 and 1.5.

Finally, we obtain the lineage-population bias by comparing eqs. (4.31) and (4.34):

$$\psi(x) \underset{x\to 0}{\sim} \begin{cases} \phi(x) & \text{if } \beta_0 < 1 \\ x^{\beta_0-1}\phi(x) & \text{if } \beta_0 \geq 1\,. \end{cases} \tag{4.35}$$

When $\beta_0 = 1$, there is no lineage-population bias as predicted in section 2.2. Indeed, $\psi^{b(x)}(x)$ is equal to $\phi^{2xb(x)}(x)/x \underset{x\to 0}{\sim} x^{\kappa_0+1}/x = \phi^{b(x)}(x)$, where the factor x and the modified kernel $2xb(x)$ that increases coefficient κ_0 by 1 exactly compensate.

Interestingly, there is no bias for any value $\beta_0 < 1$, which, following the discussion of section 2.2, implies that there is no correlation between size and divisions. Indeed, with deterministic partitioning, the daughter cell inherits half the volume of its mother, so that the only way to reach vanishing sizes it to divide a lot, whereas when all fractions of volume are allowed at division, a cell can also reach small sizes with few divisions if it inherits a small fraction of its mother volume. As a consequence, the correspondence between final size and number of divisions is blurred by the presence of noise in the volume partitioning.

On the other hand, when $\beta_0 > 1$, that is for cell growing slower than exponential in the region of small sizes, the lineage-population bias depends on β_0. Surprisingly, the lineage statistics is biased towards small cells as compared to the population statistics, unlike what we expected from the knowledge of deterministic partitioning. This suggests a correlation between small sizes and small numbers of divisions, that may be explained by the fact that cells that reach very small sizes, because of extremely-asymmetric partitioning, must grow during a very long time before reaching sizes at which they are likely to divide again, and end up with less divisions than average.

4.3 Validity for the adder model

Until now, we focused on the *sizer* model, where the division rate is only a function of the size x of the cell. Other models of cell size control have been proposed in the literature and presented in section 4.1 of chapter 1. In particular, the *adder* model which postulates that the distribution of volume added between birth and division is independent of the birth volume has been observed to account for a broad range of experimental data (Taheri-Araghi et al., 2015; Jun et al., 2018). In this section, we show that the asymptotic results derived above remain valid for the adder model. To do so, let us first re-write explicitly the population balance equation (eq. (1.54)) at the population probability level for the pair or variables (x, x_b) where x_b is the size at birth:

$$\partial_t \psi(x, x_b) = -\partial_x \left[\nu(x)\psi(x, x_b)\right] - \left[\Lambda_p(t) + \nu(x)\zeta(x - x_b)\right]\psi(x, x_b) \qquad \text{for } x > x_b \tag{4.36}$$

$$\nu(x_b)\psi(x_b, x_b) = m \int \frac{\mathrm{d}x'}{x'} \mathrm{d}x'_b \; b(x_b/x')\nu(x')\zeta(x' - x'_b)\psi(x', x'_b)\,, \tag{4.37}$$

where $\psi(x, x_b)$ is the fraction of cells of size x at time t which were born at size x_b. The lineage equation is obtained from this equation by setting $m = 1$ and $\Lambda_p(t) = 0$ again. Note already a fundamental difference between this equation and eq. (4.5): the term accounting for the birth of new cells enters as a boundary condition for the adder, because the added volume is reset to 0 at division.

For the large-size limit, a direct integration of the steady-state version of eq. (4.36) gives

$$\psi(x, x_b) = \psi(x_b, x_b)\frac{\nu(x_b)}{\nu(x)} \exp\left[-\int_{x_b}^{x} \mathrm{d}y \; \frac{\Lambda}{\nu(y)} + \int_0^{x - x_b} \mathrm{d}y \; \zeta(y)\right]. \tag{4.38}$$

For any given x_b, in the limit where $x \to +\infty$, we have $x - x_b \sim x$, so that the exponential does not depend on x_b anymore. Thus, the marginal size distribution obeys:

$$\psi(x) \underset{x\to+\infty}{\sim} C\nu(x)^{-1} \exp\left[-\int^{x} \mathrm{d}y \; \left(\frac{\Lambda}{\nu(y)} + \zeta(y)\right)\right], \tag{4.39}$$

where $C = \int_0^\infty \mathrm{d}x_b \; \psi(x_b, x_b)\nu(x_b)$ comes from the integration of the joint probability $\psi(x, x_b)$ over x_b. Finally, this large-size behavior is the same as eq. (4.26) for the sizer, with $r(x) = \nu(x)\zeta(x)$. As a consequence, the lineage-population biases in the large-size limit eqs. (4.29) and (4.30) remain valid for the adder model.

In the small-size limit, we saw before that the tail behavior was independent of the shape of the division rate, which is the only difference between the sizer and adder models, so we anticipate that the results are unchanged. To prove it, we still consider $\nu(x) \underset{x\to0}{\sim} \nu_0 x^{\beta_0}$ and $b(x) \underset{x\to0}{\sim} b_0 x^{\kappa_0}$. Multiplying eq. (4.36) by x^k and integrating over x and x_b leads after simple manipulations to:

$$\nu_0(mL_k - 1) \int \mathrm{d}x\mathrm{d}x_b \; x^{\beta_0 + k}\zeta(x - x_b)\psi(x, x_b) = \Lambda M_k - \nu_0 k M_{k+\beta_0-1}\,, \tag{4.40}$$

where $M_k = \int_0^\infty \mathrm{d}x \; x^k\psi(x) \equiv \int_0^\infty \mathrm{d}x\mathrm{d}x_b \; x^k\psi(x, x_b)$ is the k-th moment of the marginal size distribution $\psi(x)$.

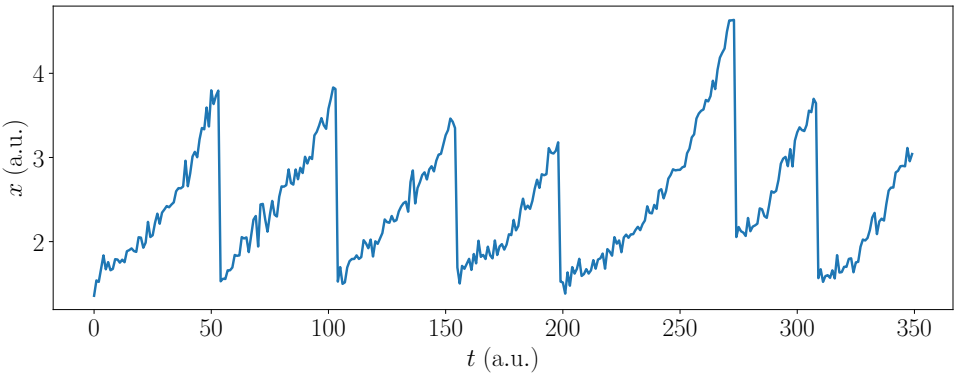


Figure 4.5: Example of size evolution versus time for a single lineage from Tanouchi et al., 2017, in the condition 27 °C.

Now we suppose that there is a $\delta_0 \geq 0$ such that $\zeta(x) = O(x^{\delta_0})$ when $x \to 0$, meaning that the division rate per unit volume ζ is growing as a power law or slower. In this case, the integral in the left hand side is smaller than $M_{k+\beta_0+\delta_0}$, and thus the integral does not diverge when $M_{k+\beta_0+\delta_0}$ does not diverge. The rest of the proof is the same as for the sizer, where $\beta_0 + \delta_0$ plays the role of α_0. Finally, the small-size lineage-population biases eq. (4.35) remain true for the adder model.

Before closing this section, we would like to draw the attention of the reader on the fact that in none of the different lineage-population biases derived in the previous sections does the division rate appear explicitly. In this section we showed that they hold for the adder model, which is a particular choice of two-variable division rate. This observation suggests that these biases could be correct for a much broader class of division rates, possibly involving other variables.

5 Noisy single-cell growth

Until now we considered that single-cell growth was deterministic. However, experimental data suggest that exponentially-growing cells are subject to stochasticity at different levels. Note that for these cells, the term 'single cell growth rate' often refers to the numerical factor ν in the function $\nu(x) = \nu x$, rather than to the function itself as we used before; therefore to be coherent with the literature on the subject we adopt this definition in this section. The variability in single cell growth rates has been mainly modeled with a Markov process, where the single cell growth rate changes from one cycle to the next one, but remains constant inside each cycle, so that the growth of a single cell is deterministic (Doumic et al., 2015; García-García et al., 2019). This kind of modeling accounts for cell-to-cell variability, which affects the population growth rate (Olivier, 2017), either increasing or decreasing it depending on mother-daughter correlations (Lin et al., 2020). On the other hand, experimental cycles exhibit fluctuations around the exponential trend, as shown on fig. 4.5 for a single cell size trajectory using E. coli data from Tanouchi et al., 2017. This prompts us to describe single-cell growth as a random process with a diffusive term accounting for in-cycle variability. Experimental data on E. coli (Kiviet et al., 2014)

suggest that both in-cycle and cell-to-cell sources of variability are present. However, a recent study (Jia et al., 2021) suggested that the cell-to-cell variability in growth has little impact on the steady-state size distribution. We thus decide to focus on the in-cycle source of variability in cell growth.

In the absence of noise, the Langevin dynamics of cell growth is given by $\mathrm{d}x = \nu x \mathrm{d}t$, and when considering noisy exponential growth, a Gaussian noise is usually put on the growth rate itself (Alonso et al., 2014): $\mathrm{d}x = (\nu \mathrm{d}t + \sqrt{2D}\mathrm{d}W)x$, where W is the Wiener process. In the corresponding Fokker-Planck representation, this leads to a new term with diffusion coefficient $D(x) = Dx^2$, which we call *multiplicative noise*. Therefore, the number $n(x,t)$ of cells of size x at time t follows

$$\partial_t n(x,t) = -\nu \partial_x [x n(x,t)] + D\partial_{x^2}[x^2 n(x,t)] - r(x)n(x,t) + m \int \frac{\mathrm{d}x'}{x'} b(x/x') r(x') n(x',t)\,, \tag{4.41}$$

supplemented with the 'no-flux' boundary conditions at $x = 0$ and $x = \infty$:

$$\lim_{x\to 0} D\partial_x[x^2 n(x,t)] - \nu x n(x,t) = \lim_{x\to\infty} D\partial_x[x^2 n(x,t)] - \nu x n(x,t) = 0\,, \tag{4.42}$$

which ensure that eq. (1.65) holds, namely that the instantaneous population growth rate is the average value of the division rate. With the stronger boundary conditions $x^2 p_{\mathrm{back}}(x,t) \to 0$ when $x \to 0$ and $x \to +\infty$, and $x^3 \partial_x p_{\mathrm{back}}(x,t) \to 0$ when $x \to 0$ and $x \to +\infty$, we recover eq. (4.18), namely that the steady-state population growth rate Λ is equal to the single cell growth rate ν. This follows again from the integration of the population balance equation at the level of the population distribution obtained from eq. (4.41), after multiplication by x. In the following we suppose that these conditions are fulfilled.

To our knowledge, exact solutions to eq. (4.41) were obtained for deterministic partitioning and only for specific growth rates $\nu(x)$, division rates $r(x)$ and diffusion coefficients $D(x)$. For instance, it was solved for constant functions $\nu(x)$, $r(x)$ and $D(x)$, both for symmetric (Efendiev, Brunt, Wake, et al., 2018) and asymmetric (Efendiev, Brunt, Zaidi, et al., 2018) partitioning; and for asymmetric partitioning, exponential growth $\nu(x) = \nu x$, multiplicative noise $D(x) = Dx^2$ and quadratic division rate $r(x) = rx^2$ (Zaidi et al., 2016). In this last case, the solution is a series of modified Bessel functions, generalizing the Dirichlet series obtained when there is no diffusion, and arising from the quadratic division rate which turns eq. (4.41) into a modified Bessel equation. Therefore, it seems difficult to generalize this method to more general power law division rates.

In this section, in order to investigate the impact of this source of stochasticity on the lineage-population bias derived in section 4.1, we seek asymptotic lineage and population distributions for large sizes and for more general divisions rates $r(x) = rx^\alpha$ and partition kernels. We show in appendix B.2 and appendix C that the lineage and population steady state distributions are equivalent in the large size limit and given by

$$\psi(x) \underset{x\to\infty}{\sim} \phi(x) \underset{x\to\infty}{\sim} x^{\frac{\nu}{2D} - \frac{3}{2} - \frac{\alpha}{4}} \exp\left[-\frac{2}{\alpha}\sqrt{\frac{r}{D}}\, x^{\frac{\alpha}{2}}\right]. \tag{4.43}$$

Just like the case of deterministic growth, the large-size behaviors of ϕ and ψ are independent of the partitioning kernel. In the case $\alpha = 2$, we recover the result obtained in Zaidi et al., 2016 (up to a missing factor $\sqrt{x}$ due to a typo).

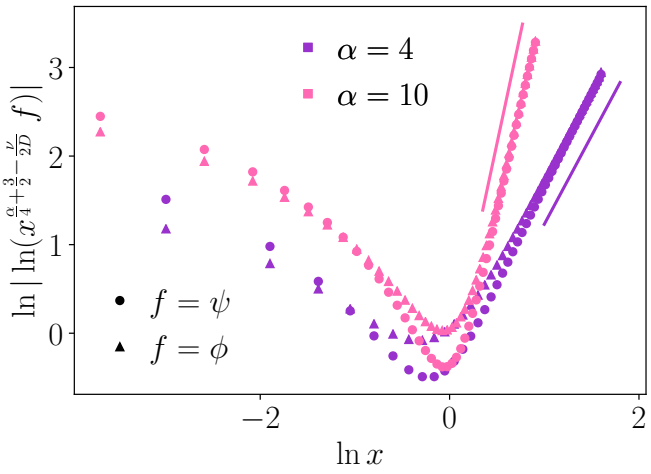


Figure 4.6: Large size asymptotic behaviors of the lineage and population distributions ϕ and ψ given by eq. (4.43). We fixed $r(x) = x^\alpha$, $\nu(x) = x$, $D = 0.4$ and $b(x) = x^2(1-x)^2/B(3,3)$ for all x, and show two different values of the strength control α. The slopes $\alpha/2$ of the solid lines are, from left to right: 5 and 2.

We numerically solve the PBE with the diffusive term using a finite difference method, both for the lineage statistics and the population statistics. The scheme is implicit in the first case and hybrid in the second: all terms are implicit except $\Lambda_p(t)$, explicitly computed using eq. (1.65). Results are shown on fig. 4.6, for two different values $\alpha = 4$ and $\alpha = 10$ of the size control strength. In both cases, the population and lineage distributions coincide in the large-size limit and align with the theoretical prediction eq. (4.43).

It may seem surprising that, unlike what happens in the case of deterministic growth discussed in section 4.1, no lineage-population bias is observed here. In fact, this can be understood by a generalization of the exact lineage-population bias obtained for deterministic exponential growth (eq. (4.21)). With the same method, we show in appendix D that the steady-state population distribution $\psi_\nu^{b(x)}$ with growth rate ν and partition kernel $b(x)$ is equal to the lineage distribution $\phi_{\nu+2D}^{mxb(x)}$ for the modified dynamics with $\hat{\nu} = \nu + 2D$ and $\hat{b}(x) = mxb(x)$, divided by the size:

$$\phi_{\nu+2D}^{mxb(x)}(x) = Kx\psi_\nu^{b(x)}(x)\,. \tag{4.44}$$

Thus, eq. (4.43) is coherent with eq. (4.44), where the bias towards smaller cells accounted for by the factor x is balanced by the larger effective growth rate $\nu + 2D$ favoring larger cells. Indeed, one easily check from eq. (4.43) that $\phi_{\nu+2D}(x)/x = \phi_\nu(x)$.

Finally, similarly to what happens for small sizes in presence of a stochastic kernel, the lineage-population bias is killed by the presence of multiplicative noise. When there is no noise, only cells that divided few times can reach large sizes, which imposes correlations between the number of divisions and the final size. Here however, this correlation is canceled because the number of divisions can be balanced by the noisy growth: large cells can come from lineages with numerous divisions if they grew faster on average than the deterministic growth at rate ν.

6 Constant populations

Before closing this chapter, we comment on some properties of constant populations in simple cases. In chapter 3, we investigated the effect of death or dilution on the phenotypic distributions and showed that they are biased when the death/dilution rate is phenotype-dependent. To obtain the explicit size distributions in presence of a non-trivial death rate is a whole new problem, but when the population is kept constant, simple relations can be derived. For consistency, in this section we adopt the notations from chapter 3, where $\psi(x,t)$ and $\psi^\circ(x,t)$ are the distributions in the presence and absence of death, respectively.

The population balance equation for the backward size distribution $\psi(x,t)$ in the presence of death reads:

$$\begin{aligned}\partial_t\psi(x,t) = &-\partial_x[\nu(x)\psi(x,t)] - [r(x)+\Lambda_p(t)+\gamma_t(x)]\,\psi(x,t)\\ &+ m\int dx'\Sigma(x|x')r(x')\psi(x',t)\,,\end{aligned} \tag{4.45}$$

where the death rate $\gamma_t(x)$ is a priori time and size dependent. For simplicity we consider deterministic growth here ($D=0$), which has no consequences on the following results. We recall that the instantaneous population growth rate is given by eq. (3.18):

$$\Lambda_t = \int dx\,[(m-1)\,r(x) - \gamma_t(x)]\,\psi(x,t)\,, \tag{4.46}$$

and can be canceled in two simple cases: when the dilution is uniform and perfectly balances the population growth:

$$\gamma_t(x) \equiv \gamma_t = (m-1)\int \mathrm{d}x\; r(x)\psi(x,t)\,, \tag{4.47}$$

and when the death rate is phenotype-dependent and balances each division on average:

$$\gamma_t(x) \equiv \gamma(x) = (m-1)\,r(x)\,. \tag{4.48}$$

We already noted that in the first case the size distribution is unchanged, which can be seen here by plugging eq. (4.47) into eq. (4.45): $\psi(x,t)$ follows the same equation eq. (4.5) as the population distribution without death $\psi^\circ(x,t)$.

In the second case, for any size x there is on average one division giving birth to m daughter cells for $m-1$ death events each killing one cell. Reporting this death rate in eq. (4.45) leads to:

$$\partial_t\psi(x,t) = -\partial_x[\nu(x)\psi(x,t)] - mr(x)\psi(x,t) + m\int dx'\Sigma(x|x')r(x')\psi(x',t)\,, \tag{4.49}$$

which is eq. (4.4) describing single lineages without death, with a rescaled division rate $\hat{r}(x) = mr(x)$, so that

$$\psi_r(x,t) = \phi^\circ_{mr}(x,t)\,. \tag{4.50}$$

We can understand qualitatively this relation as follows: in a mother machine, one mother cell is replaced by one smaller cell with rate $r(x)$. On the other hand, with the size-dependent death rate in population, at a rate $r(x)$ one mother cell gives birth to m

smaller daughter cells and $m-1$ cells of the same size as the mother are removed. Thus with rate $r(x)$, m mother-sized cells are replaced by m smaller cells, which is equivalent to say that one mother is replaced by one daughter at a rate $\hat{r}(x) = mr(x)$. We then understand easily that the protocol with the size-dependent death rate leads to statistics biased towards smaller cells as compared to the lineage statistics.

7 Conclusion

The recent development of mother machine devices revived the interest in the statistical comparison between lineage measurements and population snapshots. The unprecedented amount of single-cell data offers new insights on the way cells maintain size homeostasis. It is hence fundamental to quantify the statistics obtained in single-lineage setups and to understand how they differ from classical population snapshot.

To investigate this bias for cell size distributions, we worked on the steady-state population balance equation, separately for the population and lineage levels. We showed that for the special case of exponential growth, the population distribution is proportional to the lineage distribution with modified partitioning kernel and single cell growth rate, divided by the size x. This bias is reminiscent of the correlations between the size and the number of divisions and implies in particular than cells are on average smaller in population than in lineage. For more general power-law growth rates, with deterministic partitioning, we obtained the exact analytical expression for the lineage distribution, which is in good agreement with experimental data on E. coli, despite the slight stochasticity of the partitioning present in these data. This expression can then be used to infer the relevant parameters of the model, such as the strength of the size control, the growth rate and the asymmetry of the division.

We would like to emphasize a point here: the fact that E. coli data are in good agreement with the theoretical predictions obtained for the size-control model we considered does not mean that cells from these data follow a sizer mechanism, but only that the sizer provides a good description *for the size variable*. Indeed, we now know that E. coli follows an adder mechanism instead (Taheri-Araghi et al., 2015), and it was shown that if the sizer reproduces perfectly the size distribution for single lineage E. coli data, the sizer-simulated and experimental age distributions are indeed different (see Robert et al., 2014, fig. S8). Nonetheless, the fact that the inferred parameters are in good agreement with their values obtained by direct analysis of the data without any assumption on the model shows that the sizer description of the size distribution is meaningful, and not an artifact from over-fitting for example.

When relaxing the hypothesis of deterministic partitioning, we derived the small and large-size tails for both statistics. We showed that the large-size behavior is independent of the partitioning kernel, and that the lineage-population bias depends only on the growth rate of large cells. In the small size limit, the distributions only depend on the behavior of the partitioning kernel near 0 and of the growth rate of small cells, but is independent of the division rate. Two regimes are observed for the small-size lineage-population bias: for fast-growing small cells it is canceled, while for slow-growing small cells, it explicitly depends on the growth rate of small cells. Importantly, we showed that these asymptotic

behaviors remain valid for the adder mechanism, which is increasingly seen as the most relevant model for cell size control. When considering noisy in-cycle growth, the two large-size distributions become equal and explicitly depend on the noise.

An important result of this article is the cancellation of the lineage-population bias on cell-size when noise is introduced in the system, either on the growth rate for large sizes, or on the partitioning kernel for small sizes. Indeed, noise kills the correlations between the size and the number of divisions undergone by the cell, thus the selection of lineages with high reproductive success in population has no impact on the size distribution.

This work can be extended in several directions. First, experimental inference suggests a non-trivial behavior of the division rate at large sizes (Robert et al., 2014), even though the estimation in this region could be unreliable because of the lack of statistics. Therefore, it would be useful to relax the hypothesis of a power-law division rate. Second, we focused on a one-variable model, with the exception of the two-variable adder model, and it would be interesting to investigate more complex models with n variables. For example, modeling cell-to-cell variability in growth imposes to treat the single-cell growth rate ν as a second random variable. Finally, constant-population experiments, such as the dynamics cytometer (Hashimoto et al., 2016), may not all be well described by a uniform dilution rate, but rather by a dilution rate dependent on size, generation, ... as discussed in section 4.4 of chapter 3. In this case, the size distribution obtained would not be the same as in a freely-growing population but would bear the mark of the dilution protocol.

8 Appendices

A Exact lineage solution for deterministic partitioning

In this appendix, we seek exact solutions to eq. (4.4) for deterministic volume partitioning, either symmetric or asymmetric, and power law division and growth rates $r(x) = rx^\alpha$ and $\nu(x) = \nu x^\beta$.

A.1 Symmetric partitioning

We follow the method proposed in Hall et al., 1990 which consists in two changes of variables to reshape the PBE into an equation for constant division and growth rates. We start from the steady-state equation:

$$[\nu x^\beta \phi(x)]' = -rx^\alpha \phi(x) + m^{1+\alpha} rx^\alpha \phi(mx)\,, \tag{4.51}$$

and we define $Z(x) = x^\beta \phi(x)$, then

$$Z'(x) = \frac{r}{\nu} x^{\alpha-\beta} \left[-Z(x) + m^{1-\beta+\alpha} Z(mx)\right]\,. \tag{4.52}$$

We now define $u = rx^{\alpha-\beta+1}/[\nu(\alpha-\beta+1)]$ and $Y(u) = Z(x)$, so that the equation on Y reads:

$$Y'(u) + Y(u) = m^{1-\beta+\alpha} Y(m^{1-\beta+\alpha} u)\,. \tag{4.53}$$

The equation on Y has the same shape as eq. (4.51) for constant growth and division rates: $\alpha = \beta = 0$, and for a modified number of daughter cells $\hat{m} = m^{1-\beta+\alpha}$. Hall and Wake solved this equation in the case where $\hat{m} > 1$, which is equivalent to $1 - \beta + \alpha > 0$ since $m > 1$. The solution to this equation is (Hall et al., 1989):

$$Y(u) = C \sum_{k=0}^{\infty} c_k \exp\left[-m^{k(\alpha-\beta+1)} u\right] \tag{4.54}$$

$$c_0 = 1 \tag{4.55}$$

$$c_k = \frac{(-1)^k m^{k(\alpha-\beta+1)}}{\prod_{j=1}^{k} (m^{j(\alpha-\beta+1)} - 1)}\,. \tag{4.56}$$

Reverting to the notation in x with the function ϕ gives

$$\phi(x) = \frac{C}{x^\beta} \sum_{k=0}^{\infty} c_k \exp\left[-m^{k(\alpha-\beta+1)} \frac{r}{\nu} \frac{x^{\alpha-\beta+1}}{\alpha-\beta+1}\right]\,. \tag{4.57}$$

A.2 Asymmetric partitioning

For simplicity we consider binary fission ($m = 2$), with asymmetric partitioning: $b(x) = \delta(x - 1/\omega_1)/2 + \delta(x - 1/\omega_2)/2$, where $\omega_1 > \omega_2 > 1$ and $1/\omega_1 + 1/\omega_2 = 1$. Starting from:

$$[\nu x^\beta \phi(x)]' = -rx^\alpha \phi(x) + \frac{rx^\alpha}{2} \left[\omega_1^{1+\alpha} \phi(\omega_1 x) + \omega_2^{1+\alpha} \phi(\omega_2 x)\right]\,, \tag{4.58}$$

and making the same two changes of variables as for the symmetrical case, the equation reads

$$Y'(u) + Y(u) = \frac{1}{2}\left[\Omega_1 Y(\Omega_1 u) + \Omega_2 Y(\Omega_2 u)\right] , \tag{4.59}$$

where we defined $\Omega_i = \omega_i^{1-\beta+\alpha}$ for $i \in \{1, 2\}$. Since $1 - \beta + \alpha > 0$, the solution to this equation is (Suebcharoen et al., 2011):

$$Y(u) = C \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} c_{k,l} \exp\left[-\Omega_1^k \Omega_2^l u\right] \tag{4.60}$$

$$c_{0,0} = 1 \tag{4.61}$$

$$c_{k,0} = \frac{(-1)^k \Omega_1^k}{2^k \prod_{j=1}^{k} (\Omega_1^j - 1)} \tag{4.62}$$

$$c_{0,l} = \frac{(-1)^l \Omega_2^l}{2^l \prod_{j=1}^{l} (\Omega_2^j - 1)} \tag{4.63}$$

$$c_{k,l} = \frac{\Omega_1 c_{k-1,l} + \Omega_2 c_{k,l-1}}{2 - 2\Omega_1^k \Omega_2^l} . \tag{4.64}$$

Reverting to the original notations gives

$$\phi(x) = \frac{C}{x^\beta} \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} c_{k,l} \exp\left[-\omega_1^{k(\alpha-\beta+1)} \omega_2^{l(\alpha-\beta+1)} \frac{r}{\nu} \frac{x^{\alpha-\beta+1}}{\alpha - \beta + 1}\right] . \tag{4.65}$$

B Asymptotic lineage distribution for stochastic partitioning

In the following sub-appendices, since only the large-size limit is investigated, for simplicity we drop the subscript ∞ for r, ν, α and β, and the limit $\to \infty$ is always understood as $\to +\infty$. Moreover, we make two general comments. First, when the limit $k \to \infty$ is considered, corresponding to the large size behavior, the integrals of the type $\int_0^\infty \mathrm{d}x\ x^k f(x)$ are dominated by the behavior of the function f as $x \to \infty$. Therefore, when for example $r(x) \underset{x\to\infty}{\sim} r x^\alpha$, we write $\int_0^\infty \mathrm{d}x\ x^k r(x)\phi(x) \underset{k\to\infty}{\sim} \int_0^\infty \mathrm{d}x\ x^k r x^\alpha \phi(x) = r N_{\alpha+k}$. Second, the following transformation is used to isolate the moments L of the kernel b (Cheng et al., 1988):

$$\begin{aligned}
&\int_0^\infty \mathrm{d}x\ x^k \int_x^\infty \mathrm{d}y\ b(x/y) y^{\alpha-1} \phi(y) \\
&= \int_0^\infty \mathrm{d}y\ \ y^{\alpha-1}\phi(y) \int_0^y \mathrm{d}x\ x^k b(x/y) \\
&= \int_0^\infty \mathrm{d}y\ y^{k+\alpha}\phi(y) \int_0^1 \mathrm{d}u\ u^k b(u) \\
&= N_{\alpha+k}\ L_k ,
\end{aligned} \tag{4.66}$$

where we went from the second to the third line with the change of variable $u = x/y$.

B.1 Deterministic growth

We first consider the case of deterministic growth ($D = 0$) for an arbitrary growth rate with power β, and follow the method proposed in Cheng et al., 1988 for fragmentation processes. We multiply the steady-state version of eq. (4.4) by $x^{k-\beta+1}$ and integrate over x, to recast the PBE as a recursion relation on the moments of the distribution:

$$N_{k+\alpha-\beta+1} \underset{k\to\infty}{\sim} \frac{\nu}{r} \frac{k-\beta+1}{1-L_{k-\beta+1}} N_k \,. \tag{4.67}$$

For simplicity we define $\rho = \alpha - \beta + 1 > 0$, and n such that $k = n\rho$, leading to

$$N_{(n+1)\rho} \underset{n\to\infty}{\sim} \frac{\nu}{r} \frac{n\rho-\beta+1}{1-L_{n\rho-\beta+1}} N_{n\rho} \,. \tag{4.68}$$

Iterating this relation leads to the general term:

$$N_{n\rho} \underset{n\to\infty}{\sim} N_\rho \left(\frac{\nu}{r}\right)^{n-1} \prod_{j=1}^{n-1} \frac{j\rho-\beta+1}{1-L_{j\rho-\beta+1}} \,. \tag{4.69}$$

We compute the numerator as

$$\prod_{j=1}^{n-1} (j\rho-\beta+1) = \rho^{n-1}(n-1)! \prod_{j=1}^{n-1} \left(1-\frac{\beta-1}{j\rho}\right) \tag{4.70}$$

$$\underset{n\to\infty}{\sim} \rho^{n-1}(n-1)!(n-1)^{\frac{1-\beta}{\rho}} \,, \tag{4.71}$$

where we used that $\prod_{j=1}^{n}\left(1-\frac{a}{j}\right) \sim n^{-a}$ as $n \to \infty$.

We now show that the moments $L_{j\rho-\beta+1}$ of the partition kernel can be neglected in this limit. We consider a power-law partition kernel $b(x) = b_1(1-x)^{\kappa_1}$ in the limit $x \to 1$, but the argument can be made more general.

$$L_k \underset{k\to\infty}{\sim} b_1 \int_0^1 \mathrm{d}x \; x^k (1-x)^{\kappa_1} \tag{4.72}$$

$$\underset{k\to\infty}{\sim} k^{-(\kappa_1+1)} \,, \tag{4.73}$$

where we recognized the Beta function $B(k+1, \kappa_1+1) = \int_0^1 \mathrm{d}y \; y^k(1-y)^{\kappa_1}$, whose asymptotic behavior when only one of the two parameters tends to infinity (here k) is given by: $B(k+1, \kappa_1+1) \underset{k\to\infty}{\sim} \Gamma(\kappa_1+1)k^{-(\kappa_1+1)}$. Then

$$\ln \prod_{j=1}^{n-1} (1-L_{j\rho-\beta+1}) = \sum_{j=1}^{n-1} \ln(1-L_{j\rho-\beta+1}) \tag{4.74}$$

$$\underset{n\to\infty}{\sim} -\sum_{j=1}^{n-1} (j\rho-\beta+1)^{-(\kappa_1+1)} \,, \tag{4.75}$$

where the second line is obtained by a first-order expansion of the natural logarithm. Finally, since $\kappa_1 > 0$, this series is converging when $n \to \infty$, and so is the product in the denominator of eq. (4.69).

The general term then reads

$$N_{n\rho} \underset{n\to\infty}{\sim} \left(\frac{\rho\nu}{r}\right)^{n-1} (n-1)!(n-1)^{\frac{1-\beta}{\rho}} . \tag{4.76}$$

We next use Stirling approximation: $n! \underset{n\to\infty}{\sim} \sqrt{2\pi n}(n/e)^n$, switch back to $k = n\rho$ and replace ρ:

$$N_k \underset{k\to\infty}{\sim} \left(\frac{\nu k}{re}\right)^{\frac{k}{\alpha-\beta+1}} k^{\frac{1-\beta}{\alpha-\beta+1}-\frac{1}{2}} . \tag{4.77}$$

The inverse Mellin transform of this moment is obtained in appendix C.

B.2 Stochastic growth

We examine the case of exponential growth $\nu(x) = \nu x$ with multiplicative noise $D(x) = Dx^2$. Following the same steps as for the deterministic growth, we multiply by x^k the steady-state population balance equations at the population and lineage levels obtained from eq. (4.41), and integrate them over x:

$$M_{k+\alpha} \underset{k\to\infty}{\sim} \frac{1}{r} \frac{\nu(k-1) + Dk(k-1)}{1 - mL_k} M_k \tag{4.78}$$

$$N_{k+\alpha} \underset{k\to\infty}{\sim} \frac{1}{r} \frac{\nu k + Dk(k-1)}{1 - L_k} N_k . \tag{4.79}$$

As for the deterministic case, the moments L_k are negligible for large k, so that the moments $M_{k+\alpha}$ and $N_{k+\alpha}$ differ only by their numerators. We conduct the calculations for the lineage distribution first and then show why the difference in the numerators does not affect the general moment.

We define n such that $k = n\alpha$, and iterate the relation to obtain the general term:

$$N_{n\alpha} \underset{n\to\infty}{\sim} N_\alpha \left(\frac{1}{r}\right)^{n-1} \prod_{j=1}^{n-1} \frac{\nu j\alpha + Dj\alpha(j\alpha-1)}{1 - L_{j\alpha}} . \tag{4.80}$$

The numerator is computed as:

$$\prod_{j=1}^{n-1} [\nu j\alpha + Dj\alpha(j\alpha-1)] = (\alpha^2 D)^{n-1}(n-1)!^2 \prod_{j=1}^{n-1} \left(1 - \frac{D-\nu}{j\alpha D}\right) \tag{4.81}$$

$$\underset{n\to\infty}{\sim} (\alpha^2 D)^{n-1}(n-1)!^2 (n-1)^{\frac{\nu-D}{\alpha D}} . \tag{4.82}$$

The Stirling approximation: $n!^2 \underset{n\to\infty}{\sim} 2\pi n(n/e)^{2n}$ is used to obtain

$$N_{n\alpha} \underset{n\to\infty}{\sim} \left(\frac{\alpha^2 D}{r}\right)^{n-1} \left(\frac{n-1}{e}\right)^{2(n-1)} (n-1)^{\frac{\nu-D}{\alpha D}+1} . \tag{4.83}$$

Switching back to $k = n\alpha$ leads to:

$$N_k \underset{k\to\infty}{\sim} \left(\sqrt{\frac{D}{r}}\frac{k}{e}\right)^{\frac{2k}{\alpha}} k^{\frac{2}{\alpha}\frac{\nu-D}{2D}-1} . \tag{4.84}$$

The inverse Mellin transform of this moment is obtained in appendix C.

The numerator of the general moments $M_{n\alpha}$ is given by

$$\prod_{j=1}^{n-1}\left[\nu(j\alpha-1)+Dj\alpha(j\alpha-1)\right]=(\alpha^2D)^{n-1}(n-1)!^2\prod_{j=1}^{n-1}\left(1-\frac{1}{j\alpha}\right)\prod_{j=1}^{n-1}\left(1+\frac{\nu}{j\alpha D}\right) \tag{4.85}$$

$$\underset{n\to\infty}{\sim}(\alpha^2D)^{n-1}(n-1)!^2(n-1)^{-\frac{1}{\alpha}}(n-1)^{\frac{\nu}{\alpha D}}\,, \tag{4.86}$$

which is identical to eq. (4.82), so that the moment M_k is equal to N_k given by eq. (4.84), and leads to the same distribution for large sizes.

C Mellin transform of polynomial-exponential distribution

For a distribution y characterized by its large x behavior:

$$y(x)\underset{x\to+\infty}{\sim}x^{\eta-\lambda(\mu-1/2)-1}\exp\left[-\frac{x^\lambda}{\lambda\omega}\right]\,, \tag{4.87}$$

the moments of large order read

$$M_k\underset{k\to+\infty}{\sim}\int_0^\infty \mathrm{d}x\; x^{k+\eta-\lambda(\mu-1/2)-1}e^{-\frac{x^\lambda}{\lambda\omega}} \tag{4.88}$$

$$\underset{k\to+\infty}{\sim}\lambda^{-1}(\lambda\omega)^{(k+\eta)/\lambda-\mu+1/2}\int_0^\infty \mathrm{d}t\; t^{(k+\eta)/\lambda-\mu-1/2}e^{-t}\,, \tag{4.89}$$

where we went from the first to the second line using the change of variable $t=x^\lambda/\lambda\omega$. We recognize the function $\Gamma(z)=\int_0^\infty \mathrm{d}t\; t^{z-1}e^{-t}$ in the second line with $z=(k+\eta)/\lambda-\mu+1/2$, and we use the Stirling approximation: $\Gamma(z+1)\underset{z\to\infty}{\sim}\sqrt{2\pi z}\left(\frac{z}{e}\right)^z$. Finally, the Mellin transform reads:

$$M_k\underset{k\to+\infty}{\sim}\lambda^{-\frac{3}{2}}(\lambda\omega)^{(k+\eta)/\lambda-\mu+1/2}\sqrt{2\pi(k+\eta-\lambda(\mu+1/2))}\left(\frac{k+\eta-\lambda(\mu+1/2)}{\lambda e}\right)^{\frac{k+\eta}{\lambda}-\mu-1/2} \tag{4.90}$$

$$\underset{k\to+\infty}{\sim}\left(\frac{k\omega}{e}\right)^{\frac{k}{\lambda}}k^{\frac{\eta}{\lambda}-\mu}\,. \tag{4.91}$$

D Lineage-population bias for exponentially-growing cells

In this section, we consider the case of exponential growth $\nu(x)=\nu x$ with multiplicative noise $D(x)=Dx^2$, in steady-state so that $\Lambda=\nu$, for which eq. (4.5) and eq. (4.4) read:

$$0=-\nu\partial_x[x\psi(x)]+D\partial_{x^2}[x^2\psi(x)]-[r(x)+\nu]\,\psi(x)+m\int\frac{dx'}{x'}b(x/x')r(x')\psi(x') \tag{4.92}$$

$$0=-\nu\partial_x[x\phi(x)]+D\partial_{x^2}[x^2\phi(x)]-r(x)\phi(x)+\int\frac{dx'}{x'}b(x/x')r(x')\phi(x')\,. \tag{4.93}$$

We multiply the population equation by x, and recast it for the function $q(x) = x\psi(x)$:

$$0 = -\nu x\partial_x q(x) + Dx\partial_{x^2}[xq(x)] - [r(x) + \nu]\, q(x) + \int \frac{dx'}{x'}\frac{mx}{x'} b(x/x')r(x')q(x')\,. \quad (4.94)$$

We identify the derivative of a product $-\nu x\partial_x q(x) - \nu q(x) = -\nu\partial_x[xq(x)]$, and show straightforwardly that $Dx\partial_{x^2}[xq(x)] = D\partial_{x^2}[x^2 q(x)] - 2D\partial_x[xq(x)]$, so that the second term is absorbed in the first-order derivative describing exponential growth:

$$0 = -(\nu + 2D)\partial_x[xq(x)] + D\partial_{x^2}[x^2 q(x)] - r(x)q(x) + \int \frac{dx'}{x'}\frac{mx}{x'} b(x/x')r(x')q(x')\,. \quad (4.95)$$

This equation is eq. (4.93), obeyed by the lineage distribution ϕ with modified growth rate $\hat{\nu} = \nu + 2D$ and partition kernel $\hat{b}(x) = mxb(x)$, therefore $q(x)_\nu^{b(x)}$ is proportional to $\phi_{\nu+2D}^{mxb(x)}$:

$$\phi_{\nu+2D}^{mxb(x)}(x) = Kx\psi_\nu^{b(x)}(x)\,, \quad (4.96)$$

with $K = \left(\int_0^\infty \mathrm{d}x\; x\psi_\nu^{b(x)}(x)\right)^{-1}$ a normalization constant. Importantly, $\hat{b}$ is a proper kernel, which is normalized as a consequence of the conservation of volume of the original kernel b: $\int_0^1 \mathrm{d}x\; \hat{b}(x) = m\int_0^1 \mathrm{d}x\; xb(x) = 1$. Note however that the modified kernel $\hat{b}$ need not ensure itself the conservation of volume. In fact, imposing this property for kernel $\hat{b}$ reads: $m\int_0^1 \mathrm{d}x\; x\hat{b}(x) = m^2\int_0^1 \mathrm{d}x\; x^2 b(x) = 1$. Combining the conservation of volume for both kernels leads to $L_1^2 = L_2$, so that the variance of b is null and $\hat{b}(x) = b(x) = \delta(x - 1/m)$. Consequently, the modified kernel conserve the volume only if it is equal to the original kernel, that is in the case of equal fission in m parts

Bibliography for Chapter 4

[Alonso et al., 2014] Alonso, A. A., I. Molina, and C. Theodoropoulos (2014). Modeling Bacterial Population Growth from Stochastic Single-Cell Dynamics. *Applied and Environmental Microbiology* 80.(17), pp. 5241–5253.

[Balagué et al., 2013] Balagué, D., J. Cañizo, and P. Gabriel (2013). Fine asymptotics of profiles and relaxation to equilibrium for growth-fragmentation equations with variable drift rates. *Kinetic and Related Models* 6.(2), pp. 219–243.

[Cheng et al., 1988] Cheng, Z. and S. Redner (1988). Scaling Theory of Fragmentation. *Physical Review Letters* 60.(24), pp. 2450–2453.

[Doumic Jauffret et al., 2010] Doumic Jauffret, M. and P. Gabriel (2010). Eigenelements of a general aggregation-fragmentation model. *Mathematical Models and Methods in Applied Sciences* 20.(5), pp. 757–783.

[Doumic et al., 2021] Doumic, M. and M. Hoffmann (2021). Individual and population approaches for calibrating division rates in population dynamics: Application to the bacterial cell cycle. *arXiv:2108.13155*.

[Doumic et al., 2015] Doumic, M., M. Hoffmann, N. Krell, and L. Robert (2015). Statistical estimation of a growth-fragmentation model observed on a genealogical tree. *Bernoulli* 21.(3), pp. 1760–1799.

[Efendiev, Brunt, Wake, et al., 2018] Efendiev, M., B. van Brunt, G. C. Wake, and A. A. Zaidi (2018). A functional partial differential equation arising in a cell growth model with dispersion. *Mathematical Methods in the Applied Sciences* 41.(4), pp. 1541–1553.

[Efendiev, Brunt, Zaidi, et al., 2018] Efendiev, M., B. van Brunt, A. A. Zaidi, and T. H. Shah (2018). Asymmetric cell division with stochastic growth rate. Dedicated to the memory of the late Spartak Agamirzayev. *Mathematical Methods in the Applied Sciences* 41.(17), pp. 8059–8069.

[Friedlander et al., 2008] Friedlander, T. and N. Brenner (2008). Cellular Properties and Population Asymptotics in the Population Balance Equation. *Physical Review Letters* 101.(1), p. 018104.

[García-García et al., 2019] García-García, R., A. Genthon, and D. Lacoste (2019). Linking lineage and population observables in biological branching processes. *Physical Review E* 99.(4), p. 042413.

[Genthon, 2022a] Genthon, A. (2022a). Analytical cell size distribution: lineage-population bias and parameter inference. *arXiv:2206.06146*.

[Genthon, 2022b] Genthon, A. (2022b). Analytical cell size distribution: lineage-population bias and parameter inference. *Journal of The Royal Society Interface* 19.(196), p. 20220405.

[Hall et al., 1989] Hall, A. J. and G. C. Wake (1989). A functional differential equation arising in modelling of cell growth. *The Journal of the Australian Mathematical Society. Series B. Applied Mathematics* 30.(4), pp. 424–435.

[Hall et al., 1990] Hall, A. J. and G. C. Wake (1990). Functional differential equations determining steady size distributions for populations of cells growing exponentially. *The Journal of the Australian Mathematical Society. Series B. Applied Mathematics* 31.(4), pp. 434–453.

[Hashimoto et al., 2016] Hashimoto, M., T. Nozoe, H. Nakaoka, R. Okura, S. Akiyoshi, K. Kaneko, E. Kussell, and Y. Wakamoto (2016). Noise-driven growth rate gain in clonal cellular populations. *Proceedings of the National Academy of Sciences* 113.(12), pp. 3251–3256.

[Horváth et al., 2013] Horváth, A., A. Rácz-Mónus, P. Buchwald, and Á. Sveiczer (2013). Cell length growth in fission yeast: an analysis of its bilinear character and the nature of its rate change transition. *FEMS Yeast Research* 13.(7), pp. 635–649.

[Jia et al., 2021] Jia, C., A. Singh, and R. Grima (2021). Cell size distribution of lineage data: Analytic results and parameter inference. *iScience* 24.(3), p. 102220.

[Jia et al., 2022] Jia, C., A. Singh, and R. Grima (2022). Characterizing non-exponential growth and bimodal cell size distributions in fission yeast: An analytical approach. *PLoS Computational Biology* 18.(1), e1009793.

[Jun et al., 2018] Jun, S., F. Si, R. Pugatch, and M. Scott (2018). Fundamental principles in bacterial physiology—history, recent progress, and the future with focus on cell size control: a review. *Reports on Progress in Physics* 81.(5), p. 056601.

[Kar et al., 2021] Kar, P., S. Tiruvadi-Krishnan, J. Männik, J. Männik, and A. Amir (2021). Distinguishing different modes of growth using single-cell data. *eLife* 10, e72565.

[Kiviet et al., 2014] Kiviet, D. J., P. Nghe, N. Walker, S. Boulineau, V. Sunderlikova, and S. J. Tans (2014). Stochasticity of metabolism and growth at the single-cell level. *Nature* 514.(7522), pp. 376–379.

[Lin et al., 2020] Lin, J. and A. Amir (2020). From single-cell variability to population growth. *Physical Review E* 101.(1), p. 012401.

[Messelink et al., 2021] Messelink, J. J., F. Meyer, M. Bramkamp, and C. P. Broedersz (2021). Single-cell growth inference of Corynebacterium glutamicum reveals asymptotically linear growth. *eLife* 10, e70106.

[Michel, 2006] Michel, P. (2006). Existence of a solution to the cell division eigenproblem. *Mathematical Models and Methods in Applied Sciences* 16 (supp01), pp. 1125–1153.

[Olivier, 2017] Olivier, A. (2017). How does variability in cell aging and growth rates influence the Malthus parameter? *Kinetic and Related Models* 10.(2), pp. 481–512.

[Pesti et al., 2021] Pesti, B., Z. Nagy, L. Papp, M. Sipiczki, and Á. Sveiczer (2021). Cell Length Growth in the Fission Yeast Cell Cycle: Is It (Bi)linear or (Bi)exponential? *Processes* 9.(9), p. 1533.

[Robert et al., 2014] Robert, L., M. Hoffmann, N. Krell, S. Aymerich, J. Robert, and M. Doumic (2014). Division in Escherichia coli is triggered by a size-sensing rather than a timing mechanism. *BMC Biology* 12.(1), p. 17.

[Suebcharoen et al., 2011] Suebcharoen, T., B. van Brunt, and G. C. Wake (2011). Asymmetric cell division in a size-structured growth model. *Differential and Integral Equations* 24.(7), pp. 787–799.

[Taheri-Araghi et al., 2015] Taheri-Araghi, S., S. Bradde, J. T. Sauls, N. S. Hill, P. A. Levin, J. Paulsson, M. Vergassola, and S. Jun (2015). Cell-Size Control and Homeostasis in Bacteria. *Current Biology* 25.(3), pp. 385–391.

[Tanouchi et al., 2017] Tanouchi, Y., A. Pai, H. Park, S. Huang, N. E. Buchler, and L. You (2017). Long-term growth data of Escherichia coli at a single-cell level. *Scientific Data* 4.(1), p. 170036.

[Thomas, 2017] Thomas, P. (2017). Single-cell histories in growing populations: relating physiological variability to population growth. *BiorXiv*.

[Thomas, 2018] Thomas, P. (2018). Analysis of Cell Size Homeostasis at the Single-Cell and Population Level. *Frontiers in Physics* 6, p. 64.

[Totis et al., 2021] Totis, N., C. Nieto, A. Kuper, C. Vargas-Garcia, A. Singh, and S. Waldherr (2021). A Population-Based Approach to Study the Effects of Growth and Division Rates on the Dynamics of Cell Size Statistics. *IEEE Control Systems Letters* 5.(2), pp. 725–730.

[Zaidi et al., 2021] Zaidi, A. A. and B. van Brunt (2021). Asymmetrical cell division with exponential growth. *The ANZIAM Journal* 63.(1), pp. 70–83.

[Zaidi et al., 2016] Zaidi, A. A., B. van Brunt, and G. C. Wake (2016). Probability density function solutions to a Bessel type pantograph equation. *Applicable Analysis* 95.(11), pp. 2565–2577.

Chapter 5

Stochastic thermodynamics of cell growth and division[†]

[†]This chapter is based on the article Genthon et al., 2022, reproduced here with the authorization of all co-authors.

Contents

1 Introduction

Thermodynamics aims at determining which physical processes are possible or not, and at giving universal constraints for their speed, accuracy, power and efficiency. A clear thermodynamic description of cell growth and division would give us clues about the performances of these processes compared to their thermodynamic bounds: are cell growth and division optimized? This raises the more fundamental question of the nature of the output to be optimized in cell colonies. Ultimately, understanding the performance of these processes could help us compare the different models of cell size control and the influence of the different parameters of the models.

In this chapter, we take a first step in this direction by proposing a thermodynamic description of cell growth and division. These two processes are clearly absolutely irreversible, and must be constrained by the laws of thermodynamics. By absolutely irreversible, we mean that a process occurs only in one direction, and its time-reversed counterpart is never observed. Since in stochastic thermodynamics the notion of total entropy production quantifies the time-symmetry breaking by comparing the forward to backward (time-reversed) path probabilities (eq. (1.26)), absolutely irreversible processes are challenging to address because they lead to infinite productions of entropy. A divergent production of entropy would be associated with a diverging dissipation of heat in the environment, which we of course never observe, even for absolutely irreversible processes. This suggests that this definition of entropy production is not adapted to such processes and should be modified in a meaningful way. Different approaches have been proposed to tackle this issue, and we give here a brief summary of two of them, which may be relevant for cell division. A more complete account can be found in Busiello et al., 2020.

A first way to bypass the difficulty is to argue that absolutely irreversible processes are not actually absolutely non-reversible, but rather very rare and thus never observed. This idea has been put forward in Zeraati et al., 2012 and in England, 2013. In these works, lower bounds on the production of entropy were proposed based on physical reasoning. Of particular interest for us is England, 2013, where the author computed a lower bound for the production of entropy associated with cell division, based on the probability of going from two cells to one cell following the breaking of all peptide bonds from one cell, resulting in its dissolution in the environment. Although of conceptual significance, this result has limited impact to understand the process of cell division since it only gives a vanishingly small backward path probability, which could never be observed. A second way, proposed in Fuchs et al., 2016 in the context of stochastic resetting, is to compute the contributions of the irreversible parts of the processes to the change in the Shannon entropy of the system, without using the notion of total entropy production. This is the approach we pursue here, namely describing cell growth and division in terms of two subprocesses: branching and resetting, which can then be analyzed separately using stochastic thermodynamics.

Early works on target search (Coppey et al., 2004; Bénichou et al., 2005), followed by the seminal article Evans et al., 2011, have laid the foundations of resetting, a stochastic process involving an instantaneous transition to a pre-defined position or region of space. In the past ten years, resetting has been a very active field of research, which various

groups have extensively studied in the context of target search or in the context of non-equilibrium steady-states (NESS) and first passage times (see Evans et al., 2020 for a comprehensive review). The first study of resetting from the point of view of stochastic thermodynamics was carried out in Fuchs et al., 2016, where the first and second laws of thermodynamics for resetting to a single fixed position were derived. Later, Roldán et al. developed a path integral approach for resetting (Roldán et al., 2017), which was used by Pal et al. to derive integral fluctuation relations (Pal et al., 2017; Gupta et al., 2020) and thermodynamic uncertainty relations (Pal et al., 2021) for systems with resetting.

To our knowledge, the combination of stochastic resetting and branching processes has only been considered for the purpose of search strategies (Eliazar, 2017; Pal et al., 2019). In particular, the authors compared first passage times for branching search, a term coined in Eliazar, 2017 to describe a strategy mixing branching and resetting, and for resetting alone. However, the thermodynamics of stochastic processes involving both resetting and branching at the same time has not been studied so far.

Branching processes with resetting are not only useful in the context of target search but are fundamental in biology, in particular to describe populations of cells that grow and divide. Indeed, cell division intrinsically features branching, since one mother cell gives birth to two daughters cells. As a result of division, some traits of daughter cells are reset to an absolute value, such as the age of the cell which is reset at 0, while others are reset to a value relative to that of the mother at the moment of the division. For instance, the size of the daughter cell restarts at a value close to half the size of the mother cell for organisms undergoing symmetric division. In order to study cell division from the point of view of stochastic thermodynamics, we extend the analysis of Fuchs et al., 2016 to allow a restart at a relative position instead of an absolute one, and to incorporate branching in the formalism.

The population balance equations for the different models of cell size control at the probability level (see section 4.3 of chapter 1) are very similar in form to Fokker-Planck equations used in stochastic thermodynamics to describe the motion of overdamped Brownian particles (see section 2.1 of chapter 1). For this reason, this chapter is organized as follows: in section 2, we first present our results in the general setting of a population of overdamped Brownian particles undergoing relative resetting and branching in a 1-dimensional potential. We derive the first and second laws of thermodynamics for this model and identify the separate contributions of resetting and branching. Moreover, we propose an alternative version of the second law, which remains valid for athermal systems which are not in contact with a heat bath, and we show how our results can be generalized to the case of an n-dimensional problem. In section 3, we apply our results to the three cell size control models: timer, sizer and adder. We obtain analytical expressions for the work and entropy terms associated with resetting and branching, whose signs are indicative of the transfer of energy and entropy between the two sub-processes. Finally, stochastic thermodynamics has been often used to analyze information processing and efficiency of small biological systems. In the same spirit, we propose an analogy between cell division and stochastic heat engines, which leads us to introduce an efficiency quantifying a form of conversion of entropy production from resetting to branching.

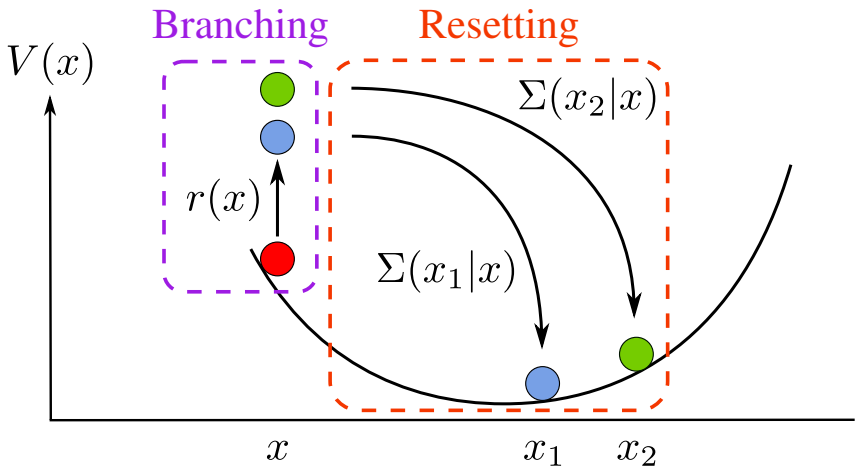


Figure 5.1: Illustration of a particle branching into $m = 2$ particles, each of which is reset to a position relative to the branching position x through the transition probability Σ. The parabolic potential shown here is for illustration only, potentials can be arbitrary in our model.

2 Thermodynamics of branching processes with stochastic resetting

2.1 Model

We consider a population of overdamped Brownian particles with mobility μ and diffusion coefficient D, in contact with a heat bath at temperature T and subject to a potential V. For simplicity, we treat the case of a 1-dimensional potential $V(x)$ in the main text and show in appendix A that our results hold in n dimensions. In the following we assume Einstein relation $D = \mu T$, with unit Boltzmann's constant $k_B = 1$. In addition to their diffusive motion, particles randomly branch at a space-dependent rate $r(x)$, leading one particle to give birth to m particles (including the original one) at the same position. Instantly after branching, all m particles are reset to new positions with the transition probability $\Sigma(x|x')$, for a restart at position x if the original particle branched at position x', as illustrated on fig. 5.1. Since the positions of the new particles depend on x', we call it *relative* resetting, as opposed to *absolute* resetting, where particles either restart at a fixed position x_0 ($\Sigma(x|x') = \delta(x - x_0)$) or restart at a random position with a probability distribution $\Sigma(x|x') = f(x)$, independently of x'. In this case, an explicit solution of the NESS is in general no longer available, unlike what happens for absolute resetting. The dynamics of the number $n(x,t)$ of particles at position x at time t is described by the following generalized Fokker-Planck equation:

$$\partial_t n(x,t) = -\partial_x \left[\mu F(x) n(x,t) - D\partial_x n(x,t)\right] - r(x)n(x,t) + m\int \mathrm{d}x' \Sigma(x|x') r(x') n(x',t)\,, \tag{5.1}$$

where $F(x) = -\partial_x V$ is the conservative force deriving from the potential V. We recast this equation at the probability level by defining the proportion of particles at position

x at time t: $p(x,t) = n(x,t)/N_t$, with $N_t = \int \mathrm{d}x\ n(x,t)$ the total number of particles at time t:

$$\partial_t p(x,t) = -\partial_x j(x,t) - [\Lambda_p(t) + r(x)]\, p(x,t) + m \int \mathrm{d}x' \Sigma(x|x') r(x') p(x',t)\,, \tag{5.2}$$

where we have defined the current

$$j(x,t) = \mu F(x) p(x,t) - D\partial_x p(x,t)\,. \tag{5.3}$$

We take vanishing boundary conditions for probability p and current j at $x = 0$ and $x \to +\infty$. Note that p is a backward probability, but since the forward probability is not used in this chapter, we drop the subscript without ambiguity.

Before going into the details of the thermodynamics of this model, let us make an important remark. In this first section, we aim to give a general thermodynamic theory of branching processes with resetting, which is framed in the context of the paradigmatic system of stochastic thermodynamic: the overdamped Brownian particle. The position of such a particle follows Gaussian fluctuations resulting from the random collisions with the medium particles, and is accounted for by the diffusive term $D\partial_{x^2}[p(x,t)]$ in the Fokker-Planck equation. However, in the previous chapter on size distributions, we introduced noise on cell growth by adding a diffusive term of the form $D\partial_{x^2}[x^2 p(x,t)]$ to the population balance equation (section 5 of chapter 4), which corresponds to a Gaussian noise on the logarithm of the size. Doing so, the size remains positive even for noises of possibly large amplitude. Therefore, the kind of noise describing Brownian particles is not the best suited to describe the noise in cell growth. This is however not a problem since we propose an alternative description for athermal systems in section 2.3 where the diffusion coefficient is set to zero. This alternative version is the only one applied in section 3 to cell growth and division, where growth is thus considered deterministic with a rate $\nu(x)$.

2.2 First and second laws of thermodynamics

We follow the approach from Fuchs et al., 2016 to derive the first and second laws of thermodynamics from eq. (5.2). First, we multiply eq. (5.2) by the potential $V(x)$ and integrate over x to obtain the time evolution of the internal energy $U = \int \mathrm{d}x\ V(x)p(x,t)$:

$$\begin{aligned} -\dot{U} = &\int \mathrm{d}x\ j(x,t)F(x) + m \int \mathrm{d}x\ r(x)p(x,t)\,[V(x) - \langle V\rangle_{\rho_{nb}}] \\ &+ \Lambda_p(t)U - (m-1)\int \mathrm{d}x\ r(x)p(x,t)V(x)\,, \end{aligned} \tag{5.4}$$

where we have introduced the newborn position distribution

$$\rho_{nb}(x,t) = \frac{\int \mathrm{d}x'\ \Sigma(x|x')r(x')p(x',t)}{\int \mathrm{d}x'\ r(x')p(x',t)}\,, \tag{5.5}$$

defined as the ratio of the rate of birth of new particles at position x to the rate of birth of the total number of newborn particles. The notation $\langle\cdot\rangle_{\rho_{nb}}$ indicates the average value with respect to the distribution ρ_{nb}, while by default, averages values, variances and covariances are implicitly computed with the probability distribution p.

Like in the introduction to stochastic thermodynamics in section 2 of chapter 1, we identify the first term as the rate at which heat is transferred from the system to the thermostat:

$$\dot{Q} = \int \mathrm{d}x \; j(x,t)F(x) \,. \tag{5.6}$$

The second term is the rate at which work is extracted from the system due to the resetting of particles to their new positions:

$$\dot{W}_{\mathrm{rst}} = m \int \mathrm{d}x \; r(x)p(x,t) \left[V(x) - \langle V \rangle_{\rho_{nb}}\right] \,. \tag{5.7}$$

The last two terms are interpreted as the work extraction rate from the system due the branching of particles. They can be written more explicitly by using the expression of the population growth rate as the average value of the division rate: $\Lambda_p(t) = (m-1)\int \mathrm{d}x \; r(x)p(x,t)$ (see eq. (1.65), which is still valid for eq. (5.2) involving a diffusive term, because of the vanishing boundary conditions for current j):

$$\begin{aligned} \dot{W}_{\mathrm{brc}} &= \Lambda_p(t)U - (m-1)\int \mathrm{d}x \; r(x)p(x,t)V(x) && (5.8)\\ &= -(m-1)\mathrm{Cov}(r,V) \,. && (5.9) \end{aligned}$$

This contribution is null if $r(x)$ and $V(x)$ are independent, which is trivially the case when at least one of the two functions is constant. Indeed, the average internal energy is not affected by the apparition of newborn particles at any position if the energy landscape $V(x)$ is flat; nor if the branching rate is constant, so that branching affects equally all particles regardless of their positions and thus has no impact on $p(x,t)$.

Finally, the first law for branching processes with relative resetting reads:

$$-\dot{U} = \dot{Q} + \dot{W}_{\mathrm{rst}} + \dot{W}_{\mathrm{brc}} \,, \tag{5.10}$$

where we count positively work extracted from the system and heat dissipated into the environment. Note that unlike in section 2 of chapter 1, there is no external protocol λ here, and thus no work associated to it. On the other, the changes in energy due to branching and resetting are categorized as works because they are not due to the interactions with the heat bath.

The non-equilibrium entropy of the system is defined by the following Shannon entropy:

$$S_{\mathrm{sys}} = -\int \mathrm{d}x \; p(x,t)\ln p(x,t) \,. \tag{5.11}$$

To derive the second law, we take the time derivative of this entropy and use eq. (5.2), which gives:

$$\begin{aligned} \dot{S}_{\mathrm{sys}} = &-\int \mathrm{d}x \; \frac{j(x,t)\partial_x p(x,t)}{p(x,t)} - \Lambda_p(t)S_{\mathrm{sys}} - (m-1)\int \mathrm{d}x \; r(x)p(x,t)\ln p(x,t) \\ &+ m\int \mathrm{d}x \; r(x)p(x,t)\left[\ln p(x,t) - \langle \ln p \rangle_{\rho_{nb}}\right] \,, \end{aligned} \tag{5.12}$$

where the term due to the current $j(x,t)$ can be split into two contributions if the temperature is non-zero (the case $T = 0$ is discussed in section 2.4):

$$-\int \mathrm{d}x\, \frac{j(x,t)\partial_x p(x,t)}{p(x,t)} = \int \mathrm{d}x\, \frac{j^2(x,t)}{Dp(x,t)} - \int \mathrm{d}x\, \frac{\mu j(x,t)F(x)}{D}\,. \tag{5.13}$$

We identify four contributions to the rate of change in the entropy of the system, respectively the entropy production rate due to the heat exchange with the thermostat $\dot{S}_\mathrm{m}$, the entropy production rate of non-equilibrium current $j(x)$: $\dot{S}_\mathrm{c}$, the branching entropy production rate $\dot{S}_\mathrm{brc}$ and the resetting entropy production rate $\dot{S}_\mathrm{rst}$:

$$\dot{S}_\mathrm{m} = \frac{\dot{Q}}{T} = \int \mathrm{d}x\, \frac{\mu j(x,t)F(x)}{D}\,, \tag{5.14}$$

where we used Einstein's relation $D = \mu T$,

$$\dot{S}_\mathrm{c} = \int \mathrm{d}x\, \frac{j^2(x,t)}{Dp(x,t)} \geq 0 \tag{5.15}$$

$$\dot{S}_\mathrm{brc} = -\Lambda_p(t) S_\mathrm{sys} - (m-1)\int \mathrm{d}x\; r(x)p(x,t)\ln p(x,t) \tag{5.16}$$

$$= -(m-1)\mathrm{Cov}(r, \ln p) \tag{5.17}$$

$$\dot{S}_\mathrm{rst} = m\int \mathrm{d}x\; r(x)p(x,t)\left[\ln p(x,t) - \langle \ln p\rangle_{\rho_{nb}}\right]\,. \tag{5.18}$$

Using the positivity of the entropy production rate of non-equilibrium currents, the second law for branching processes with relative resetting in steady state ($\dot{S}_\mathrm{sys} = 0$) reads:

$$\dot{S}_\mathrm{rst} + \dot{S}_\mathrm{brc} \leq \dot{S}_\mathrm{m} \tag{5.19}$$

The first and second laws we derived reduce to the ones obtained in Fuchs et al., 2016 if there is no branching and if the particle is reset to a fixed position x_0. Indeed, setting $m = 1$ and thus $\Lambda = 0$, leads to $\dot{W}_\mathrm{brc} = 0$ and $\dot{S}_\mathrm{brc} = 0$, and therefore eq. (5.10) and eq. (5.19) read respectively $-\dot{U} = \dot{Q} + \dot{W}_\mathrm{rst}$ and $\dot{S}_\mathrm{rst} \leq \dot{S}_\mathrm{m}$. In our framework, absolute resetting to fixed position x_0 is obtained by setting $\Sigma(x|x') = \delta(x - x_0)$, then $\rho_{nb}(x) = \delta(x - x_0)$ and thus $\langle V\rangle_{\rho_{nb}} = V(x_0)$ in eq. (5.7) and $\langle \ln p\rangle_{\rho_{nb}} = \ln p(x_0)$ in eq. (5.18), in agreement with Fuchs et al., 2016.

Instead, when there is branching but no resetting, i.e. when particles randomly multiply and then continue to diffuse from the same position, the transition probability is given by $\Sigma(x|x') = \delta(x - x')$, and thus $\rho_{nb}(x,t) = r(x)p(x,t)/\int \mathrm{d}x'\; r(x')p(x',t)$. In this case, $\int \mathrm{d}x\; r(x)p(x,t)V(x) = \int \mathrm{d}x\; r(x)p(x,t)\langle V\rangle_{\rho_{nb}}$ in eq. (5.7), leading to $\dot{W}_\mathrm{rst} = 0$, and similarly $\dot{S}_\mathrm{rst} = 0$. Without resetting, the first and second laws finally read $-\dot{U} = \dot{Q} + \dot{W}_\mathrm{brc}$ and $\dot{S}_\mathrm{brc} \leq \dot{S}_\mathrm{m}$ respectively.

2.3 Alternative form of the second law

To derive the second law in the previous section, we decomposed eq. (5.13) into two terms: a rate of change in medium entropy $\dot{S}_{\rm m}$ due to the heat exchange with surrounding heat bath, and a positive entropy production rate $\dot{S}_{\rm c}$ due to non-equilibrium currents.

Alternatively, another decomposition is obtained by replacing the current $j(x)$ by its definition:

$$-\int \mathrm{d}x\ \frac{j(x,t)\partial_x p(x,t)}{p(x,t)} = \mu \int \mathrm{d}x\ p(x,t)\partial_x F(x) + \mu T \int \mathrm{d}x\ p(x,t)\left(\partial_x \ln p(x,t)\right)^2\ , \quad (5.20)$$

which leads to:

$$\dot{S}_{\rm sys} = \dot{S}_{\rm brc} + \dot{S}_{\rm rst} + \dot{S}_{\rm fd} + \mu T \int \mathrm{d}x\ p(x,t)\left(\partial_x \ln p(x,t)\right)^2\ , \quad (5.21)$$

where we introduced the average force divergence $\dot{S}_{\rm fd} = \mu\langle \partial_x F\rangle$.

To establish a closer connection with the discussion of the athermal case (see next section), it is useful to combine the last term in eq. (5.21) with $\dot{S}_{\rm fd}$ by using the notion of entropic force, $F_{\rm ent}(x,t) = -T\partial_x \ln p(x,t)$. With this, one can introduce a generalized force $\tilde{F}(x,t) = F(x) + F_{\rm ent}(x,t) \equiv F(x) - T\partial_x \ln p(x,t)$, and the corresponding generalized average force divergence contribution to the entropy production rate, $\dot{\tilde{S}}_{\rm fd} = \mu\langle \partial_x \tilde{F}\rangle$. Note that we have:

$$\begin{aligned}
\mu T \int \mathrm{d}x\ p(x,t)\left(\partial_x \ln p(x,t)\right)^2 &= -\mu \int \mathrm{d}x\ p(x,t)\partial_x \ln p(x,t)\ F_{\rm ent}(x,t) \\
&= -\mu \int \mathrm{d}x\ \partial_x p(x,t)\ F_{\rm ent}(x,t) \\
&= \mu \int \mathrm{d}x\ p(x,t)\partial_x F_{\rm ent}(x,t) \\
&\equiv \mu\langle \partial_x F_{\rm ent}\rangle, \qquad (5.22)
\end{aligned}$$

which leads to the result

$$\dot{S}_{\rm fd} + \mu T \int \mathrm{d}x\ p(x,t)\left(\partial_x \ln p(x,t)\right)^2 = \mu\langle \partial_x [F + F_{\rm ent}]\rangle = \dot{\tilde{S}}_{\rm fd}. \quad (5.23)$$

We then rewrite eq. (5.21) in the following equivalent form:

$$\dot{S}_{\rm sys} = \dot{S}_{\rm brc} + \dot{S}_{\rm rst} + \dot{\tilde{S}}_{\rm fd}. \quad (5.24)$$

Using the positivity of the last term in the right hand side of eq. (5.21), we obtain an alternative version of the second law in steady state:

$$\dot{S}_{\rm brc} + \dot{S}_{\rm rst} \leq -\dot{S}_{\rm fd}\,. \quad (5.25)$$

The two versions of the second law eqs. (5.19) and (5.25) provide different bounds for the entropy production rate due to branching and resetting, and by combining them we have:

$$\dot{S}_{\rm brc} + \dot{S}_{\rm rst} \leq \min\left(-\dot{S}_{\rm fd}, \dot{S}_{\rm m}\right)\,. \quad (5.26)$$

In the limit of large temperatures, the current $j(x,t) = \mu F(x)p(x,t) - D\partial_x p(x,t)$ becomes purely diffusive $j(x,t) \simeq -D\partial_x p(x,t)$, and as a result, the medium entropy $\dot{S}_\mathrm{m} = \mu\langle \partial_x F\rangle = \dot{S}_\mathrm{fd}$. Thus, in this limit the upper bound reads $\dot{S}_\mathrm{brc} + \dot{S}_\mathrm{rst} \leq -|\dot{S}_\mathrm{fd}|$. The behavior in the opposite limit of vanishing temperature is examined in more details in the next section.

2.4 Athermal systems

Until now, we considered particles in contact with a heat bath, which allowed us to use standard stochastic thermodynamics. However, athermal systems where $T = 0$ are important to describe situations where particles move deterministically between branching events. In that case, there is no diffusion, only the deterministic force $F(x)$ is present, but branching and resetting events are still stochastic. Such situations are biologically relevant as we discuss in section 3.

The first law is still mathematically valid, even if the current $j(x,t) = \mu F(x)p(x,t)$ is only convective. We find $\dot{Q} = \mu\langle F^2\rangle$, and we call the corresponding quantity Q an athermal heat, although it is important to emphasize that this quantity cannot be interpreted as an exchange of heat with a thermostat.

The second law in the form of section 2.2 is no longer valid, since $\dot{S}_\mathrm{m}$ and $\dot{S}_\mathrm{c}$ diverge in the case $T = 0$. However, the other decomposition of the Shannon entropy (section 2.3) remains defined because the last term in eq. (5.21) (entropic contribution) tends to 0 as $T \to 0$, since the integral is not singular in practice. We propose to call the following equality

$$\dot{S}_\mathrm{sys} = \dot{S}_\mathrm{brc} + \dot{S}_\mathrm{rst} + \dot{S}_\mathrm{fd} \tag{5.27}$$

the second law for athermal systems (despite the absence of a corresponding inequality) because it corresponds to the $T = 0$ version (i.e., in absence of entropic forces) of eq. (5.24).

3 Application to models of cell size control

As stated in the introduction, branching processes with resetting appear in biology, for example in the context of cell division. Indeed, when cells divide they give birth to m daughter cells (usually 2), and many cell properties are affected by division. Cell division provides examples of absolute resetting, for instance age which is reset at 0 for both daughter cells independently of the age of the dividing mother cell, and of relative resetting, such as for the volume or the number of proteins for example, that are split at division between the two daughter cells.

In this section, we use the results from section 2 in the context of growing colonies of cells, focusing in particular on the three most common division strategies discussed in section 4.1 of chapter 1: the sizer, the timer and the adder.

3.1 Sizer

Size-controlled populations are described by the Fokker-Planck equation eq. (5.2), where the position x of the Brownian particle is understood as the cell size. In a rich medium,

most bacteria follow an exponential growth between divisions: $\dot{x} = \nu x$, with ν the single-cell growth rate which is non-fluctuating in our model and strictly positive. If this single-cell growth can be noisy as discussed in section 5 of chapter 4, these fluctuations are very often ignored in the literature. We also ignore them here, as they do not play a role in the thermodynamics of cell division, which is the main focus of this section. Therefore, the cell population is considered an athermal system. The above law of exponential growth is a Langevin equation with unit mobility $\mu = 1$, and a force $F(x) = \nu x$ deriving from a potential $V(x) = -\nu x^2/2$, which is non-confining, unlike the example shown in fig. 5.1. Let us now comment on immediate consequences of the sizer hypothesis for the thermodynamics quantities.

First, $V(x)$ is a decreasing function of x while the size regulation imposes that the division rate $r(x)$ be an increasing function of x. Thus, the covariance between those two functions is negative, which results in a positive production of work due to branching by eq. (5.9): $\dot{W}_{\rm brc} \geq 0$. In contrast, the decrease of $V(x)$ with x leads to a negative production of work due to resetting. This can be seen easily from eq. (5.7) when replacing the average value computed with the newborn distribution by its definition:

$$\dot{W}_{\rm rst} = m \int \mathrm{d}x \; r(x) p(x,t) \left[V(x) - \int \mathrm{d}x' \; \Sigma(x'|x) V(x') \right] , \tag{5.28}$$

where the term in the bracket is the difference between the potential for a cell of given size x and the average value of the potential for a new cell, born from the division of a cell of size x. Due to the non-confining shape of potential $V(x)$, the restarting size x has a higher internal energy than dividing sizes x', thus this difference is negative for any x and $\dot{W}_{\rm rst} \leq 0$. Note that with the sign convention chosen in the first law eq. (5.10), $\dot{W}_{\rm rst} \leq 0$ corresponds indeed to an increase of the internal energy of the population because of resetting.

Second, athermal systems are covered by section 2.4 and the steady-state second law equality reads

$$\Lambda = -\dot{S}_{\rm rst} - \dot{S}_{\rm brc} , \tag{5.29}$$

where we used that $\dot{S}_{\rm fd} = \langle \partial_x F \rangle = \nu$, and the property that in a steady-state, the population growth rate Λ must be equal to the single cell growth rate ν (eq. (4.18)). Moreover, the rate of athermal heat introduced in section 2.4 is now given by $\dot{Q} = \nu^2 \langle x^2 \rangle$.

Third, when following a single lineage of cells, for example in a mother machine setup (Wang et al., 2010), then $\dot{S}_{\rm brc} = 0$. As a result, the second law in steady-state (eq. (5.29)) reduces to $\dot{S}_{\rm rst} = -\nu$. Interestingly, in this case, the entropy production rate due to resetting depends only on the single cell growth rate, but not on the division rate $r(x)$ nor on the kernel Σ.

Finally, the signs of the branching and resetting entropy production rates can be obtained with reasonable assumptions. We know from the mathematical literature (Hall et al., 1990) and from chapter 4 that cell size distributions are analytical only in some simple cases. Even in those cases, they take the form of infinite series and computing the values of the branching and resetting entropy production rates for these distributions is out of reach for the moment. To make further progress, we use a log-normal ansatz for the population size distribution, which, even though it is not rigorously a solution of eq. (5.2),

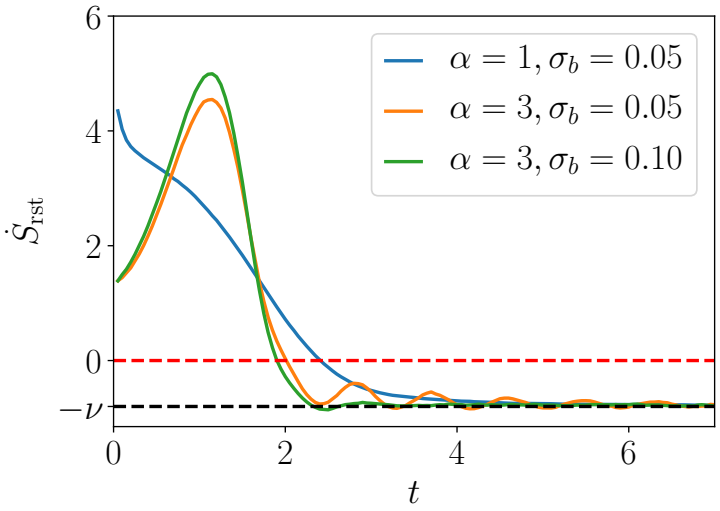


Figure 5.2: Time evolution of the entropy production rate due to resetting $\dot{S}_{\mathrm{rst}}$, in the case of a size-control mechanism with rate $r(x) = x^\alpha$, deterministic growth between division at a rate $\nu = 0.8$ and Gaussian kernel $\Sigma(\cdot|x') = \mathcal{N}(x'/2, (\sigma_b x')^2)$ for the partition of volume between the $m = 2$ daughter cells at division. Each curve corresponds to a choice of parameters $(\alpha,\ \sigma_b)$, and is the result of 100000 single lineage simulations (without branching so $\dot{S}_{\mathrm{brc}} = 0$), starting at size $x_0 = 0.5$. All curves converges to the black dashed line at $\dot{S}_{\mathrm{rst}} = -\nu$ in the long time limit.

is known to be a good fit of experimental data as long as the transition probability Σ is peaked around symmetric division (see Hosoda et al., 2011 for a data collapse of the log size distributions from various datasets on a Gaussian function). We also assume that the division rate is well captured by a power law, as discussed in section 4.2 of chapter 1. Assuming that $p = \mathrm{Lognormal}(\mu, \sigma^2)$ and $r(x) = x^\alpha$, we show in appendix B that the covariance in eq. (5.17) is computable and given by:

$$\dot{S}_{\mathrm{brc}} = (m-1)\alpha\sigma^2 \left(1 + \frac{\alpha}{2}\right) \exp\left[\alpha\mu + \frac{\alpha^2\mu^2}{2}\right] \geq 0\,, \tag{5.30}$$

which is positive regardless of the values of $\alpha \geq 0$, μ and σ. Using this inequality, eq. (5.29) imposes that $\dot{S}_{\mathrm{rst}} \leq 0$, meaning that resetting is a way to avoid extremely large cells and thus to reduce the Shannon entropy of the size distribution at steady-state. However, this constraint no longer exists in the transient dynamics, where resetting can increase the Shannon entropy of the size distribution: $\dot{S}_{\mathrm{rst}} \geq 0$.

To illustrate the last two points, we show on fig. 5.2 the evolution of $\dot{S}_{\mathrm{rst}}$ with time. We simulated 10^5 independent single lineages ($\dot{S}_{\mathrm{brc}} = 0$) each starting with a cell of initial size $x_0 = 0.5$ growing at a rate $\nu = 0.8$. We chose a power law for the division rate $r(x) = x^\alpha$ and a transition probability $\Sigma(x|x') = b(x/x')/x'$ depending on the ratio of the daughter to mother volumes through the Gaussian distribution $b = \mathcal{N}(1/2, \sigma_b^2)$, so that $\Sigma(\cdot|x') = \mathcal{N}(x'/2, (\sigma_b x')^2)$. For three different couples (α, σ_b), the curves exhibit a positive region in the transient dynamics (above the red dashed line), corresponding to a

widening of the distribution of sizes due to resetting events from regions of high to low probabilities in size space, and a resulting increase of the Shannon entropy. In the long time limit, they all converge to $-\nu$, independently of α and σ_b.

3.2 Timer

The timer mechanism is not described by eq. (5.2), but by a similar equation for which the source term (apparition of new cells) enters as a boundary condition, as detailed in section 4.1.1 of chapter 1. We recall the dynamical equation for the (backward/population) distribution of ages:

$$\partial_t p(a,t) = -\partial_a p(a,t) - [\Lambda_p(t) + r(a)]\, p(a,t) \tag{5.31}$$

$$p(0,t) = m \int \mathrm{d}a\; r(a) p(a,t) = \frac{m}{m-1}\, \Lambda_p(t)\,. \tag{5.32}$$

The boundary condition appears separated because age is defined on $[0,\infty[$ and the resetting age $a = 0$ is on the boundary of the domain. This is the first difference with the case treated in Fuchs et al., 2016, in which the resetting position x_0 was inside the domain of definition of x. The second difference is that age is by definition the time elapsed since birth and cannot undergo thermal fluctuations, therefore its dynamics between divisions is deterministic, which corresponds to a temperature $T = 0$, a mobility $\mu = 1$ and a force $F(a) = 1$, deriving from a potential $V(a) = -a$. The signs of the different contributions to the first law are the same as those found for the sizer case. Indeed, because the potential $V(a)$ is decreasing with age, $\dot{W}_{\mathrm{rst}} \leq 0$, and since $r(a)$ is generally an increasing function of age for non extreme-ages (Robert et al., 2014), the covariance between V and r is negative, and so $\dot{W}_{\mathrm{brc}} \geq 0$. We also find that $\dot{Q} = 1 \geq 0$.

Let us now derive the time-evolution of the Shannon entropy of the age distribution. We multiply eq. (5.31) by $-\ln p(a,t)$ and integrate over a:

$$\dot{S}_{\mathrm{sys}} = p(0,t) - p(0,t)\ln p(0,t) - \Lambda_p(t) S_{\mathrm{sys}} + \int \mathrm{d}a\; r(a) p(a,t) \ln p(a,t)\,, \tag{5.33}$$

where we used an integration by part and the boundary condition $\lim_{a\to\infty} p(a,t) = 0$. We now use the boundary condition eq. (5.32) to express $p(0,t)$:

$$\begin{aligned}\dot{S}_{\mathrm{sys}} =& \frac{m\Lambda_p(t)}{m-1} - \Lambda_p(t) S_{\mathrm{sys}} - (m-1)\int \mathrm{d}a\; r(a)p(a,t)\ln p(a,t) \\ &+ m\int \mathrm{d}a\; r(a)p(a,t)\ln\left(\frac{p(a,t)}{p(0,t)}\right)\,,\end{aligned} \tag{5.34}$$

where we identify

$$\dot{S}_{\mathrm{brc}} = -\Lambda_p(t) S_{\mathrm{sys}} - (m-1)\int \mathrm{d}a\; r(a)p(a,t)\ln p(a,t)\,, \tag{5.35}$$

$$= -(m-1)\mathrm{Cov}(r, \ln p)\,, \tag{5.36}$$

and

$$\dot{S}_{\mathrm{rst}} = m\int \mathrm{d}a\; r(a)p(a,t)\ln\left(\frac{p(a,t)}{p(0,t)}\right)\,. \tag{5.37}$$

Finally, the steady-state second law for the timer reads:

$$\frac{m\Lambda}{m-1} = -\dot{S}_{\mathrm{rst}} - \dot{S}_{\mathrm{brc}}\,. \tag{5.38}$$

Note that unlike what happened for the sizer, the term proportional to Λ does not arise from the force divergence, which is null in this case (since $F(a) = 1$), but from the boundary condition eq. (5.32).

Using the analytical steady-state solution to eq. (5.31) given in section 4.5 of chapter 1, we can compute explicitly the resetting and branching entropy production rates:

$$\dot{S}_{\mathrm{rst}} = m\Lambda\left[S_{\mathrm{sys}} + \ln\left(\frac{m\Lambda}{m-1}\right) - \frac{m}{m-1}\right] \tag{5.39}$$

$$\dot{S}_{\mathrm{brc}} = -m\Lambda\left[S_{\mathrm{sys}} + \ln\left(\frac{m\Lambda}{m-1}\right) - 1\right]\,, \tag{5.40}$$

which are functions of Λ and S_{sys}, themselves only depending on the branching rate $r(a)$. Moreover, the steady-state age distribution is a decreasing function of age, which implies that (i) the branching entropy production rate is positive: $\dot{S}_{\mathrm{brc}} \geq 0$ by eq. (5.36), (ii) the resetting entropy production rate is negative: $\dot{S}_{\mathrm{rst}} \leq 0$ by eq. (5.37).

3.3 Adder

In the adder theory, the distribution of added volume between birth and death is independent of the volume at birth (see section 4.1.3 of chapter 1). Two variables are required to model the adder: for example the size x, which undergoes relative resetting, and the volume added since birth $\Delta = x - x_b$, which undergoes absolute resetting to value 0 at division, similarly to the age in the timer. In this sense, the increment of volume can be seen as a physiological age.

Conducting the same analysis as before on eq. (1.54), we derive the equation for the evolution of the Shannon entropy $S_{\mathrm{sys}} = -\int \mathrm{d}x\mathrm{d}\Delta\; p(x,\Delta,t)\ln p(x,\Delta,t)$ of the joint distribution $p(x,\Delta,t)$ of volume and added volume:

$$\dot{S}_{\mathrm{sys}} = \dot{S}_{\mathrm{rst}} + \dot{S}_{\mathrm{brc}} + \dot{S}_{\mathrm{fd}} + \frac{m\Lambda_p(t)}{m-1}\,. \tag{5.41}$$

As for the sizer, if we consider exponential growth: $F(x) = \nu x$, then $\dot{S}_{\mathrm{fd}} = \nu$, and in steady state $\Lambda = \nu$. Finally, the steady-state second law equality for the adder reads

$$\frac{2m-1}{m-1}\,\Lambda = -\dot{S}_{\mathrm{rst}} - \dot{S}_{\mathrm{brc}}\,. \tag{5.42}$$

The left hand side is the sum of the contributions $m\Lambda/(m-1)$ (similar to that of the timer), coming from the boundary condition at $\Delta = 0$, and Λ (similar to that of the sizer), arising from the force-divergence entropy due to the growth of cells at a rate ν (equal to Λ in steady-state). Note that since the adder is a multivariate model with two variables, one could expect the force-divergence entropy to be a sum of two terms, as detailed in

appendix A. However, the force F here is supposed to be a function of the size x only and independent of the added volume Δ, thus $\partial_\Delta F(x) = 0$.

In the spirit of the last part of appendix A, one can coarse-grain the variable Δ by integrating eq. (4.36) on Δ, leading to an equation of the form of a sizer (eq. (5.2)) with marginal probability $\hat{p}(x,t) = \int \mathrm{d}\Delta\ p(x,\Delta,t)$ and coarse-grained branching rate $\hat{r}(x,t) = F(x)\int \mathrm{d}\Delta\, p(\Delta,t|x)\zeta(\Delta)$. Thus, the steady-state second law for the marginal size distribution obeys eq. (5.29) with the coarse-grained branching rate. Without branching, this second law reduces to $\dot{S}_{\mathrm{rst}} = -\nu$. In the third remark of section 3.1 we already noted that for single lineages the entropy production rate due to resetting was independent of the division rate, we see now that is also independent of the size control mechanism, namely sizer or adder.

3.4 Analogy with heat engines

For each of the division-control models studied above, we obtained a second law in the form of $\beta\Lambda = -\dot{S}_{\mathrm{rst}} - \dot{S}_{\mathrm{brc}}$ (eqs. (5.29), (5.38) and (5.42)), with β an integer equals to 1 for the sizer, 2 for the timer, and 3 for the adder for cells obeying binary fission ($m = 2$). The form of this common second law suggests the definition of the efficiency

$$\eta = \frac{-\dot{S}_{\mathrm{brc}}}{\dot{S}_{\mathrm{rst}}}\,. \tag{5.43}$$

This definition is inspired by thermodynamic machines, which operate with a driving process and an output process, respectively associated with the entropy production rates $\sigma_1 \geq 0$ and $\sigma_2 \leq 0$. In that case, the thermodynamic efficiency reads $\eta = -\sigma_2/\sigma_1 \geq 0$. The second law $\sigma_{tot} = \sigma_1 + \sigma_2 \geq 0$, where σ_{tot} is the total entropy production rate, further implies that $\eta \leq 1$. In our case, despite the absence of a thermostat and therefore the absence of a first law, by analogy, $\beta\Lambda \geq 0$, $-\dot{S}_{\mathrm{rst}}$ and $-\dot{S}_{\mathrm{brc}}$ play the roles of σ_{tot}, σ_1 and σ_2, respectively. Indeed, we proved for the sizer and the timer that $-\dot{S}_{\mathrm{rst}}$ and $-\dot{S}_{\mathrm{brc}}$ have the same sign as σ_1 and σ_2, respectively. This analogy in which resetting is the driving process that enables branching, which is the output process, can also be understood intuitively at the level of energies. Populations of cells thrive by dividing, a process for which the creation of a new cell is made possible by the size reduction of both the mother cell and the newborn cell. Indeed, we proved for both the sizer and timer that $\dot{W}_{\mathrm{rst}} \leq 0$ and $\dot{W}_{\mathrm{brc}} \geq 0$, which fundamentally comes from the non-confining shapes of the potentials $V(x)$ and $V(a)$. This implies that branching has an energetic cost for the colony, which is covered by the energetic gain due to resetting.

For the timer, we proved that $\dot{S}_{\mathrm{rst}}$ and $\dot{S}_{\mathrm{brc}}$ are only functions of the branching rate $r(a)$, which we describe by a power law $r(a) = a^\alpha$. Thus the strength α of the age control is the only parameter in the model. On fig. 5.3 (left), we plot the evolution of the Shannon entropy of the age distribution, the population growth rate and the efficiency against α. The population growth rate Λ is obtained by plugging the steady-state solution eq. (1.75) into the boundary condition eq. (5.32) and solving numerically for Λ. Knowing Λ, the Shannon entropy of the age distribution is numerically evaluated, and so is the efficiency using eqs. (5.39) and (5.40). The efficiency is an increasing function of α, and converges

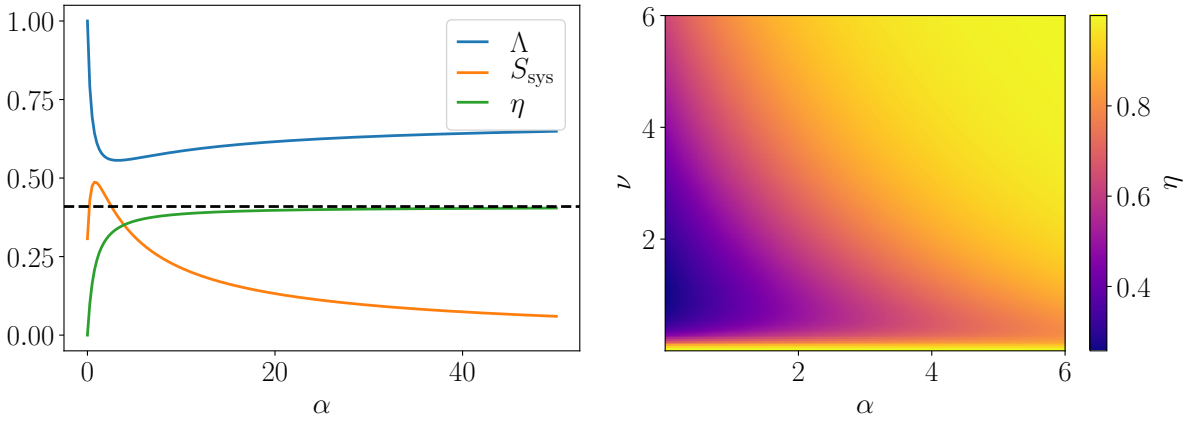


Figure 5.3: Numerical evaluation of the efficiency η when varying parameters of the models for the timer (left) and sizer (right) when $m = 2$. Left: When increasing the strength α of the age control, abnormally old cells disappear and the age distribution narrows around young cells, leading to a decrease in the Shannon entropy (orange). The efficiency (green) monotonously increases with α, up to the limit $\ln 2/(1 + \ln 2)$ (black dashed line). Right: For any ν ($= \Lambda$), the efficiency is a monotonously increasing function of the size control α, up to the maximum value 1. For any α, the efficiency tends to 1 in both the $\nu \to 0$ (Carnot) and $\nu \to \infty$ limits, with a single minimum in between.

to the asymptotic value computed in appendix C:

$$\lim_{\alpha \to +\infty} \eta = \frac{\ln 2}{1 + \ln 2} \approx 0.41 \,, \tag{5.44}$$

as the strength of the control increases, leading to a synchronized population where all cells deterministically divide at age 1.

More parameters are required to describe the sizer: the strength α of the control in the branching rate $r(x) = x^\alpha$, and also the single cell growth rate ν and the parameters of the kernel Σ. Here, we illustrate our result in the case of symmetric division, thus our model has two parameters: α and ν. We plot on fig. 5.3 (right) the efficiency against α (x-axis) and ν (y-axis). We compute the branching entropy production rate using eq. (5.30) and the approximate parameters $\mu = \ln \nu/\alpha - \ln 2/4$ and $\sigma^2 = \ln 2/2\alpha$, that can be obtained following the method proposed in Hosoda et al., 2011 as detailed in appendix D. For any ν, the efficiency is an increasing function of α, and converges to 1, the maximal efficiency. This comes from the fact that Λ is independent of α, and $\dot{S}_{\text{brc}} \to \infty$ as $\alpha \to \infty$ from eq. (5.30). For any α, the efficiency starts at 1 when $\nu = 0$ (analogous to Carnot efficiency when the entropy production is null), decreases until reaching a minimum and then tends to 1 (obtained from the rate of increase of $\dot{S}_{\text{brc}}$ with ν) as ν is varied from 0 to ∞.

For both the timer (fig. 5.3 left) and the sizer (not shown here), the Shannon entropy decreases as the strength of control measured by α is increased in the region where α is large. This means that the diversity in the controlled trait is reduced across the population. We suspect that such a lack of diversity might be harmful for the population at some level, which could be one of the reasons why cells need not implement such a strong

and efficient (in the sense of η) control.

We remind the reader that the two plots cannot be compared directly, in order to find the most efficient control strategy for example, since the efficiencies plotted are related to different distributions. Indeed, the efficiency for the timer is the ratio of the branching to resetting entropy production rates associated with the age distribution, versus the size distribution for the sizer.

4 Conclusion

The stochasticity in the thermodynamics of small systems usually comes from thermal fluctuations, which are suppressed when considering macroscopic systems. Such systems are described by classical thermodynamics instead, where observables are equal to their average values. We show however that using the framework of stochastic thermodynamics for macroscopic systems with stochastic events, such as division, can shed a new light on the thermodynamic constraints of the systems. In our case, the stochasticity of the thermal fluctuations is replaced by that of the divisions.

By describing cell growth and division in terms of two subprocesses, branching and resetting, we find that resetting is a process which rises the internal energy of the system, thereby allowing the system to pay the energy cost associated with branching. Branching is required by cells to self-replicate, and in doing so, to maintain certain traits within lineages, an essential feature for the survival of the cell colony. In exponentially growing colonies, the population growth rate emerges from a carefully controlled balance between resetting and branching. We introduce an efficiency, akin to the thermodynamic efficiency of thermal machines, which quantifies the entropy conversion between these two processes.

With recent progresses on calorimetric measurements at the level of the single cell (Rodenfels et al., 2019; Song et al., 2019; Hong et al., 2020), it is natural to ask if the thermodynamic quantities we defined in this chapter are linked to those obtained in experiments. Trying to map the two lines of work would be a fascinating project in the future, but we can already give some comments. First, the models of cell size control we considered are coarse-grained and described by one or two variables only, therefore we do not expect them to offer a good description of the true thermodynamics of the cell a priori. As explained in the introduction on stochastic thermodynamics in section 2 of chapter 1, the mesoscopic description of a system results from the coarse-graining of equilibrium degrees of freedom that do not participate in the production of entropy. This is the reason why we extended our formalism to the case of n-variable models in appendix A. Our formulas are thus applicable to much more complex models than the three cell size control models we used as illustrations. The challenge is then to determine which non-equilibrium bio-molecular mechanisms and variables are relevant and sufficient to describe the thermodynamic state of a cell. Moreover, the fact that the signs of the works and entropies associated with branching and resetting are identical for the timer and the sizer may suggest a conversion mechanism between resetting and branching that is general beyond these two simple models. Second, the discussion on the efficiency of cell division is based on an athermal description, and is essentially an information-theoretic result. It tells us how expanded or compressed are the age and size distributions because

of cell division, and should therefore remain relevant regardless of any consideration on the correspondence between our model and thermal measurements.

The present work could be extended in several future directions. The assumption that the colony is in an exponentially growing phase could be relaxed, and additional sources of stochasticity affecting single-cell growth rate could be included (Thomas et al., 2018). To improve the description of the adder model, it would be interesting to obtain analytical solutions for simple choices of the resetting rate. Recently, several experimental works have suggested that the adder model can be justified microscopically using incremental threshold models (Pandey et al., 2020). Along the same line, some recent works inspired by Nieto et al., 2020 proposed a unified model for cell size distributions accounting for the sizer, timer and adder behaviors, using an N-stages description of the cell cycle, which is applicable to bacterial exponential growth (Jia et al., 2021) and to yeast growth (Jia et al., 2022). We note that the incremental model and the N-stages model both fall into the class of resetting models studied here, and for that reason it would be interesting to adapt our approach to these more recent models.

In the three mechanisms of cell size control we studied, the opposite of the sum of the resetting and branching entropy production rates equals the population growth rate times a prefactor which seems to depend on the number of key variables of the model. Indeed, this factor is one for the sizer, two for the timer and three for the adder. In principle, the prefactor could take other values depending on the number of variables entering in the growth function. Interestingly, this prefactor might therefore give insight into a particular mechanism of control and possibly be related to the latent variables, which are detected in Bayesian approaches of lineage data (Nakashima et al., 2020).

5 Appendices

A Multi-dimensional systems

For simplicity, in the main text we presented our results in one dimension, but we show in this appendix that they hold in n dimensions. Let $\boldsymbol{x} = (x_1, ..., x_n)$ be a n-dimensional vector, and let the branching rate $r(\boldsymbol{x})$ and the resetting kernel $\Sigma(\boldsymbol{x}|\boldsymbol{x'})$ depend on these n variables. The bold notation is used in this appendix for vectors in order to highlight the differences with scalar quantities from the main text, unlike in the rest of the thesis where bold symbols indicate trajectories. For more generality, and regarding the discussion of section 2.4, we consider that k variables $(x_1, ..., x_k)$ are in contact with different heat baths $\{T_i\}_{i=1,...,k}$, and that $n-k$ variables $(x_{k+1}, ..., x_n)$ are not linked to any heat bath and thus do not undergo thermal fluctuations. Among those $n-k$ athermal degrees of freedom, some are subjected to deterministic forces F, while others stay constant between resetting events ($F = 0$). We label $(x_{k+1}, ..., x_{k+l})$ the l variables for which $F \neq 0$, and $(x_{k+l+1}, ..., x_n)$ the $n-k-l$ variables for which $F = 0$. In this setting, the potential $V(\boldsymbol{x})$ only depends on the $k+l$ first variables, which defines the same number of forces: $F_i(\boldsymbol{x}) = -\partial_{x_i} V(\boldsymbol{x})$. For these degrees of freedom, we define the mobilities μ_i and the currents $j_i(\boldsymbol{x}) = \mu_i F_i(\boldsymbol{x}) p(\boldsymbol{x}) - \mu_i T_i \partial_{x_i} p(\boldsymbol{x})$ for $i = 1, ..., k$, and $j_i(\boldsymbol{x}) = \mu_i F_i(\boldsymbol{x}) p(\boldsymbol{x})$ for $i = k+1, ..., k+l$, and recall that for the variables $(x_{k+l+1}, ..., x_n)$ there is current. For better legibility, we drop the time-dependence of the probabilities in this section, and we define the vectors of currents $\boldsymbol{j}(\boldsymbol{x})$, forces $\boldsymbol{F}(\boldsymbol{x})$, and mobilities $\boldsymbol{\mu}$, whose components are respectively $j_i(\boldsymbol{x})$, $F_i(\boldsymbol{x})$ and μ_i for $i = 1, ..., k+l$.

Finally the equation for the evolution of $p(\boldsymbol{x})$ reads:

$$\partial_t p(\boldsymbol{x}) = -\boldsymbol{\nabla} \cdot \boldsymbol{j}(\boldsymbol{x}) - [\Lambda + r(\boldsymbol{x})]\, p(\boldsymbol{x}) + m \int \mathrm{d}\boldsymbol{x}' \Sigma(\boldsymbol{x}|\boldsymbol{x}') r(\boldsymbol{x}') p(\boldsymbol{x}')\,. \tag{5.45}$$

The first law of thermodynamics is obtained following the same steps as in the 1d case, and when replacing x by $\boldsymbol{x}$, the expressions of $\dot{W}_{\mathrm{rst}}$ (eq. (5.7)) and $\dot{W}_{\mathrm{brc}}$ (eq. (5.9)) are unchanged. The heat is replaced by a sum of heats associated with each degree of freedom:

$$\dot{Q} = \int \mathrm{d}\boldsymbol{x}\ \boldsymbol{j}(\boldsymbol{x}) \cdot \boldsymbol{F}(\boldsymbol{x}) \tag{5.46}$$

$$= \sum_{i=1}^{k} \int \mathrm{d}\boldsymbol{x}\ j_i(\boldsymbol{x}) F_i(\boldsymbol{x}) + \sum_{i=k+1}^{k+l} \int \mathrm{d}\boldsymbol{x}\ \mu_i p(\boldsymbol{x}) F_i^2(\boldsymbol{x})\,, \tag{5.47}$$

where for $i = 1, ..., k$, $\dot{Q}_i = \int \mathrm{d}\boldsymbol{x}\ j_i(\boldsymbol{x}) F_i(\boldsymbol{x})$ is the heat exchange rate with the i^{th} thermostat, and for $i = k+1, ..., k+l$, $\mu_i \langle F_i^2 \rangle$ is the rate of change of the athermal heat discussed in section 2.4, associated with the i^{th} degree of freedom.

The second law also follows from the same calculation and the entropy production rates due to branching (eq. (5.17)) and resetting (eq. (5.18)) are unchanged as compared to the 1d case. The term due to currents can be split into two contributions corresponding

to the thermal and athermal degrees of freedom:

$$
\begin{aligned}
&-\int \mathrm{d}\boldsymbol{x}\ \frac{\boldsymbol{\nabla}(p(\boldsymbol{x}))\cdot\boldsymbol{j}(\boldsymbol{x})}{p(\boldsymbol{x})} \\
&= -\int \mathrm{d}\boldsymbol{x}\ \left[\sum_{i=1}^{k}\frac{\left(F_i(\boldsymbol{x})p(\boldsymbol{x})-\mu_i^{-1}j_i(\boldsymbol{x})\right)j_i(\boldsymbol{x})}{T_i p(\boldsymbol{x})}-\sum_{i=k+1}^{k+l}\mu_i F_i(\boldsymbol{x})\partial_{x_i}p(\boldsymbol{x})\right] \\
&= \sum_{i=1}^{k}\left[\int \mathrm{d}\boldsymbol{x}\left(\frac{j_i^2(\boldsymbol{x})}{\mu_i T_i p(\boldsymbol{x})}-\frac{F_i(\boldsymbol{x})j_i(\boldsymbol{x})}{T_i}\right)\right]+\sum_{i=k+1}^{k+l}\mu_i\int \mathrm{d}\boldsymbol{x}\ p(\boldsymbol{x})\partial_{x_i}F_i(\boldsymbol{x})\,,
\end{aligned}
\tag{5.48}
$$

where $\dot{S}_{\mathrm{c}}^{(i)}=\int \mathrm{d}\boldsymbol{x}\ j_i^2(\boldsymbol{x})/\mu_i T_i p(\boldsymbol{x})\geq 0$ is the entropy production rate due to the non-equilibrium current associated with the degree of freedom i, $\dot{S}_{\mathrm{m}}^{(i)}=\int \mathrm{d}\boldsymbol{x}\ F_i(\boldsymbol{x})j_i(\boldsymbol{x})/T_i=\dot{Q}_i/T_i$ is the entropy exchange rate with thermostat i, and $\dot{S}_{\mathrm{fd}}^{(i)}=\mu_i\int \mathrm{d}\boldsymbol{x}\ p(\boldsymbol{x})\partial_{x_i}F_i(\boldsymbol{x})$ is the force-divergence entropy production rate associated with the athermal degree of freedom i. Finally the second law for n-dimensional systems reads:

$$
\dot{S}_{\mathrm{sys}}=\dot{S}_{\mathrm{brc}}+\dot{S}_{\mathrm{rst}}+\sum_{i=1}^{k}\left(\dot{S}_{\mathrm{c}}^{(i)}-\dot{S}_{\mathrm{m}}^{(i)}\right)+\sum_{i=k+1}^{k+l}\dot{S}_{\mathrm{fd}}^{(i)}\,. \tag{5.49}
$$

Note that with the unification of deterministic and entropic forces, we obtain again eq. (5.24), now with the identification

$$
\dot{\tilde{S}}_{\mathrm{fd}}=-\sum_{i=1}^{k}\mu_i T_i\langle\partial_{x_i}^2\ln p\rangle+\sum_{i=1}^{k+l}\mu_i\langle\partial_{x_i}F_i\rangle\,, \tag{5.50}
$$

encompassing the entropic contribution of the degrees of freedom which are in contact with a thermal reservoir. We note that the contribution of the deterministic forces (last term in the right hand side of eq. (5.50)) takes the form $\sum_{i=1}^{k+l}\mu_i\langle\partial_{x_i}F_i\rangle=\langle\mathrm{div}(\boldsymbol{\mu}\circ\boldsymbol{F})\rangle$, which is the average value of the divergence of the Hadamard product $\boldsymbol{\mu}\circ\boldsymbol{F}$, defined by the components $(\boldsymbol{\mu}\circ\boldsymbol{F})_i=\mu_i F_i$.

The multivariate model is particularly interesting when the dynamics of resetting and branching is controlled by a set of hidden variables, while we have access to only one observable, let us say x_1. In this case, the natural quantity to look at is the Shannon entropy of the marginal distribution of x_1: $\hat{p}(x_1)=\int \mathrm{d}x_2...\mathrm{d}x_n\ p(x_1,...,x_n)$, and we show that the computation of $\dot{S}_{\mathrm{brc}}$ and $\dot{S}_{\mathrm{rst}}$ reduces to a 1d problem with coarse-grained branching rate and resetting kernel. To obtain the evolution of the Shannon entropy associated with $\hat{p}(x_1)$, we integrate eq. (5.45) over $x_2,...,x_n$, multiply it by $\ln\hat{p}(x_1)$ and integrate it over x_1. The contribution of resetting it given by:

$$
\dot{S}_{\mathrm{rst}}=m\int \mathrm{d}\boldsymbol{x}\ r(\boldsymbol{x})p(\boldsymbol{x})\left[\ln\hat{p}(x_1)-\langle\ln\hat{p}\rangle_{\rho_{nb}}\right] \tag{5.51}
$$

$$
=m\int \mathrm{d}x_1\ \hat{r}(x_1)\hat{p}(x_1)\left[\ln\hat{p}(x_1)-\langle\ln\hat{p}\rangle_{\hat{\rho}_{nb}}\right]\,, \tag{5.52}
$$

where we defined the coarse-grained branching rate

$$
\hat{r}(x_1)=\int \mathrm{d}x_2...\mathrm{d}x_n\ p(x_2,...,x_n|x_1)r(x_1,...,x_n)\,, \tag{5.53}
$$

and where the newborn distribution is also marginalized:

$$\hat{\rho}_{nb}(x_1) = \frac{\int \mathrm{d}\boldsymbol{x}'\ \hat{\Sigma}(x_1|\boldsymbol{x}')r(\boldsymbol{x}')p(\boldsymbol{x}')}{\int \mathrm{d}x_1\ \hat{r}(x_1)\hat{p}(x_1)}\,, \tag{5.54}$$

with the coarse-grained kernel $\hat{\Sigma}(x_1|\boldsymbol{x}') = \int \mathrm{d}x_2 ... \mathrm{d}x_n\ \Sigma(x_1, ..., x_n|\boldsymbol{x}')$. Similarly, the branching entropy production rate reduces to

$$\dot{S}_{\mathrm{brc}} = -(m-1)\mathrm{Cov}_{\hat{p}}(\hat{r}, \ln \hat{p})\,. \tag{5.55}$$

Before closing the present discussion on multi-dimensional systems, it is worth mentioning a subtlety of the coarse-grained description introduced above. Note that when reducing a multivariate problem by eliminating degrees of freedom, the dynamics associated to the remaining variables is non-Markovian in general. Thus, as far one is concerned with one-time observables, the above description is correct because the Markovian dynamics that is obtained when integrating out the unobserved degrees of freedom corresponds to a *substitute* Markov process having the same one-time statistics as the underlying marginal non-Markov dynamics (Hänggi et al., 1977; Hänggi et al., 1982; García-García, 2012). However, multi-time observables (e.g., correlation functions) are not well captured by such effective Markov dynamics unless some sort of Markov approximation makes sense. In particular, using the path integral representation of the substitute process to compute generating functionals of the marginal non-Markovian degrees of freedom would be misleading, typically yielding incorrect results.

B Branching entropy production rate for the sizer in steady-state

In this appendix, we show that if the steady state size distribution is log-normal:

$$p(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[\frac{-(\ln x - \mu)^2}{2\sigma^2}\right]\,, \tag{5.56}$$

with parameters μ, $\sigma > 0$, and that the division rate is a power law $r(x) = x^\alpha$, where $\alpha \geq 0$ is the strength of the size control, then the entropy production rate due to branching, given by eq. (5.17), is positive. To do so, one needs to compute the covariance between the division rate and the logarithm of the size distribution:

$$\mathrm{Cov}(r, \ln p) = \langle r \ln p\rangle - \langle r\rangle\langle \ln p\rangle\,, \tag{5.57}$$

where the moments of the log-normal distribution are known:

$$\langle x^\alpha\rangle = \exp\left[\alpha\mu + \frac{\alpha^2\mu^2}{2}\right]\,, \tag{5.58}$$

as well as its entropy:

$$\langle \ln p\rangle = -\frac{1}{2} - \mu - \ln(\sigma\sqrt{2\pi})\,. \tag{5.59}$$

Therefore, one only needs to compute the term:

$$\begin{aligned}\langle r \ln p\rangle &= -\int_0^\infty \mathrm{d}x\ \frac{x^\alpha}{x\sigma\sqrt{2\pi}}\left[\frac{(\ln x-\mu)^2}{2\sigma^2}+\ln x+\ln(\sigma\sqrt{2\pi})\right]\exp\left[\frac{-(\ln x-\mu)^2}{2\sigma^2}\right]\\ &= \frac{\exp\left[\frac{-\mu^2}{2\sigma^2}\right]}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty}\mathrm{d}u\left[\frac{-u^2}{2\sigma^2}+u\left(\frac{\mu}{\sigma^2}-1\right)\right]\exp\left[\frac{-u^2}{2\sigma^2}+u\left(\frac{\mu}{\sigma^2}+\alpha\right)\right]\\ &\quad-\left(\ln(\sigma\sqrt{2\pi})+\frac{\mu^2}{2\sigma^2}\right)\langle x^\alpha\rangle\\ &= -\langle x^\alpha\rangle\left[\ln(\sigma\sqrt{2\pi})+\mu+\frac{1}{2}+\alpha\sigma^2+\frac{\alpha^2\sigma^2}{2}\right],\end{aligned} \tag{5.60}$$

where the second line is obtained with the change of variable $u=\ln x$. The third line comes from the resolution of the integral, which is given by the following Gaussian formulas:

$$\int_{-\infty}^{+\infty}\mathrm{d}u\ u\exp\left[-au^2+bu\right]=\frac{\sqrt{\pi}b}{2a^{3/2}}\exp\left(\frac{b^2}{4a}\right) \tag{5.61}$$

$$\int_{-\infty}^{+\infty}\mathrm{d}u\ u^2\exp\left[-au^2+bu\right]=\frac{\sqrt{\pi}(2a+b^2)}{4a^{5/2}}\exp\left(\frac{b^2}{4a}\right). \tag{5.62}$$

Finally, combining eq. (5.17) and eqs. (5.57) to (5.60) leads to

$$\dot{S}_{\mathrm{brc}}=(m-1)\alpha\sigma^2\left(1+\frac{\alpha}{2}\right)\exp\left[\alpha\mu+\frac{\alpha^2\mu^2}{2}\right]\geq 0\,, \tag{5.63}$$

which is positive regardless of the actual values of μ, σ and $\alpha\geq 0$.

C Asymptotic efficiency for the timer

In this appendix, we demonstrate eq. (5.44) which gives the value of the timer efficiency in the limit of strong age-control ($\alpha\to\infty$), and when $m=2$. Note that for $\alpha=+\infty$, no steady-state is reached since cells divide deterministically when reaching age 1, so that the age distribution is given the delta function $p(a,t)=\delta(a-(t-\lfloor t\rfloor))$ where $\lfloor t\rfloor$ is the integer part of t. Therefore, we consider that α is large enough so that the power-law branching rate $r(a)=a^\alpha$ can be approximated by a step, taking value 0 between 0 and 1 and diverging for $a>1$, and the population grows as $N(t)=N_0 2^{\lfloor t\rfloor}$, but not infinite so that the steady-state age distribution exists and the population growth rate $\Lambda_t=1/t\ln\left(N(t)/N_0\right)$ tends to the steady-state growth rate Λ. Thus,

$$\begin{aligned}\Lambda &= \lim_{t\to\infty}\frac{\lfloor t\rfloor}{t}\ln 2\\ &= \ln 2\,.\end{aligned} \tag{5.64}$$

Using this result, we can compute the Shannon entropy of the steady-state age distribution (eq. (1.75)):

$$\begin{aligned} S_{\text{sys}} &= -\int_0^\infty \mathrm{d}a\ p(a)\ln p(a) \\ &= -\ln(2\Lambda)\int_0^\infty \mathrm{d}a\ p(a) + 2\Lambda\int_0^\infty \mathrm{d}a\left(\Lambda a + \int_0^a \mathrm{d}a'\ r(a')\right)\exp\left[-\Lambda a - \int_0^a \mathrm{d}a'\ r(a')\right] \\ &= -\ln(2\Lambda) + 2\Lambda\int_0^1 \mathrm{d}a\ \Lambda a\ \exp\left[-\Lambda a\right] \\ &= 1 - 2\ln 2 - \ln(\ln 2)\,. \end{aligned} \tag{5.65}$$

We went from line 2 to line 3 using the normalization of p for the first integral, and we decomposed the second integral into two parts: from 0 to 1 and from 1 to ∞. Between 1 and ∞, the contribution of the branching rate diverges and thus the integral is nullified by the exponential function, and only the integral from 0 to 1 remains, for which $r(a) \approx 0$.

Finally, we plug eqs. (5.64) and (5.65) into eqs. (5.39) and (5.40), and plug them into the definition of the efficiency eq. (5.43) to obtain:

$$\eta = \frac{\ln 2}{1+\ln 2} \approx 0.41\,. \tag{5.66}$$

D Parameters of the log-normal steady-state size distribution

We follow in this appendix the analysis conducted in Hosoda et al., 2011 to determine the parameters (μ, σ) of the log-normal ansatz for the steady-state size distribution.

We consider the steady-state sizer model with deterministic exponential growth, corresponding to eq. (5.2) with $F(x) = \nu x$, $\mu = 1$ and $D = 0$. Moreover, we focus on the class of homogeneous kernels for which the probability for a daughter cell to inherit a certain volume at birth depends only on the ratio of daughter to mother volumes. In this case, $\Sigma(x|y) = b(x/y)/y$, and we choose a Gaussian distribution $b = \mathcal{N}(1/2, \sigma_b^2)$ so that the transition probability is Gaussian and centered around symmetric division: $\Sigma(\cdot|y) = \mathcal{N}(y/2, (y\sigma_b)^2)$. Note that this makes sense only for small values of σ_b such that negative ratios of volume have negligible probability. In the main text, simulations are done for deterministic and symmetric partitioning of volume: $b(x) = \delta(x - 1/2)$. We recast this equation into an equation on the moments of the distribution by multiplying it by x^k and integrating over x:

$$\nu k\langle x^k\rangle - \langle x^{k+\alpha}\rangle - \langle x^k\rangle\langle x^\alpha\rangle + 2\langle x^{k+\alpha}\rangle m_k = 0\,, \tag{5.67}$$

where the integral term corresponding to cell births has been transformed like in eq. (4.66), and where m_k are the non-central moments of the Normal law.

Evaluating this expression for $k = 1$, knowing that $m_k = 1/2$ gives:

$$\langle x^\alpha\rangle = \nu\,, \tag{5.68}$$

which is independent of the actual distribution p. Indeed, this result is the combination of two results: the equality between the population growth rate and the mean division

rate eq. (1.65), and the equality between population growth rate and single cell growth rate for exponentially-growing cells eq. (4.18), in steady-state.

Experimental data are well accounted for by log-normal distributions, so we plug the ansatz $p = \text{Lognormal}(\mu, \sigma^2)$ in eq. (5.67), in order to deduce the values of the parameters μ and σ. The moments of p are given by eq. (5.58), and we use eq. (5.68) to re-write eq. (5.67) as

$$(2m_k - 1)\, e^{k\alpha\sigma^2} + k - 1 = 0\,. \tag{5.69}$$

Then, we evaluate this expression for $k = 2$, given that $m_2 = 1/4 + \sigma_b^2$:

$$\sigma^2 = \frac{1}{2\alpha} \ln\left(\frac{2}{1 - 4\sigma_b^2}\right)\,. \tag{5.70}$$

Reporting this in eq. (5.68), we finally obtain

$$\mu = \frac{\ln \nu}{\alpha} - \frac{1}{4} \ln\left(\frac{2}{1 - 4\sigma_b^2}\right)\,. \tag{5.71}$$

We emphasize the fact that the log-normal distribution is not a solution to the population balance equation, but only a good fit in some ranges of parameters. This can be seen by evaluating eq. (5.69) for higher moments: for example when $k = 3$ and $m_3 = 1/8 + 3\sigma_b^2/2$, this relation gives $\sigma^2 = \ln\left[8/(3 - 12\sigma_b^2)\right]/(3\alpha)$, which is inconsistent with eq. (5.70).

Bibliography for Chapter 5

[Bénichou et al., 2005] Bénichou, O., M. Coppey, M. Moreau, P.-H. Suet, and R. Voituriez (2005). Optimal Search Strategies for Hidden Targets. *Physical Review Letters* 94.(19), p. 198101.

[Busiello et al., 2020] Busiello, D. M., D. Gupta, and A. Maritan (2020). Entropy production in systems with unidirectional transitions. *Physical Review Research* 2.(2), p. 023011.

[Coppey et al., 2004] Coppey, M., O. Bénichou, R. Voituriez, and M. Moreau (2004). Kinetics of Target Site Localization of a Protein on DNA: A Stochastic Approach. *Biophysical Journal* 87.(3), pp. 1640–1649.

[Eliazar, 2017] Eliazar, I. (2017). Branching Search. *EPL (Europhysics Letters)* 120.(6), p. 60008.

[England, 2013] England, J. L. (2013). Statistical physics of self-replication. *The Journal of Chemical Physics* 139.(12), p. 121923.

[Evans et al., 2011] Evans, M. R. and S. N. Majumdar (2011). Diffusion with Stochastic Resetting. *Physical Review Letters* 106.(16), p. 160601.

[Evans et al., 2020] Evans, M. R., S. N. Majumdar, and G. Schehr (2020). Stochastic resetting and applications. *Journal of Physics A: Mathematical and Theoretical* 53.(19), p. 193001.

[Fuchs et al., 2016] Fuchs, J., S. Goldt, and U. Seifert (2016). Stochastic thermodynamics of resetting. *EPL (Europhysics Letters)* 113.(6), p. 60009.

[García-García, 2012] García-García, R. (2012). Nonadiabatic entropy production for non-Markov dynamics. *Physical Review E* 86.(3), p. 031117.

[Genthon et al., 2022] Genthon, A., R. García-García, and D. Lacoste (2022). Branching processes with resetting as a model for cell division. *Journal of Physics A: Mathematical and Theoretical* 55, p. 074001.

[Gupta et al., 2020] Gupta, D., C. A. Plata, and A. Pal (2020). Work Fluctuations and Jarzynski Equality in Stochastic Resetting. *Physical Review Letters* 124.(11), p. 110608.

[Hall et al., 1990] Hall, A. J. and G. C. Wake (1990). Functional differential equations determining steady size distributions for populations of cells growing exponentially. *The Journal of the Australian Mathematical Society. Series B. Applied Mathematics* 31.(4), pp. 434–453.

[Hänggi et al., 1977] Hänggi, P. and H. Thomas (1977). Time evolution, correlations, and linear response of non-Markov processes. *Zeitschrift für Physik B* 26.(1), pp. 85–92.

[Hänggi et al., 1982] Hänggi, P. and H. Thomas (1982). Stochastic processes: Time evolution, symmetries and linear response. *Physics Reports* 88.(4), pp. 207–319.

[Hong et al., 2020] Hong, S., E. Dechaumphai, C. R. Green, R. Lal, A. N. Murphy, C. M. Metallo, and R. Chen (2020). Sub-nanowatt microfluidic single-cell calorimetry. *Nature Communications* 11.(1), p. 2982.

[Hosoda et al., 2011] Hosoda, K., T. Matsuura, H. Suzuki, and T. Yomo (2011). Origin of lognormal-like distributions with a common width in a growth and division process. *Physical Review E* 83.(3), p. 031118.

[Jia et al., 2021] Jia, C., A. Singh, and R. Grima (2021). Cell size distribution of lineage data: Analytic results and parameter inference. *iScience* 24.(3), p. 102220.

[Jia et al., 2022] Jia, C., A. Singh, and R. Grima (2022). Characterizing non-exponential growth and bimodal cell size distributions in fission yeast: An analytical approach. *PLoS Computational Biology* 18.(1), e1009793.

[Nakashima et al., 2020] Nakashima, S., Y. Sughiyama, and T. J. Kobayashi (2020). Lineage EM algorithm for inferring latent states from cellular lineage trees. *Bioinformatics* 36.(9), pp. 2829–2838.

[Nieto et al., 2020] Nieto, C., J. Arias-Castro, C. Sánchez, C. Vargas-García, and J. M. Pedraza (2020). Unification of cell division control strategies through continuous rate models. *Physical Review E* 101.(2), p. 022401.

[Pal et al., 2019] Pal, A., I. Eliazar, and S. Reuveni (2019). First Passage under Restart with Branching. *Physical Review Letters* 122.(2), p. 020602.

[Pal et al., 2017] Pal, A. and S. Rahav (2017). Integral fluctuation theorems for stochastic resetting systems. *Physical Review E* 96.(6), p. 062135.

[Pal et al., 2021] Pal, A., S. Reuveni, and S. Rahav (2021). Thermodynamic uncertainty relation for systems with unidirectional transitions. *Physical Review Research* 3.(1), p. 013273.

[Pandey et al., 2020] Pandey, P. P., H. Singh, and S. Jain (2020). Exponential trajectories, cell size fluctuations, and the adder property in bacteria follow from simple chemical dynamics and division control. *Physical Review E* 101.(6), p. 062406.

[Robert et al., 2014] Robert, L., M. Hoffmann, N. Krell, S. Aymerich, J. Robert, and M. Doumic (2014). Division in Escherichia coli is triggered by a size-sensing rather than a timing mechanism. *BMC Biology* 12.(1), p. 17.

[Rodenfels et al., 2019] Rodenfels, J., K. M. Neugebauer, and J. Howard (2019). Heat Oscillations Driven by the Embryonic Cell Cycle Reveal the Energetic Costs of Signaling. *Developmental Cell* 48.(5), 646–658.e6.

[Roldán et al., 2017] Roldán, É. and S. Gupta (2017). Path-integral formalism for stochastic resetting: Exactly solved examples and shortcuts to confinement. *Physical Review E* 96.(2), p. 022130.

[Song et al., 2019] Song, Y., J. O. Park, L. Tanner, Y. Nagano, J. D. Rabinowitz, and S. Y. Shvartsman (2019). Energy budget of Drosophila embryogenesis. *Current Biology* 29.(12), R566–R567.

[Thomas et al., 2018] Thomas, P., G. Terradot, V. Danos, and A. Y. Weiße (2018). Sources, propagation and consequences of stochasticity in cellular growth. *Nature Communications* 9.(1), p. 4528.

[Wang et al., 2010] Wang, P., L. Robert, J. Pelletier, W. L. Dang, F. Taddei, A. Wright, and S. Jun (2010). Robust Growth of Escherichia coli. *Current Biology* 20.(12), pp. 1099–1103.

[Zeraati et al., 2012] Zeraati, S., F. H. Jafarpour, and H. Hinrichsen (2012). Entropy production of nonequilibrium steady states with irreversible transitions. *Journal of Statistical Mechanics: Theory and Experiment* 2012.(12), p. L12001.

General conclusion

In a perfectly balanced population tree where all lineages are equivalent, for example because of a very strong control on division, the behavior of any lineage is representative of that of the population. However, such a tree is never observed since all the bio-molecular mechanisms involved in the cell cycle are stochastic, which results in fluctuations between the lineages. The relation between the single cell stochasticity and the deterministic growth of the population is thus in general non trivial. While biologists would seek a complete description of the molecular mechanisms controlling a biological process, physicists often try to reduce the complexity of the problem by proposing mesoscopic descriptions in terms of few coarse-grained variables that capture the main features of the process. In the context of bacterial cells, these variables are typically the size and age of the cell, or the quantity of a key protein; and the cell cycle is reduced to a few processes like cell elongation and aging, and the partition of resources and volume between the daughter cells at division. Mapping the microscopic and mesoscopic descriptions is an exciting research program, which we did not pursue in this thesis.

Starting from the existence of the fluctuations between lineages, the main focuses of this thesis were to relate the scales of the single cell and the population, to understand the role of the fluctuations in this relation, and to show how lineage data can be used in practice. Since new experiments on finite populations in confined geometries have greatly enhanced our capacity to probe cell colonies for long times in well controlled environment, we also address the question of the treatment of data from these experiments, where lineages are continuously flushed away.

In the first two chapters of the thesis, we sought universal principles and we hope to have passed on the following messages:

- There is a statistical bias between the experiments carried out in bulk and in mother-machine. For example, it takes less time for the population to double than for an isolated cell in a single-lineage to divide on average, which may seem counter-intuitive at first.
- This bias can be used to our advantage and provides possible uses of single-lineage data, like the inference of the population growth rate from mother-machine data.
- There is an extra 'survivor bias' in finite populations, where early-ending lineages have to be carefully taken into account when sampling the population tree.
- The phenotypic variability in freely-growing populations is the result of the selection of the fittest lineages, while no such selection is present in single-lineage experiments. Therefore, the lineage-population bias lies at the heart of the strength of selection, which is limited by the variability of fitness in the population.

Importantly, these results are universal in the sense that they only rely on the structure of the branching tree and not on a particular model or dynamics. As such, they could be used in other areas of biophysics where similar trees are present, like species trees or stem cell differentiation trees. It remains open to determine in what way they could be useful in these contexts, but we hope that the formalism of selection will find applications outside bacterial evolution. In the context of cell colonies, universality implies that the results are independent of the dynamics, namely the model of cell size control, the presence of mutations, the fluctuations of environment, ... Instead, they are implied by the topology of the population tree, where one cell asexually reproduces to give birth to two daughter cells. In particular, lineages that do not make it to the end of the experiment can be the result not only of dilution but also of cellular death or any other phenomenon.

A significant limitation of our approach is that the history of the tree with the number of divisions along each lineage is necessary, although in a lot of situations this information is not available. In ecology or for in-vivo experiments for example, available data are mostly snapshot data. Showing how to use these snapshot data to infer the history of the tree would be a major step in the understanding of evolution for more realistic systems.

In the last two chapters, we shifted the perspective and adopted a model-based approach. The modeling approach is complementary to the search for universal principles: it provides testable predictions that can help validate or invalidate hypotheses and proposes precise and quantitative descriptions of specific systems. The works from these two chapters have been designed to join the effort in the understanding of cell size homeostasis, and in the discrimination between the different models of cell size control, even though they represent only a first step un that direction.

We derived steady-state cell size distributions for lineages of size-regulated cells, which were successfully compared to experimental data. This comparison can be seen as a test of the sizer model, or as a way to use single lineage data to infer the parameters of the cell cycle. We showed how fluctuations in volume at division, single cell growth and volume partitioning shape the size distribution, and thus directly impact the lineage-population bias.

Independently, we proposed a simple thermodynamic description of cell growth and division for both age and size-regulated populations. Although we did not compare the different mechanisms of size regulation, we hope this formalism could give arguments in favor of certain models against others regarding their thermodynamic efficiencies. We extended our formalism from simple coarse grained models to general n-variable models, and it would be interesting to identify the set of relevant variables necessary to describe the thermodynamic state of a cell, so that the theory matches recent calorimetric experiments at the scale of the single cell.

RÉSUMÉ

Les expériences sur les cellules en croissance peuvent être menées soit en croissance libre, soit dans des géométries qui confinent la colonie et limitent sa croissance. Comment échantillonner les arbres généalogiques de ces populations pour chaque configuration ? Y a-t-il des biais statistiques entre eux ? Comment quantifier la sélection naturelle pour ces populations ? Ce sont les principales questions que nous abordons dans cette thèse. Dans une première partie, nous étudions le biais statistique entre le niveau lignée unique et le niveau population, qui présente des similitudes avec les théorèmes de fluctuation en thermodynamique stochastique. Pour cela, nous développons un formalisme basé sur les histoires des lignées d'une même population, et nous obtenons des contraintes universelles qui sont exploitées dans deux directions. Premièrement, ce biais nous renseigne sur la force de la sélection, qui quantifie les corrélations entre les valeurs d'un trait cellulaire et le succès reproductif de la lignée. Cette sélection résulte de la variabilité des lignées dans la population, que nous analysons en utilisant la théorie de la réponse linéaire. Nous généralisons le formalisme pour autoriser les situations où les lignées se terminent avant la fin de l'expérience, en raison de la mort cellulaire et de la dilution. Nous montrons comment les lignées mortes doivent être prises en compte dans les différents échantillonnages, et comment la mort affecte la variabilité phénotypique et donc la force de la sélection. Deuxièmement, nous montrons comment les données de lignées uniques peuvent être utilisées pour inférer des propriétés au niveau de la population, comme le taux de croissance de la population, ou paramètre de Malthus. En nous concentrant sur les populations de cellules régulées en taille, nous obtenons les distributions de taille à l'équilibre pour les expériences de lignées uniques, qui peuvent aussi être utilisées pour inférer les paramètres du cycle cellulaire tels que le taux d'élongation des cellules et l'asymétrie de la division. En outre, nous montrons comment différentes sources de stochasticité peuvent modifier le biais lignée-population pour les statistiques de taille. Dans une deuxième partie indépendante, nous proposons une description thermodynamique de la croissance et de la division cellulaire à l'aide de modèles macroscopiques simples de contrôle de la taille. Cette question est importante pour comprendre comment les colonies de cellules sont contraintes par la thermodynamique. En décomposant la division cellulaire en deux sous-processus : le branchement (création d'une nouvelle cellule identique), et la réinitialisation, ou *resetting* (modification des propriétés des deux cellules), nous dérivons les deux lois de la thermodynamique pour une colonie de cellules, et nous identifions la contribution de chaque processus au changement d'énergie moyenne et d'entropie de Shannon. Cela nous permet de comprendre comment les distributions d'âge et de taille sont affectées par la division cellulaire du point de vue de la théorie de l'information.

MOTS CLÉS

Populations de cellules, lignées de cellules, sélection naturelle, distribution de taille des cellules, thermodynamique stochastique, théorèmes de fluctuation.

ABSTRACT

Experiments on growing cells can be carried out either in bulk or in confined geometries that constrain the growth of the colony. How should the population trees be sampled in each setup? Are there statistical biases between them? How to quantify natural selection in these trees? These are the main questions we address in this thesis. In a first part, we study the statistical bias between the single-lineage and population levels, which has similarities with fluctuation theorems in stochastic thermodynamics. To do so, we develop a theoretical framework based on lineage histories within population trees, and obtain universal constraints that are exploited in two directions. First, this bias informs on the strength of selection, that quantifies the correlations between the value of a cell trait and the reproductive success of the lineage. This selection results from the variability of lineages in the population, which we analyze using linear response theory. We also extend our framework to allow situations where lineages end before the end of the experiment, due to cell death or dilution. We show how dead lineages should be taken into account in the statistics, and how death impacts the phenotypic variability and therefore the strength of selection. Second, we show how single-lineage data can be used to infer population-level quantities like the population growth rate, also called Malthus parameter. Focusing on size-regulated populations, we derive steady-state cell size distributions for single-lineage experiments, that can also be used to infer cell cycle parameters such as the single cell elongation rate and the asymmetry of division. In addition, we explore how the lineage-population bias for size statistics is affected by different sources of stochasticity. In a second independent part, we propose a thermodynamic description of cell growth and division using simple coarse-grained models of cell size control. This question is important to understand how cell colonies are constrained by thermodynamics. Using a decomposition of cell division in two sub-processes: branching (creation of an identical new cell), and resetting (restart of the properties of the two cells), we derive the first and second laws of thermodynamics for a colony of cells, and identify the contribution of each process to the change in average energy and Shannon entropy. This allows us to understand how the distributions of age and size are affected by cell division from an information-theoretic point of view.

KEYWORDS

Cell populations, cell lineages, natural selection, cell size distribution, stochastic thermodynamics, fluctuation theorems.