# Modélisation de l'évolution de la taille des génomes et de leur densité en gènes par mutations locales et grands réarrangements chromosomiques

Stephan Fischer

▶ **To cite this version:**

Stephan Fischer. Modélisation de l'évolution de la taille des génomes et de leur densité en gènes par mutations locales et grands réarrangements chromosomiques. Evolution [q-bio.PE]. INSA de Lyon, 2013. Français. NNT : . tel-00924831v1

**HAL Id: tel-00924831**

**https://theses.hal.science/tel-00924831v1**

Submitted on 7 Jan 2014 (v1), last revised 27 Oct 2014 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Numéro d'ordre 2013-xxx
Année 2013

# Modélisation de l'évolution de la taille des génomes et de leur densité en gènes par mutations locales et grands réarrangements chromosomiques

**Thèse présentée par**
Stephan Fischer

**Devant**
L'Institut National des Sciences Appliquées de Lyon

**Pour obtenir**
Le grade de docteur

**Formation doctorale**
Mathématiques et Informatique (InfoMaths)

**Spécialité**
Modélisation mathématique et informatique du vivant

Soutenance prévue le 2 décembre 2013 devant le jury composé de :

| | |
|---|---|
| Samuel Bernard | Chargé de recherche, CNRS, Université Lyon 1, co-encadrant |
| Guillaume Beslon | Professeur, INSA de Lyon, co-directeur de thèse |
| Carole Knibbe | Maître de conférences, Université Lyon 1, co-directrice de thèse |
| Amaury Lambert | Professeur, Collège de France, Université Paris 6, rapporteur |
| Benoît Perthame | Professeur, Université Paris 6, examinateur |
| Eduardo Rocha | Directeur de recherche, Institut Pasteur, examinateur |
| Olivier Tenaillon | Chargé de recherche HDR INSERM, Université Paris 7, rapporteur |

| SIGLE | ECOLE DOCTORALE | NOM ET COORDONNEES DU RESPONSABLE |
|---|---|---|
| CHIMIE | **CHIMIE DE LYON**<br>**http://www.edchimie-lyon.fr**<br><br>Insa : R. GOURDON | M. Jean Marc LANCELIN<br>Université de Lyon – Collège Doctoral<br>Bât ESCPE<br>43 bd du 11 novembre 1918<br>69622 VILLEURBANNE Cedex<br>Tél : 04.72.43 13 95<br>directeur@edchimie-lyon.fr |
| E.E.A. | **ELECTRONIQUE, ELECTROTECHNIQUE, AUTOMATIQUE**<br>**http://edeea.ec-lyon.fr**<br><br>Secrétariat : M.C. HAVGOUDOUKIAN<br>eea@ec-lyon.fr | M. Gérard SCORLETTI<br>Ecole Centrale de Lyon<br>36 avenue Guy de Collongue<br>69134 ECULLY<br>Tél : 04.72.18 65 55 Fax : 04 78 43 37 17<br>Gerard.scorletti@ec-lyon.fr |
| E2M2 | **EVOLUTION, ECOSYSTEME, MICROBIOLOGIE, MODELISATION**<br>**http://e2m2.universite-lyon.fr**<br><br>Insa : H. CHARLES | Mme Gudrun BORNETTE<br>CNRS UMR 5023 LEHNA<br>Université Claude Bernard Lyon 1<br>Bât Forel<br>43 bd du 11 novembre 1918<br>69622 VILLEURBANNE Cédex<br>Tél : 06.07.53.89.13<br>e2m2@ univ-lyon1.fr |
| EDISS | **INTERDISCIPLINAIRE SCIENCES-SANTE**<br>**http://www.ediss-lyon.fr**<br><br>Sec : Samia VUILLERMOZ<br>Insa : M. LAGARDE | M. Didier REVEL<br>Hôpital Louis Pradel<br>Bâtiment Central<br>28 Avenue Doyen Lépine<br>69677 BRON<br>Tél : 04.72.68.49.09 Fax :04 72 68 49 16<br>Didier.revel@creatis.uni-lyon1.fr |
| INFOMATHS | **INFORMATIQUE ET MATHEMATIQUES**<br>**http://infomaths.univ-lyon1.fr**<br><br>Sec :Renée EL MELHEM | Mme Sylvie CALABRETTO<br>Université Claude Bernard Lyon 1<br>INFOMATHS<br>Bâtiment Braconnier<br>43 bd du 11 novembre 1918<br>69622 VILLEURBANNE Cedex<br>Tél : 04.72. 44.82.94 Fax 04 72 43 16 87<br>infomaths@univ-lyon1.fr |
| Matériaux | **MATERIAUX DE LYON**<br>**http://ed34.universite-lyon.fr**<br><br>Secrétariat : M. LABOUNE<br>PM : 71.70 –Fax : 87.12<br>Bat. Saint Exupéry<br>Ed.materiaux@insa-lyon.fr | M. Jean-Yves BUFFIERE<br>INSA de Lyon<br>MATEIS<br>Bâtiment Saint Exupéry<br>7 avenue Jean Capelle<br>69621 VILLEURBANNE Cedex<br>Tél : 04.72.43 83 18 Fax 04 72 43 85 28<br>Jean-yves.buffiere@insa-lyon.fr |
| MEGA | **MECANIQUE, ENERGETIQUE, GENIE CIVIL, ACOUSTIQUE**<br>**http://mega.ec-lyon.fr**<br><br>Secrétariat : M. LABOUNE<br>PM : 71.70 –Fax : 87.12<br>Bat. Saint Exupéry<br>mega@insa-lyon.fr | M. Philippe BOISSE<br>INSA de Lyon<br>Laboratoire LAMCOS<br>Bâtiment Jacquard<br>25 bis avenue Jean Capelle<br>69621 VILLEURBANNE Cedex<br>Tél :04.72 .43.71.70 Fax : 04 72 43 72 37<br>Philippe.boisse@insa-lyon.fr |
| ScSo | **ScSo***<br>**http://recherche.univ-lyon2.fr/scso/**<br><br>Sec : Viviane POLSINELLI<br>Brigitte DUBOIS<br>Insa : J.Y. TOUSSAINT | M. OBADIA Lionel<br>Université Lyon 2<br>86 rue Pasteur<br>69365 LYON Cedex 07<br>Tél : 04.78.77.23.86 Fax : 04.37.28.04.48<br>Lionel.Obadia@univ-lyon2.fr |

*ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

# Remerciements

Je suis parfaitement conscient que la partie remerciements est la partie la plus lue d'une thèse. Je pense que, malheureusement, à peu près tout a été fait et il devient très difficile d'être original. Au risque de décevoir, je me contenterai donc de proposer un élément original puis je remercierai – de façon assez classique – les personnes qui ont contribué de manière décisive à la réussite de ma thèse.

En premier lieu, je souhaite remercier les personnes qui ont eu la charge de lire plus que les remerciements. D'abord, Amaury Lambert et Olivier Tenaillon, qui ont accepté le rôle de rapporteur de mon manuscrit. Merci pour les critiques qu'ils m'ont transmises. Je remercie également Benoît Perthame et Eduardo Rocha qui ont accepté de venir à Lyon pour assister à ma soutenance en tant que membres du jury.

Je pense que le quotidien est un aspect négligé dans la plupart des lieux de travail. Le choix de ma thèse n'a pas été motivé uniquement par la thématique scientifique mais aussi par les personnes qui composent l'équipe Beagle. Le partage des locaux avec Dracula[1] n'ayant pas rompu cette harmonie, cela fait de très nombreuses personnes avec qui j'ai partagé d'excellents moments. La liste est cependant bien trop longue puisqu'il faut remercier toutes les personnes m'ayant supporté, les joueurs de go, les fléchiverbistes, les partenaires de discussion, les personnes qui m'ont aidé au long de ces 3 ans ou pendant mon stage de master, la personne qui fait inlassablement des blagues douteuses, etc. J'ai été très heureux de partager tout ce temps avec vous, j'ai appris beaucoup de choses grâce à vous (sur des plans plus ou moins scientifiques) et je pense que vous méritez bien l'honneur de chercher votre nom dans la grille ci-dessous :

*grille en cours de construction*

J'aimerais remercier en particulier mes collègues de bureau, Bérénice et Jules, qui sont coupables de nombreuses fautes[2] mais je leur pardonne bien volontiers. Étant plutôt timide, je peux aujourd'hui avouer que j'avais un peu d'appréhension avant de les connaître.

[1] (*sic*) c'est le nom d'une équipe Inria en biomaths (dont fait partie Samuel)

[2] Connaissant leur mauvaise foi, je me vois ici dans l'obligation de citer, en exemple, les attaques répétées sur mon poste de travail. Trop peu de gens savent que la customisation de mon bureau est due à Bérénice ou connaissent les attaques musclées ou copées de Jules.

Mais il faut se rendre à l'évidence : les parents ont bien choisi leurs enfants. Partager avec vous ce bureau pendant plus 2 ans aura été source de tellement de discussions passionnantes qu'il m'est difficile de l'exprimer correctement. J'espère que vous êtes conscients que j'ai été heureux de travailler avec vous.

Il y a aussi les personnes qui ont contribué indirectement à cette thèse puisque, paradoxalement, je les ai vus moins souvent pendant la thèse. Je pense ici à ma famille que j'ai quittée il y a maintenant 8 ans pour mieux découvrir la France, ce qui a été une bonne idée puisque ça m'a permis de rencontrer des fausses françaises mais on y reviendra. Je pense évidemment également à Tigrou, qui n'a elle pas eu le bénéfice de l'éloignement et qui a su me supporter avec une extrême patience relative tous les jours, notamment aux heures de repas. Mes grands-parents m'auront toujours soutenu et transmis tout ce qu'ils avaient de meilleur, je leur en suis très reconnaissant. J'aimerais remercier ma famille élargie, que je n'ai malheureusement pas l'occasion de voir très souvent. En particulier, merci Laurence pour tes visites et pour avoir pensé à moi en chaque occasion.

Mes collègues comprendront que je m'attarde sur le sort de mes deux frères qui ont dû me supporter au quotidien pendant près de 18 ans. Ça y est vous êtes enfin débarrassés de moi ! Merci à Pierre et Louise, Julien, Maryline, Loan et Tom, qui me font bon accueil à chaque fois que je rentre au foyer familial et qui m'auront apporté beaucoup de réconfort quand j'ai pu les voir.

Enfin, merci maman et papa. J'apprécie beaucoup que vous veniez assister à ma soutenance mais c'est surtout pour tout ce que vous avez fait pour moi tout au long de mon éducation et encore maintenant que je veux vous remercier. Il est difficile de condenser en quelques lignes ce que vous avez fait pour moi, mais il est évident que vous avez contribué directement à ma réussite dans tout ce que j'ai entrepris, y compris ce manuscrit !

L'ordre est un peu bizarre et ils désespèrent peut-être maintenant d'être remerciés, mais un grand merci à mes directeurs de thèse. En même temps, vous l'avez cherché, vous êtes tellement différents et avez contribué de manière tellement diversifiée – tout en étant complémentaires – à ma thèse que ça rend les remerciements difficiles sans que j'ai l'air de léser l'un d'entre vous.

Je commencerai par Guillaume, dont l'action sur ma thèse a été la plus discrète mais constante. J'ai déjà souligné ici – et ailleurs – la qualité humaine et scientifique de l'équipe, j'en tiens Guillaume pour principal initiateur. Merci pour ça mais aussi tous les services, scientifiques et humains là aussi, que tu m'as rendus tout au long de la thèse.

Ensuite, j'aimerais remercier ensemble Samuel et Carole, tant la complémentarité du couple est étonnante. Samuel, bien plus en retrait, mais quand on lit la thèse, on retrouve en fond tous les éléments que tu as proposés, souvent bien avant que je les aies réellement adoptés. Un peu comme pour Guillaume, j'ai peur que tu t'imagines que je mésestime ce que tu as fait pour moi car j'ai peu eu l'occasion de t'en remercier et ces lignes sont bien

peu de choses en regard de ce que tu as fait pour moi tout au long de ma thèse. Merci !

Carole, j'ai tout de même une pensée particulière pour toi. Peu importe si les examinateurs ne comprennent pas ce que je dis, mais je pense que tu es bien la seule à avoir fait les démonstrations complètes des parties les plus mathématiques de ce manuscrit. Pour une biologiste et informaticienne de formation, ce n'est déjà pas peu dire. Mais surtout, je pense qu'aux moments où j'avais le plus de doutes, tu étais bien présente, ce qui à nouveau est un euphémisme, je pense que les gens qui ne te connaissent pas ne se représentent pas le degré d'abnégation dont tu as fait preuve. Je ne peux pas m'empêcher de penser que tu en fais trop, mais je pense que si je ne devais retenir qu'une chose de cette thèse (disons au moins sur le plan humain), c'est bien ce que tu as fait pour moi au moment où j'en avais le plus besoin.

Merci à vous trois !

Enfin, j'aimerais remercier Hanna. On a suivi le même parcours depuis qu'on se connaît, encore aujourd'hui : on a des intérêts communs mais bien plus encore. J'ai été surpris par le nombre de personnes qui s'imaginent qu'un étudiant en sciences qui réussit à peu près passe ses journées ou ses week-ends à parler science ou travailler. Heureusement que la réalité est éloignée de cet *a priori*. Peu de gens connaissent la richesse de ta pensée et de ta personnalité. Tu t'intéresses virtuellement à tout et c'est une source d'émulation constante. On a encore beaucoup à faire, à exprimer et à explorer et, avec toi, cette perspective ne laisse aucune appréhension.

# Résumé

Bien que de nombreuses séquences génomiques soient maintenant connues, les mécanismes évolutifs qui déterminent la taille des génomes, et notamment leur part d'ADN non codant, sont encore débattus. Ainsi, alors que de nombreux mécanismes faisant grandir les génomes (prolifération d'éléments transposables, création de nouveaux gènes par duplication, ...) sont clairement identifiés, les mécanismes limitant la taille des génomes sont moins bien établis. La sélection darwinienne pourrait directement défavoriser les génomes les moins compacts, sous l'hypothèse qu'une grande quantité d'ADN à répliquer limite la vitesse de reproduction de l'organisme. Cette hypothèse étant cependant contredite par plusieurs jeux de données, d'autres mécanismes non sélectifs ont été proposés, comme la dérive génétique et/ou un biais mutationnel rendant les petites délétions d'ADN plus fréquentes que les petites insertions.

Dans ce manuscrit, nous montrons à l'aide d'un modèle matriciel de population que la taille du génome peut aussi être limitée par la dynamique spontanée des duplications et des grandes délétions, qui tend à raccourcir les génomes même si les deux types de réarrangements se produisent à la même fréquence. En l'absence de sélection darwinienne, nous prouvons l'existence d'une distribution stationnaire pour la taille du génome même si les duplications sont deux fois plus fréquentes que les délétions. Pour tester si la sélection darwinienne peut contrecarrer cette dynamique spontanée, nous simulons numériquement le modèle en choisissant une fonction de fitness qui favorise directement les génomes contenant le plus de gènes, tout en conservant des duplications deux fois plus fréquentes que les délétions. Dans ce scénario où tout semblait pousser les génomes à grandir infiniment, la taille du génome reste pourtant bornée. Ainsi, notre étude révèle une nouvelle force susceptible de limiter la croissance des génomes. En mettant en évidence des comportements contre-intuitifs dans un modèle pourtant minimaliste, cette étude souligne aussi les limites de la simple « expérience de pensée » pour penser l'évolution.

Nous proposons un modèle mathématique de l'évolution structurelle des génomes en mettant l'accent sur l'influence des différents mécanismes de mutation. Il s'agit d'un modèle matriciel de population, à temps discret, avec un nombre infini d'états génomiques possibles. La taille de population est infinie, ce qui élimine le phénomène de dérive génétique. Les mutations prises en compte sont les mutations ponctuelles, les petites insertions et délétions, mais aussi les réarrangements chromosomiques induits par la recombinaison ectopique de l'ADN, comme les inversions, les translocations, les grandes délétions et les duplications. Nous supposons par commodité que la taille des segments réarrangés suit

une loi uniforme, mais le principal résultat analytique est ensuite généralisé à d'autres distributions. Les mutations étant susceptibles de changer le nombre de gènes et la quantité d'ADN intergénique, le génome est libre de varier en taille et en compacité, ce qui nous permet d'étudier l'influence des taux de mutation sur la structure génomique à l'équilibre.

Dans la première partie de la thèse, nous proposons une analyse mathématique dans le cas où il n'y a pas de sélection, c'est-à-dire lorsque la probabilité de reproduction est identique quelle que soit la structure du génome. En utilisant le théorème de Doeblin, nous montrons qu'une distribution stationnaire existe pour la taille du génome si le taux de duplications par base et par génération n'excède pas 2.58 fois le taux de grandes délétions. En effet, sous les hypothèses du modèle, ces deux types de mutation déterminent la dynamique spontanée du génome, alors que les petites insertions et petites délétions n'ont que très peu d'impact. De plus, même si les tailles des duplications et des grandes délétions sont distribuées de façon parfaitement symétriques, leur effet conjoint n'est, lui, pas symétrique et les délétions l'emportent sur les duplications. Ainsi, si les tailles de délétions et de duplications sont distribuées uniformément, il faut, en moyenne, plus de 2.58 duplications pour compenser une grande délétion. Il faut donc que le taux de duplications soit quasiment trois fois supérieur au taux de délétions pour que la taille des génomes croisse à l'infini. L'impact des grandes délétions est tel que, sous les hypothèses du modèle, ce dernier résultat reste valide même en présence d'un mécanisme de sélection favorisant directement l'ajout de nouveaux gènes. Même si un tel mécanisme sélectif devrait intuitivement pousser les génomes à grandir infiniment, en réalité, l'influence des délétions va rapidement limiter leur accroissement. En résumé, l'étude analytique prédit que les grands réarrangements délimitent un ensemble de tailles stables dans lesquelles les génomes peuvent évoluer, la sélection influençant la taille précise à l'équilibre parmi cet ensemble de tailles stables.

Dans la deuxième partie de la thèse, nous implémentons le modèle numériquement afin de pouvoir simuler l'évolution de la taille du génome en présence de sélection. En choisissant une fonction de fitness non bornée et strictement croissante avec le nombre de gènes dans le génome, nous testons le comportement du modèle dans des conditions extrêmes, poussant les génomes à croître indéfiniment. Pourtant, dans ces conditions, le modèle numérique confirme que la taille des génomes est essentiellement contrôlée par les taux de duplications et de grandes délétions. De plus, cette limite concerne la taille totale du génome et s'applique donc aussi bien au codant qu'au non codant. Nous retrouvons en particulier le seuil de 2.58 duplications pour une délétion en deçà duquel la taille des génomes reste finie, comme prévu analytiquement. Le modèle numérique montre même que, dans certaines conditions, la taille moyenne des génomes diminue lorsque le taux de duplications augmente, un phénomène surprenant lié à l'instabilité structurelle des grands génomes. De façon similaire, augmenter l'avantage sélectif des grands génomes peut paradoxalement faire rétrécir les génomes en moyenne. Enfin, nous montrons que si les petites insertions et délétions, les inversions et les translocations ont un effet limité sur la taille du génome, ils influencent très largement la proportion d'ADN non codant.

**Mots-clés** : Taille du génome, densité en gènes, évolution moléculaire, réarrangements chromosomiques, duplications, délétions, chaîne de Markov, processus stochastiques.

# English title, abstract and keywords

## Modelling of the evolution of genome size and gene density by local mutations and large chromosomal rearrangements

Even though numerous genome sequences are now available, evolutionary mechanisms that determine genome size, notably their fraction of non-coding DNA, are still debated. In particular, although several mechanisms responsible for genome growth (proliferation of transposable elements, gene duplication and divergence, *etc.*) were clearly identified, mechanisms limiting the overall genome size remain unclear. Darwinian selection could directly disadvantage less compact genomes, under the hypothesis that a larger quantity of DNA could slow down the speed of reproduction of the organism. Because this hypothesis was proven wrong by several datasets, non selective mechanisms have been proposed, *e.g.* genetic drift and/or a mutational bias towards small DNA deletions compared to small DNA insertions.

In this manuscript, we use a matrix model to show that genome size can also be limited by the spontaneous dynamics of duplications and large deletions, which tends to decrease genome size even if the two types of rearrangements occur at the same rate. In the absence of Darwinian selection, we prove the existence of a stationary distribution of genome size even if duplications are twice as frequent as large deletions. To test whether selection can overcome this spontaneous dynamics, we simulate our model numerically and choose a fitness function that directly favors genomes containing more genes, while keeping duplications twice as frequent as large deletions. In this scenario where, at first sight, everything seems to favor infinite genome growth, genome size remains nonetheless bounded. As a result, our study reveals a new pressure that could be responsible for limiting genome growth. By illustrating counter-intuitive behaviors in a minimal model, this study also underlines the limits of simple “thought experiments" to understand evolution.

We propose a mathematical model of the structural evolution of genomes that focuses on the influence of several mutation mechanisms. It is a matrix model of a population evolving on a discrete time scale in a space encompassing an infinity of possible genome structures. Population size is also infinite, which eliminates genetic drift. Mutations taken

into account are point mutations, small insertions and deletions, but also chromosomal rearrangements induced be ectopic recombination of DNA, namely inversions, translocations, large deletions and duplications. For simplicity, we suppose that rearrangement sizes are distributed uniformly, but the main analytical result is generalized to other distributions. As the mutations may change the number of genes or the length of intergenic DNA, genomes can vary in size and gene density, which enables us to study the influence of mutation rates on the genome structure obtained at steady-state.

In the first part of the thesis, we propose a mathematical analysis in the absence of selection, *i.e.* when the probability of reproduction does not depend on the genome structure. By using Doeblin's condition, we show that a stationary distribution of genome size exists as soon as the duplication rate per base per generation is lower than 2.58 times the large deletion rate. Indeed, under the hypotheses of the model, these two types of mutations determine the spontaneous dynamics of genome size, while small insertions and small deletions have a limited impact. What is more, even if the sizes of duplications and large deletions are distributed symmetrically, the overall effect on size is in fact not symmetrical, as large deletions have a higher impact than duplications. Thus, if the sizes of duplications and large deletions are distributed uniformly, more than 2.58 duplications are needed, on average, to compensate for a large deletion. In order to achieve infinite growth of genome size, the duplication rate must thus be nearly three times higher than the deletion rate. Under the hypotheses of the model, the impact of deletions is such that the latter result remains true even if a selection mechanism that directly favors retention of new genes is applied. Intuitively, such a selective pressure should lead to infinite growth by accumulation of new genes but, in fact, the influence of large deletions rapidly limits this growth. To sum up, the analytical study predicts that large rearrangements delimit a subspace of stable genome sizes in which genomes can evolve, selection influencing the precise size towards which the population will converge among this subset of stable sizes.

In the second part of the thesis, we implement the model numerically in order to simulate the evolution of genome size in the presence of selection. By choosing an unbounded fitness function that strictly increases with the number of genes, we test the behaviour of the model in extreme conditions, pushing genomes to grow indefinitely. However, in these conditions, the numerical model confirms that genomes size is essentially controlled by the duplication and large deletion rates. What is more, this limit applies to the overall genome size, that is to coding as well as non-coding DNA. Simulations display the threshold at 2.58 duplications for one deletions below which genome size remains finite, as predicted analytically. The numerical study shows that, under some conditions, the average genome size shrinks when the duplication rate increases, a surprising phenomenon linked with the structural instability of large genomes. Similarly, increasing the selective advantage of large genomes can lead to an average genome reduction. Finally, we show that, even if small insertions, small deletions, inversions and translocations have a limited impact on genome size, they have an important influence on the fraction of non-coding DNA of the genome.

**Keywords** : Genome size, gene density, molecular evolution, chromosomal rearrangements, duplications, deletions, Markov chain, stochastic processes.

# Table des matières

# Chapitre I

# Introduction

> Who could be certain that the lower forms did not in fact require more genes to conduct their dreary affairs ?
>
> C.A. Thomas Jr

L'évolution des espèces est basée sur la transmission d'information génétique via des polymères d'ADN ou d'ARN (pour certains virus) qui constituent le génome. L'ADN/ARN est un support universel d'information génétique, mais son organisation globale peut fortement varier d'une espèce à l'autre. Chez les virus, le génome est composé d'une ou plusieurs molécules d'ARN, d'ADN à simple brin ou d'ADN à double brin. Chez les procaryotes, bactéries et archées, le génome est en général composé d'un chromosome circulaire d'ADN à double brin, accompagné de plusieurs plasmides, non essentiels à la survie. Chez les eucaryotes, le génome désigne habituellement l'ADN nucléaire, présent dans le noyau des cellules, généralement répartis sur un ou plusieurs chromosomes linéaires. Les organites intracellulaires comme la mitochondrie ou les chloroplastes contiennent également de l'ADN, dit non-nucléaire, qui apporte des fonctionnalités essentielles à la survie de l'organisme eucaryote.

Il n'y a pas que la façon de stocker l'ADN ou l'ARN qui varie en fonction des espèces. Pour affiner les distinctions, on peut prendre en compte le nombre de chromosomes qui composent le génome ainsi que le nombre de copies de chaque chromosome présent dans la cellule. Dans ce dernier cas, on parle de ploïdie. Par exemple, de nombreuses bactéries, comme *Escherichia coli*, sont haploïdes : il n'y a qu'une seule copie du chromosome. Chez l'humain, une cellule contient généralement deux copies de chaque chromosome : on parle de diploïdie. En revanche, à l'issue de la méiose, les gamètes humains ne possédent qu'un exemplaire de chaque chromosome : les gamètes sont haploïdes. Enfin, on peut caractériser un génome en utilisant sa longueur en nombre de paires de bases (pour les organismes à double brin) ou sa densité en gènes. Dans son usage le plus commun, la notion de gène

désigne un segment d'ADN contenant l'information nécessaire pour produire une protéine ou un ARN fonctionnel.

La taille et la composition du génome sont le résultat de nombreuses pressions évolutives qui s'exercent sur les populations. À court terme, les fluctuations de la taille du génome sont relativement faibles : il faut s'intéresser à des périodes d'évolution très longues pour constater des changements importants, mais parfois spectaculaires, notamment chez les plantes. La question occupe une place ambivalente dans la communauté scientifique : d'une part, elle fascine à l'idée de trouver une explication universelle et élégante applicable à toutes les espèces, d'autre part, l'absence relative d'impact direct à court terme rend cette question secondaire.

Dans cette introduction, nous allons montrer pourquoi la taille du génome n'est pas facile à appréhender. Dans une première section, nous donnerons un aperçu de la diversité des mécanismes mutationnels qui peuvent faire varier la taille du génome, et nous rappellerons comment les mesures et observations ont déjoué l'intuition et les prévisions simples pour être à l'origine de quelques paradoxes célèbres, notamment celui dit de « la valeur C »[1], suite à l'observation que la taille du génome n'est pas simplement corrélée à la complexité apparente d'un organisme. L'amélioration des techniques de séquençage a permis d'avoir accès au contenu exact des génomes, ce qui a mené à la découverte de l'étendue de l'ADN intergénique et intronique, que nous rassemblons ici sous le terme d'ADN non codant[2]. Cet ADN apparemment sans utilité pour l'organisme est souvent peuplé d'éléments transposables[3], qualifiés « d'égoïstes », qui prolifèreraient à l'insu de l'organisme. Ces nouvelles données ont permis de mieux comprendre le paradoxe de la valeur C et de le résoudre. On verra cependant que ces résolutions ne sont que partielles et se heurtent à de nouveaux paradoxes et difficultés. Plusieurs mécanismes ont alors été proposés pour arriver à une explication cohérente, voire universelle, de la taille des génomes, mais les données ne permettent pas de confirmer une de ces théories en particulier. Cette contradiction apparente entre l'universalité des explications proposées et la difficulté à les vérifier vient du fait qu'elles sont souvent formulées de manière qualitative. Les différentes théories ne s'opposent pas réellement entre elles et peuvent être ajustées aux données en ajoutant l'effet de certains facteurs de confusion. Il s'agit ici de « modèles de l'esprit » qui mènent à des « expériences de pensée ».

Les expériences de pensée sont essentielles pour aboutir aux théories, mais reposent fortement sur notre intuition et, par extension, sur l'historique du domaine de recherche. Les

[1] Ce paradoxe sera défini en page 27.

[2] Le terme d'ADN non codant est un raccourci de langage, puisque certains gènes ne codent pas pour des protéines mais pour des ARN fonctionnels (comme les ARN ribosomiques), et ne sont donc pas « codants » dans le sens de « traduits en protéine ». Il serait en fait plus correct – mais moins pratique – de parler à chaque fois d'« ADN intergénique ou intronique ». Par ailleurs, les éléments transposables sont en général comptés dans la part d'ADN non codant d'un génome même s'ils portent des gènes nécessaires à leur transposition.

[3] Un élément transposable est une séquence d'ADN reconnue par des enzymes spécifiques capables de la déplacer (comme un « couper-coller ») ou de la multiplier (comme un « copier-coller »). Souvent, la séquence de l'élément transposable contient la séquence du ou des gènes des enzymes nécessaire à son déplacement ou à sa copie. L'élément est alors dit autonome.

traduire en modèles quantitatifs, les mettre en équation, permet de vérifier s'il n'existe pas des phénomènes qui auraient échappé à notre intuition. Comme les modèles de l'esprit sont souvent flous, il est difficile d'en donner une version quantifiable, car de nombreux paramètres ne sont pas explicités. De ce fait, la plupart des modèles qui implémentent des mécanismes similaires à ceux proposés par les expériences de pensée ne prétendent pas résoudre la question de la taille du génome dans toute sa complexité. Certains s'attachent à expliquer le contenu en gènes dits essentiels, nécessaires à la survie, certains le nombre total de gènes, d'autres encore le nombre d'éléments répétés ou d'éléments transposables dans des zones précises du non-codant, sans préjuger de l'évolution du reste du génome ou en supposant qu'il est fixe. La deuxième section de l'introduction s'attachera à présenter les différents mécanismes représentés par ces modèles.

La dernière partie de l'introduction présentera la démarche générale de la thèse. Nous nous attacherons à un mécanisme particulier. Il s'agira des réarrangements chromosomiques, dont nous étudierons l'impact relativement à celui des mutations locales, en l'absence puis en présence de sélection sur le nombre de gènes. Il ne s'agit donc pas de trouver une explication définitive et universelle à la taille du génome, mais de comprendre les pressions exercées par un mécanisme en particulier. Nous expliquerons comment cette idée a émergé et comment nous proposons d'illustrer au mieux les différents impacts des réarrangements. Nous présenterons le modèle que nous avons conçu, ainsi que le découpage du manuscrit qui correspond à des études de ce modèle dans différentes conditions.

# 1 Mécanismes mutationnels, paradoxe de la taille du génome et hypothèses explicatives

## 1.1 Mécanismes mutationnels faisant varier la taille du génome

Les mutations de l'ADN les plus connues sont les mutations dites ponctuelles, qui ne font pas changer la taille du génome puisqu'elles correspondent au remplacement d'une base par une autre : par exemple, une adénine (A) est remplacée par une guanine (G). Cela peut se produire lors de la réplication de l'ADN, ou sous l'effet d'agents mutagènes chimiques ou physiques comme les radiations ultraviolettes. Cependant, les mutations causées par les erreurs de réplication ou les mutagènes ne sont pas toutes des mutations ponctuelles : il peut également s'agir de petites insertions ou délétions. Ainsi, le complexe enzymatique de réplication peut insérer des bases supplémentaires dans le brin d'ADN néosynthétisé, ou bien ne pas copier certaines bases du brin d'ADN matrice. Ces insertions et délétions de quelques bases peuvent se produire dans toutes les parties du génome, mais sont particulièrement fréquentes lorsque le brin matrice contient des séquences répétées courtes (typiquement 2 à 10 paires de bases) situées « en tandem », c'est-à-dire directement adjacentes. Par exemple, la séquence ATTCATTCATTCATTCATTC contient 5 fois le motif ATTC « en tandem ». Les séquences répétées de ce type sont nommées « microsatellites » chez les eucaryotes et « loci de contigence » chez les procaryotes. Elles peuvent induire la

formation d'une boucle dans le brin matrice ou dans le brin synthétisé, ce qui cause l'ajout ou la suppression d'une ou plusieurs copies du motif, un phénomène appelé « replication slippage » ou « slipped-strand mispairing » (voir figure I.1).

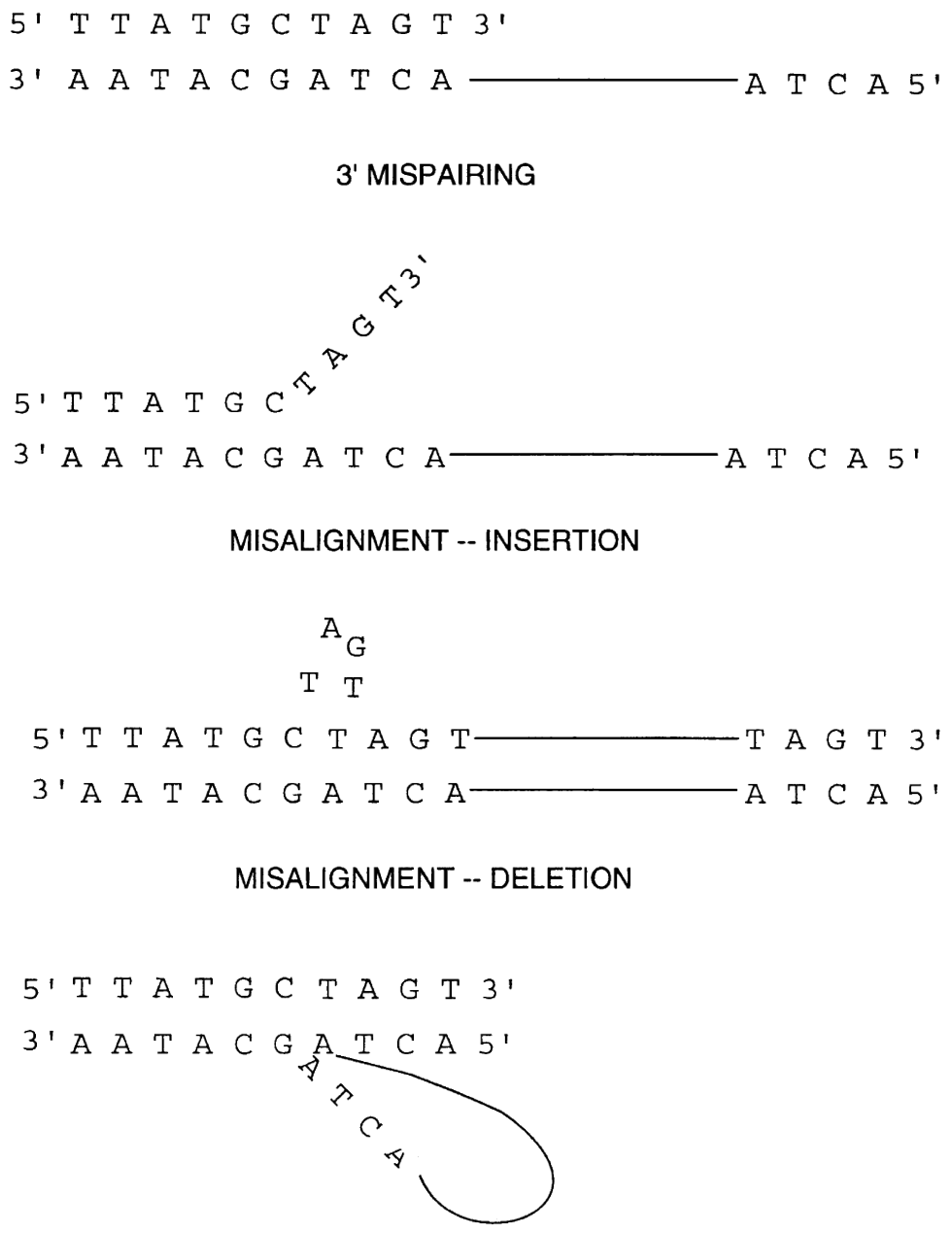


Figure I.1 – Mécanisme pouvant conduire à une petite insertion ou à une petite délétion lors de la réplication (image tirée de Petrov (2002)). La polymérase peut se décrocher de l'ADN et se repositionner sur une partie de l'ADN déjà répliquée, menant à une petite insertion. Elle peut également se repositionner plus loin sur l'ADN, ici grâce à la complémentarité entre les séquences du brin décroché (TAGT) et l'endroit où la polymérase se fixe (ACTA), ce qui produit une petite délétion.

Par ailleurs, l'ADN peut aussi subir des mutations à plus grande échelle appelées réarrangements chromosomiques (voir figure I.2). Des segments d'ADN de centaines, milliers

voire millions de paires de bases peuvent être dupliqués, excisés, inversés ou déplacés, sous l'effet de différents mécanismes moléculaires comme la recombinaison homologue non allélique (aussi appelée recombinaison homologue ectopique), la recombinaison dite illégitime par NHEJ (non homologous end-joining) ou la recombinaison site-spécifique mise en jeu dans la transposition. La recombinaison homologue permet normalement de réparer une cassure double-brin dans une molécule d'ADN en utilisant une autre molécule d'ADN de séquence identique ou fortement similaire : la chromatide sœur si le chromosome cassé venait d'être répliqué, ou bien le chromosome homologue dans le cas des cellules diploïdes. Mais si les chromosomes contiennent des séquences répétées, l'alignement peut se faire de façon dite ectopique, c'est-à-dire entre des positions différentes, et provoquer alors des duplications, des délétions ou des inversions de segments de chromosomes (voir Chen (2011); Gu *et al.* (2008) pour plus de détails). La recombinaison NHEJ est une autre famille de mécanismes de réparation des cassures double-brin. Elle est nommée « non homologue » car les extrémités des molécules cassées sont directement réassemblées sans nécessiter une molécule identique ou similaire pour servir de guide à la réparation. Elle utilise en fait les microhomologies de quelques paires de bases souvent présentes au niveau des extrémités simple brin de la molécule cassée. Outre la réparation normale des cassures, elle peut aussi occasionner des délétions, des inversions, des duplications ou des translocations de segments de chromosomes (voir Chen (2011); Gu *et al.* (2008) pour plus de détails). Les éléments transposables sont aussi responsables de nombreux réarrangements chromosomiques, soit parce qu'ils servent de substrat à la recombinaison homologue ectopique, soit parce qu'ils peuvent entraîner d'autres segments d'ADN dans leurs déplacements (voir figure I.2). Ainsi, plusieurs mécanismes moléculaires sont susceptibles de provoquer des duplications et délétions de longs segments d'ADN, modifiant donc significativement la taille du génome et ayant parfois des conséquences phénotypiques importantes, selon les gènes contenus dans le segment.

Enfin, des chromosomes ou même des génomes entiers peuvent être dupliqués en raison d'accidents de ségrégation des chromosomes lors de la division cellulaire. Dans le cas de la trisomie 21 par exemple (environ 1 naissance sur 700), le chromosome 21 est présent en 3 exemplaires au lieu de 2. D'autres anomalies chromosomiques peuvent se produire, rendant par exemple l'œuf fécondé tétraploïde au lieu de diploïde pour tous les chromosomes, ce qui le rend généralement non viable. Cependant, dans certains embranchements de l'arbre de la vie, de tels événements de duplication du génome complet ont été conservés par l'évolution. Ainsi, les génomes de la plante *Arabidopsis thaliana*, de la paramécie *Paramecium tetraurelia*, de la levure *Saccharomyces cerevisiae* ou encore du poisson tétraodontiforme *Tetraodon nigroviridis* portent les traces d'événements de duplication complète (Jaillon *et al.*, 2009).

## 1.2 Paradoxe de la « valeur C »

Si les mutations peuvent faire varier la taille du génome, une population peut contenir des individus ayant des tailles différentes : quelle taille va alors être conservée à l'échelle du temps évolutif ? La réponse est-elle la même pour toutes les espèces ? Ces questions

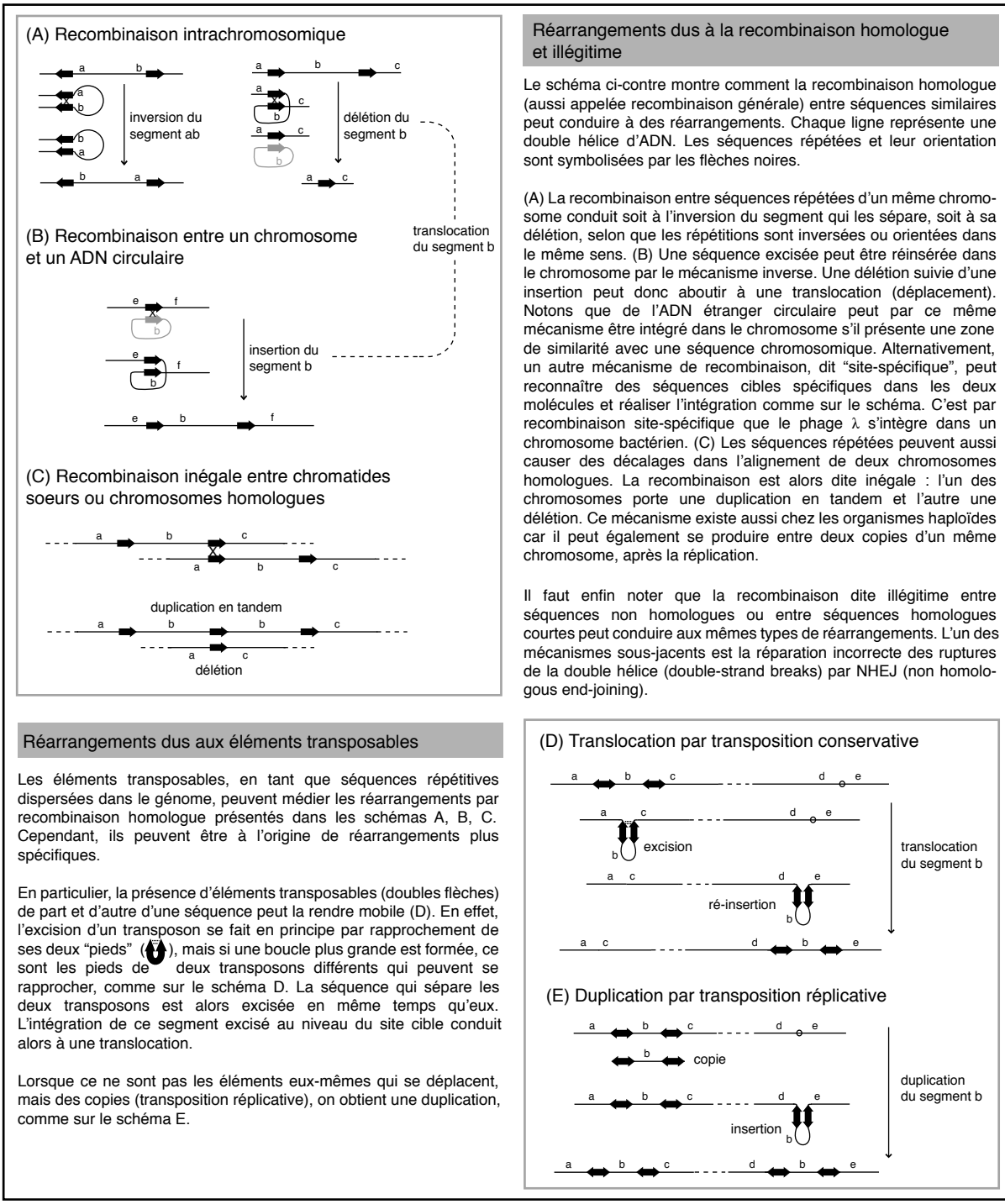


FIGURE I.2 – Mécanismes permettant des réarrangements à grande échelle du génome (reproduit avec permission de Knibbe (2006)).

n'ont pas attendu les techniques de séquençage pour être posées et pour donner lieu à des premières expériences de pensée, reposant sur l'intuition. La découverte des chromosomes et de l'ADN a montré que tous les êtres vivants partagent une même base biochimique. Les êtres vivants étant en apparence différents, on pouvait s'attendre à ce qu'il y ait des

différences dans la *quantité* d'ADN. Avant le développement du séquençage, on s'attend à ce que la taille du génome reflète directement son contenu en gènes et en information. Des organismes en apparence complexes, comme les humains, auraient besoin de plus d'information pour leur développement et leur fonctionnement que des microbes. De même, on peut facilement envisager que deux organismes proches d'un point de vue phylogénétique et morphologique aient des génomes comparables.

Pour tester ce type d'hypothèses, sans connaître la séquence du génome, on utilise le poids de l'ADN comme prédicteur de la taille du génome. C'est ce qu'on appelle la valeur C, exprimée en picogrammes, qui correspond au poids du contenu haploïde (par opposition à polyploïde) d'une cellule. Par exemple, l'humain est polyploïde, plus exactement diploïde : chaque chromosome est présent en deux exemplaires. Pour éliminer cette redondance, le poids final est divisé par deux, ou annoncé comme étant la valeur 2C. Précisons aussi que le terme valeur C porte souvent à confusion. Il se rapporte bien au poids de l'ADN et non à la complexité apparente, comme on pourrait le penser en apprenant que le paradoxe de la valeur C met en relation la complexité et le contenu en ADN. La lettre C viendrait du fait que les premières expériences s'attachaient non pas à comparer les différences entre espèces, mais la remarquable constance (d'où le C) de la taille du génome au sein d'une espèce[1].

En compilant les valeurs C issus de différentes espèces, Thomas a conclu en 1971 (Thomas, 1971) qu'il fallait bien admettre que la complexité des organismes n'était pas corrélée de manière évidente à la valeur C, avec quelques exemples surprenants à la clé, comme le fait que certains amphibiens et poissons auraient une valeur C 20 fois supérieure à celle des humains. Thomas résume alors en quelques phrases l'idéologie de l'époque :

> It was argued that mammals display a greater developmental complexity than primitive fish, therefore, they must have more genes, yet why should the lower forms have more DNA, if DNA is the chemical basis of the gene? Early opponents of the DNA theory of heredity drew strength from the misinterpretation of these observations and they continue to inspire some. Others contented themselves with interim interpretations such as: maybe those animals and plants having huge amounts of DNA per nucleus were highly polyploid. In this way the c-value (the amount of DNA per haploid set of chromosomes) could still be a reasonable number.

Thomas (1971) rappelle aussi que deux espèces proches peuvent avoir des valeurs C variant de l'ordre de 80x (pour les *Ranunculaceae* qui ont pourtant le même caryotype) voire 2000x (pour certaines algues unicellulaires). Ces observations sont accompagnées de remarques assez sarcastiques, qui donnent à l'article un ton particulier.

> Here the matter rested, for who could be certain that the lower forms did not in fact require more genes to conduct their dreary[2] affairs. [...] Surely the

[1] Source : Wikipedia, article *C-value* (en anglais).

[2] dreary: boring and making you feel unhappy (Cambridge Online Dictionary)

> genetic information content of these closely related species could not be too different. Otherwise, they would differ in morphology, which, after all, is an extremely sensitive expression of genetic potentiality.

Cet article est à l'origine du terme de « paradoxe de la valeur C », symbolique du moment à partir duquel il a fallu chercher une explication moins triviale à la taille du génome. En utilisant les données disponibles aujourd'hui, on se rend compte que le paradoxe a très bien survécu, même si on peut le nuancer (figure I.3).

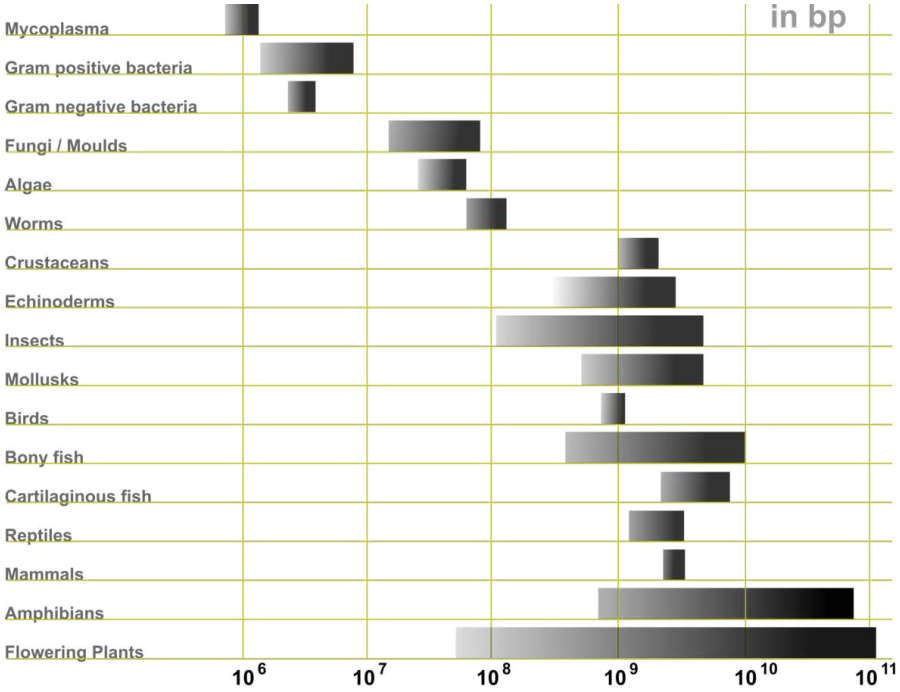


Figure I.3 – Distribution de la taille des génomes connus pour différente familles d'organismes (image Wikipedia).

Les variations de taille de génome dépendent fortement des grandes familles de l'arbre du vivant. Chez les bactéries ou chez les mammifères, les variations ne sont pas si grandes, alors que les plantes s'étendent sur de nombreux ordres de grandeur, y compris dans des espèces de la même sous-famille. Même si les organismes pluricellulaires ont globalement des génomes plus longs que les organismes unicellulaires, certaines familles balayent des tailles de génomes très larges, avec certains cas qui peuvent paraître inattendus, comme les salamandres, dont les génomes sont parmi les plus grands génomes amphibiens, bien plus grands que le génome humain.

Notons pour conclure que l'opportunité du terme « paradoxe » semble discutable. Il ne s'agit pas d'un paradoxe au sens logique du terme, mais souligne en fait l'échec de l'intuition et, pourrait-on dire, des premières expériences de pensée. Gregory (2001) suggère d'ailleurs qu'il faudrait arrêter d'utiliser ce terme, mais pour des raisons différentes. Il juge que le paradoxe est résolu (comme nous allons le voir tout de suite), mais son opinion ne semble pas partagée, puisque le terme continue d'être utilisé (par attachement

historique ? parce que les autres auteurs considèrent que la question reste ouverte ? un peu des deux ?).

## 1.3 Résolution partielle du paradoxe : composition des génomes en éléments codants et non-codants

Pour résoudre un paradoxe, on peut remonter aux axiomes. Rappelons l'idée principale : la taille du génome reflète le contenu en information, la complexité requiert plus d'information. Il faut donc que l'une des deux hypothèses au moins soit fausse. La valeur C élimine la redondance liée à la polyploïdie mais pas celle présente au sein d'un chromosome. Au début des années 1970, une notion importante fait son apparition : « l'ADN poubelle » ou *junk DNA* en anglais (Ohno, 1972). Si on suppose que la complexité requiert plus d'information, alors nécessairement les organismes à faible complexité et grand génome ont beaucoup d'ADN qui ne porte pas d'information utile au développement et à la maintenance de l'organisme. Plusieurs possibilités sont envisagées pour expliquer l'existence de cet ADN poubelle. D'une part, on sait alors que de nouveaux gènes sont créés par duplication (Ohno, 1970). Or, un gène récemment dupliqué n'apporte pas d'information nouvelle pour l'individu, la séquence étant initialement redondante. Si la duplication n'augmente pas l'expression du gène de manière avantageuse, on peut alors imaginer que sous l'effet d'une mutation, cet ADN perde sa fonction et devienne non-codant : il devient alors un pseudogène. Une pseudogénisation de ce type se fait sans diminuer la fitness de l'individu, mais en augmentant sa part d'ADN poubelle. D'autre part, on découvre l'existence de morceaux d'ADN qui se répliquent indépendamment des individus et capables de se dupliquer et de se réinsérer dans le génome : les éléments transposables. Ces éléments sont en apparence purement égoïstes, dans le sens où ils semblent ne pas contribuer au développement ou à la survie de l'organisme. La part d'ADN due aux éléments transposables est qualifiée de *selfish DNA* (Doolittle et Sapienza, 1980; Orgel et Crick, 1980; Orgel *et al.*, 1980). Certains auteurs font clairement la distinction entre *junk DNA* (pseudogènes) et *selfish DNA* (éléments transposables), d'autres considèrent que le *selfish DNA* est inclus dans le *junk DNA*. Dans la suite du manuscrit, on prendra les acceptions au sens le plus strict.

Une résolution du paradoxe semble donc se profiler : une fois connue la quantité précise d'ADN codant, on pourra la relier à la complexité apparente. Malheureusement, cela implique de connaître les détails de la séquence d'ADN, la valeur C n'est plus d'aucune utilité. Cette perspective de résolution entraîne de nouvelles questions : pourquoi le contenu en ADN non codant est-il aussi variable et, si ce n'est pas la complexité, qu'est-ce qui détermine la taille du génome ? Comme il faut attendre encore quelques décennies avant que le séquençage ne devienne performant, c'est surtout ces dernières questions qui ont été soulevées dans les années 1970, avec déjà de nombreuses hypothèses entrant en compétition. On peut distinguer deux courants de pensée. D'un côté, ceux qui proposent que la quantité d'ADN influence la fitness des individus et a donc été sélectionnée. De l'autre, ceux qui proposent que l'aléatoire et la contingence ont une place forte dans l'évolution, y compris dans le cas du contenu total en ADN. Nous reviendrons vers ces hypothèses,

qui existent encore aujourd'hui, dans la sous-section suivante.

Le séquençage des génomes a été une étape importante et le répertoire de génomes séquencés disponible aujourd'hui permet de revenir au lien entre complexité et contenu du génome. Après l'introduction du non codant dans le paradoxe, on s'est logiquement reporté à une explication de la complexité par le nombre de gènes, ce qui a mené au paradoxe de la valeur G (Hahn et Wray, 2002). Comme le terme de paradoxe l'indique, cette prédiction échoue également, dans le sens où le nombre de gènes ne semble pas non plus bien corréler avec la complexité apparente des organismes. On peut alors continuer à affiner les hypothèses. Comme le lien entre complexité et quantité d'information n'est pas remis en cause, c'est tout naturellement qu'on accuse le nombre de gènes d'être un mauvais prédicteur du contenu en information. On distingue couramment 3 raisons à cela : (i) les gènes dupliqués, donc redondants en terme d'information, (ii) l'épissage alternatif des introns, qui crée plusieurs transcrits (ARN messagers), donc porteurs d'information différente, avec un seul gène et (iii) l'ADN non codant qui peut participer en réalité au développement et au fonctionnement de la cellule (voir par exemple Alexander *et al.*, 2010; Touzain *et al.*, 2011).

On recense donc les variants d'épissage, en gardant en réserve l'idée de redondance qui peut s'exprimer à différents niveaux (génétique, transcriptomique, métabolique, etc.). Les données actuelles ne sont pas assez riches pour permettre des comparaisons à grande échelle, mais notons que ces mesures offrent un meilleur lien avec la complexité apparente que la taille du génome (Schad *et al.*, 2011), sans être totalement probantes. En effet, même si on offre une résolution du paradoxe initial, on s'éloigne de l'idée élégante de trouver un prédicteur simple de la complexité apparente. En intégrant la transcriptomique et vraisemblablement, à terme, des éléments de régulation et de métabolisme, on s'oriente plutôt vers une tentative d'explication de la complexité elle-même, plutôt que la recherche d'un élément qui corrèle avec elle. Il peut paraître plus pertinent dans ce cas de partir du concept de complexité, en se détachant *a priori* de la problématique de la taille du génome (Adami, 2002; Tenaillon *et al.*, 2007).

Il existe donc dans la littérature deux façons d'aborder la question : soit approfondir le paradoxe en intégrant des nouvelles données (celles qui finissent en -omique évoquées en partie dans le paragraphe précédant, mais aussi sur les découvertes liées au rôle du non-codant pour l'individu) en visant à terme une explication de la complexité, ou revenir en arrière et chercher à comprendre ce qui détermine la taille du génome, en laissant de côté la complexité. Dans la suite de ce manuscrit, c'est la deuxième question — celle de la détermination de la taille du génome — qui va nous intéresser.

## 1.4 Mécanismes proposés actuellement pour expliquer la taille du génome

Dans la recherche d'une solution du paradoxe de la valeur C, le génome a été progressivement disséqué dans le but de comprendre ce que chaque séquence, codante ou non codante

pouvait apporter à l'ADN. Finalement, il n'y a pas de lien simple entre la complexité apparente et le nombre de gènes et, même en invoquant le non codant comme une solution au paradoxe, il faut expliquer pourquoi certains organismes n'en contiennent pas tandis que d'autres en ont près de 99 %. Le non codant peut être fonctionnel, il peut proliférer à l'insu de l'individu (éléments transposables) ou être présent à cause de pseudogènes et autres événements de réarrangements. Quelle que soit la vue qu'on adopte à propos de son rôle, la compréhension de la taille du génome passe forcément par l'étude de la dynamique du codant et du non codant. De fait, plusieurs mécanismes ont été proposés pour expliquer les variations de la taille de ces deux compartiments génomiques :

**Lien direct entre la taille du génome et la vitesse de reproduction de l'organisme** Une différence importante entre les génomes des bactéries et des eucaryotes est le nombre d'origines de réplication. Les bactéries n'ont qu'une origine de réplication de l'ADN : on peut donc imaginer que le temps nécessaire pour copier l'ADN est proportionnel à la taille du génome. Chez les eucaryotes, on peut argumenter que les augmentations de taille sont compensées par l'apparition de nouvelles origines de réplication. De plus, chez les bactéries, le taux de croissance est souvent utilisé comme proxy de la fitness. Il paraît alors naturel de conclure que, chez les bactéries, avoir un génome plus grand et donc plus long à répliquer diminue directement la fitness. Cette interprétation peut éventuellement être vraie dans certaines conditions de laboratoire (Poole *et al.*, 2003), mais il n'y a globalement pas de lien entre vitesse de reproduction de l'organisme et taille du génome (Mira *et al.*, 2001), voire une corrélation négative (Touchon et Rocha, 2007). Lors de la reproduction, la réplication de l'ADN est nécessaire, mais elle ne représente qu'une partie du processus et n'est donc pas nécessairement l'étape limitante.

**Quantité optimale d'ADN pour la taille de la cellule : théories nucléosquelettique (Cavalier-Smith) et nucléotypique (Gregory)** Cavalier-Smith (1978) défend l'idée selon laquelle, même si l'ADN non codant est inutile en terme d'information génétique, il influence le développement dans le mesure où le volume total occupé par l'ADN dépend de l'ADN non codant. L'ADN non codant permettrait donc de contrôler de manière purement physique la taille du noyau, tandis que les gènes contrôlent la taille de la cellule. Il y aurait donc coévolution entre l'ADN codant (qui contrôle le volume de la cellule) et l'ADN non-codant (qui contrôle le volume du noyau). Cavalier-Smith base son approche sur des corrélations entre différentes grandeurs physiques de la cellule, notamment une corrélation entre la taille du génome et le volume de la cellule (Cavalier-Smith, 1985; Cavalier-Smith et Beaton, 1999; Beaton et Cavalier-Smith, 1999).

Il existe une variante de la théorie de la quantité optimale d'ADN, baptisée nucléotypique, défendue initialement par Bennett (1972) mais étendue et mise en avant aujourd'hui par Gregory. Cette théorie cherche un lien différent entre l'ADN, la taille du noyau et le volume de la cellule, sans passer par l'information génétique. Il n'y a en effet pas d'explication consensuelle sur le mécanisme qui permettrait ce lien. Gregory (2001) met en avant le lien avec la durée du cycle cellulaire et propose une explication possible. En bref, l'idée est la suivante. Le noyau, dont le volume est déterminé par la quantité d'ADN total, donne

une taille minimale à la cellule. La cellule grossit au cours du temps jusqu'à sa division, ce qui donnerait le lien entre volume et temps de division. Des marqueurs, par exemple les cyclines, donneraient le signal de la division en se fixant à des endroits précis dans le noyau. Leur expression influencerait alors le temps de division et la taille des cellules : plus les cyclines sont exprimées, plus le cycle est court et la cellule est petite. D'autre part, la quantité d'ADN augmentant le volume du noyau, on peut imaginer qu'à expression égale, il faut plus de temps pour que les cyclines se fixent, ce qui augmente la taille de la cellule. Comparé à la théorie nucléosquelettique, le but est donc de faire le lien directement avec le volume de la cellule.

Si on suppose par ailleurs qu'à un environnement donné correspond une taille de cellule optimale, il y aurait alors une quantité d'ADN optimale en fonction de l'environnement dans lequel se trouve la cellule. Il y aurait donc une adaptation de la cellule au mode de vie ou aux conditions extérieures, qui imposerait une quantité d'ADN optimale. L'avantage de ces théories dites « adaptatives » vient du fait que si elles sont vraies, de nombreux effets tels que la dérive ou les biais spontanés tendent à s'effacer, puisque leurs fluctuations sont limitées par l'intensité de la sélection. La sélection contrôlerait la prolifération d'éléments transposables ou d'ADN poubelle (Gregory, 2001). Tous les mécanismes passifs, biaisés ou non, pourraient être mis à profit pour atteindre la quantité optimale d'ADN (Gregory, 2003, 2004).

**Prolifération de l'ADN égoïste (Doolittle et Sapienza, Orgel et Crick)** Selon Doolittle et Sapienza (1980) et Orgel et Crick (1980), jouer un rôle dans le phénotype de l'organisme et lui conférer un avantage sélectif n'est pas le seul moyen pour une séquence de se maintenir dans un génome : une séquence même légèrement délétère pourrait se maintenir si elle se réplique suffisamment fréquemment. En effet, on peut distinguer dans un génome différents segments en fonction de leurs taux de réplication, le taux basal étant celui d'une réplication par division cellulaire, mais pouvant être plus élevé pour certains segments comme les éléments transposables, capables de produire des copies supplémentaires d'eux-mêmes entre deux divisions cellulaires. Doolittle et Sapienza (1980) définissent ainsi une sélection dite « non phénotypique » qui sélectionne les segments d'ADN qui se répliquent le plus fréquemment sans affecter le phénotype de l'organisme :

> Natural selection does not operate on DNA only through organismal phenotype. Cells themselves are environments in which DNA sequences can replicate, mutate, and so evolve. Although DNA sequences which contribute to organismal phenotypic fitness or evolutionary adaptability indirectly increase their own chances of preservation, and may be maintained by classical phenotypic selection, the only selection pressure which DNAs experience directly is the pressure to survive within cells. If there are ways in which mutation can increase the probability of survival within these cells without effect on the organismal phenotype, then sequences whose only 'function' is self-preservation will inevitably arise and be maintained by what we call 'non-phenotypic selection'. Furthermore, if it can be shown that a given gene (region of DNA) or

> class of genes (regions) has evolved a strategy which increases its probability of survival within cells, then no additional (phenotypic) explanation for its origin or continued existence is required.

Cette « sélection non phénotypique » tend à faire grandir les génomes, mais les auteurs suggèrent que l'ADN en excès a tout de même un coût énergétique pour l'organisme. La taille d'un génome résulterait donc d'une tension entre les deux niveaux de sélection. Pour expliquer que les génomes procaryotes soient globalement plus courts et plus compacts que les génomes eucaryotes, les auteurs suggèrent que l'intensité de la sélection phénotypique contre l'ADN en excès serait plus forte chez les procaryotes :

> The intensity of non-phenotypic pressure on DNA to survive even without function should be independent of organismal physiology. The intensity of phenotypic selection pressure to eliminate excess DNA is not, this being greatest in organisms for which DNA replication comprises the greatest fraction of total energy expenditure. Prokaryotes in general are smaller and replicate themselves and their DNA more often than eukaryotes (especially complex multicellular eukaryotes). Phenotypic selection for small 'streamlined' prokaryotic genomes with little excess DNA may be very strong.

Cependant, cette hypothèse physiologique ne permet pas de comprendre les grandes variations de tailles de génomes observées au sein des amphibiens ou des plantes par exemple.

**Sensibilité des petites populations à la prolifération de l'ADN non codant (Lynch)** Avant de présenter cette nouvelle hypothèse sur l'origine évolutive de l'ADN non codant, il nous faut faire une parenthèse pour présenter la théorie neutraliste de l'évolution, qui a été développée dans les années 1970. Par opposition à une vision qui consiste à voir l'évolution comme une succession d'adaptations (événements qui accroissent la fitness), la théorie neutraliste de l'évolution affirme que les changements évolutifs moléculaires sont essentiellement dus à des mutations qui n'ont pas de grand effet sur la survie et la reproduction de l'organisme, et dont la fréquence dans la population dépend plus des effets de taille de population finie (dérive génétique) que de la sélection naturelle (Kimura, 1968, 1983). Beaucoup d'éléments du génome auraient donc été fixés dans la population[1] de manière neutre et non pas parce qu'ils représentaient un avantage en soi.

La taille de la population et le mode de reproduction jouent à cet égard un rôle crucial. Lors du passage d'une génération à l'autre, certains allèles peuvent être perdus : seule une sous-partie du matériel génétique de la génération parentale servira effectivement à constituer les génomes des descendants. Si un allèle est favorable, il a de plus grandes chances d'être

[1] En génétique des population, la fixation désigne la transition entre une population où plusieurs allèles (variantes d'un gène, d'une séquence) coexistent pour un locus donné à une population qui ne contient pour ce locus plus qu'un seul des allèles, qui est dit fixé.

sélectionné et donc d'arriver à fixation. Si deux allèles en compétition ont la même valeur sélective, chacun a une chance sur deux d'arriver à fixation au bout d'un temps assez long. Les cas intéressants sont ceux où un allèle est légèrement favorable comparé à un autre : la sélection introduit un biais en sa faveur, mais comme la taille finie de la population introduit des effets stochastiques, l'autre allèle peut très bien arriver à fixation malgré tout. C'est ce qu'on appelle la dérive génétique : des mutations neutres, voire légèrement défavorables peuvent être fixées dans la population, tandis que des mutations légèrement favorables ne seront même pas « vues » par la sélection. Plus la population est petite, plus la probabilité est grande qu'une mutation défavorable l'emporte. En présence de mutations effectivement neutres ou légèrement délétères, on s'attend donc à ce que de nombreuses fixations aient lieu, sans que cela apporte d'avantage aux individus ou à la population.

Lynch et Conery (2003) ont proposé une variante de la théorie de l'ADN égoïste, dans laquelle les différences entre espèces ne sont pas expliquées par des différences de physiologie mais par des différences de taille efficace[1] de population : si l'ADN en excès est légèrement délétère, les grandes populations pourront l'éliminer par sélection naturelle, mais pas les petites populations, dans lesquelles la sélection n'est pas très efficace en raison de la forte dérive génétique qu'elles subissent. Ainsi, Lynch considère que les éléments transposables se dupliquent indépendamment des individus, mais qu'ils ont un effet légèrement délétère pour la fitness de l'individu, notamment à cause du risque d'insertion dans les gènes lors de la transposition (Lynch et Conery, 2003; Lynch, 2006). S'il y a très peu de dérive, on aura donc tendance à éliminer tous les éléments transposables. S'il y a beaucoup de dérive, ils prolifèrent plus facilement, on peut donc faire des prédictions liées à leur transposition à différents niveaux structurels. On peut s'attendre à ce que le nombre ou la longueur des introns soient plus importants quand la population est petite et à ce que la taille du génome liée au non-codant intergénique augmente également à cause de la fixation de transpositions. Les bactéries, qui ont une taille de population efficace très élevée, ont relativement peu d'éléments transposables et de non codant (moins de 10% pour la plupart). À l'inverse, les mammifères ont des tailles de population efficaces très faibles : leurs génomes contiennent de nombreuses familles d'éléments transposables et une fraction de non codant énorme (97% chez l'humain). Entre les deux, on trouve les eucaryotes unicellulaires, qui ont une fraction de non codant beaucoup plus modérée et relativement peu d'introns.

En poussant cet argument jusqu'au bout, Lynch met de nombreuses innovations majeures de l'évolution sur le compte de la dérive, et propose même la dérive comme étant à l'origine de la complexité (Lynch et Conery, 2003). Par exemple, les eucaryotes utilisent la technique d'épissage des introns[2]. Ce mécanisme permet de faire de l'épissage alternatif, c'est-à-dire de fabriquer plusieurs protéines différentes à partir d'un même gène, comme

[1] La taille de population efficace n'est pas la taille de population réelle. Elle est généralement définie de telle sorte que la dérive observée soit la même que pour une population idéale qui serait « parfaitement mixée ». La taille $N_e$ de cette population idéale est la taille efficace, elle est généralement beaucoup plus faible que la taille réelle (Charlesworth, 2009).

[2] L'épissage consiste à éliminer des séquences non codantes présentes au sein de gènes (introns) lors de la formation des ARN messagers.

par exemple dans le système immunitaire. Il est tentant de conclure que les introns et les mécanismes complexes d'épissage apportent un avantage et ont donc été sélectionnés. Au contraire, Lynch voit la dérive comme l'explication la plus parcimonieuse. Selon lui, la complexité ne serait pas le résultat de la sélection mais de l'absence de sélection et du combat contre les tares apportées par la dérive. Pour comprendre comment ce raisonnement s'applique aux introns, il faut revenir aux éléments transposables. En cas de dérive importante, ceux-ci prolifèrent et font croître la part de non-codant, mais cassent également les gènes non essentiels, provoquant l'apparition d'introns. Dans ce contexte, si un individu arrive à détourner une machinerie existante pour les exciser, il peut réactiver tous les gènes qui contiennent des introns. Un tel mécanisme apporterait un avantage direct en terme de réparation d'insertions existantes et de protection des descendants vis-à-vis d'autres insertions. Après spécialisation progressive de la machinerie, on peut plus facilement envisager l'apparition de l'épissage alternatif et la stabilisation de l'utilisation des introns : c'est ce qu'on appelle une exaptation. Cet argumentaire peut être repris dans différents contextes et permet de faire des prédictions générales en lien avec la dérive (Lynch, 2010).

**Réduction via un biais spontané vers les petites délétions (Petrov, Mira, Moran, Kuo, Ochman)** Alors que pour Lynch et Conery (2003), c'est la contre-sélection des éléments transposables (efficace seulement dans les grandes populations) qui freine la croissance des génomes, certains auteurs ont proposé un autre mécanisme : des petites délétions plus fréquentes et/ou plus longues que les petites insertions, induisant un biais spontané vers la réduction du matériel génétique. Ce biais a d'abord été établi chez la drosophile (Petrov *et al.*, 1996; Petrov et Hartl, 1998) puis chez le criquet (Petrov, 2000) et chez la sauterelle (Bensasson *et al.*, 2001). Comparés à la drosophile, ces deux derniers organismes ont des tailles de non codant plus importantes et le biais vers les petites délétions est plus faible. Il y aurait donc un lien entre l'ampleur du biais spontané vers les petites délétions et la taille du non codant.

À la base, Petrov étudie d'ailleurs l'effet des délétions sur des zones considérées comme neutres du génome : éléments transposables *dead-on-arrival* (ayant perdu la capacité de transposer) et pseudogènes. En effet, la séquence de ces éléments n'est *a priori* pas conservée par la sélection, ce qui permet l'accumulation de mutations spontanées. Il voit ce biais comme un mécanisme qui compense l'expansion du *junk DNA* et du *selfish DNA*. Il y aurait alors un point d'équilibre entre l'accumulation de séquences non codantes et leur érosion par petites délétions. Petrov insiste sur le fait que le biais mutationnel est un mécanisme spontané, qui peut être filtré par la sélection et serait donc visible essentiellement pour les séquences non codantes. La quantité d'ADN non codant évoluerait donc par dérive jusqu'à atteindre un équilibre dynamique dans lequel les différentes forces mutationnelles s'équilibrent (Petrov, 2001) et propose un modèle phénoménologique qui fait le lien entre le biais mesuré et la taille de non codant observée (Petrov, 2002). Ce biais a été étudié en détail chez de nombreux organismes et la méthodologie utilisée à l'époque par Petrov a été remise en cause car les zones étudiées sur le génome n'étaient pas complètement neutres et ne représentait pas exactement le biais spontané. Néanmoins, des études récentes confirment ce biais chez de nombreuses bactéries (Kuo et Ochman,

2009) et chez la drosophile (Leushkin *et al.*, 2013).

Le biais spontané vers les délétions élimine la nécessité de la sélection pour justifier que la taille du non codant soit limitée. Si la part de non codant du génome s'érode petit à petit sous l'effet de délétions plus nombreuses que les insertions, il n'y a plus besoin de supposer un effet légèrement délétère de l'ADN en excès (comme dans la théorie de l'ADN égoïste ou dans celle de Lynch et Conery (2003)).

En ce qui concerne la partie codante, les choses sont plus compliquées puisque les petits indels sont contre-sélectionnés s'ils inactivent des gènes, indépendamment du biais. Néanmoins, l'effet du biais devient plus fort s'il est couplé avec une dérive génétique importante. Ce type d'arguments est notamment développé pour expliquer l'évolution réductive des génomes de certaines espèces, dont bon nombre de bactéries endocytobiotes, bactéries qui se sont associées à un hôte et vivent au sein de cet hôte tout en lui fournissant des éléments importants pour sa survie. Du fait des goulets d'étranglements subis par la population bactérienne à chaque reproduction de l'hôte, la taille efficace des populations de bactéries endosymbiotes est bien inférieure à celles des bactéries libres. De plus, comme l'hôte leur fournit un environnement stable ainsi que des nutriments pour la survie, le nombre de gènes essentiels à la survie est faible. Mira *et al.* (2001) suggèrent qu'à cause de la dérive, les gènes non-essentiels sont inactivés (donc pseudogénisés), puis l'ADN non codant érodé à cause du biais vers la délétion, résultant en une réduction progressive du génome jusqu'à ne conserver que les parties essentielles. Cependant, on peut noter que l'argumentation de Mira et al. est un peu moins radicale que celle de Petrov, dans le sens où ils estiment que les petites délétions ne suffiraient pas à expliquer la vitesse d'érosion : il faudrait aussi prendre en compte les grandes délétions.

**Théorie neutraliste : consensus actuel** Les théories développées par Lynch et par Petrov, Mira, Moran, Kuo et Ochman se basent donc sur la même idée mais se focalisent sur des biais mutationnels différents, ce qui explique la différence dans leurs prédictions. D'après les seconds, une petite taille de population entraînera la perte de gènes par dérive puis, par biais mutationnel vers les délétions, l'érosion du non-codant. Pour Lynch, l'accent est mis sur les insertions d'éléments transposables, en supposant par ailleurs que les petites délétions ne sont pas spécialement plus fréquentes que les petites insertions, d'où un grossissement du génome par dérive à cause du non-codant. On remarque ici l'importance des *a priori* : selon le poids accordé aux mécanismes de création ou de destruction de matière, on arrive à des conclusions opposées. Les bases du raisonnement sont cependant les mêmes, ce qui a permis l'aboutissement à une forme de consensus.

Kuo et Ochman (2009) ont étudié l'existence de biais mutationnels chez de nombreuses espèces, révélant un biais pour les 12 bactéries et les 2 archées étudiées, mais pour seulement 2 des 3 eucaryotes étudiés. Le biais chez la levure et la drosophile est plus faible que pour les procaryotes étudiés et il est en faveur des insertions pour les primates. Ce genre d'observations a mené à un traitement différencié des procaryotes et des eucaryotes. Le raisonnement de Lynch est accepté globalement pour expliquer que les procaryotes sont plus petits que les eucaryotes ainsi qu'au sein des eucaryotes, où la fraction de non-codant

est grande et les éléments transposables courants, tandis que le biais à la délétion est faible ou inexistant. Il s'agirait donc d'une tendance globale entre différentes branches de l'arbre du vivant, mais qui ne serait pas forcément valable pour des espèces relativement proches (Whitney et Garland, 2010). Notamment, pour comparer les bactéries entre elles, on s'appuiera plutôt sur les conclusions de Mira *et al.* (2001). Kuo *et al.* (2009) théorisent cette conception et vérifient statistiquement que la dérive entraîne une baisse de la taille du génome chez les bactéries.

**Plantes : nécessité de l'existence d'un mécanisme de délétion à grande échelle**
Comme les plantes font partie des eucaryotes et que leurs tailles de génomes sont très variables à cause de la prolifération d'éléments transposables, le mécanisme décrit par Lynch semble parfaitement convenir. Les chercheurs dans ce domaine n'auront d'ailleurs pas attendu les contributions de Lynch pour expliquer les variations de taille de génome par la prolifération d'éléments transposables, en se référant au principe du *selfish DNA*. La quantité d'éléments transposables était souvent expliquée par la capacité plus ou moins bonne d'une plante à se protéger contre leur prolifération via des mécanismes spécifiques. La taille de génome ne refléterait donc pas la taille de la population mais l'efficacité de la suppression de la prolifération des éléments transposables (mécanismes qui se mettraient en marche quand la transposition atteint un certain seuil (Rabinowicz, 2000)).

La prépondérance accordée aux éléments transposables atteint un point culminant avec un article intitulé *Do plants have a one-way ticket to genomic obesity ?* (Bennetzen et Kellogg, 1997). Probablement à cause de son titre, cet article est cité comme étant un défenseur caricatural de la croissance infinie des génomes. Son contenu est un peu plus mesuré et intéressant. Les auteurs font remarquer que l'accent mis sur les éléments transposables laisse entendre que les plantes font face à un afflux constant d'ADN qu'elles doivent combattre via des mécanismes spécifiquement dédiés à contrer la transposition. Dit de cette manière, on pourrait être amené à croire que leur génome ne fait que croître puisqu'il n'est jamais question d'élimination d'ADN. Pour illustrer l'absurdité de cette proposition, ils appliquent cette hypothèse à une famille de plantes herbacées et en arrivent à la conclusion que l'ancêtre commun devait alors avoir un génome minuscule et que, de temps en temps, la taille du génome a dû être multipliée par des facteurs énormes en un temps très court pour expliquer les tailles de génome actuels. Ils reprennent ensuite les données en supposant que les gains et les pertes d'ADN sont aussi probables les unes que les autres. Ils arrivent à des conclusions plus plausibles, avec des alternances de croissances et de décroissances modérées et un ancêtre commun de taille intermédiaire. Les auteurs préfèrent cette deuxième option mais regrettent l'absence d'un mécanisme de délétion d'ADN qui soit suffisamment performant pour expliquer des phases de réduction qui soient aussi rapides. Pour eux, aucun des phénomènes connus à l'époque n'est crédible, il faut donc chercher à découvrir un mécanisme inconnu ou se résigner à accepter que celui-ci n'existe pas et que les génomes des plantes sont condamnés à gonfler éternellement (d'où le titre de l'article).

Étant donné les variations spectaculaires de taille chez les plantes, les biais vers les petites délétions ne semblent être pas des candidats appropriés. Une partie des chercheurs

du domaine se tournent donc vers les grandes délétions, mais les données actuelles ne permettent que de confirmer partiellement cette hypothèse (nous y reviendrons dans la discussion). De plus, il n'existe pas à l'heure actuelle de modèle que prenne en compte des événements aussi importants, rendant l'interprétation des données plus complexe.

Ainsi, alors que les mécanismes qui poussent les génomes à grandir sont bien identifiés, il n'y a pas encore de consensus sur les mécanismes qui s'opposent à cette croissance.

# 2 Modèles pour la régulation de la taille du génome, du nombre de gènes ou de la quantité d'éléments répétés

Nous allons à présent présenter les différents types de modèles susceptibles d'apporter un éclairage sur les pressions évolutives qui s'exercent sur la taille des génomes.

## 2.1 Modèles liés au problème de la transmission d'information : limitation de la taille de codant

Au sein du dogme central de la biologie moléculaire, l'ADN est vu comme le porteur d'information. Cette information est utilisée de deux façons. Elle est transmise à la descendance via la réplication des chromosomes, ce qui est une des bases du processus d'évolution : tout n'est pas perdu à chaque génération et, à l'inverse, tout n'est pas conservé strictement, ce qui permet de générer de la variabilité. L'information de l'ADN est également utilisée pour produire des ARN messagers puis des protéines, qui sont essentielles pour la construction et le fonctionnement des organismes. De nos jours, avec la montée de l'épigénétique, la vision est moins caricaturale et le support de l'information est plus difficile à identifier clairement. Cependant, le dogme central reste très présent. C'est donc sans réelle surprise que des modèles qui étudient le problème de la transmission d'information se sont placés au niveau de la réplication de l'ADN, puis au niveau de la transmission de l'ADN vers les ARN messagers ou vers les protéines. Dans tous les cas, la transmission imparfaite de l'information entraîne des limites pratiques à la quantité d'ADN codant ou de gènes essentiels au fonctionnement de l'organisme.

### 2.1.1 Limitation de la taille du codant à cause de la réplication de l'ADN

Le travail entrepris par Eigen est l'un des plus importants pour donner une base mathématique à l'étude de l'évolution à moyen terme (Eigen, 1971). Les modèles proposés dans son article séminal ne s'intéressent d'ailleurs pas à des organismes existants ou à l'ADN à proprement parler, mais plutôt à des proto-ARN et des proto-enzymes qui auraient pu être à la base du vivant. Il ne cherche pas à reproduire des données mais à établir des

conditions minimales dans lesquelles l'évolution peut avoir lieu.

Le modèle initial est un modèle général qui permet de prendre en compte la reproduction, la dégradation et la fidelité de la réplication. Eigen étudie des polymères de taille $\nu < \nu_{max}$ où chaque base peut être choisie parmi $\lambda$ élements (par exemple $\lambda = 4$ pour des ARN, $\lambda = 20$ pour des protéines). Il y a alors $\lambda^\nu$ séquences de taille $\nu$ et $\lambda + ... + \lambda^{\nu_{max}}$ séquences possibles en tout. Chaque séquence $i$ est dégradée à un taux $k_0\mathcal{D}_i$ et se reproduit à taux $k_0\mathcal{A}_i$. Le facteur $k_0$ est une constante en $\text{sec}^{-1}$ qui rend les deux autres paramètres sans dimension. La reproduction se fait correctement à fréquence $\mathcal{Q}_i$ et donne par erreur la séquence $l$ à taux $\varphi_{li}$. L'évolution de la concentration de l'espèce $i$ s'écrit alors

$$\dot{x_i} = k_0 \left[\mathcal{A}_i\mathcal{Q}_i - \mathcal{D}_i\right] x_i + \sum_{l\neq i} \varphi_{il} x_l - \varphi_{0i} x_i$$

Le dernier terme ($\varphi_{0i}x_i$) est un terme de dilution : pour que la sélection ait lieu, Eigen introduit des facteurs limitants. On peut limiter le nombre de monomères accessibles (comme les nucléotides) et diluer de manière à conserver un nombre de polymères qui soit constant. Cela permet d'éviter une croissance exponentielle des espèces et d'assurer la convergence vers un polymère maître, baptisé *master sequence* par Eigen. En effet, on peut attribuer à chaque séquence $i$ a une productivité $E_i = \mathcal{A}_i - \mathcal{D}_i$ (liée à la quantité de séquences produites) et une fitness $W_i = \mathcal{A}_i\mathcal{Q}_i - \mathcal{D}_i$ (liée à la quantité de séquences identiques produites). En milieu limitant, Eigen réécrit l'équation ci-dessus sous la forme

$$\dot{x_i} = k_0 \left[W_i - \overline{E}\right] x_i + \sum_{l\neq i} \varphi_{il} x_l$$

La dilution est intégrée dans la production moyenne $\overline{E}$ : à un temps donné, un polymère maintient ou étend sa concentration uniquement si sa production dépasse la production moyenne. Progressivement, toutes les espèces sont éliminées jusqu'à ce qu'il n'en reste plus qu'une $i_m$, qui a la fitness la plus élevée $W_m$. Dans le cas où $\mathcal{Q}_m$ est strictement inférieur à 1, le système se stabilise vers une population stationnaire de séquences identiques $i_m$, accompagnées de mutants qui apparaissent à chaque génération à partir de la séquence $i_m$. Eigen considère que la probabilité que les mutants reviennent vers $i_m$ est négligeable : individuellement, les mutants sont voués à l'extinction, mais à l'échelle de la population, ils sont générés à chaque mutation à partir de $i_m$, ce qui permet éventuellement la convergence vers une distribution stationnaire. On parle souvent de structure en quasispecies (Eigen *et al.*, 1988).

Le facteur $\mathcal{Q}_m$ joue un rôle important. Si $\mathcal{Q}_m = 1$, l'évolution est bloquée car la population n'est asymptotiquement composée que de la séquence maîtresse $i_m$. Quand on diminue $\mathcal{Q}_m$ on augmente la proportion de mutants et on offre donc plus de possibilités pour une évolution future. Cependant, comme $i_m$ est la séquence maîtresse, elle respecte forcément la relation

$$\mathcal{A}_m\mathcal{Q}_m - \mathcal{D}_m > \overline{\mathcal{A}}_{k\neq m} - \overline{\mathcal{D}}_{k\neq m}$$

À tout instant, il y a donc une qualité de réplication minimale requise $\mathcal{Q}_m > \mathcal{Q}_{min} = (\overline{\mathcal{A}}_{k\neq m} + \mathcal{D}_m - \overline{\mathcal{D}}_{k\neq m})/\mathcal{A}_m$ pour garantir la sélection.

Pour rendre les calculs faisables et adaptés au cas de proto-ARN, Eigen suppose ensuite que la taille des séquences est fixe et regarde quelles tailles peuvent être sélectionnées quand il y a des mutations ponctuelles qui surviennent avec une probabilité $1 - q$ au niveau de chaque base. Même si un polymère parvient à se répliquer indéfiniment et se réplique plus rapidement que d'autres, s'il ne parvient pas à se répliquer à l'identique ou presque, il y a un risque que l'information soit perdue à long terme. On a alors un processus de sélection sans réelle transmission d'information : les polymères se répliquent sans jamais pouvoir acquérir de caractéristique supplémentaire. Eigen calcule alors une valeur seuil, *error threshold* en anglais, du taux de mutation ponctuelle au delà de laquelle la transmission d'information n'est plus possible, pour une longueur de séquence donnée. La probabilité qu'une séquence de longueur $\nu$ se reproduise à l'identique est $q^\nu$. Pour que la séquence soit sélectionnée, il faut alors que $q^\nu > \mathcal{Q}_{min}$, soit

$$\nu < \frac{|\ln \mathcal{Q}_{min}|}{|\ln q|} \simeq \frac{|\ln \mathcal{Q}_{min}|}{1-q}$$

On peut alors inverser le raisonnement : connaissant le taux de mutation ponctuelle, on peut chercher la plus longue séquence pouvant être transmise. Eigen argumente qu'en l'absence de conservation d'information, le processus d'évolution ne peut avoir lieu. Plus la séquence est longue, plus la fidélité de la réplication doit être grande. Le taux de mutation fixe une quantité maximale d'information qui peut être transmise.

Notons que dans ce premier modèle, toutes les bases sont en quelque sorte prises en compte comme étant porteuses d'information, il n'y est donc pas question de non-codant. Le reste de l'article est consacré à la complexification progressive du processus d'évolution, notamment pour passer de l'état de proto-ARN, qui code essentiellement pour une protéine, à plusieurs ARN fonctionnant ensemble. C'est la théorie de l'hypercycle, qui n'a plus de lien direct avec la question étudiée ici.

Le modèle d'Eigen est souvent appliqué à l'évolution des virus, qui ont des tailles de génomes très petites et des taux de mutation ponctuelle très élevés (Eigen et Schuster, 1979; Nowak, 1992; Wilke, 2003). Des études postérieures ont assoupli certaines hypothèses comme la taille infinie de la population (Nowak et Schuster, 1989) ou l'homogénéité des taux du mutation le long de la séquence (Barbosa *et al.*, 2012), mais dans tous les cas les mutations considérées restent locales, quoique l'importance des duplications et des délétions soit discutée dans Eigen *et al.* (1988).

### 2.1.2 Limitation du nombre de gènes à cause de l'expression des gènes

Plus tard, la transcription a été prise en compte et ajoutée à la problématique de la transmission d'information. Il faut faire attention ici à une différence fondamentale. Eigen considérait des erreurs transmises aux générations futures et donnant lieu à un phénomène d'accumulation qui est à l'origine de l'error threshold. Quand on prend en compte la transcription ou la traduction, on parle d'erreurs non héritables : ce type d'erreurs n'empêche pas l'évolution à long terme mais le fonctionnement de l'organisme en temps réel.

L'idée est la suivante : en théorie, les gènes (et l'épigénétique) organisent un réseau qui permet à l'organisme de fonctionner si on considère que leur transformation en protéines se fait sans problèmes mais en pratique, la machinerie de transcription et de traduction fait des erreurs qui peuvent déstabiliser ce réseau théorique.

L'idée originale est souvent attribuée à Bird (1995). En se basant sur les données disponibles à l'époque, il identifie 3 catégories d'organismes : procaryotes, eucaryotes non vertébrés, eucaryotes vertébrés. À chaque catégorie semble être associée une plage de gènes bien précise qu'il associe à un processus de complexification au cours de l'évolution. Les procaryotes ont moins de 10 000 gènes, les eucaryotes ont réussi à repousser la limite à environ 40 000 gènes et les vertébrés à 100 000[1]. Selon lui, une différence de taux de mutation n'est pas suffisante pour expliquer cette limite (il rejette donc une interprétation type *error threshold*). Il s'agit selon lui d'un problème de régulation, plus précisément de *silencing*[2] des gènes, qui ont toujours tendance à s'exprimer à un certain niveau basal, ce qui perturbe le fonctionnement de l'organisme. Plus il y a de gènes, plus il y a d'erreurs et de protéines non désirables. Il faut donc avoir un meilleur système de régulation pour réussir à augmenter le nombre de gènes sans augmenter les protéines superflues. Dans la transition procaryote vers eucaryote, l'encapsulation par le noyau et l'enroulement de l'ADN autour d'histones aurait permis de lever une première barrière. Dans la transition vers les vertébrés, il invoque l'utilisation massive de la méthylation comme moyen de régulation supplémentaire.

L'approche de Bird est qualitative, mais on peut en faire un modèle, notamment pour faire la distinction entre erreurs de réplication et erreurs de régulation. Pál et Hurst (2000) proposent un modèle basé sur la génétique des populations qui distinguent les erreurs héritables (liées aux mutations ponctuelles) et les erreurs non héritables. Ces dernières comprennent les problèmes de *silencing*, une stochasticité d'expression trop importante, les problèmes de méthylation transmis à la lignée cellulaire, les problèmes de biosynthèse des protéines (repliement ou modifications post-traductionnelles). Le modèle distingue le taux de mutation $\mu$, la probabilité qu'un gène soit traduit avec/par erreur $p$ et l'impact sur la fitness de cette erreur $\delta$, en supposant un effet multiplicatif des erreurs. D'après le modèle, l'effet des erreurs non héritables est plus important que celui des mutations dès que $p\delta > \mu$ et pourrait donc limiter de fait le nombre de gènes utilisables avant que l'*error threshold* ne s'applique. Cependant, $p$ et $\delta$ sont difficiles à estimer, on peut simplement s'attendre à ce que $\delta$ soit très faible. Ils utilisent également le modèle pour justifier la différence qualitative entre vertébrés et unicellulaires eucaryotes identifiée par Bird. Si on suppose que la méthylation a un coût supplémentaire mais permet de réduire la valeur de $p$, elle peut être sélectionnée à condition que l'impact des erreurs non héritables $\delta$ soit assez élevé. Les auteurs pensent que pour les organismes multicellulaires et à longue durée de vie, cela sera le cas, mais pas pour les organismes unicellulaires et/ou à courte durée de vie. De plus pour les vertébrés, où les cellules sont fréquemment renouvelées, l'impact de certaines erreurs non héritables peut être important si elles touchent les cellules à la base du renouvellement.

[1] Les valeurs données sont celles de l'article de Bird (1995). Elles ont été, depuis, revues à la baisse.

[2] La traduction de *silencing* en français existe : il s'agit d'« extinction ». Ce mot étant trop ambigu, nous avons préféré utiliser la version anglaise.

Dans la même ligne, Zeldovich *et al.* (2007) proposent un modèle qui vise un problème plus précis, mais avec une représentation un peu moins phénoménologique que celle de Pál et Hurst. Il s'agit de prendre en compte les problèmes de repliement pour des protéines essentielles à l'organisme. Zeldovich considère que les protéines fonctionnent correctement pour une gamme de repliements donnée délimitée par les énergies $E_{min}$ et $E_{max}$ et qu'à chaque mutation les $\Gamma$ protéines essentielles d'un individu changent d'énergie selon un noyau de transition $W$ déterminé approximativement d'après des données expérimentales. Dès qu'une protéine atteint la valeur $E_{max}$, l'individu meurt. La fitness est binaire : tant qu'il survit, l'individu a une fitness indépendante du niveau d'énergie des protéines. En supposant que les protéines évoluent indépendamment, on a une population dont le nombre d'individus est donné par une exponentielle. Cette exponentielle est croissante si le nombre d'individus se reproduisant est plus élevé que le nombre d'individus qui meurent à chaque génération parce qu'une protéine atteint $E_{max}$. C'est le cas uniquement si le nombre de gènes essentiels est plus petit que $\Gamma^*$, qui dépend du taux de croissance des survivants, du taux du mutation et du noyau de transitions. Le modèle a l'avantage de pouvoir être relié à des expériences. D'abord, Zeldovich *et al.* (2007) comparent la distribution stationnaire des niveaux d'énergie théoriques aux distributions mesurées sur certaines protéines issues de la base de données ProTherm puis déduisent que la limite de viabilité est de 6 mutations par gène essentiel, qui serait une valeur retrouvée dans des expériences avec des virus à ARN. D'autre part, $E_{max} - E_{min}$ diminue quand la température augmente, ce qui accélère l'inactivation des protéines et la limitation du nombre de gènes essentiels, ce qui semble également vérifié en comparant les bactéries mésophiles aux thermophiles et hyperthermophiles.

## 2.2 Modèles pour l'évolution des séquences non-codantes

Les modèles ci-dessus étudient des limites de transmission d'informations qui rendent impossible l'augmentation du nombre de gènes au-delà d'un seuil donné. Leurs résultats ne s'appliquent donc pas aux séquences non codantes qui ont des dynamiques propres. Parmi les séquences non codantes, deux types semblent se développer en partie indépendamment des séquences codantes : les séquences répétées en tandem et les éléments transposables (rappelons que bien qu'ils portent des gènes nécessaires à leur transposition, les éléments transposables sont généralement comptés dans la part d'ADN non codant d'un génome). Comme elles semblent avoir la faculté de se développer très rapidement, elles ont attiré l'attention de nombreux modélisateurs qui cherchent à comprendre leur dynamique mais aussi pourquoi leur nombre reste borné. Les mécanismes utilisés pour modéliser ces deux types d'éléments sont assez proches : Charlesworth *et al.* (1994) proposent une review où les deux approches sont traitées et comparées. Un certain nombre des modèles présentés ci-dessous sont tirés de cette review.

### 2.2.1 Modèles pour les séquences répétées en tandem

Chez les eucaryotes, on classe les séquences répétées en tandem selon la longueur du motif et le nombre de répétitions (Charlesworth *et al.*, 1994). Les microsatellites ont un motif de 2-5 paires de bases (bp) et atteignent une taille d'environ 100 bp. Les minisatellites ont un motif d'environ 15 bp répétés sur 0.5 à 30 kb. Les satellites ont un motif de 2 à 100 bp et sont localisés dans des clusters qui peuvent atteindre jusqu'à 100 Mb, situés dans l'hétérochromatine (dans les télomères, centromères et le chromosome Y notamment). Le nombre de motifs dans ces séquences évolue très rapidement, soit par recombinaison chromosomique, soit par glissement de la polymérase de réplication (Bhargava et Fuentes, 2009). La recombinaison est invoquée comme pouvant être à l'origine des séquences répétées (Smith, 1976), puis responsables d'échanges des motifs entre chromatides par recombinaison inégale (section 1.1) : le nombre total de copies est conservé, une chromatide en gagne, l'autre en perd. En comparaison, le glissement de la polymérase a un effet plus limité mais change le nombre total de copies. La nature de ces mutations est fondamentalement différente : le nombre de copies échangées par recombinaison dépend du nombre de copies présentes tandis que les glissements ont une nature additive (indépendante de la taille de départ).

Plusieurs modèles ont été proposés, en s'appuyant sur l'un ou l'autre des mécanismes. Krüger et Vogel (1975) ont proposé un modèle basé sur la recombinaison inégale de chromosomes ou de matériel génétique en général qui s'inscrit dans ce cadre. La taille des recombinaisons est limitée à un élément seulement dans le modèle analysé, elles deviennent donc purement additives. Le modèle adopté utilise un formalisme à base de gamètes qui ne change pas le fait que, pour une taille donnée, on gagne ou on perd un élément à chaque mutation, avec un léger biais toutefois : quand un allèle de taille quelconque doit recombiner avec un allèle comportant un élément, soit il conserve sa taille, soit il perd un élément. La conservation privilégiée des allèles comportant un élément assure la convergence du système. On obtient à la fin une distribution géométrique avec un maximum pour les allèles à un élément. Pour sortir de cette distribution, les auteurs utilisent la fitness, qui contrôle facilement le comportement étant donné que le biais du système spontané est relativement faible. Ils testent deux fonctions de fitness, l'une évolue de manière quasi-exponentielle avec le nombre d'éléments, l'autre atteint un maximum pour un nombre d'éléments $n_{opt}$. La distribution suit exactement la fonction de fitness : si l'exponentielle croît, elle ne se stabilise pas, si elle décroît, elle est encore plus piquée en 1, s'il y a un nombre d'éléments optimal $n_{opt}$, elle atteint un maximum en cette valeur. La nature additive des mutations est donc totalement contrôlée par la force de la sélection.

Walsh (1987) prend en compte deux types de recombinaisons : recombinaisons inégales entre chromatides ou recombinaison intrachromosomique menant à une délétion. Comme Krüger, il limite les échanges ou les pertes à un élément mais ils sont effectués avec une probabilité qui croît avec le nombre d'éléments. L'état avec un élément est considéré comme absorbant. À nouveau, le système est biaisé vers une baisse et converge logiquement vers l'état absorbant. En théorie, ce qui change par rapport à Krüger et Vogel (1975), c'est que les vitesses de changements sont plus rapides quand le nombre d'éléments est grand,

soit une forme de diffusion biaisée dont le coefficient augmente quand on s'éloigne de l'origine. Cependant, l'auteur utilise une modélisation en chaîne de Markov pour estimer la vitesse de convergence et, pour obtenir des probabilités de transition, renormalise les taux et élimine cette propriété : on revient à une forme de diffusion biaisée homogène. Même en autorisant les duplications pour l'état à un élément, la distribution converge rapidement vers une distribution avec un mode en un ou deux éléments. Pour déplacer le mode plus amplement, l'auteur introduit l'effet de la sélection, qui permet de contrôler assez facilement le comportement du système. On retrouve donc les résultats de Krüger et Vogel (1975), mais avec un formalisme légèrement différent.

Les résultats deviennent plus complexes quand on prend en compte la nature non-additive de la recombinaison. Stephan (1987) conçoit un modèle de génétique des populations où la taille des changements n'est plus limitée à un élément. Plus précisément, il prend en compte deux événements multiplicatifs : une amplification de taux $\mu$ et une recombinaison inégale liée à la méiose de taux $\gamma$. Dans ce cas, le système est en apparence plutôt biaisé vers les gains, puisque les recombinaisons permettent des gains et des pertes de manière symétrique. L'auteur ajoute toutefois un facteur, qui s'apparente à de la fitness, qui permet de moduler les transitions. L'analyse utilise une mise à l'échelle linéaire, qui ne permet pas prendre en compte proprement les effets multiplicatifs et d'aboutir à une prédiction du comportement du système. L'auteur utilise alors des simulations dans un domaine fini. Il note que le nombre d'éléments est bas puis explose tout à coup quand le rapport $\mu/\gamma$ devient plus élevé. Le système se comporte donc de manière assez binaire : soit le nombre d'éléments est très faible, soit très élevé.

On peut considérer que le modèle le plus abouti est celui présenté par Falush et Iwasa (1999) puisqu'il mêle les effets additifs et les effets multiplicatifs. Les auteurs notent que chez certaines espèces, il pourrait y avoir un biais dans l'acquisition des éléments qui augmente avec le nombre d'éléments initialement présents (figure I.4A). Ils suggèrent donc un modèle qui prend en compte des effets additifs liés aux problèmes de réplication et des effets multiplicatifs biaisés vers le gain dus, selon eux, à des échanges entre chromatides (figure I.4B). Biologiquement, l'origine de ce biais n'est pas claire, puisqu'on s'attend à une conservation globale du nombre d'éléments en cas d'échange, alors qu'ici il augmente en moyenne. Mathématiquement, on a une situation proche de celle de Stephan (1987) : le système est *a priori* biaisé vers le gain. Un paramètre $a$ permet de contrôler le biais : le système est biaisé vers les gains dès que $a > 0$. Les auteurs proposent une approximation diffusive pour calculer les moments de la distribution stationnaire. Ils montrent que, malgré le biais à l'augmentation, le système peut converger pour une gamme de valeurs de $a$ assez large, avec des moments qui divergent progressivement. Pour $a < 0.268$, les deux premiers convergent, pour $0.268 < a < 1$, seule la moyenne converge. Le paramètre $a$ permet donc d'augmenter la valeur moyenne et l'éclatement de la distribution sans que la sélection soit nécessaire. On peut obtenir des trajectoires individuelles très saccadées (I.4C) et une distribution pour la population avec un mode à une position plus intéressante que 1 ou 2 (I.4D). Grâce à la structure multiplicative, on obtient un résultat non-intuitif selon lequel le nombre d'éléments ne diverge pas malgré le biais vers les gains apparents. Grâce à la structure additive, on a un mode avec un nombre d'éléments qui correspond à des valeurs plausibles. Il existe donc possible d'obtenir une distribution de

tailles de microsatellites proche de la réalité en utilisant uniquement des mécanismes qui ne dépendent pas de la sélection.

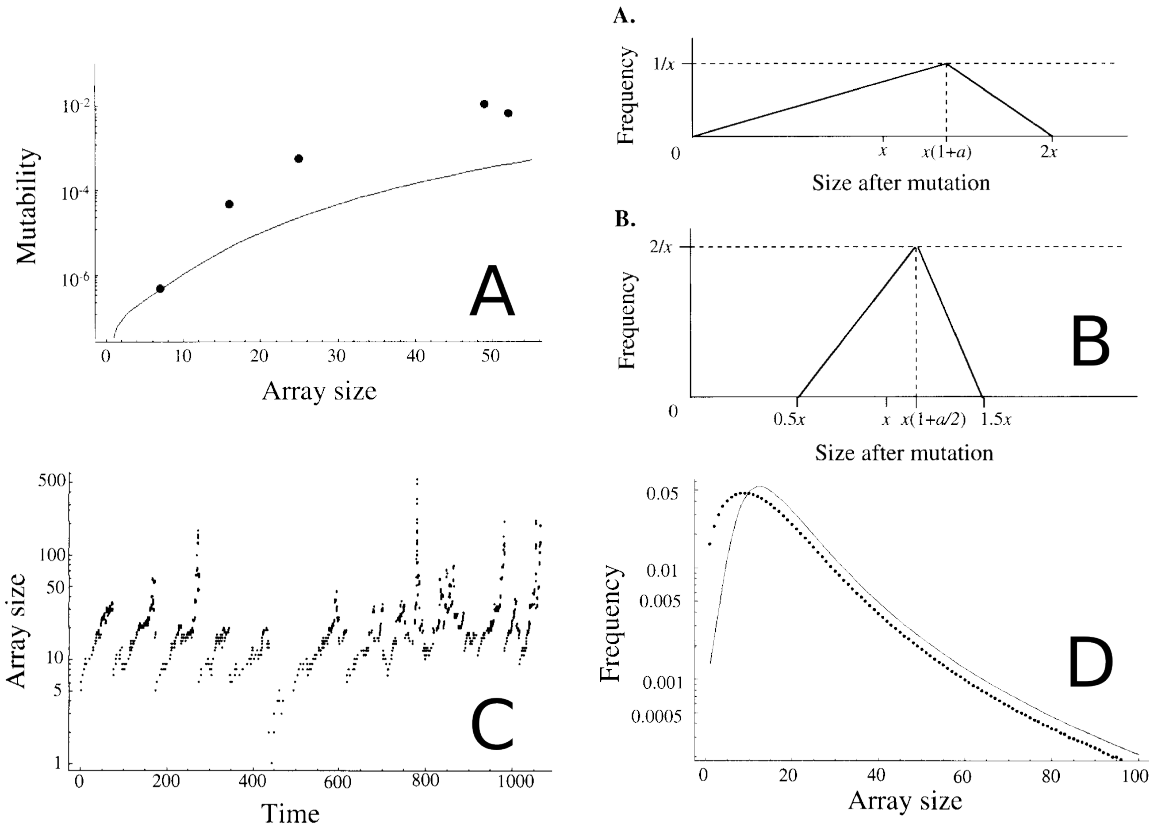


Figure I.4 – Images tirées de Falush et Iwasa (1999). Les auteurs constatent que l'instabilité des microsatellites croît avec le nombre d'éléments (A). Pour l'expliquer, ils proposent deux modèles d'échanges de copies entre chromatides contrôlées par un biais $a$ (B). Le modèle A. suppose que toutes les copies peuvent être échangées, le modèle B. que seules la moitié est concernée. En bas, on montre la trajectoire du nombre de copies pour un individu (C) et la distribution stationnaire pour l'ensemble de la population (D – simulée en pointillés, approximation analytique en trait plein).

### 2.2.2 Modèles pour l'évolution des éléments transposables

Les éléments transposables sont classés selon leur composition et leur mode de transposition. On distingue généralement les rétrotransposons (dits de classe I) qui utilisent un intermédiaire ARN pour se tranposer, des transposons (dits de classe II). Chez les rétrotransposons, on classe les éléments selon la présence ou l'absence d'une longue séquence répétée (LTR, qui influence le mode de transposition) et selon la capacité à coder l'enzyme de transposition. Les éléments transposables peuvent subir des événements de même nature que les microsatellites. L'excision des éléments et leur amplification est souvent considérée comme étant additive, ce qui correspondrait aux glissements de polymérase pour les séquences répétées en tandem. La recombinaison peut également agir sur plu-

sieurs éléments à la fois et avoir un effet qui dépend fortement du nombre d'éléments déjà présents.

Pour ces raisons, les modèles sont assez proches de ceux présentés pour les microsatellites. On peut par exemple citer le modèle d'Ohta et Kimura (1981), qui prend en compte la transposition, des effets de recombinaison et de la dérive génétique méiotique. Cela se traduit notamment par la prise en compte d'effets additifs via ce que les auteurs appellent la *single replication* et des effets multiplicatifs via la *cluster replication*. On est donc dans un modèle proche de Falush et Iwasa (1999) mais avec des effets de génétique des populations de tailles finies qui rendent les équations plus compliquées. Les auteurs étudient uniquement les deux premiers moments (sans mise à l'échelle) et remarquent que ces deux premiers moments peuvent diverger assez facilement, et augmentent dans tous les cas tous les deux rapidement, ce qui ne permet pas de localiser la population en fonction des paramètres.

On trouve également des modèles qui ne prennent en compte que les événements additifs. Moody (1988), puis Basten et Moody (1991), utilisent un formalisme emprunté aux processus de branchements pour étudier des éléments transposables qui évoluent avec des variations de $\pm 1$ copie et éventuellement de la sélection. La différence principale est l'introduction de probabilités de transitions qui dépendent du nombre de copies et une probabilité de rester sur place, avec un nombre d'éléments borné. On retrouve globalement les mêmes résultats que pour les microsatellites. Initialement, les auteurs considèrent le cas particulier où les probabilités de transition ne variant pas avec le nombre d'éléments : on a une distribution quasi-géométrique avec un mode en 0. Les auteurs considèrent alors des variantes où le mode de la distribution est contrôlé via la fitness ou en modifiant les valeurs de transitions, qui introduisent un biais vers un état précis. Dans ce dernier cas, les auteurs invoquent une transposition dépendant du nombre d'éléments déjà présents, qui peut être réglée de manière à favoriser la transposition sous le mode désiré et à favoriser l'excision au-dessus. Ce type d'arguments se base sur l'existence de mécanismes pour limiter la transposition chez certains organismes(voir par exemple Lisch, 2009).

Langley *et al.* (1988) adoptent un point de vue légèrement différent concernant le rôle de la recombinaison dans la régulation des éléments transposables. Pour les microsatellites, l'échange entre chromatides est un mécanisme plausible, alors que pour les éléments transposables, l'échange ne peut se produire de manière neutre que si les éléments sont regroupés en clusters ne contenant pas de gènes, comme suggéré par Ohta et Kimura (1981). Cependant, on peut s'attendre à ce que les éléments transposables s'insèrent de manière dispersée dans le génome et que, s'il y a recombinaison entre deux éléments quelconques, on ait de grandes chances que la recombinaison se produise entre deux chromosomes ou deux zones de chromosomes différentes : on parle de recombinaison ectopique. On risque alors de perdre la structure du chromosome ou d'échanger des gènes en plus des éléments transposables. Si une telle recombinaison se produit au moment de la méiose, on obtient deux gamètes avec des chromosomes aberrants, probablement non viables. Langley *et al.* (1988) voient donc la recombinaison comme un moyen de limiter l'expansion des éléments transposables à cause de la non viabilité des gamètes produits. Ils conçoivent un modèle de génétique des populations avec les effets additifs classiques (transposition et excision)

et une fitness qui ne dépend que de la recombinaison ectopique. Ce modèle permet d'inclure des taux de recombinaisons qui varient par segment de chromosome, ce qui permet de faire des vérifications expérimentales en comparant des zones qui recombinent peu à des zones qui recombinent fréquemment. Ils prévoient que le nombre d'éléments transposables reste faible et une tolérance plus grande dans les zones qui recombinent peu, ce qui semble corroboré en partie par les données expérimentales. Le nombre d'éléments serait donc limité par l'effet délétère des recombinaisons ectopiques plus que par les taux d'excision.

D'autres modèles issus de la génétique des populations prennent en compte ce type d'effets. Pour plus de détails, on se réfèrera à Charlesworth *et al.* (1994) et Rouzic et Deceliere (2005). Pour conclure cette sous-section, on remarquera que différent types de formalismes sont utilisés pour étudier les éléments transposables et les séquences répétées mais qu'on retrouve toujours les mêmes ingrédients : effets additifs, effets multiplicatifs, fitness et biais des mutations. Pour obtenir une distribution intéressante, ces modèles utilisent souvent un biais ou une fitness orientés vers une valeur précise, mais on a vu que le mélange d'effets multiplicatifs et additifs permet également d'obtenir des valeurs intéressantes de manière beaucoup moins intuitive.

## 2.3 Modèles pour l'évolution d'un trait quelconque

Nous avons détaillé jusqu'ici des modèles qui mettent l'accent sur des parties bien identifiées du génome, souvent centrées sur l'étude d'un individu en particulier ou prenant en compte les effets de population via le formalisme de la génétique des populations. Dans cette sous-section, nous mettons en avant les modèles qui cherchent un formalisme adapté pour décrire l'évolution d'un ou plusieurs traits quelconques (souvent abstraits) à l'échelle de la population. Ces modèles ne s'intéressent pas spécifiquement à la question de la taille du génome ; il s'agit ici de donner un bref aperçu.

On peut commencer par distinguer les modèles qui cherchent un formalisme général qui englobe les systèmes étudiés classiquement en génétique des populations. Il s'agit de donner une représentation de la population à l'aide d'une mesure puis de caractériser le processus qui agit sur cette mesure. Fleming et Viot (1979) proposent un formalisme valable pour un espace « génomique » compact (au sens mathématique), basé initialement sur des modèles à allèles multiples, comme le modèle de Wright-Fisher. Il ne s'agit donc pas de faire varier la structure du génome mais d'étudier la convergence d'un processus de population dans l'espace d'allèles en prenant en compte le plus de mécanismes possibles, comme la sélection, la dérive génétique et les mutations. La dérive génétique s'apparente à de la diffusion et les mutations peuvent être interprétées comme un opérateur linéaire. L'introduction de la sélection complique l'analyse, mais cette étude offre un cadre stochastique formel qui permet d'articuler toutes ces notions.

Plus récemment, on trouve des modèles introduisant des processus qui permettent d'enrichir des modèles classiques prenant en compte une mesure de fitness. On peut citer

par exemple la coévolution de traits vers un certain optimum, habituellement étudiée dans un contexte de « paysage évolutif ». Si les processus de mutations permettent des déplacements faibles dans l'espace des traits, la population suit le gradient qui mène à l'optimum de fitness (le sommet) le plus proche. Dieckmann et Law (1996) reviennent sur les hypothèses qui mènent à ce comportement macroscopique, décrit par une équation canonique, et propose des modèles microscopiques qui permettent de le reproduire mais aussi de l'étendre. Grâce aux modèles microscopiques, on peut délimiter les conditions dans lequel le comportement moyen donné classiquement peut être observé et donner en plus une description plus riche de la dynamique adaptative de la population. En particulier, il permet de prendre en compte, théoriquement, les interactions entre plusieurs traits présents simultanément qui peuvent perturber la convergence.

Cette idée d'interaction entre individus est absente dans beaucoup de modèles classiques, notamment de génétique des populations. Dans un contexte écologique, cette idée est assez naturelle. Par exemple, si une source de nourriture est moyennement abondante, la quantité de ressources qu'une espèce pourra en tirer dépend de la présence d'autres espèces qui puisent dans la même ressource. L'opportunité d'une stratégie évolutive dépend donc de la présence d'autres individus adoptant la même stratégie ou des stratégies similaires. Geritz *et al.* (1998) illustrent le rôle de « stratégies évolutivement singulières », en se focalisant sur la topologie des points où le gradient est nul, qui peuvent mener à l'apparition de dimorphisme dans une population initialement monomorphe (voir aussi Diekmann *et al.*, 2005). Champagnat *et al.* (2006) étudient l'influence des trajectoires individuelles dans l'évolution de trait basée sur un modèle de mutation-sélection. De la même façon, l'introduction d'interactions avec les autres individus dans les processus de naissances, mutations et de morts permet de renouer avec des modèles classiques (comme les équations de Kimura (voir par exemple Kimura, 1964)) mais également de les étendre, selon la renormalisation utilisée pour passer des processus microscopiques aux processus de population (voir aussi Champagnat et Lambert, 2007).

La prise en compte d'effets de population et d'interactions peut également se faire via les processus de branchements. En introduisant une dépendance entre les différents processus, on peut introduire des effets de sélection ou d'interactions écologiques. Lambert (2005) propose cette formalisation dans le cadre du modèle logistique, qui décrit l'évolution d'une population dans un environnement avec des ressources limitées, mais ne prend classiquement pas en compte l'interaction entre les individus. Cette vision microscopique d'individus qui se reproduisent en interagissant dans un environnement à ressources éventuellement limitées peut ensuite être intégrée à d'autres modèles, qui prennent en compte les mutations et l'évolution de traits soumis à la sélection. Par exemple, Lambert (2006) incorpore progressivement ce type de processus dans le modèle à deux allèles de Haldane et Fisher. Dans le modèle original, il s'agit de calculer la probabilité de fixation d'un allèle dans une populations de taille $N_e$, sachant que l'allèle est présent à fréquence $p$ initialement et apporte un gain de fitness $s$. L'ajout de compétition, de robustesse et d'effets de variations de la population permettent d'enrichir le résultat original mais aussi de montrer qu'il est biaisé quand la population globale est dans une phase de croissance ou de décroissance.

Dans un registre différent, plusieurs modèles cherchent à rendre compte d'une dynamique évolutive spécifique. Par exemple, de nombreux articles étudient comment la population trouve un compromis entre la nécessité de robustesse et d'adaptabilité. Pour avoir un espace de phénotypes de topologie assez arbitraire, plusieurs études considèrent une population qui évolue sur des réseaux mutationnels et montrent que, selon la géométrie de l'espace, les deux notions ne sont pas nécessairement opposées (Van Nimwegen *et al.*, 1999; Wagner, 2008). Plus récemment, Draghi *et al.* (2010) utilisent une approximation continue d'une chaîne de Markov pour représenter les phénotypes et arrivent à une conclusion similaire. Enfin, une autre difficulté consiste à prendre en compte les variations de l'environnement. Ancel Meyers *et al.* (2005) étudient l'impact de la fréquence de variation de l'environnement entre deux états A et B sur la sélection de gènes spécifiques à un environnement A ou B ou de gènes polyvalents, mais moins performants. Les gènes spécifiques ne sont sélectionnés que si la variation de l'environnement est assez lente. Dans ce cas, à chaque variation environnementale de A vers B, les gènes correspondant à l'environnement A sont éliminés au profit des gènes correspondant à l'environnement B, et réciproquement. Si la fréquence de variation est trop rapide, ce renouvellement ne peut pas avoir lieu, ce qui favorise le gène polyvalent (il y a un coût à maintenir les deux allèles spécialisés). Leibler et Kussell (2010) utilisent un formalisme de physique statistique pour étudier l'évolution d'individus dans un environnement variable. Pour un environnement donné, ils étudient l'ensemble des trajectoires individuelles au cours du temps (obtenues à partir de branchements) et cherchent à quantifier la variation de fitness due aux variations environnementales de la variation de fitness due aux mutations. Ils montrent que les variations environnementales perturbent les mesures classiques de la sélection mais que l'évolution d'une lignée permet, en théorie, d'obtenir les informations nécessaires pour calculer la sélection effective.

Tous ces modèles s'intéressent à des traits phénotypiques : prise en compte de la fitness en se basant sur la structure de la population, robustesse, adaptabilité. Dans l'évolution de la structure du génome, ces éléments vont jouer un rôle important, mais il est difficile de les prendre en compte tout en se rattachant à des mécanismes détaillés de l'évolution du génome. Dans les modèles présentés dans cette sous-section, le génome n'est pas représenté explicitement : on définit la portée d'une mutation dans un espace de phénotypes, avec une portée relativement limitée dans cet espace de phénotype. Dans le modèle que nous voulons développer, les réarrangements auront régulièrement des effets assez forts, que ce soit en terme de fitness, d'adaptabilité ou de robustesse. Cependant, ce sont des premières approches qui vont vers l'unification des processus mutationnels (pour l'instant basées sur des approximations diffusives), la sélection, les effets de petite population et l'interaction entre individus. Dans notre modèle, ces derniers aspects seront absents ou simplifiés. Même si nous nous attacherons à justifier les choix simplificateurs que nous feront pour la structure de population ou pour la fitness, il serait intéressant d'intégrer, à terme, des effets de petites populations ou de sélection plus réalistes que ceux présentés dans ce manuscrit.

## 2.4 aevol

Pour finir le survol des démarches de modélisation, nous introduisons un modèle particulier, qui n'est pas un modèle mathématique, mais qui prend en compte différents aspects explorés par les modèles précédents. Le logiciel présenté dans Knibbe *et al.* (2007), baptisé *aevol* pour *a*rtificial *evol*ution, prend en compte les aspects liés aux mutations dans le codant et le non codant, à la taille de la population et à la fitness. Ce modèle individu-centré est trop compliqué pour pouvoir être analysé mathématiquement mais suffisamment simple pour étudier comment la structure du génome évolue quand on fait varier indépendamment les effets liés aux mutations, à la taille de la population ou à la fitness.

La description complète est donnée dans Knibbe (2006). On se limitera ici aux aspects pertinents pour la taille et la structure du génome. Dans aevol, la séquence génomique de chaque individu de la population est explicitement simulée. Le nombre et la taille des gènes peut varier grâce à une structure proche de la biochimie des procaryotes. Pour être traduit en protéine, un gène doit être transcrit et traduit, ce qui nécessite un certain nombre de séquences signal (prédéfinies) en amont et en aval du gène. Les gènes sont localisées sur un chromosome circulaire double brin avec des bases $\{0, 1\}$ au lieu de {A, C, G, T}. La biochimie est donc simplifiée et adaptée au système binaire. Le reste du génome est non codant. Ce système permet une grande flexibilité de structure. La longueur du génome, la longueur des gènes, le pourcentage de codant peuvent varier, la façon de réguler et transcrire les gènes peuvent varier.

Les mutations prennent en compte les mutations locales (mutations ponctuelles, petites insertions, petites délétions) et les réarrangements chromosomiques (inversions, translocations, grandes délétions et duplications). Les réarrangements chromosomiques sont essentiels pour l'évolution des génomes. Sans duplication, il est presque impossible d'augmenter le nombre de gènes, et les autres réarrangements peuvent modifier la répartition des gènes le long du chromosome ou la quantité de non codant. Les mutations simulées sont les mutations spontanées : elles peuvent être neutres, avantageuses ou délétères voire létales. Seuls les individus les plus adaptés se reproduiront. Ainsi, les mutations fixées dans la population ne sont qu'un petit sous-ensemble des mutations spontanées qui se produisent à chaque génération. Les propriétés des mutations fixées peuvent être très différentes de celles des mutations spontanées. Par exemple, les grandes délétions fixées sont rares comparées aux petites, alors que toutes les tailles sont équiprobables dans le mécanisme spontané simulé. Nous reprendrons les mutations prises en compte dans aevol pour le modèle étudié dans cette thèse, nous donnerons le détail des processus plus bas (section 3).

La sélection utilise un principe volontairement simple pour garantir que la structure des génomes ait un sens. L'individu doit effectuer un certain nombre de processus biologiques abstraits qui doivent être remplis avec le bon dosage. Il ne suffit donc pas de réaliser les processus, mais d'ajuster précisément le niveau d'expression. La population est de taille finie, ce qui ajoute des effets de dérive génétique. La sélection permet le développement

d'un répertoire génique complet (les génomes étant initialisés avec un seul gène).

Dans les simulations, les populations convergent vers des quantités d'ADN codant et de non codant assez stables, qui sont très intimement liées à certains paramètres de mutation. Quand on augmente tous les taux de mutations conjointement, on observe un rétrécissement des parties codantes et des parties non codantes du génome (figure I.5). Une analyse plus approfondie montre que ce comportement est essentiellement contrôlé par les grandes délétions et les duplications, de nature multiplicative dans le modèle. Même en adoptant une distribution de pertes et de gains parfaitement symétrique, la taille du génome converge vers une taille indépendante de la taille de départ, à condition que la taille des délétions et des duplications augmente avec la taille du génome. Dans de nombreuses simulations, le nombre de gènes et la taille du non codant sont tous les deux limités en premier lieu par le processus de duplications et de délétions. En effet, l'ADN non codant, bien qu'il n'impacte pas directement la fitness de l'individu, est mutagène pour les gènes dans le sens où il fournit des breakpoints pour des délétions et des duplications. C'est ce processus, mis en évidence grâce à aevol (Knibbe *et al.*, 2007), que nous allons étudier plus en détails dans ce manuscrit.

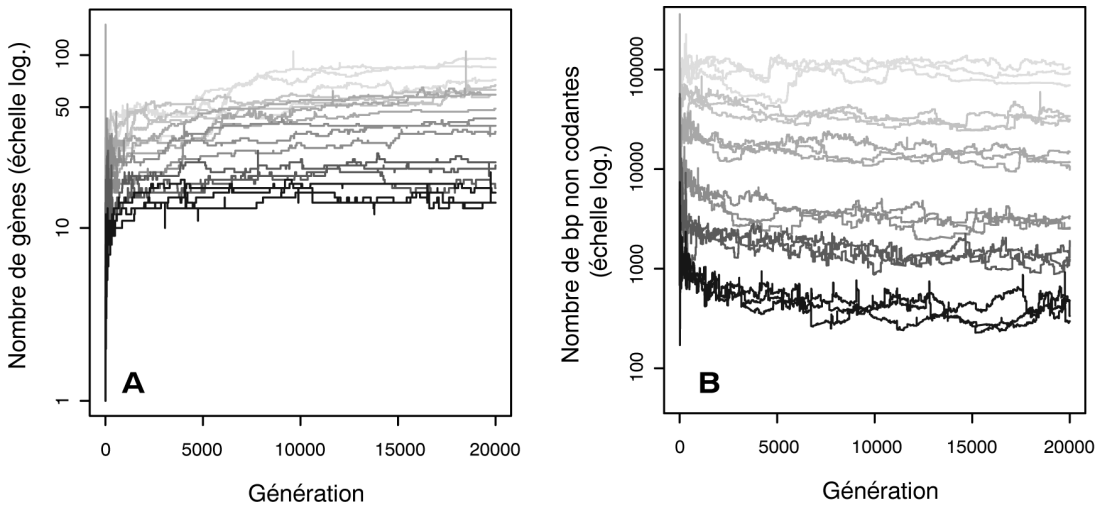


Figure I.5 – Évolution du codant (A) et du non codant (B) au cours du temps lors de simulations d'aevol pour différents taux de mutations, mais identiques pour tous les types de mutations (mutations ponctuelles et réarrangements). Plus la courbe est foncée, plus les taux de mutations sont élevés, ils vont de $5.10^{-6}$ à $2.10^{-4}$ mutations par paire de bases par génération (images tirées de Knibbe (2006))

# 3 Mise en place d'un modèle pour l'étude des pressions mutationnelles sur la taille du génome

## 3.1 Idée générale

Les modèles présentés jusqu'ici utilisent en général des mécanismes assez simples et font souvent appel à la contre-sélection directe des grands génomes pour éviter des effets indésirables, comme la croissance infinie de la taille ou du nombre d'éléments considérés. Le risque avec ce type d'approche, c'est d'utiliser la sélection directe pour ignorer tout ce qui ne correspond pas aux données : il suffit de définir une fonction de fitness qui suppose non viables les individus porteurs d'une certaine quantité d'ADN. Même s'il n'est pas impossible que la quantité d'ADN fasse en soi l'objet d'une telle sélection directe, il est cependant nécessaire d'explorer théoriquement les effets opérant même en l'absence de sélection, comme l'ont fait Falush et Iwasa (1999), Knibbe *et al.* (2007) ou ceux qui s'attachent à une vision neutraliste de l'évolution en étudiant les biais mutationnels. Ce manuscrit propose d'établir un cadre de modélisation à partir duquel on pourra incorporer au mieux une partie des effets identifiés dans aevol. Pour être générateur d'idées, ce cadre se limite à ces seuls effets, ce qui permet d'éviter les facteurs de confusion et simplifie l'analyse mathématique et l'implémentation informatique. En effet, dans le cas d'aevol, il est difficile d'établir des liens de cause à effet très clairs à cause du nombre de paramètres. Un modèle mathématique avec moins de paramètres, car restreint à l'étude d'un phénomène en particulier, permet une analyse plus poussée de ce phénomène. Les liens de causalité sont plus faciles à identifier, en contrepartie certains phénomènes biologiques restent flous. D'après les données et modèles existants, il faudra prendre en compte les deux manières de faire varier la taille d'un génome : par ADN codant ou non-codant. Il faudra également prendre en compte la possibilité d'inclure des biais pour les petites insertions et délétions. Enfin, il faudra prendre en compte les réarrangements à plus grande échelle et la duplication de matériel existant.

## 3.2 Présentation du modèle

**Hypothèses sur la structure des génomes** Nous allons considérer un espace d'états $X$ contenant toutes les états possibles pour un génome. À chaque génome est associé une taille $s \in \mathbb{N}$. $X$ est un espace de cardinalité infinie, dénombrable ou non, bien que le cas dénombrable paraisse plus naturel. Par exemple, $X$ peut être défini comme toutes les séquences formées par les lettres {A,C,G,T} de toutes les tailles possibles dans $\mathbb{N}$. Nous supposerons par commodité qu'un génome est circulaire, comme chez la plupart des bactéries. Cette hypothèse n'est pas centrale mais permet d'éviter des effets de bord.

**Hypothèses sur les mutations** Nous considérons différents types de mutations : mutations ponctuelles, petites insertions, petites délétions, inversions, translocations, grandes

délétions et duplications. Leur impact sur la taille du génome ne dépend pas de la location de la mutation et est défini comme suit :

- Pour les mutations ponctuelles, les inversions et les translocations, la taille de génome ne change pas. Si la taille de départ est $s_0$, la taille après mutation reste $s_0$ avec probabilité 1.
- Pour les petites insertions, 1 à $l_{ins}$ bases sont ajoutées au génome. La taille après l'insertion appartient donc à $\{s_0 + 1, ..., s_0 + l_{ins}\}$, y compris si $s_0 = 0$ qui n'est donc pas un état absorbant. Initialement, nous supposons que le nombre de bases gagnées suit une loi indépendante de $s_0$, mais arbitraire. Notons que dans ce cadre, la transposition réplicative d'un élément transposable peut être vue comme une de ces "petites" insertions, dans le sens où le nombre de bases ajoutées par cet événement ne dépend pas de la taille initiale du génome.
- Pour les petites délétions, 1 à $l_{sdel}$ bases sont retirées au génome. La taille après la délétion appartient donc à $\{s_0 - l_{sdel}, ..., s_0 - 1\}$. Initialement, nous supposons que le nombre de bases perdues suit une loi indépendante de $s_0$, mais arbitraire. Si $s_0 < l_{sdel}$, les pertes de plus de $s_0$ bases sont converties en pertes de $s_0$ bases. En particulier, si $s_0 = 0$, la taille après la petite délétion est 0 avec probabilité 1.
- Pour les duplications, si $s_0 > 0$ le nombre de bases copiées varie de 1 à $s_0$. La taille après mutation appartient à $\{s_0 + 1, ..., 2s_0\}$. Initialement, nous supposons que chaque taille finale possible peut être atteinte avec probabilité uniforme $1/s_0$. Si $s_0 = 0$, la taille après duplication est 0 avec probabilité 1.
- Pour les grandes délétions, si $s_0 > 0$ le nombre de bases supprimées varie de 1 à $s_0$. La taille finale appartient à $\{0, ..., s_0 - 1\}$. Initialement, nous supposons que chaque taille finale possible peut être atteinte avec une probabilité uniforme $1/s_0$. Si $s_0 = 0$, la taille reste 0 avec probabilité 1.

Notons qu'il s'agit ici des hypothèses générales et uniquement de l'impact sur la taille. Nous laissons de côté pour l'instant la question de la répartition entre codant et non codant, qui n'a d'importance que lorsque la sélection est considérée, dans la deuxième partie du manuscrit. La distribution uniforme pour les délétions et les duplications correspond au cas où les points de cassure sont tirés uniformément sur le génome. Par exemple, si on suppose que les réarrangements sont dus à des éléments transposables, cela revient à prendre le cas où les éléments sont disposés régulièrement le long du génome, en supposant qu'un réarrangement associe deux éléments transposables quelconques. Le fait d'associer deux éléments quelconques plutôt que deux éléments proches sur la séquence pourrait être justifié par le repliement tridimensionnel de l'ADN : la proximité physique n'est pas uniquement due à la proximité sur la séquence. Cette distribution est la distribution *spontanée* des réarrangements. Quand on ajoute de la sélection, les événements délétères ont peu de chances d'être fixés par la sélection, notamment les très grandes délétions. Elle ne reflète donc pas du tout la distribution fixée. Éliminer dès le départ les très grandes délétions sous prétexte qu'elles sont létales introduirait un biais important dans le modèle,

comme nous allons le voir. Nous reviendrons sur ce point dans la discussion (chapitre VI) et proposerons une généralisation de nos résultats à d'autres distributions spontanées (chapitre II, section 5).

Pour chaque mutation, on spécifie un taux d'occurrence par base par génération : $\mu_{switch}$ pour les mutations ponctuelles, $\mu_{ins}$ pour les petites insertions, $\mu_{sdel}$ pour les petites délétions, $\mu_{inv}$ pour les inversions, $\mu_{trans}$ pour les translocations, $\mu_{ldel}$ pour les grandes délétions et $\mu_{dup}$ pour les duplications. Nous appelons $\mu = \mu_{switch} + \mu_{ins} + \mu_{sdel} + \mu_{inv} + \mu_{trans} + \mu_{ldel} + \mu_{dup}$ le taux de mutation total par base et par génération. Notons que comme le taux de petites insertions est, comme les autres taux, exprimé par paire de base et non par élément transposable, chaque base peut provoquer une insertion de $l_{ins}$ bases. Nous modélisons donc un cas pire que celui où le nombre d'insertions est donné par le nombre d'éléments, en termes de pression vers la croissance du génome.

Les processus de mutation vérifient deux hypothèses centrales :

H1 Nous supposons qu'il existe une projection $\varphi : X \to \mathbb{N}$ compatible avec les mutations. Dans notre cas, il s'agit de la projection size : $X \to \mathbb{N}$ qui associe un génome à son nombre de paires de base. Rappelons que les génomes se déplacent dans l'espace des structures $X$, qui ne contient pas seulement la taille, mais toute sorte d'information, comme par exemple le positionnement des séquences codantes. Pour une taille donnée, de nombreuses architectures sont possibles. La compatibilité implique que pour deux génomes $x, y \in X$ qui ont la même longueur ($\text{size}(x) = \text{size}(y) = s_0$), la probabilité que $x$ ou $y$ atteignent une certaine taille $s_f \in \mathbb{N}$ après une mutation est exactement la même, même si $x$ et $y$ ont une architecture détaillée différente. En d'autres termes, les probabilités de variation de taille dépendent uniquement de la taille initiale, ce qui est le cas avec les mutations présentées ci-dessus.

H2 À chaque génération, les différents types de mutation suivent des processus ponctuels de Poisson indépendants de paramètre $\mu_{type}$, où $\mu_{type}$ est le taux de la mutation considéré. Le support de ce processus est le génome, le nombre de mutation total est donné par une loi de Poisson de paramètre $\mu s_0$, où $s_0$ est la taille du génome considéré au début de la génération. Comme les processus sont indépendants, on peut tirer *a priori* le nombre total de mutations et déterminer ensuite la séquence de mutations. La probabilité qu'une mutation quelconque de la séquence soit une petite insertion (par exemple) est $\mu_{ins}/\mu$.

Quand nous disons que les mutations sont indépendantes, il faut comprendre cela en termes de leur occurrence. Nous supposons qu'il n'y a aucun lien causal entre les mutations : il n'y a pas de cascades de mutation, comme une petite insertion qui favoriserait ou défavoriserait directement une délétion. Cela revient à dire qu'il n'y a pas d'*ordre privilégié* dans les mutations. Par contre, l'ordre a une importance : les distributions de tailles de génomes après 2 mutations ne sont pas exactement les mêmes si on effectue d'abord une petite délétion puis une grande délétion ou d'abord la grande délétion, puis la petite délétion. Les génomes seront légérèment plus petits dans le deuxième cas car la grande

délétion sera en moyenne plus grande, alors que l'impact de la petite délétion est le même dans les deux cas.

**Déroulement d'une génération** Nous considérons l'évolution d'une population d'individus haploïdes asexués, chaque individu étant représenté par l'état de son génome dans l'espace $X$ des états génomiques possibles. À chaque génération, on s'intéresse à la distribution des génomes dans l'espace $X$. L'évolution d'une génération à l'autre se fait en 2 étapes : une étape de sélection et une étape de reproduction, durant laquelle les mutations ont lieu. Dans ce travail, les populations sont en général infinies, mais nous proposerons des variantes qui sont valables dans le cas où la population est finie. Pour repérer les individus à chaque génération, nous définissons un vecteur $(\nu_t(x))_{x\in X}$ qui, pour une génération donnée $t \in \mathbb{N}$, donne la densité de la population pour chaque état génomique $x$ de $X$. La population initiale est spécifiée par le vecteur $\nu_0$.

Nous supposons que les processus de sélection et de mutation sont séparés. Dans le cadre le plus général possible, la sélection est un opérateur $\mathrm{Sel}_t$. Quand la sélection opère, $\mathrm{Sel}_t(\nu_t)$ détermine le nombre de descendants des individus répertoriés dans $\nu_t$ : il renvoie un vecteur de densité de même support mais en appliquant un filtre qui peut être défini librement et qui varie éventuellement avec le temps (d'où le $t$ en indice).

Dans un deuxième temps, les descendants sont obtenus par réplication du parent via le processus de mutation, donné par un opérateur $M$. Dans ce modèle on considérera que les lois de mutation ne varient pas avec le temps. $M$ ne peut pas être défini complètement librement : il doit renvoyer un vecteur de densité et respecter les contraintes dictées par les hypothèses faites pour les mutations ci-dessus.

L'équation générale du modèle donne le vecteur de densité au temps $t+1$ en fonction de celui au temps $t$ d'après l'équation suivante :

$$\nu_{t+1} = M \circ \mathrm{Sel}_t(\nu_t) \tag{I.1}$$

## 3.3 Cas particuliers étudiés

Au cours de cette thèse nous avons étudié la dynamique du modèle décrit ci-dessus dans deux cas particuliers :

**Première partie : étude sans sélection** Le but de ce modèle étant d'étudier les phénomènes non adaptatifs, la première étape consiste à analyser la dynamique spontanée des génomes, c'est-à-dire sans sélection. Cela permet de se concentrer sur les seuls mécanismes de mutation, tout en conservant la généralité la plus grande possible. En effet, en l'absence de sélection, il ne sera pas nécessaire de préciser l'architecture détaillée et la manière dont agit la sélection. Ces deux questions sont difficiles à résoudre et ne pourront qu'être abordées de manière schématique dans la deuxième partie.

Comme le nombre de paramètres est faible (ce sont les taux de mutation), ce cadre est idéal pour l'analyse mathématique de l'impact des mutations. Nous nous concentrons sur l'interaction entre les différentes mutations. Lesquelles prennent le dessus ? Quel est l'impact d'un biais des petits indels ou des duplications et grandes délétions ? Les résultats restent-ils valables pour des processus de mutations un peu plus généraux ?

**Deuxième partie : évolution de génomes simplifiés sous une sélection favorisant les grands génomes** Dans la deuxième partie, nous introduisons la sélection, mais sous une forme particulière. Dans la plupart des modèles, la sélection empêche la croissance infinie des génomes. Ici, nous l'utiliserons pour pousser les génomes à grossir le plus possible en sélectionnant directement les génomes ayant le plus grand nombre de gènes et en laissant le non-codant évoluer sans impact direct sur la fitness.

L'introduction de la fitness a un double but. Même si la mise en place est simplifiée, et par ce fait caricaturale, elle permet de confirmer une prédiction analytique, selon laquelle la force spontanée des processus de mutation ne peut pas être surmontée par la sélection, du moins sous nos hypothèses. Les réarrangements imposent une taille limite que la sélection ne peut pas repousser. De plus, l'introduction de la sélection permet de donner un sens au codant et au non-codant. En faisant varier les paramètres, on pourra ainsi observer l'impact qualitatif de chaque type de mutation sur la taille finale du génome et sur son pourcentage de codant.

# Première partie

# Évolution de la taille du génome sans sélection

# Chapitre II

# Étude de l'évolution spontanée du génome

> Le paradoxe de la science est qu'il n'y a qu'une réponse à ses méfaits et à ses périls : encore plus de science.
>
> Romain Gary, *Charge d'âme*

## 1 Présentation du problème : le paradoxe de la médiane

Le raisonnement mathématique présenté dans ce chapitre peut paraître par bien des aspects compliqué et technique, alors que l'idée derrière ce raisonnement est relativement simple. Il s'agit de lever un paradoxe lié à l'intuition qu'on peut développer si on néglige certains aspects mathématiques du problème. Ce paradoxe n'est pas nouveau et se résout facilement, il est juste appliqué à un problème posé de manière à induire en erreur (comme l'évolution de la taille du génome ou dans l'exemple donné ci-dessous). Dans cette section, nous présentons schématiquement l'origine du paradoxe.

### 1.1 Le dilemme du correcteur

Imaginons la situation suivante : on a aménagé une grande salle dotée de tables alignées côte-à-côte portant chacune un identifiant. En entrant dans la salle, on voit la table 0, à côté se trouve la table 1, puis la table 2 et ainsi de suite en continuant jusqu'à la table 102400. À chaque table s'assoit un correcteur qui doit corriger les copies du baccalauréat

qu'on lui transmet. La subtilité réside dans la distribution des copies qui ne se fait pas uniformément mais suit une procédure un peu compliquée supervisée par le ministre en personne.

Le ministre entre dans la salle puis pose toutes les copies sur la table 100 (figure II.1A). Au premier coup de sifflet, le correcteur à cette table, appelé dans la suite correcteur 100, coupe le paquet en 2, mettant 1/3 des copies d'un côté, 2/3 de l'autre côté. Il distribue le premier 1/3 aux cent personnes situées aux tables 0 à 99, en veillant à les répartir équitablement entre ces cent personnes. Il prend ensuite les 2/3 restant et les distribue équitablement aux cent personnes de l'autre côté, numérotées de 101 à 200. Remarquons qu'après cette première étape, le correcteur 100 s'est débarrassé de toutes ses copies, et que la majorité des copies qu'il a distribuées se trouvent du côté « plus grand » comparé à lui (figure II.1B).

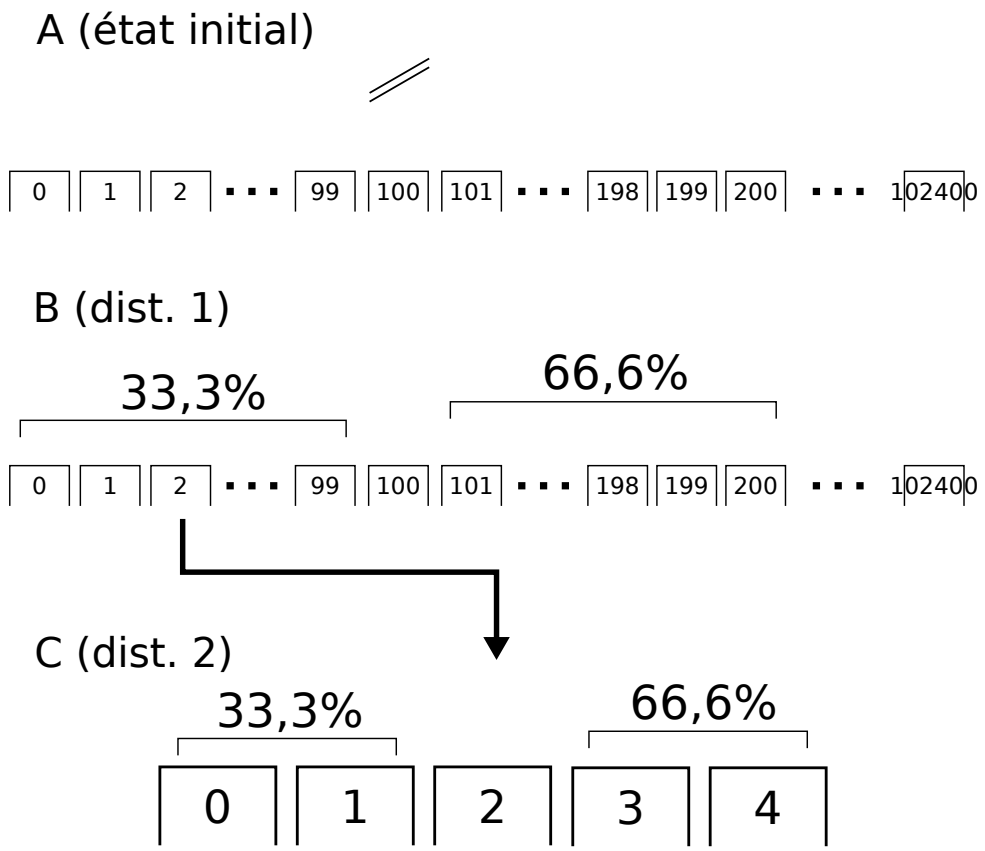


Figure II.1 – Initialement, toutes les copies du baccalauréat sont données au correcteur assis à la table 100 (A). Heureusement pour lui, à la première redistribution, il s'en débarasse : 2/3 des copies vont vers des tables avec des numéros plus grands et 1/3 vont vers des tables avec des numéros plus petits (B). Lors de la deuxième redistribution, chaque correcteur applique le même principe, mais à son échelle : il sert autant de tables à droite et à gauche, tout en conservant le biais 1/3 pour les tables plus petites et 2/3 pour les tables plus grosses (C).

Au coup de sifflet suivant, chaque correcteur qui a des copies sur sa table fait de même que

le correcteur 100 précédemment. Il coupe le paquet en 2 tas de 1/3 et 2/3. Si le correcteur est à la table $n$, il distribue le premier tiers équitablement aux tables 0 à $n-1$ et les deux autres tiers aux tables $n+1$ à $2n$ (figure II.1C). Le correcteur situé à la table 0 donne toutes ses copies au correcteur 100. Pour rendre le mélange plus efficace, cette étape est répétée plusieurs fois (avec un maximum de 10), les fonctionnaires s'exécutant avec zèle à chaque coup de sifflet. Quand le ministre en a marre de souffler, chaque correcteur doit corriger les copies actuellement sur sa table.

La question est la suivante : vous ne savez pas encore combien il y aura de coups de sifflet, mais vous devez choisir où vous asseoir, sachant qu'on vous laisse le choix entre les tables 50 à 200 (les autres tables sont déjà réservées pour des proches du ministre ou occupées par des lève-tôt). On sait que les copies partent de la table 100, vaut-il mieux choisir de s'asseoir à une table plus petite ou plus grande que 100 ? Que faire si au lieu de couper le tas en 1/3 et 2/3 à chaque fois, on coupait en 1/2 et 1/2 ? en 1/4 et 3/4 ?

La résolution de ce problème est mathématiquement très simple, l'idée ici est plutôt de confronter l'intuition. Peut-on, sans calcul, se persuader de la solution à ce problème ? Décider à partir de quel découpage se placer sur les tables à petit chiffre devient favorable ? Étudions maintenant deux raisonnements qui ne contiennent pas de calcul et qui répondent à ce problème.

Premièrement, réfléchissons en humain. Je suis un correcteur, il est donc normal que je décrive les choses de mon point de vue. À chaque coup de sifflet, je reçois des copies de ma droite et de ma gauche, tandis que je distribue la grande majorité des copies vers les tables plus grandes. Il est difficile de prévoir combien de copies je reçois, mais je sais que ceci est valable pour n'importe quel correcteur. À chaque coup de sifflet, tous les correcteurs, sans exception, distribuent la grande majorité des copies vers les tables plus grandes. Si localement, à partir de chaque table de départ, le mouvement se fait vers des tables plus grandes, on a du mal à imaginer comment cela ne se retrouverait pas globalement. Il vaudrait donc mieux s'asseoir sur une des tables de 50 à 100.

Deuxièmement, adoptons le point de vue de l'autre acteur important : la copie. Imaginons que je suis une copie et étudions ma trajectoire. Si je suis sur la table de numéro $n$, j'ai 2 chances sur 3 d'aller vers une table plus grande, au maximum celle avec le numéro $2n$, et j'ai 1 chance sur 3 d'aller vers une table plus petite. Si je suis allée vers une table plus petite, un phénomène particulier intervient. Imaginons que je parte de la table 100. Si j'atterris sur une des tables de 0 à 49, je ne pourrais par revenir à la table 100 à l'étape d'après, je suis en quelque sorte piégée au niveau des petites tables. À l'inverse, si je vais vers une table plus grande, je pourrais toujours revenir vers la table 100 et même vers des tables encore plus petites. Autrement dit, du point de vue de la copie, le fait d'aller vers des nombres plus grands ou plus petits n'est pas symétrique : on peut facilement revenir en arrière après avoir atteint une table plus grande mais pas si on est allé sur une table plus petite. Ceci étant dit, comme l'ampleur du mouvement dépend de la table initiale, on a du mal à visualiser le mouvement global de la copie. Sans calcul, on peut affirmer que si le découpage est 1/2 et 1/2, la copie va probablement rester piégée au niveau de petites tables qu'il faudra alors éviter. Pour les autres découpages, on peut difficilement

conclure.

Ces deux raisonnements peuvent paraître justes *a priori*, bien que seul le deuxième soit correct en totalité. La confusion vient ici du fait que le processus étudié n'est pas invariable par translation et que le premier raisonnement est valable pour la moyenne, mais pas pour la médiane, qui est en fait la grandeur d'intérêt dans ce problème. Ces deux éléments rendent l'intuition inefficace, il faut transformer le problème avant de pouvoir trouver la solution.

## 1.2 Résolution du paradoxe : rôle de la moyenne et de la médiane, invariance par translation

Aller du local vers le global est dangereux dans ce type de problème. Si 2/3 des copies vont vers des tables plus grandes, on voit que pour chaque table de départ, aussi bien la moyenne que la médiane augmentent. Plus précisément, si la table de départ est $n$, un calcul simple permet d'établir que l'espérance du numéro de table vaut $1/(3n) \times n(n-1)/2 + 2/(3n) \times n(3n+1)/2 = n + (n+1)/6$ et la médiane vaut approximativement $n + 1/6(3n/2) = n + n/4$. La médiane augmente donc localement plus que la moyenne et pourtant, contrairement à la moyenne, elle n'est pas additive. Globalement, la moyenne est la somme pondérée des moyennes locales. L'espérance au $k$-ème coup de sifflet est égale à la somme des espérances sachant que la copie se trouvait sur la table $n$ (qu'on vient de calculer ci-dessus) après le $k-1$-ème coup de sifflet, pondérées par la probabilité que la copie se trouvait effictivement sur la table $n$. Comparé au calcul de l'espérance à l'étape $k-1$, on échange $n$ par $n + (n+1)/6$ : l'espérance augmente à chaque coup de sifflet. Par contre, la médiane globale ne respecte pas forcément les tendances locales, elle peut se mettre à diminuer. En simulant quelques étapes du processus, on peut voir qu'à part au premier pas de temps, c'est ce qui se passe (figure II.2). Visuellement, c'est toute la densité qui semble diminuer, bien que la moyenne augmente réellement, ce qui indique que la queue de la densité a des propriétés particulières.

Notre intuition peut avoir tendance à vouloir passer du local au global mais cela pose un problème si le processus n'est pas invariant par translation : on a tendance à se focaliser sur le schéma lié à un point de départ (une observation en biologie) et faire comme s'il était valable tout du long. Autrement dit, on sera tenté de penser que parce que, pour un point de départ donné, la probabilité d'aller vers des tables plus grandes est plus grande, le biais est à l'augmentation constante, donc les sauts qui vont vers des tables plus petites seront automatiquement compensés. Ce raisonnement serait vrai si le biais était le même tout au long de la chaîne, ce qui n'est évidemment pas le cas ici. La vraie question vient d'être soulevée : comment les augmentations et les diminutions se compensent-elles ? En adoptant le point de vue de la copie tout à l'heure, on a vu qu'en réalité, une augmentation est en quelque sorte plus faible qu'une diminution. Intuitivement, on risque de devoir subir beaucoup d'augmentations pour compenser une diminution. Si on part de la table 100 et qu'on atterrit sur la table 10, il faudra au moins 4 redistributions pour espérer retrouver la table 100. Pour trouver exactement combien d'augmentations sont nécessaires pour

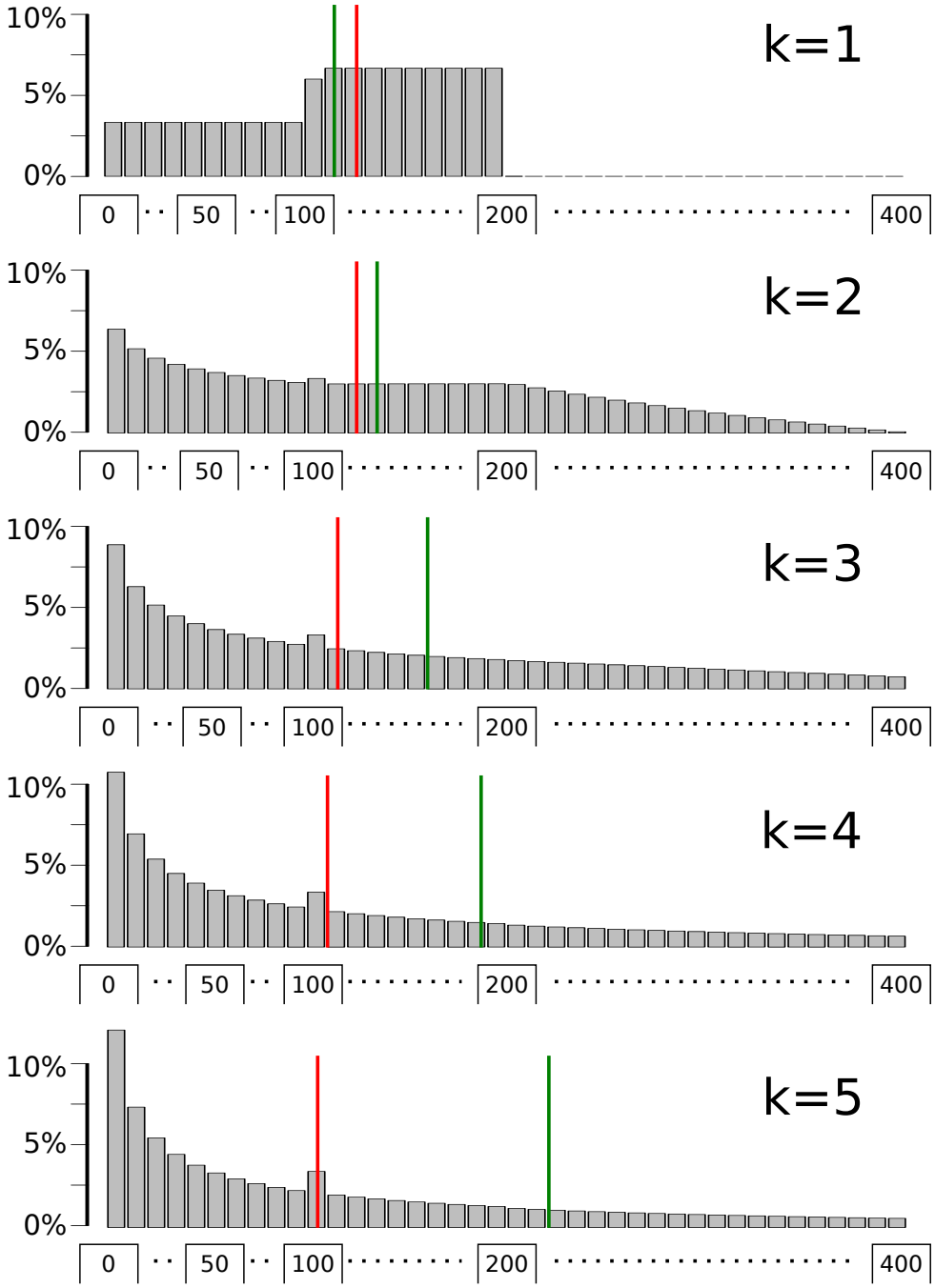


Figure II.2 – On simule les cinq premiers coups de sifflets ($k \in \{1, ..., 5\}$). L'intuition a tort et raison à la fois si elle extrapole à partir des cas locaux : la moyenne (en vert) augmente bien, mais la médiane (en rouge) diminue dès le troisième coup de sifflet.

compenser une diminution, l'intuition est bloquée par le fait qu'à chaque redistribution, l'ampleur des mouvements de la copie dépend du point de départ.

On va donc essayer d'adopter un point de vue où l'ampleur des mouvements ne varie plus avec le point de départ. Si on y arrive, on pourra effectivement se restreindre à ce qui se passe au niveau d'un seul point de départ pour comprendre ce qui se passe

globalement. Le problème fait intervenir un processus de nature multiplicative : le support de la distribution des copies à partir d'une table donnée augmente linéairement avec le point de départ tandis que la hauteur de la distribution est simplement mise à l'échelle via une renormalisation. Dans ce cas, si on place les tables sur une échelle logarithmique en réexprimant les densités sous forme d'histogramme, on voit apparaître une distribution qui tend à être la même partout : les copies allant de la table 100 vers les tables 181 à 200 couvrent la même distance que les copies allant de la table 50 vers les tables 91 à 100 et représentent la même fraction de la distribution (figure II.3). La distribution en échelle logarithmique est biaisée vers la diminution, du moins en moyenne (la médiane n'a pas changé après le changement d'échelle) (figure II.4). Cette fois, la moyenne revêt un sens différent, puisqu'elle correspond mieux à notre intuition. Si on se met du point de vue d'une copie, le biais ne dépend plus du point de départ, les augmentations et diminutions sont les mêmes partout. En regardant la distribution locale, elle a toutes les informations sur la totalité du processus, elle voit exactement comment les augmentations et les diminutions vont se compenser au bout d'un nombre arbitraire de sauts, sans avoir à regarder « à côté » ce qui se passe. Elle peut donc cette fois passer du local au global sans hésiter.

Après changement d'échelle, on peut donc calculer combien d'augmentations sont nécessaires pour compenser une diminution. Si la copie va vers une table plus grande, son avancée en échelle logarithmique sera en moyenne d'environ $2\log 2 - 1 \simeq 0.39$, alors qu'en allant vers une table plus petite, elle recule en moyenne d'environ $-1$ (on en fera la démonstration ultérieurement). Il faut donc à peu près $1/(2\log 2 - 1) = 2.6$ augmentations pour compenser une diminution. On peut maintenant répondre à la question du découpage critique. Si le découpage est $1/2$ et $1/2$, les diminutions vont l'emporter, comme on l'avait déjà compris intuitivement. Si le découpage est $1/3$ et $2/3$, il y a 2 augmentations pour une diminution, la tendance est toujours à la diminution : il faudra donc éviter les tables 50 à 100, et choisir par exemple la table 200. Si le découpage est $1/4$ et $3/4$, il y a 3 augmentations pour une diminution, il sera plus raisonnable de s'asseoir de l'autre côté, à la table 50.

Cet exemple n'est pas particulièrement intéressant du point de vue scientifique et le lecteur aura pu avoir des réactions différentes. Peut-être la solution lui paraissait évidente depuis le début, car il avait reconnu la structure multiplicative et se demande encore pourquoi la résolution ne parle pas du théorème de la limite centrale, qui propose une solution bien plus poussée pour connaître la position des copies. Peut-être estime-t-il qu'il aurait pu trouver la solution si le problème n'avait pas été posé de manière aussi étrange. Le but de cet exercice n'était pas réellement de proposer un paradoxe profond, mais de montrer comment le fait de mal poser un problème peut orienter l'intuition vers de mauvaises pistes. L'objectif ici n'est pas tant de faire des mathématiques de haut niveau que d'appliquer un principe relativement simple à un problème où ce mécanisme n'a jamais vraiment été considéré.

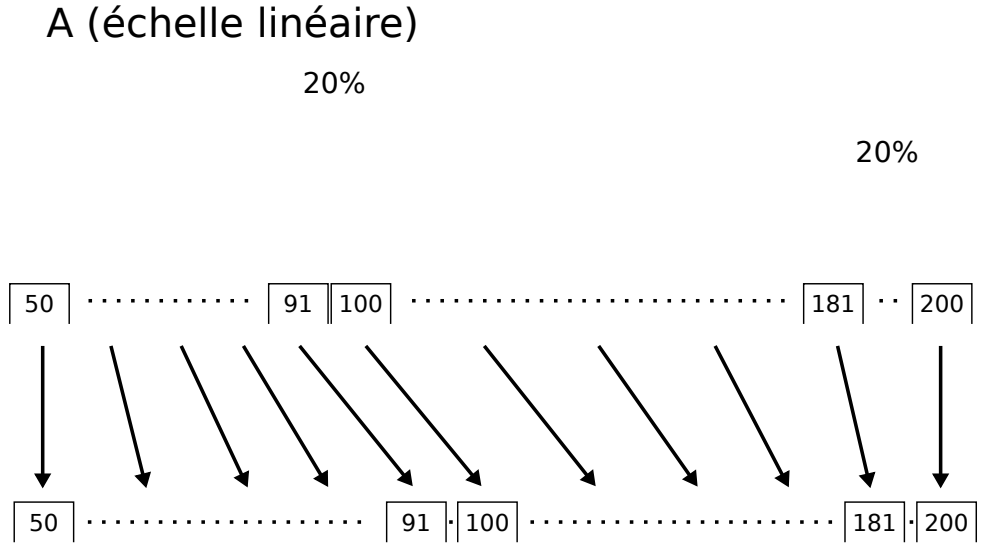


Figure II.3 – On étudie le processus « de loin » : on représente ici globalement la répartition des copies qui viennent de la table 50 (en noir) et de la table 100 (en gris) à destination de tables plus grandes, par blocs de 20 %. En échelle linéaire (A), quand les tables sont régulièrement espacées dans la salle, les blocs de 20 % provenant de la table 50 sont plus étroits et plus proches de leur table de départ que pour la table 200. On décide alors de déplacer les tables, sans changer leur ordre : au lieu d'être régulièrement espacées, on les place sur une échelle logarithmique. Deux tables dont les numéros diffèrent du même *facteur* doivent se trouver à la même distance. Par exemple, la distance entre les tables 50 et 100 (facteur 2) est la même que celle entre les tables 100 et 200. En échelle logarithmique (B), les propriétés locales deviennent (asymptotiquement) invariantes par translation : les blocs de 20 % ont exactement la même largeur et sont situés à la même distance de la table de départ.

## 2 Modèle de l'évolution de la taille du génome : définitions supplémentaires

L'exemple présenté ci-dessus est très proche du problème biologique qui nous intéresse. Les numéros des tables deviennent des tailles de génome et le nombre de copies sur une table le nombre d'individus d'une taille donnée. Le nombre de processus étudiés est plus grand et le découpage en générations, qui peuvent contenir plusieurs mutations, ajoutent des difficultés supplémentaires, mais le schéma global et la méthode de résolution sont les mêmes. La mise à l'échelle est la clé pour résoudre ce type de problème et se placer « du point de vue de la copie » revient à étudier une trajectoire individuelle dans une

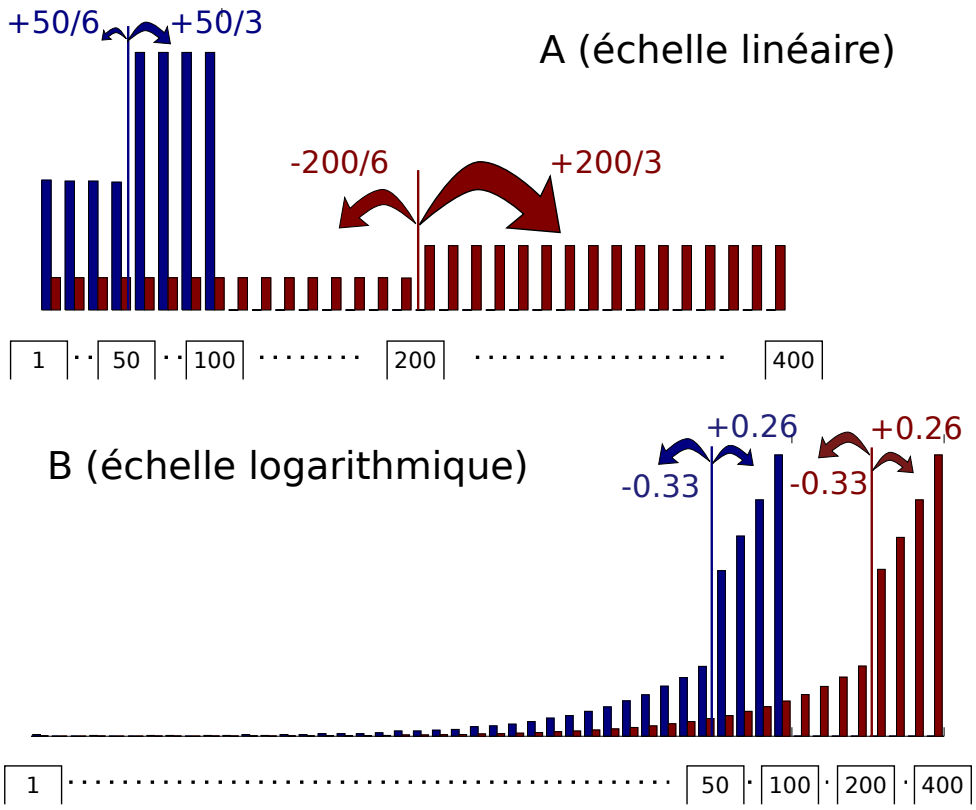


Figure II.4 – On considère la distribution complète de délétions et de duplications « de loin » pour les copies partant des tables 50 et 200. En échelle linéaire (A), les distributions dépendent du point de départ : en particulier, le déplacement moyen dû aux délétions et aux duplications (indiqué par les flèches devient de plus en plus grand, indiquant un déplacement vers les gains de plus en plus fort. En échelle logarithmique (B), les distributions sont très similaire (à translation près) : le déplacement moyen est constant et il est orienté vers les pertes.

chaîne d'états qui, en l'absence de sélection darwinienne, correspond mathématiquement à une chaîne de Markov. On verra que les tailles de génomes convergent dans des cas où l'espérance de la taille de génome augmente. Dans le cas des génomes, le paradoxe est même encore plus troublant, car cela signifie qu'à chaque mutation, il y a création de matière. On ajoute en permanence des nouvelles paires de bases au système, et pourtant les génomes vont avoir majoritairement tendance à diminuer.

Le modèle étudié ici est celui présenté dans l'introduction (à la section 3). Pour l'analyse, on aura cependant besoin de quelques définitions et concepts supplémentaires. On rappelle que l'espace des génomes est $X$, mais que chaque génome peut être projeté sur sa taille. Grâce à la projection en taille, il n'est pas utile de spécifier l'architecture détaillée du génome (position, longueur, chevauchement, sens des gènes, etc.) puisqu'on ne s'intéresse pour l'instant qu'à la dynamique mutationnelle sans considérer l'action de la sélection darwnienne (celle-ci sera incorporée dans la deuxième partie de ce manuscrit).

**Définition 2.1.** Par analogie avec $X$, l'espace des structures des génomes, on appelera $X_S$ l'espace des tailles des génomes, bien qu'en réalité on ait simplement $X_S = \mathbb{N}$.

Comme les transitions (mutations) dans l'espace des génomes $X$ ne dépendent que de la taille initiale du génome occupée dans $X_S$, on peut se contenter de caractériser l'évolution de la densité marginale correspondant aux tailles de génomes, c'est-à-dire l'évolution de la densité des tailles de génome dans $X_S$.

**Définition 2.2.** $P_S = ((P_S)_{ij})_{i,j\in X_S}$, où $(P_S)_{ij}$ est la probabilité qu'un génome de taille initiale $i$ arrive à une taille $j$ après *une seule mutation*. Comme $X_S = \mathbb{N}$ est infini dénombrable, $P_S$ est une matrice stochastique de taille infinie. Les valeurs des transitions sont données dans la définition des processus de mutation en introduction. $P_S$ décrit l'évolution de la taille des génomes mutation après mutation.

Lors de la reproduction du génome (une génération), plusieurs mutations peuvent se produire suivant un processus de Poisson. $P_S$ est la matrice qui permet de traduire mathématiquement nos hypothèses sur chaque type de mutation (y compris les taux de mutation), mais il nous faut une matrice qui prenne en compte les mutations multiples correspondant aux processus ponctuels de Poisson.

**Définition 2.3.** $M_S = ((M_S)_{ij})_{i,j\in X_S^*}$ représente l'impact des mutations locales et des réarrangements à l'échelle d'une génération. $(M_S)_{ij}$ est la probabilité qu'un génome de taille initiale $i$ finisse à la taille $j$ après *une génération*. $M_S$ est une matrice stochastique de taille infinie.

Le lien entre $M_S$ et $P_S$ n'est pas directement explicité, on peut néanmoins noter que la donnée de $P_S$ suffit à caractériser totalement les transitions de $M_S$ via les processus de Poisson. Par contre, contrairement à $P_S$, $M_S$ est définie sur $X_S^* = X_S\backslash\{0\}$ au lieu de $X_S$. Tous les individus qui finissent à la longueur 0 après avoir subi leurs mutations sont automatiquement réassignés à un état correspondant à la longueur 1. En effet, alors que 0 n'est pas un état absorbant à l'échelle des processus de mutations à cause des petites insertions, il devient absorbant à l'échelle des générations parce que le nombre de mutations par génération donné par la loi de Poisson vaut 0. En présence de sélection, les génomes de taille 0 sont éliminés donc ils ne posent pas de problèmes particuliers. Par contre, si nous conservions cet état dans le système lors de l'analyse sans sélection, l'existence d'une distribution stationnaire ne refléterait pas nécessairement la dynamique spontanée des génomes : même si les génomes avaient tendance à croître, ils risqueraient d'être « piégés » dans l'état absorbant. Comme nous allons le voir, en l'absence de sélection, l'évolution de la taille des génomes est donnée par la chaîne de Markov $(X_S^*, M_S)$. Contrairement à $(X_S, M_S)$, $(X_S^*, M_S)$ ne contient pas d'état absorbant (il y a une probabilité non nulle de quitter chaque état), donc n'admet pas de distribution stationnaire triviale.

En introduction, nous avons défini une projection size qui associe à une structure génomique sa taille, ainsi que le vecteur $\nu_t$ qui répertorie toutes les architectures génomiques présentes à la génération $t$. En généralisant la projection en taille à ce vecteur, on peut définir la projection $\mathbf{size}_\nu$ qui, à partir de la densité $\nu_t$ définie sur $X$, renvoie la densité des tailles de génomes correspondant dans l'espace $X_S^* = \mathbb{N}^*$ :

$$\forall s \in X_S^*, \quad \mathbf{size}_\nu(\nu_t)(s) := \int_X \mathbf{1}_{\{x\in X, \mathrm{size}(x)=s\}} d\nu_t$$

On peut alors réécrire l'équation de l'introduction directement en termes de taille de génome. En notation matricielle, la densité de population à l'étape $t+1$ dans l'espace $X_S^*$ des tailles de génomes est donnée par

$$\mathbf{size}_\nu(\nu_{t+1}) = \mathbf{size}_\nu(\mathrm{Sel}_t(\nu_t))M_S \tag{II.1}$$

Comme on étudie dans ce chapitre le cas sans sélection, tous les états génomiques confèrent la même probabilité de reproduction et l'opérateur de sélection est simplement la fonction identité. L'équation générale sur $X$ (I.1) se réduit à $\nu_{t+1} = M(\nu_t)$. La projection sur la taille dans $X_S^*$ donne alors, en notation matricielle :

$$\mathbf{size}_\nu(\nu_{t+1}) = \mathbf{size}_\nu(\nu_t)M_S \tag{II.2}$$

**Propriété 2.4.** Comme $M_S$ est stochastique, l'équation (II.2) peut être interprétée comme représentant l'évolution de la chaîne de Markov ($X_S^*$, $M_S$) dans l'espace des tailles de génomes. La théorie de Markov s'intéresse au déplacement d'un élément dans un espace d'états pour lequel les transitions ne dépendent que de l'état occupé et pas de la trajectoire antérieure (dans notre cas, uniquement la taille de génome actuelle). $\mathbf{size}_\nu(\nu_t)$ est alors vu comme la densité de probabilité pour un génome donné de se trouver dans un des états de $X_S^*$. Comme l'espace des tailles de génomes est dénombrable, le processus de Markov devient une chaîne de Markov. Comme $M_S$ ne varie pas avec le temps, la chaîne de Markov est homogène en temps. Une chaîne de Markov est caractérisée par son espace d'états, ici $X_S^*$, et une matrice de transitions d'état à état, ici $M_S$, d'où la notation ($X_S^*$,$M_S$).

La prochaine section (section 3) est consacrée à l'étude des propriétés de la chaîne de Markov ($X_S^*$, $M_S$). Nous donnons une condition pour l'existence d'une distribution asymptotique des tailles des génomes qui ne dépend que des taux des mutations dont l'effet varie avec la taille du génome (duplications et grandes délétions). Dans la section 4, nous reprenons le raisonnement avec une approximation continue qui capture l'essentiel de la dynamique pour les grands génomes. Cette approximation permet de faire les calculs plus simplement et plus rapidement, notamment pour accéder au comportement détaillé de la dynamique via les moments d'ordre 2 des distributions. La dynamique globale est conservée : la condition d'existence d'une distribution asymptotique reste la même que dans la section 3. En prenant en compte les moments d'ordre 2, on peut par contre dériver des bornes sur les quantiles de la distribution asymptotique et donc sur les tailles de génome réellement viables. Dans la section 5, nous généraliserons les résultats obtenus. Nous montrerons que la démarche utilisée peut être étendue à d'autres familles de distributions de tailles de mutations, tant qu'il est possible de trouver une mise à l'échelle rendant le système invariant par translation. Nous verrons ensuite que les bornes dérivées via l'approximation continue sont des bornes fortes sur les tailles, puisqu'elles s'appliquent non seulement asymptotiquement, mais aussi au niveau de chaque génération, de telle manière que l'ajout de la sélection ou la modification de la taille de la population ne peut pas être utilisé pour les surmonter. Ces deux éléments ne remettent donc pas en cause

l'existence des bornes, mais ils déterminent comment la population va se stabiliser par rapport à ces bornes, en particulier à quelle distance des bornes la viabilité à long terme est envisageable.

# 3 Existence et unicité d'une distribution stationnaire pour la taille de génome

Nous allons montrer l'existence et l'unicité d'une distribution stationnaire pour la chaîne de Markov $(X_S^*, M_S)$ en utilisant une extension de la condition de Doeblin, qui s'applique à la matrice $M_S$ du processus.

**Théorème 3.1** (Condition de Doeblin en $k$ étapes)**.** *Soit* $\mathbf{M}$ *une matrice de transition définie sur un espace d'état* $S$ *telle qu'il existe un* $k \geq 1$*, un état* $i_f \in S$ *et* $\varepsilon > 0$ *vérifiant* $(\mathbf{M}^k)_{ii_f} \geq \varepsilon$ *pour tout* $i \in S$*. Alors* $\mathbf{M}$ *admet un unique vecteur stationnaire* $\pi$ *tel que* $(\pi)_{i_f} \geq \varepsilon$ *et, pour toute distribution initiale* $\mu$*,*

$$\|\mu\mathbf{M}^t - \pi\| \leq 2(1-\varepsilon)^{\lfloor \frac{t}{k} \rfloor}, \quad t \geq 0.$$

Il faut noter que dans cette définition et dans le reste du chapitre, la norme considérée est la norme 1. Par exemple, $\|\mu\| = \sum_{i \in X} |\mu_i|$. Une démonstration de ce théorème est donnée dans Stroock (2005). Nous utilisons la condition de Doeblin en 2 étapes pour prouver le théorème suivant.

**Théorème 3.2** (Distribution stationnaire pour la taille des génomes sans sélection)**.** *Si* $(2\log 2 - 1)\mu_{dup} < \mu_{ldel}$*, la chaîne de Markov* $(X_S^*, M_S)$ *a une unique distribution stationnaire asymptotique* $\nu_\infty^S$*. Pour toute condition initiale* $\mathbf{size}_\nu(\nu_0)$*,*

$$\lim_{t \to \infty} \|\mathbf{size}_\nu(\nu_0) M_S^t - \nu_\infty^S\| = 0$$

La condition de Doeblin s'applique à la chaîne de Markov générationnelle, c'est-à-dire après calcul du nombre de mutations à effectuer dans une génération. Pour évaluer les transitions au niveau d'une génération, il faut donc commencer par s'intéresser aux mutations, c'est-à-dire aux transitions données dans $P_S$, associée à la chaîne de Markov mutationnelle $(X_S, P_S)$. D'un point de vue topologique, on peut noter que tous les états de $(X_S, P_S)$ communiquent avec leurs voisins. En effet, la perte ou le gain d'une paire de base est possible avec probabilité non nulle (via les petits indels ou même les duplications et grandes délétions). En combinant ces transitions, il est possible de passer d'un état de $X_S$ à n'importe quel autre état en un nombre fini de mutations. Au niveau d'une génération, tous ces chemins peuvent se produire avec une probabilité non nulle donnée par le processus de Poisson. Il est donc possible, en une génération, de transiter de n'importe quel état à n'importe quel autre état, ce qui signifie que tous les coefficients de $M_S$ sont non nuls, et que tous les états de la chaîne générationnelle $(X_S^*, M_S)$ communiquent. Cependant, cela ne suffit pas à démontrer le théorème ci-dessus. Comme l'espace d'états et

la matrice sont infinis, les probabilités de transition associées à certains chemins peuvent (et même doivent) être arbitrairement proche de 0. Il n'y a donc pas de borne inférieur triviale $\varepsilon > 0$ comme exigé dans la condition de Doeblin. De plus, à cause des petites insertions, il n'existe pas d'état absorbant. L'existence d'une distribution stationnaire n'est donc pas évidente *a priori*.

**Cheminement de la démonstration** Pour trouver une borne inférieure, nous allons découper l'espace des génomes en 2. D'un côté, l'espace $X_{small}$, qui contient les génomes plus petits qu'une taille spécifique $\tilde{s}$, qui sera précisée ultérieurement. De l'autre, l'espace $X_{large}$, qui contient les génomes plus grands que $\tilde{s}$. Nous allons montrer qu'il existe une taille $s_f \leq \tilde{s}$ qui peut être atteinte en $k = 2$ générations avec une probabilité plus grande qu'un certain $\varepsilon > 0$, quelle que soit la taille $s_0$ du génome au début de la génération.

- Si le génome appartient à $X_{small}$, cette condition est facilement remplie puisque le sous-espace $\{s \leq \tilde{s}\}$ est fini (lemme 3.3 ci-dessous).
- Si le génome appartient à $X_{large}$, la probabilité d'atteindre $s_f$ en deux générations vaut au moins la probabilité d'atteindre un état appartenant à $X_{small}$ lors de la première génération, puis de rejoindre $s_f$ lors de la deuxième. Pour la première génération, nous allons montrer que si les duplications ne sont pas beaucoup plus fréquentes que les grandes délétions ($\mu_{dup} < 2.59\mu_{ldel}$ approximativement), les grands génomes ont tendance à diminuer jusqu'à atteindre une taille seuil puis à conserver cette taille (lemme 3.6). En nous basant sur l'inégalité de Tchebychev, nous montrerons que le nombre de mutations est effectivement suffisant pour rétrécir en dessous d'un seuil $\tilde{s}$, quelle que soit la taille initiale. Pour la seconde génération, on se ramène au point précédent (lemme 3.3).

**Lemme 3.3.** *Supposons que nous avons un sous-espace non-vide $X' \subset X$ auquel correspond un sous-ensemble $X'_S \subset X^*_S$ fini de tailles de génomes. Alors il existe $s_f \in X'_S$ et $\varepsilon_1 > 0$ tels que $(M_S)_{is_f} \geq \varepsilon_1$ pour tout $i \in X'_S$.*

*De plus, soit $M'_S$ une matrice stochastique sur $X'_S$ telle que*

$$\forall i, j \in X'_S,\ (M'_S)_{ij} \geq (M_S)_{ij}$$

*Alors $(X'_S, M'_S)$ est une chaîne de Markov pour laquelle la condition de Doeblin s'applique en une étape.*

*Démonstration.* Prenons un état quelconque $s_f$ de $X'_S$. Comme $X'_S$ est fini, l'ensemble des probabilités de transition vers $s_f$ admet un minimum $\varepsilon_1$. Comme nous venons de le voir, tout état de $(X^*_S, M_S)$ peut atteindre n'importe quel autre état avec une probabilité non nulle, donc $\varepsilon_1 > 0$. De plus, $(M'_S)_{is_f} \geq (M_S)_{is_f} \geq \varepsilon_1$, donc la condition de Doeblin en une étape s'applique à $M'_S$. □

Les conditions sur $M'_S$ sont remplies si on prend par exemple $X'$ comme étant les génomes plus petits qu'une certaine taille en spécifiant des conditions aux bords qui recâblent les

transitions qui sortaient de $X'$ dans le processus original vers n'importe quel autre état dans $X'$. Cela nous servira dans la deuxième partie, où les simulations nécessiteront de délimiter un espace de simulation fini. D'après ce lemme, en restreignant l'espace de calcul et quelles que soient les conditions aux bords, on est assuré de converger vers une distribution stationnaire. Si on peut déterminer la valeur de $\varepsilon_1$, on peut d'ailleurs précalculer la vitesse de convergence vers la distribution stationnaire en se reportant à l'énoncé complet du théorème de Doeblin.

Le lemme montre aussi que la difficulté vient de la taille infinie de l'espace, qui ne garantit pas l'existence d'un minimum. Pour prouver le théorème, il faut donc montrer qu'en prenant un génome aussi grand qu'on veut, on revient asymptotiquement vers le même sous-ensemble d'états. Pour analyser plus efficacement le processus, il faut appliquer un changement d'échelle adapté aux transitions qui nous intéressent, comme on l'a fait dans l'exemple en début de chapitre. L'exemple était un peu extrême puisqu'après changement d'échelle, on avait quasiment une invariance par translation de toute la distribution. Ici, on se contentera d'avoir une invariance par translation des pertes et des gains moyens. On a vu que, dans ce cas, la moyenne avait un sens et permettait de donner la tendance de croissance ou de décroissance à long terme.

Notre modèle contient deux types de processus : les petits indels, à petite échelle, et les duplications et les grandes délétions, à grande échelle. Les duplications et les grandes délétions ont les mêmes propriétés que les copies dans l'exemple. En apparence, les pertes et les gains sont symétriques pour tout point de départ mais il est difficile de juger les compensations entre pertes et gains à l'échelle de plusieurs mutations (figure II.5A). En échelle logarithmique, on obtient l'invariance par translation souhaitée et l'impact plus lourd des délétions apparaît (figure II.5B). On retrouve le fait que 2.59 duplications sont nécessaires en moyenne pour compenser une délétion. Dans le cas des petits indels, l'échelle normale est déjà adaptée, puisque nous avons supposé qu'ils sont invariants par translation (figure II.5C) mais, asymptotiquement, leur impact est plus faible que les délétions et les duplications (figure II.5D), il convient donc de choisir l'échelle logarithmique plutôt que l'échelle normale.

**Définition 3.4.** $S_n$ est la variable aléatoire qui donne l'état de $(X_S, P_S)$ après $n$ mutations. Dans les notations de probabilités, le point de départ de la chaîne $s_0 \in X_S$ est indiqué en indice, comme dans $\Pr_{s_0}[S_n = k] = (P_S^n)_{s_0 k}$, la probabilité d'atteindre l'état $k \in X_S$ en $n$ mutations, partant de $s_0$. Par simplicité, quand les probabilités ne dépendent pas de l'état initial $s_0$, l'indice est omis, comme dans $\Pr_{s_0}[S_{n+1} = j | S_n = i] = \Pr[S_{n+1} = j | S_n = i] = (P_S)_{ij}$.

**Propriété 3.5.** Soit $\Delta(s) = \mathbb{E}[\log(S_{n+1}) | S_n = s] - \mathbb{E}[\log(S_n) | S_n = s]$, la taille moyenne des sauts (pertes ou gains) en échelle logarithmique, sachant que la taille de départ est $s$.

- si la (n+1)-ème mutation est une grande délétion, $\Delta(s) \underset{s \to +\infty}{\longrightarrow} -1$.
- si la (n+1)-ème mutation est une duplication, $\Delta(s) \underset{s \to +\infty}{\longrightarrow} 2\log 2 - 1$.

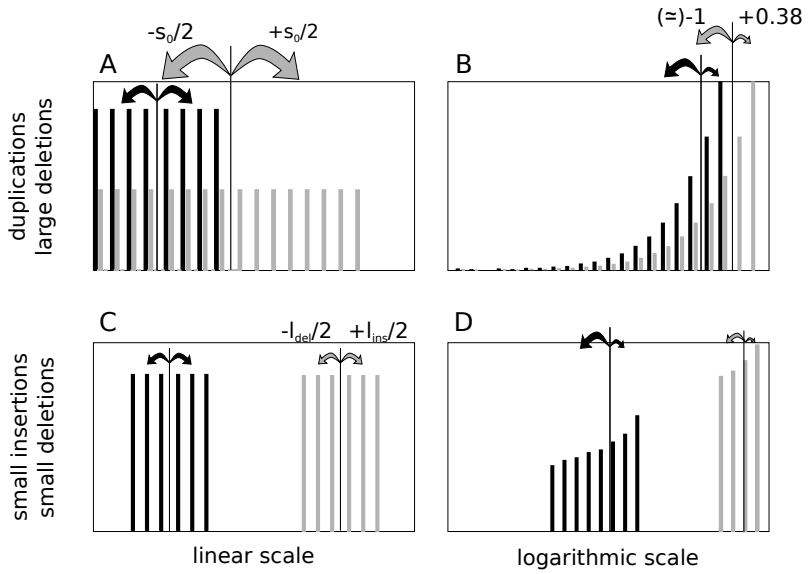


Figure II.5 – Densités de transition schématiques en échelles linéaire et logarithmique. Les flèches indiquent les pertes et les gains moyens pour chaque type de mutation. A : en échelle linéaire, à taux égaux, les processus de duplications et de grandes délétions semblent symétriques, mais les sauts moyens dépendent de la taille de départ. B : en échelle logarithmique, la symétrie apparente est brisée, mais le saut moyen tend à être le même pour toutes les tailles de départ : il y a une nette tendance à la diminution. C : l'échelle linéaire est parfaitement adaptée pour les petits indels dont l'impact ne dépend pas de la taille initiale. D : l'échelle logarithmique casse les bonnes propriétés des indels, mais leur impact moyen tend vers 0 quand la taille initiale augmente, ce qui les rend négligeables comparés aux réarrangements.

- si la (n+1)-ème mutation est un petit indel, $\Delta(s) \underset{s\to+\infty}{\longrightarrow} 0$.

Comme nous nous intéressons à des propriétés asymptotiques, nous introduisons la convention $\log 0 = 0$ pour éliminer le problème de définition en 0.

Cette propriété découle de la figure II.5, la preuve formelle est donnée dans la section 6.1 sous une forme plus détaillée (propriété 6.2). La difficulté (essentiellement technique) de la preuve du théorème 3.2 vient du fait que ce comportement est asymptotique (valable pour les grands génomes). Nous ne contrôlons pas le comportement des petits génomes, où l'impact des petits indels n'est plus négligeable comparé aux duplications et délétions. Le lemme 3.6 sert à contourner ce problème. Tant que les duplications ne sont pas beaucoup plus fréquentes que les délétions ($\mu_{dup} < 2.59\mu_{ldel}$ approximativement), les grands génomes tendent à devenir plus petits, arrivent dans la zone où cette tendance est éventuellement perturbée par les indels mais tendent à y rester.

**Lemme 3.6.** *Si* $(2\log 2 - 1)\mu_{dup} < \mu_{ldel}$*, il existe* $\delta > 0$ *et une taille limite* $\tilde{s} \in X_S$ *tels que*

*(a)* $\forall n \geq 0, \forall s \geq \tilde{s}, \quad \mathbb{E}\left[\log(S_{n+1})|S_n = s\right] \leq \mathbb{E}\left[\log(S_n)|S_n = s\right] - \delta.$

*(b)* $\exists \varepsilon' > 0,\ \forall n \geq 0,$

$$\Pr_{s_0}\left[S_n \leq \tilde{s}\right] \geq \begin{cases} \varepsilon' & si\ s_0 \leq \tilde{s} \\ \varepsilon' \left(1 - \frac{\log s_0}{\log \tilde{s} + n\delta}\right) & si\ s_0 > \tilde{s} \end{cases}$$

*(où* log *est prolongée par* $\log 0 = 0$*)*

Les détails de la démonstration du lemme 3.6 sont donnés dans la section 6.1. Schématiquement, la preuve s'appuie sur la tendance à diminuer pour expliquer le comportement des grands génomes. Pour les petits génomes, l'impossibilité à croître durablement est montrée en étudiant les temps de premier retour. Nous réaffirmons que l'échelle logarithmique est l'échelle adaptée et qu'asymptotiquement, le comportement moyen est dicté par les duplications et les grandes délétions. Si $(2\log 2 - 1)\mu_{dup} < \mu_{ldel}$, la tendance est à la diminution et on peut trouver un seuil $\tilde{s}$ au-delà duquel on peut s'assurer qu'en moyenne, les génomes vont diminuer d'au moins une certaine quantité $\delta$ à chaque mutation. Pour un génome donné initialement au-dessus du seuil, on peut alors montrer qu'en augmentant le nombre de mutations, la probabilité que ce génome passe sous le seuil $\tilde{s}$ tend progressivement vers 1 (partie entre parenthèses dans la relation $(b)$).

Nous montrons ensuite que parmi les génomes se situant sous le seuil $\tilde{s}$, on peut être sûr qu'une certaine fraction y reste en permanence. Pour éviter d'étudier les transitions détaillées, nous agrégeons tous les états situés sous $\tilde{s}$ de manière à créer un scénario défavorable qui sous-estime la fraction de génomes situés sous $\tilde{s}$. Nous montrons que lorsqu'un génome quitte la zone sous $\tilde{s}$, l'espérance du temps de premier retour vers cette zone est finie, ce qui permet de justifier la relation $(b)$. Cela montre également que si la population devait partir d'une taille initiale plus grande que $\tilde{s}$, une fois qu'une partie de la population atteint ce seuil, au moins une certaine fraction y reste durablement (expliquant la présence de $\varepsilon'$ dans les deux parties de la relation $(b)$).

Nous pouvons désormais terminer la démonstration du théorème 3.2. Il reste à montrer que le nombre de mutations permises par le processus de Poisson est suffisant pour permettre à un génome de n'importe quelle taille de réduire sa taille sous $\tilde{s}$ en une génération. Pour cela il faut extrapoler les relations obtenues pour la chaîne mutationnelle $(X_S, P_S)$ à la chaîne générationnelle $(X_S^*, M_S)$.

*Démonstration (du théorème 3.2).* Soit $G_t$ la variable aléatoire qui donne l'état de $(X_S^*, M_S)$ à la génération $t$.

Soit $\tilde{s} \in X_S$ la taille critique donnée par le lemme 3.6. Nous partitionnons l'espace des génomes en $X_{small} = \{i \in X : \text{size}(i) \leq \tilde{s}\}$ et $X_{large} = \{i \in X : \text{size}(i) > \tilde{s}\} = X \backslash X_{small}$. Il s'agit de montrer qu'il existe une taille $s_f \leq \tilde{s}$ qui peut être atteinte en $k = 2$ générations avec une probabilité supérieure à un certain $\varepsilon > 0$, indépendamment de la taille de départ.

**Cas $s_0 \leq \tilde{s}$ ($i_0 \in X_{small}$) :** La probabilité d'atteindre la taille $s_f$ en 2 générations vaut au moins la probabilité d'atteindre $s_f$ à la première génération puis d'y rester. Prenons

un quelconque $s_f \leq \tilde{s}$. On peut appliquer la première partie du lemme 3.3 à $X_{small}$.

$$\exists s_f \leq \tilde{s}, \exists \varepsilon_1 > 0, \forall s_0 \leq \tilde{s}, \quad \Pr\left[G_{t+1} = s_f | G_t = s_0\right] \geq \varepsilon_1. \tag{II.3}$$

Cette relation est vraie en particulier si $s_0 = s_f$.

$$\forall s_0 \leq \tilde{s}, \quad \Pr\left[G_{t+2} = s_f | G_t = s_0\right] \geq (\varepsilon_1)^2.$$

**Cas $s_0 > \tilde{s}$ ($i_0 \in X_{large}$) :** La probabilité d'atteindre la taille $s_f$ identifiée ci-dessus en 2 générations vaut au moins la probabilité d'atteindre un état dans $X_{small}$ à la première génération, puis de rejoindre $s_f$ à la deuxième. Nous commençons par la première étape, qui consiste à calculer $\Pr\left[G_{t+1} \leq \tilde{s} | G_t = s_0\right]$. La probabilité est obtenue en sommant les probabilités de se retrouver sous $\tilde{s}$ en $n$ mutations, données par $(S_n)_{n\in\mathbb{N}}$, en pondérant par la probabilité qu'il y ait effectivement $n$ mutations. Soit $N$ la variable aléatoire qui donne le nombre de mutations. Elle suit une loi de Poisson de paramètre $\mu s_0$. On a donc $\Pr_{=}\left[N = n\right](\mu s_0)^n e^{-\mu s_0}/n!$, d'où

$$\Pr\left[G_{t+1} \leq s | G_t = s_0\right] = \sum_{n\geq 0} \Pr_{s_0}\left[S_n \leq s\right] \frac{(\mu s_0)^n}{n!} e^{-\mu s_0}.$$

D'après le lemme 3.6,

$$\exists \varepsilon' > 0, \Pr_{s_0}\left[S_n \leq s\right] \geq \varepsilon' \left(1 - \frac{\log s_0}{\log \tilde{s} + n\delta}\right).$$

Pour que cette relation ait un sens, posons $n^*(s_0)$ tel que $\forall n \geq n^*(s_0), \Pr_{s_0}\left[S_n \leq s\right] \geq \varepsilon'/2$. On obtient $n^*(s_0) = (2\log s_0 - \log \tilde{s})/\delta$. C'est le nombre de mutations qui sont nécessaires pour être sûr que la probabilité d'aller sous $\tilde{s}$ au moins une fois est plus grande que $1/2$. En éliminant les premiers termes de la somme, on obtient

$$\Pr\left[G_{t+1} \leq s | G_t = s_0\right] \geq \sum_{n\geq n^*(s_0)} \Pr\left[S_n \leq s | S_0 = s_0\right] \frac{(\mu s_0)^n}{n!} e^{-\mu s_0}$$

d'où

$$\Pr\left[G_{t+1} \leq \tilde{s} | G_t = s_0\right] \geq \frac{\varepsilon'}{2} \Pr\left[N \geq (2\log s_0 - \log \tilde{s})/\delta\right].$$

On cherche à quantifier l'écart entre $(2\log s_0 - \log \tilde{s})/\delta$ et $\mathbb{E}\left[N\right]$ en nombre d'écarts-types pour appliquer l'inégalité de Tchebychev. Si $s_0$ est suffisamment grand, on a $(2\log s_0 - \log \tilde{s})/\delta \ll \mathbb{E}\left[N\right] = \mu s_0$. De plus, $\sigma\left[N\right] = \sqrt{\mu s_0}$. Le nombre d'écarts-types qui sépare les deux grandeurs tend donc vers l'infini : on est loin du mode et même de l'essentiel de la distribution. En appliquant une variante de l'inégalité de Tchebychev, on traduit cela mathématiquement en montrant que $\Pr\left[N \geq (2\log s_0 - \log \tilde{s})/\delta\right]$ tend vers 1. Comme cette probabilité est positive pour tout $n$ et qu'elle tend vers une limite non nulle, elle est nécessairement bornée inférieurement par un nombre strictement positif. Le facteur $\varepsilon'/2$ ne change pas cette propriété, on peut donc conclure

$$\exists \varepsilon_2 > 0, \Pr\left[G_{t+1} \leq \tilde{s} | G_t = s_0\right] \geq \varepsilon_2.$$

Une fois qu'un état $i \in X_{small}$ est atteint, nous pouvons appliquer la première partie du lemme 3.3 pour la deuxième relation avec le $s_f$ utilisé dans l'équation (II.3)

$$\forall s_0 > \tilde{s}, \quad \Pr\left[G_{t+2} = s_f | G_t = s_0\right] \geq \varepsilon_2 \varepsilon_1.$$

Le choix de $\varepsilon = \varepsilon_1 \times \min\{\varepsilon_1, \varepsilon_2\}$ remplit la condition de Doeblin pour la chaîne de Markov $(X_S^*, M_S)$. □

Le théorème que nous venons de démontrer montre que sous les hypothèses du modèle et en l'absence de sélection, même si les duplications sont deux fois plus fréquentes que les délétions et même si les petites insertions sont 100 fois plus fréquentes que les petites délétions (à cause de l'activité des éléments transposables par exemple), la distribution de taille des génomes ne va pas partir à l'infini. Dans la section suivante, nous allons utiliser une approximation continue pour mieux caractériser les quantiles de cette distribution de taille.

# 4 Détermination des valeurs maximales pour les quantiles de la distribution stationnaire via une approximation continue

Le théorème 3.2 montre que sans sélection la distribution des tailles de génomes converge vers une distribution stationnaire $\nu_\infty^S$. D'un point de vue théorique, les quantiles de cette distribution donnent des bornes qui indiquent où la population se trouvera asymptotiquement. Cependant, la preuve utilisée justifie l'existence de ces bornes mais ne donne pas d'information plus précise sur leur positionnement. Pour être plus précis, on ne peut pas se contenter de suivre l'évolution de la valeur moyenne pour chaque type de mutation : il faut prendre en compte les moments d'ordre supérieur. Nous avons vu dans la preuve du théorème 3.2 qu'asymptotiquement, les petits indels deviennent négligeables et n'influencent plus la dynamique globale (propriété 3.5) déjà pour le premier moment. Nous utilisons cette remarque pour simplifier le calcul du second moment et des bornes sur les quantiles en prenant un modèle simplifié dans un espace continu.

**Définition 4.1.** Nous considérons un génome de taille initiale $s_0 \in \mathbb{R}_+^*$ qui subit des duplications et des grandes délétions indépendantes. Nous appelons $\hat{S}_n \in \mathbb{R}_+^*$ la variable aléatoire qui donne la taille du génome après $n$ mutations. L'évolution en taille est donnée par $\hat{S}_{n+1} = \lambda_n \hat{S}_n$, où $\lambda_n \hookrightarrow \mathcal{U}([0,1])$ si la mutation $n$ est une délétion et $\lambda_n \hookrightarrow \mathcal{U}([1,2])$ si c'est une duplication. En échelle logarithmique, on a $\log \hat{S}_{n+1} = \log \lambda_n + \log \hat{S}_n$. Soit $J_n = \log \lambda_n$. Comme dans le modèle discret, nous supposons que le nombre de mutations de chaque type est donné par des lois de Poisson de paramètres $\mu_{ldel} s_0$ et $\mu_{dup} s_0$ et que les mutations suivent des processus ponctuels de Poisson indépendants. $\hat{S}_f$ est la taille du génome à la fin de la génération.

Nous commençons par analyser les propriétés des sauts (gains et pertes en échelle logarithmique).

**Propriété 4.2.** Comme les mutations suivent des processus indépendants, les $(J_n)_{n\in\mathbb{N}}$ sont indépendants et identiquement distribués. La distribution des sauts ne dépend pas

de la taille du génome $\hat{S}_n$. De plus

$$\begin{aligned}\mathbb{E}\left[J_n\right] &= \mathbb{E}\left[\log \lambda_n | \text{del.}\right] \Pr\left[\text{deletion}\right] + \mathbb{E}\left[\log \lambda_n | \text{dup.}\right] \Pr\left[\text{duplication}\right] \\ &= -1 \times \frac{\mu_{ldel}}{\mu_{ldel} + \mu_{dup}} + (2\log 2 - 1)\frac{\mu_{dup}}{\mu_{ldel} + \mu_{dup}}\end{aligned}$$

De la même manière, on peut calculer le deuxième moment (dont on pourra déduire l'écart-type $\sigma\left[J_n\right] = \sqrt{\mathbb{E}\left[J_n^2\right] - \mathbb{E}^2\left[J_n\right]}$).

$$\mathbb{E}\left[J_n^2\right] = \frac{2(\mu_{ldel} + (1 - \log 2)^2 \mu_{dup})}{\mu_{ldel} + \mu_{dup}}.$$

*Démonstration.* $\mathbb{E}\left[\log \lambda_n | \text{del.}\right] = \int_0^1 \log x dx = -1$, $\mathbb{E}\left[\log \lambda_n | \text{dup.}\right] = \int_1^2 \log x dx = 2\log 2 - 1$ et pour le deuxième moment $\int_0^1 \log^2 x dx = 2$, $\int_1^2 \log^2 x dx = 2(1 - \log 2)^2$. □

**Propriété 4.3.** $\mathbb{E}_{s_0}\left[\log \hat{S}_n\right] = \log s_0 + n\mathbb{E}\left[J_n\right]$ et $\sigma_{s_0}\left[\log \hat{S}_n\right] = \sqrt{n}\sigma\left[J_n\right]$ par l'indépendance des sauts. D'après le théorème central limite, $\log \hat{S}_n$ suit asymptotiquement une distribution normale.

En passant à une approximation continue, nous obtenons un simple processus de sauts homogène en espace (à condition de prendre l'échelle logarithmique). Comme les deux premiers moments sont finis, le système se comporte asymptotiquement comme un processus de diffusion biaisée. L'écart-type augmente moins rapidement que la moyenne ne se déplace : la moyenne donne une bonne description de la localisation de la distribution. La condition de biais est simplement donnée par le signe de $\mathbb{E}\left[J_n\right]$, la valeur absolue donne la force du biais. On retrouve la même condition que dans le cas discret : la taille des génomes a tendance à diminuer si et seulement si $\mu_{ldel} > (2\log 2 - 1)\mu_{dup}$.

La différence principale avec le système discret est l'homogénéité parfaite en espace, qui reste valable pour les petits génomes. $\log \hat{S}_n$ peut adopter des valeurs négatives, ce qui n'était pas possible dans le cas discret. Quand les génomes deviennent petits (dans le sens de la preuve du théorème 3.2) ils continuent à devenir encore plus petits, il n'y a donc pas de difficulté à montrer qu'ils restent petits, comme on l'a fait dans le lemma 3.6 pour le cas discret. Rappelons que nous avons supprimé l'effet des petits indels, dont l'impact sur les petits génomes est important (s'il y a un fort biais vers les petites insertions par exemple, les petits génomes auront tendance à augmenter).

L'approximation continue permet d'illustrer aisément le comportement global de la chaîne de Markov $(X_S, P_S)$. Pour les grands génomes, le comportement de $(X_S, P_S)$ est proche de l'approximation continue : les génomes suivent alors une diffusion biaisée, la distribution reste donc relativement compacte, sauf pour des taux de mutations près du seuil $\mu_{ldel} = (2\log 2 - 1)\mu_{dup}$. Quand les génomes deviennent petits, le biais devient éventuellement plus faible à cause du caractère discret du processus et des petits indels. Tout au bord, l'état correspondant à la taille $s = 0$ est un mur qui ne peut être franchi. En résumé, le système discret est composé d'un mur d'un côté, de la diffusion biaisée du côté infini

et un comportement moins bien caractérisé entre les deux. Si la diffusion est biaisée vers le mur, la population se dirige donc vers celui-ci jusqu'à atterrir dans la zone floue où son positionnement exact est déterminé par d'autres paramètres, notamment les taux de petits indels. En particulier, la présence des petits indels fait que l'état $s = 0$ n'est pas un état absorbant.

Nous calculons maintenant la distribution après une génération en pondérant les $(\log \hat{S}_n)_{n\in\mathbb{N}}$ par la distribution de Poisson.

**Propriété 4.4.** Par définition, pour tout $x \in \mathbb{R}$,

$$\Pr_{s_0}\left[\log \hat{S}_f \leq x\right] = \sum_{n\geq 0} \Pr_{s_0}\left[\log \hat{S}_n \leq x\right] \frac{((\mu_{ldel}+\mu_{dup})s_0)^n}{n!} e^{-(\mu_{ldel}+\mu_{dup})s_0}.$$

L'espérance est $\mathbb{E}_{s_0}\left[\log \hat{S}_f\right] = \log s_0 + s_0((2\log 2-1)\mu_{dup}-\mu_{ldel})$ et l'écart-type $\sigma_{s_0}\left[\log \hat{S}_f\right] = \sqrt{s_0}\sqrt{2(\log 2-1)^2\mu_{dup}+2\mu_{ldel}}$.

La preuve est donnée dans la section 6.2. Nous introduisons maintenant un paramètre $k$ qui va nous permettre de quantifier la part de la population qui se trouve au-delà de la moyenne plus $k$ écarts-types. Avant de donner une approximation de cette quantité, nous commençons par chercher la position de la valeur correspondant à la moyenne plus $k$ écarts-types en fonction de la taille de départ $s_0$.

**Lemme 4.5.** *Soit $k \geq 1$ et $Q_k(s_0) = \exp\left(\mathbb{E}_{s_0}\left[\log \hat{S}_f\right] + k\sigma_{s_0}\left[\log \hat{S}_f\right]\right)$. On pose $A = \mu_{ldel} - (2\log 2-1)\mu_{dup}$ et $B = \sqrt{2(\log 2-1)^2\mu_{dup}+2\mu_{ldel}}$, de manière à ce que $Q_k(s_0) = \exp(\log s_0 - As_0 + kB\sqrt{s_0})$. Si $(2\log 2-1)\mu_{dup} < \mu_{ldel}$,*

1. *$Q_k(s_0)$ atteint un maximum pour*

$$s_0^{max} = \frac{1}{A} + k^2\frac{B^2}{8A^2}\left(1+\sqrt{1+\frac{16A}{B^2}}\right)$$

2. *$Q_k(s_0) = s_0$ pour une unique valeur $s_{fixed} = k^2B^2/A^2 \geq s_0^{max}$.*

La preuve est assez immédiate, elle est détaillée dans la section 6.2. L'allure générale de la courbe et les points $s_0^{max}$ et $s_{fixed}$ sont illustrés sur la figure II.6 pour le cas où $k = 1$ et $\mu_{dup} = \mu_{ldel} = 10^{-6}$. Grâce à l'inégalité de Tchebychev, $Q_k$ permet de borner les quantiles de la distribution de taille de génome après une génération.

**Proposition 4.6.** *Supposons que $(2\log 2-1)\mu_{dup} < \mu_{ldel}$ et soit $k \geq 1$.*

1. *Il existe une borne $\tilde{s}_{max}$, dépendante de $k$, telle que*

$$\forall s_0 \in \mathbb{R}_+^*, \quad \Pr_{s_0}\left[\hat{S}_f \leq \tilde{s}_{max}\right] \geq 1 - \frac{1}{1+k^2}$$

*En d'autres termes, on peut déterminer un seuil $\tilde{s}_{max}$ indépendant de $s_0$ en dessous duquel on peut capturer une fraction arbitrairement grande de la population après seulement une génération du processus de délétions/duplications.*

2. *Soit*

$$\tilde{s}_{fixed} = k^2 \frac{2(\log 2 - 1)^2 \mu_{dup} + 2\mu_{ldel}}{(\mu_{ldel} - (2\log 2 - 1)\mu_{dup})^2}$$

*on a*

$$\forall s_0 \geq \tilde{s}_{fixed}, \quad \Pr_{s_0}\left[\hat{S}_f \leq \tilde{s}_{fixed}\right] \geq 1 - \frac{1}{1+k^2}$$

*Démonstration.* La proposition est une reformulation du lemme 4.5 en utilisant l'inégalité de Cantelli (ou inégalité unilatérale de Tchebychev). L'inégalité stipule que

$$\Pr_{s_0}\left[\log \hat{S}_f \leq \mathbb{E}_{s_0}\left[\log \hat{S}_f\right] + k\sigma_{s_0}\left[\log \hat{S}_f\right]\right] \geq 1 - \frac{1}{1+k^2}$$

d'où

$$\Pr_{s_0}\left[\hat{S}_f \leq Q_k(s_0)\right] \geq 1 - \frac{1}{1+k^2}$$

La première partie de la proposition s'obtient en posant $\tilde{s}_{max} = Q_k(s_0^{max})$ où $s_0^{max}$ est défini dans le lemme 4.5. La seconde partie s'obtient en prenant $\tilde{s}_{fixed} = s_{fixed}$ du lemme 4.5 et en notant que $Q_k'(s_0) < 0$ pour tout $s_0 \geq s_{fixed}$, d'où $Q_k(s_0) \leq Q_k(s_{fixed})$. □

Le lemme 4.5 et la proposition 4.6 définissent deux suites de bornes qui dépendent du paramètre $k$. $\tilde{s}_{max}$ donne une suite de bornes pour les quantiles de la distribution à la génération $t+1$ quelle que soit la distribution à la génération $t$. Pour $k = 1$, on déduit que la probabilité de se trouver sous $\tilde{s}_{max}$ est au moins 0.5 à n'importe quelle génération. Si les probabilités sont vues comme la densité d'une population infinie, $\tilde{s}_{max}$ donne une borne supérieure de la taille médiane de la population à n'importe quel moment (excepté la population initiale). En augmentant la valeur de $k$, on obtient une zone bornée de l'espace où on peut trouver une fraction arbitrairement proche de 100% de la population.

Il faut noter que le raisonnement appliqué ici resterait valable si un mécanisme de la sélection était appliqué. Dans l'équation générale du modèle (I.1), les individus sont sélectionnés avant mutation. D'après la proposition 4.6, quelle que soit la condition initiale, autrement dit quels que soient les individus sélectionnés, ils se retrouveront sous $\tilde{s}_{max}$ avec une probabilité plus grande que 0.5. La figure II.6 peut être utilisée pour prévoir l'impact de la stringence de la sélection sur la distribution des tailles de génome. Contrairement à ce qu'on pourrait penser naïvement, il semble que pour obtenir les plus grandes tailles de génomes, il faut sélectionner les génomes aux alentours de $s_0^{max}$, alors qu'une sélection qui ne garderait que les génomes les plus grands sélectionnerait surtout les moins robustes : après les mutations, il ne resterait que des petits génomes, d'autant plus petits que le génome initial est grand.

Nous illustrons également la suite de bornes $\tilde{s}_{fixed}$, qui n'est pas une borne valable pour toutes les conditions initiales mais dont l'expression est plus simple que $\tilde{s}_{max}$. 50% des génomes partant de $\tilde{s}_{fixed}$ vont décroître, ce qui indique qu'ils sont déjà fortement instables.

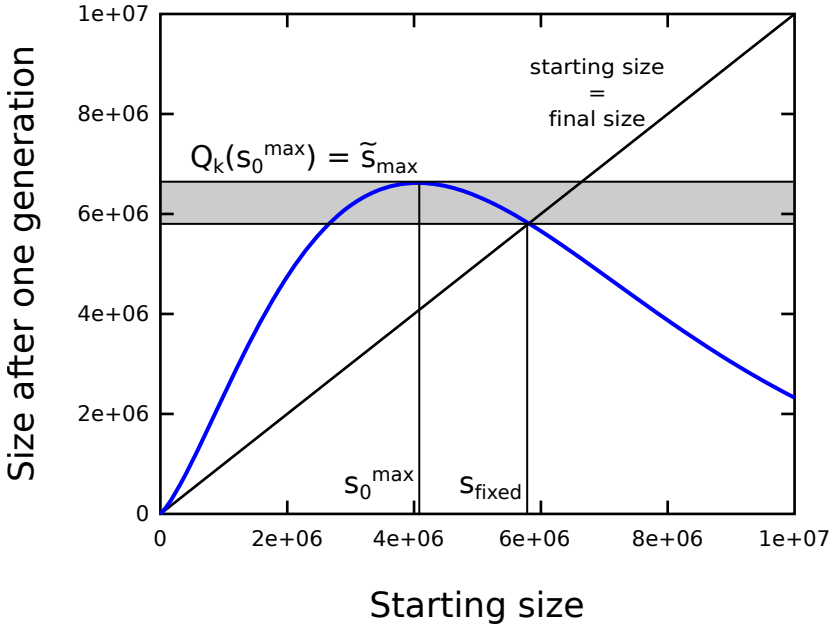


Figure II.6 – Borne supérieure pour la médiane de la distribution (illustration de $Q_k$ avec $k = 1$, $\mu_{dup} = \mu_{ldel} = 10^{-6}$). L'axe des $x$ donne la taille au début de la génération et l'axe des $y$ donne une borne supérieure pour la taille médiane à la fin de la génération. $s_{fixed}$ est le point qui délimite la taille limite à partir de laquelle la probabilité de rétrécir est au moins 0.5, $s_0^{max}$ est la taille qui semble autoriser la croissance la plus forte. Un génome partant de $s_0^{max}$ peut alors croître au-delà de $s_{fixed}$, mais il a plus d'une chance sur deux de réduire à nouveau à la génération suivante. L'aire grise délimite la zone d'états qui semblent accessibles mais ne peuvent en réalité être occupés que de manière transitoire.

La décroissance ne fait qu'empirer si on s'éloigne de $\tilde{s}_{fixed}$. L'analyse suggère qu'il est possible pour les génomes ayant une taille initiale aux alentours de $s_0^{max}$ d'acquérir, avec probabilité supérieure à 0.5, une taille plus élevée que $\tilde{s}_{fixed}$. Cependant, ce comportement ne peut être que transitoire. $\tilde{s}_{fixed}$ est donc un bon candidat pour borner la viabilité à long terme des génomes, donc du comportement moyen, même si de la sélection devait être appliquée. Par exemple, dans le cas simple où $\mu_{ldel} = \mu_{dup} = \mu_{dupdel}$, on obtient

$$\tilde{s}_{fixed} = \frac{k^2}{\mu_{dupdel}} \frac{2(\log 2 - 1)^2 + 2}{(1 - (2\log 2 - 1))^2} \simeq 5.81 \frac{k^2}{\mu_{dupdel}}. \tag{II.4}$$

Cette relation suggère une borne sur la taille du génome qui serait inversement proportionnelle au taux de grandes délétions et de duplications. Cette relation qui rappelle fortement celle de Drake (1991), qui a observé une relation inversement proportionnelle entre la taille de génomes de différents microbes et une estimation de leurs taux de mutation spontané « global ». Cependant, pour que la comparaison soit pertinente, il faudrait montrer que l'estimation des taux de mutation par Drake corrèle avec les taux spontanés de réarrangements, qui sont difficiles à estimer. Les mutations détectées par Drake sont les mutations qui inactivent un gène précis : cela inclut les mutations ponctuelles, les petites insertions et délétions qui causent un frameshift, l'insertion d'un élément transposable dans le gène, la délétion partielle ou totale du gène, les duplications dont le point d'insertion se trouve dans le gène, et les inversions dont l'un des points de cassure se trouve dans le gène (Drake,

1991). Ainsi, le taux de mutation spontané estimé dans ces expériences inclut donc bien des événements de duplication et de délétion, mais il inclut aussi d'autres événements. Selon Lynch (2010) , la relation de Drake n'est valable que pour les virus et les procaryotes, car chez les eucaryotes, la relation serait en sens inverse, les taux de mutations spontanés les plus élevés correspondant aux génomes les plus longs. Cependant, les estimations du taux de mutation prises en compte par Lynch (2010) sont des estimations du taux de mutation ponctuelle.

L'analyse proposée dans cette section a été faite dans un cadre simplifié (approximation continue), mais c'est vraisemblablement une bonne approximation pour les grands génomes dans le modèle discret qui comprend tous les types de mutations. Nous nous attendons à ce qu'une proposition équivalente puisse être obtenue dans le cadre plus général, en utilisant un mélange des raisonnements utilisés dans le cas discret et continu. Il s'agirait d'étendre l'analyse de la section 3 en incorporant les deuxièmes moments, en montrant que l'écart-type est asymptotiquement négligeable comparé aux variations de la moyenne. Cela nécessiterait néanmoins de maintenir le découpage entre grands et petits génomes, pour délimiter le domaine où l'approximation continue est valide. Ce découpage dépend en premier lieu des taux des petits indels, qui doivent être négligeables, ce qui complique un peu les calculs et certainement l'expression des bornes elles-mêmes si les taux d'indels sont tellement forts qu'ils perturbent le comportement global là où la borne était censée se trouver d'après l'approximation continue. Les impacts des indels sont cependant généralement suffisamment faibles pour que l'approximation continue soit bonne, on pourra donc essayer de tester la validité des bornes obtenues par cette approximation, y compris en ajoutant de la sélection.

# 5 Généralisations

Le théorème 3.2 montre qu'il existe une distribution stationnaire de la taille des génomes en l'absence de sélection à la seule condition que $(2\log 2 - 1)\mu_{dup} < \mu_{ldel}$. Cela signifie que sous les hypothèses du modèle, même si les duplications sont deux fois plus fréquentes que les délétions et même si les petites insertions sont par exemple 100 fois plus fréquentes que les petites délétions (du fait de l'activité d'éléments transposables par exemple), la distribution de taille des génomes ne va pas partir à l'infini. La preuve a été obtenue dans le cas où les grandes délétions et les duplications sont distribuées uniformément. Cependant, cette condition n'est pas nécessaire. Dans la première partie de cette section, nous donnons une formulation plus générale du théorème pour prendre en compte d'autres types de distributions. Dans le reste de la section, nous étudions les implications des résultats obtenus quand on applique de la sélection et selon la taille de la population.

## 5.1 Extension du théorème 3.2 à des distributions plus générales de duplications et de délétions

Initialement, nous avons supposés que les gains et pertes dus aux délétions et aux duplications suivaient une distribution uniforme pour illustrer la preuve, mais celle-ci reste valable tant que des conditions similaires à la propriété 3.5 sont remplies. Tous les calculs intermédiaires restent valables si les changements moyens de la taille du génome pour les petits indels, les duplications et les grandes délétions tendent vers une constante dans une échelle spécifiée par une fonction positive et croissante $f$. On retrouve l'idée de trouver une échelle pour laquelle les sauts moyens sont constants, comme dans la figure II.5. Dans ce cas, l'existence de la distribution stationnaire est simplement déterminée par la valeur des sauts moyens.

**Corollaire 5.1.** *(Généralisation du théorème 3.2) Supposons que les distribution de taille des duplications, grandes délétions et indels soient telles qu'il existe une fonction d'échelle $f$ positive et croissante vérifiant les conditions suivantes.*

*Pour* $\Delta(s) = \mathbb{E}\left[f(S_{n+1}) - f(S_n)|S_n = s\right]$ *:*

- *si la (n+1)e mutation est une délétion,* $\Delta(s) \underset{s\to+\infty}{\longrightarrow} \delta_{ldel}$.
- *si la (n+1)e mutation est une duplication,* $\Delta(s) \underset{s\to+\infty}{\longrightarrow} \delta_{dup}$.
- *si la (n+1)e mutation est une petite insertion,* $\Delta(s) \underset{s\to+\infty}{\longrightarrow} \delta_{ins}$.
- *si la (n+1)e mutation est une petite délétion,* $\Delta(s) \underset{s\to+\infty}{\longrightarrow} \delta_{sdel}$.

*où* $\delta_{ldel} \leq 0$, $\delta_{dup} \geq 0$, $\delta_{ins} \geq 0$ *et* $\delta_{sdel} \leq 0$ *sont des constantes dont une au moins est non nulle.*

*Alors la chaîne de Markov* $(X_S^*, M_S)$ *admet un vecteur de probabilité stationnaire asymptotique* $\nu_\infty^S$ *si*

$$\mu_{ldel}\delta_{ldel} + \mu_{dup}\delta_{dup} + \mu_{ins}\delta_{ins} + \mu_{sdel}\delta_{sdel} < 0 \tag{II.5}$$

*Si la taille des duplications et des grandes délétions grandit significativement plus vite avec la taille du génome que celle des petits indels, c'est-à-dire si* $\delta_{ins} = \delta_{sdel} = 0$*, la condition devient simplement*

$$\frac{\mu_{dup}}{\mu_{ldel}} < \frac{|\delta_{ldel}|}{\delta_{dup}}$$

*sauf si* $\delta_{dup} = 0$*, auquel cas l'existence de la distribution stationnaire est inconditionnelle.*

La preuve est la même que pour celle du théorème 3.2, en remplaçant la condition $(2\log 2 - 1)\mu_{dup} - \mu_{ldel}$ par la condition (II.5), en particulier dans la définition de $\delta$, qui incorpore l'impact moyen de toutes les mutations et définit la force du biais.

S'il existe une mise à l'échelle avec la taille du génome, l'existence d'une distribution stationnaire ne dépend pas du détail des distributions de taille des mutations mais uniquement du premier moment dans la nouvelle échelle. Si les processus sont de nature multiplicative (pas forcément uniforme), $f = \log$ est la mise à l'échelle naturelle, comme illustré jusqu'ici. Si par contre la largeur des distributions de pertes et de gains ne grossit pas proportionnellement à la taille initiale, un autre choix de $f$ sera probablement plus approprié. À l'extrême, si les pertes et les gains moyens tendent déjà vers une constante en échelle linéaire ($f = id_{\mathbb{N}}$) pour les grandes délétions et les duplications, l'impact des petits indels devra être incorporé dans la condition d'existence de la distribution stationnaire. Cette hypothèse nous paraît toutefois peu plausible, car elle implique que la taille des délétions et des duplications ne dépend quasiment pas de la taille du génome.

Si on choisit une distribution de pertes et de gains relative (comme par exemple c'était le cas pour Falush et Iwasa (1999) ou avec la distribution uniforme), elle est identique (au moins asymptotiquement) pour toutes les tailles de départ en échelle logarithmique, donc la moyenne tend vers une constante. On a alors une condition simple qui permet de s'assurer que les génomes restent bornés spontanément. La distribution relative des pertes et des gains peut être fixée librement : elle peut être uniforme, exponentiellement décroissante, multimodale, etc.

Si jamais le second moment converge vers une valeur finie dans l'échelle donnée par $f$, comme c'est le cas pour les processus de nature multiplicative, l'analyse de la section 4 s'appliquera. On peut donc trouver des limites à la taille du génome qui sont valables à chaque génération. Comme discuté ci-dessous, il sera alors impossible pour la sélection de surmonter les bornes sur la taille du génome. Il est probable que la démonstration puisse être étendue au cas où le second moment ne converge pas dans l'échelle donnée par $f$, mais reste borné. La preuve risquerait d'être plus calculatoire et les bornes encore plus faibles, mais qualitativement on retrouverait l'idée que la sélection ne peut pas surmonter la déstabilisation par les réarrangements.

## 5.2 Prédictions pour le modèle général avec sélection

Il faut ici rappeler que la distribution $\nu_t$ peut être interprétée de deux façons différentes. Dans le cadre markovien, elle représente la probabilité d'occupation de chaque état pour un individu unique. Cependant, si on suppose qu'on a une population infinie d'individus indépendants, alors tous les états sont occupés selon les proportions indiquées : les probabilités deviennent une densité d'individus. Dans cette dernière interprétation, l'existence d'une distribution stationnaire en l'absence de sélection implique qu'on peut indiquer des seuils sous lesquels une partie arbitrairement grande de la population va se trouver asymptotiquement. D'un point de vue biologique, cela peut être vue comme la dynamique spontanée des génomes. Le théorème 3.2 montre que la condition à laquelle la taille des génomes diminue spontanément est déterminée seulement par les grandes délétions et les duplications, avec des valeurs plutôt non-intuitives ($\mu_{dup} < 2.59\mu_{ldel}$ approximativement). On pourrait montrer que dans le cas contraire, la chaîne de Markov est transiente, c'est-

à-dire qu'il n'y a pas de distribution stationnaire, les génomes croissent spontanément.

Dans la section 3, nous avons identifié un seuil $\tilde{s}$ sous lequel les génomes sont forcés de revenir périodiquement, sans avoir accès à des valeurs plus précises. Dans la section 4, les seuils $\tilde{s}_{max}$ et $\tilde{s}_{fixed}$ apportent des informations plus précises : on peut quantifier la proportion de la population située sous ces seuils et, grâce au paramètre $k$ dans la proposition 4.6, on peut rendre cette proportion arbitrairement grande. Cette proposition n'a pas été prouvée pour le modèle complet, mais il est probable que, tant que les seuils sont placés assez hauts pour s'assurer que l'impact des petits indels est négligeable, on puisse trouver des expressions similaires. Comme ces seuils sont valables à chaque génération, ils ne peuvent pas être surmontés par la sélection. Si on reprend le modèle général et l'équation $I.1$ ou sa projection sur la taille II.1, on peut noter que la sélection a lieu avant les mutations. Quelle que soit la taille des individus renvoyés par la sélection dans le vecteur $\mathrm{Sel}_t(\nu_t)$, les seuils de la proposition 4.6 s'appliqueront à ce vecteur. Par exemple, en variant $k$ dans $\tilde{s}_{max}$, on peut garantir qu'au plus $100/(1+k^2)\%$ des descendants contenus dans $\nu_{t+1}$ auront une taille plus grande que $\tilde{s}_{max}$, quelle que soit l'action de la sélection.

Les seuils imposent donc une limite supérieure de stabilité à la taille des génomes mais n'indiquent pas où la population va converger exactement dans le domaine de stabilité. L'analyse se concentre ici sur la dynamique globale, alors que la dynamique plus locale dépend de l'opérateur de sélection et des détails des processus de mutation. Cependant, la figure II.6 donne une idée de l'interaction entre les opérateurs de sélection et de mutation, étant donné l'impact fort de la taille initiale sur la taille finale. La force de la sélection peut donc déterminer à quelle distance des bornes la population peut se stabiliser. Paradoxalement, plus la sélection en faveur des grands génomes est forte, plus la probabilité de se retrouver dans la zone instable à la génération suivante est forte. La population risque donc de se stabiliser plutôt loin des seuils pour garantir sa robustesse. Dans ce cas, une forte pression à l'augmentation des génomes risque de conduire les génomes à se stabiliser sur de plus petites tailles.

## 5.3 Généralisation à une population finie

Les remarques pour une population infinie tiennent en partie pour une population de taille finie d'individus mutant indépendamment. En effet, l'interprétation des chaînes de Markov est valable pour un seul individu dans un cadre probabiliste. En découplant les étapes de sélection et de mutation, on peut déterminer les probabilités de se trouver sous les seuils donnés par la proposition 4.6. L'idée est la même que pour la population infinie, sauf que l'évolution des individus est stochastique et ne peut donc pas être décrite par une équation aussi simple que l'équation (I.1). Cependant, si on suppose qu'à la génération $t$, $I_t$ individus appartenant à l'espace $X$ ont survécu après l'étape de sélection, nous pouvons utiliser les conclusions de la proposition 4.6. Il existe un seuil $\tilde{s}_{max}$ tel que la probabilité qu'un quelconque des $I_t$ génomes se trouve au-dessus de $\tilde{s}_{max}$ après mutation est au plus $1/(1+k^2)$. Comme les individus mutent indépendamment, on peut donner une borne supérieure du nombre de génomes au-dessus de $\tilde{s}_{max}$ à chaque génération via

une distribution binomiale $\mathcal{B}(I_t, 1/(1+k^2))$. La proportion de génomes au-dessus de $\tilde{s}_{max}$ prédite par la distribution binomiale est la même quelle que soit la taille de la population $I_t \in \mathbb{N} : 100/(1+k^2)\%$, celle-ci change l'écart-type autour de cette valeur, qui diminue quand la taille de la population augmente. Quand la taille tend vers l'infini, on retrouve bien le résultat du cas infini.

# 6 Détails des preuves

## 6.1 Preuve du lemme 3.6

*Remarque* 6.1. Dans cette preuve, nous considérerons la fonction log comme une fonction définie sur $\mathbb{N}$ par $\log 0 = 0$ et les valeurs usuelles du logarithme néperien pour $n \geq 1$. De cette façon, log est une fonction positive et croissante sur $\mathbb{N}$.

Avant de commencer la preuve à proprement parler du lemme 3.6, nous allons lister des propriétés importantes de la chaîne de Markov $(X_S, P_S)$. Nous allons distinguer le comportement des petits génomes de celui des grands génomes. Nous commençons par montrer que, statistiquement, la taille qu'un génome peut adopter après une mutation est d'autant plus petite que la taille initiale est petite.

**Propriété 6.2.** $\forall s_1, s_2 \in X_S, \quad s_1 \leq s_2 \Rightarrow \Pr[S_{n+1} \leq k | S_n = s_1] \geq \Pr[S_{n+1} \leq k | S_n = s_2]$

*Démonstration.* Soit $F_s(k) = \Pr[S_{n+1} \leq k | S_n = s]$. Nous notons $\mathbf{1}_{\{s,\dots\}} := \mathbf{1}_{\{k \in \mathbb{N}, k \geq s\}}$.

Nous conditionnons sur la nature de la $(n+1)$-ème mutation. Si la mutation est une grande délétion, $F_0 = \mathbf{1}_{\mathbb{N}}$ et si $s > 0$, $F_s(k) = (k+1)\mathbf{1}_{\{0,\dots,s-1\}}(k)/s + \mathbf{1}_{\{s,\dots\}}(k)$. Pour une duplication, $F_0 = \mathbf{1}_{\mathbb{N}}$, si $s > 0$, $F_s(k) = (k-s)\mathbf{1}_{\{s+1,\dots,2s\}}(k)/s + \mathbf{1}_{\{2s+1,\dots\}}(k)$. Pour les petits indels, la distribution est invariante par translation, on a donc $F_{s_1}(k) = F_{s_2}(k + s_2 - s_1) \geq F_{s_2}(k)$ comme $s_2 - s_1 \geq 0$.

Soit $s_1 \leq s_2 \in X_S$. Pour chaque type de mutation, $\forall k \in X_S$, $F_{s_1}(k) \geq F_{s_2}(k)$. Déconditionner en multipliant par la probabilité de chaque mutation ne change pas cette relation, puisque cela revient à prendre une moyenne pondérée. □

Cette propriété n'est pas essentielle mais sera utile pour montrer que les petits génomes ont tendance à rester petits. En effet, si on peut monter qu'un génome de taille $s_{min}$ tend à rester petit, alors cela sera nécessairement vrai pour les tailles de départ plus petites. On pourrait également utiliser une condition moins restrictive (voir remarque 6.8). Pour donner un cadre formel à cette utilisation d'un état « pire cas », nous introduisons la définition suivante.

**Définition 6.3.** Soit $s_{min} > 0$. Nous définissons un sous-processus $S_n^{\triangleright}$ pour la chaîne $(X_S^{\triangleright}, P_S^{\triangleright}))$ de la manière suivante. Nous gardons uniquement les états plus grands ou

égaux à $s_{min}$ dans $X_S$, soit $X_S^{\triangleright} = \{s \in X_S : s \geq s_{min}\} \subset X_S$. Les transitions dans $P_S^{\triangleright}$ sont les mêmes que celles de $P_S$, sauf celles qui allaient vers des états plus petits que $s_{min}$ qui sont recâblées vers $s_{min}$. Formellement,

$$\forall i \geq s_{min}, \forall j > s_{min}, \quad \Pr\left[S_{n+1}^{\triangleright} = j | S_n^{\triangleright} = i\right] = (P_S^{\triangleright})_{ij} := (P_S)_{ij} = \Pr\left[S_{n+1} = j | S_n = i\right]$$

et

$$\forall i \geq s_{min}, \quad (P_S^{\triangleright})_{is_{min}} := \sum_{0 \leq k \leq s_{min}} (P_S)_{ik} = \Pr\left[S_{n+1} \leq s_{min} | S_n = i\right]$$

Commençons par lister quelques propriétés simples de la chaîne de Markov $(X_S^{\triangleright}, P_S^{\triangleright})$.

**Propriété 6.4.** Si $\max\{\mu_{sdel}, \mu_{ldel}\} > 0$ et $\max\{\mu_{ins}, \mu_{dup}\} > 0$, $(X_S^{\triangleright}, P_S^{\triangleright})$ est irréductible et apériodique.

*Démonstration.* L'irréductibilité siginifie qu'il est possible d'aller de n'importe quel état $s_1 \in X_S^{\triangleright}$ à n'importe quel autre état $s_2 \in X_S^{\triangleright}$ en un nombre fini de mutations. Sous les conditions de la propriété, chaque état communique avec son voisin direct, la chaîne est bien irréductible. La périodicité d'un état $s \in X_S^{\triangleright}$ est donné par $\text{pgcd}\{n \in \mathbb{N} : \Pr_s\left[S_n^{\triangleright} = s\right] > 0\}$. Comme la chaîne est irréductible, elle est apériodique s'il existe un état $s \in X_S^{\triangleright}$ pour lequel $(P_S^{\triangleright})_{ss} > 0$ (voir par exemple Woess, 2009). C'est le cas pour $s_{min}$ si $\max\{\mu_{sdel}, \mu_{ldel}\} > 0$ puisque toutes les transitions liées à une perte sont recâblées vers $s_{min}$. □

**Propriété 6.5.** Si $s_0 \geq s_{min}$, $\forall n \geq 1$,

$$\Pr_{s_0}\left[\bigcap_{k=1}^{n}(S_k^{\triangleright} > s_{min})\right] = \Pr_{s_0}\left[\bigcap_{k=1}^{n}(S_k > s_{min})\right]$$

Cette propriété est importante dans la suite mais relativement évidente à montrer en étudiant le recâblage : on calcule les probabilités d'un chemin qui n'emprunte que des transitions qui sont identiques dans les deux processus. Pour les grands génomes, $(X_S^{\triangleright}, P_S^{\triangleright})$ se comporte exactement comme $(X_S, P_S)$.

Nous arrivons enfin à la propriété de « pire cas » qui découle de la propriété 6.2.

**Propriété 6.6.** Soit $s_0 \in X_S$ et $s_0^{\triangleright} = \max\{s_0, s_{min}\}$.

$$\forall n \geq 0, \forall s \geq s_{min}, \quad \Pr_{s_0}\left[S_n \leq s\right] \geq \Pr_{s_0^{\triangleright}}\left[S_n^{\triangleright} \leq s\right]$$

En particulier $\Pr_{s_0}\left[S_n \leq s_{min}\right] \geq \Pr_{s_0^{\triangleright}}\left[S_n^{\triangleright} = s_{min}\right]$.

*Démonstration.* La preuve se fait par récurrence. Comme $s_0^{\triangleright} \geq s_0$, la propriété est évidente pour $n = 0$. Supposons qu'elle soit vraie pour $n \geq 0$ et prenons $s \geq s_{min}$. Il s'agit de montrer que $\Pr_{s_0}\left[S_{n+1} \leq s\right] - \Pr_{s_0^{\triangleright}}\left[S_{n+1}^{\triangleright} \leq s\right] \geq 0$. Pour cela, nous allons calculer les

probabilités en conditionnant par la distribution après $n$ mutations. Pour que la comparaison entre les deux processus soit intéressante, il faut comparer les probabilités données par une même fraction de la population, autrement dit s'assurer d'avoir pris les mêmes quantiles pour $S_n$ et $S_n^{\triangleright}$. $\forall x \in [0,1]$, soit $q_x = \min\{k \in X_S : \Pr_{s_0}[S_n \leq k] \geq x\}$, la fonction qui donne la position des quantiles de $S_n$. Pour chaque $n$, la taille de génome reste finie (les duplications doublent au plus la taille du génome), $q_1$ est bien défini et, comme $X_S$ est discret, $q_x$ est constant par morceaux. Le conditionnement par la fonction des quantiles s'écrit

$$\begin{aligned}\forall x \in [0,1], \quad h(x) = &\Pr_{s_0}[S_{n+1} \leq s | S_n < q_x] \Pr_{s_0}[S_n < q_x] \\ &+ \Pr[S_{n+1} \leq s | S_n = q_x] (x - \Pr_{s_0}[S_n < q_x])\end{aligned}$$

Il s'agit d'ajouter toutes les contributions provenant des états inférieurs à l'état donné par $q_x$, tandis que $x$ contrôle la proportion des contributions provenant de $q_x$ de manière linéaire. Quand $x$ atteint la valeur $\Pr_{s_0}[S_n \leq q_x]$, toutes les contributions sont épuisées, $q_x$ pointe alors vers une nouvel état. De cette sorte, $h(0) = 0$, $h(1) = \Pr_{s_0}[S_{n+1} \leq s]$ et $h$ est une fonction continue et linéaire par morceaux. Les points où $h$ n'est pas dérivable sont les valeurs $\{\Pr_{s_0}[S_n \leq k], k \in X_S\}$ pour lesquelles la valeur de $q_x$ change. Comme la taille du génome est finie pour tout $n$, cet ensemble est fini. On définit de la même façon $q_x^{\triangleright}$ et $h^{\triangleright}$, en utilisant $S_n^{\triangleright}$ and $X_S^{\triangleright}$. Par hypothèse de récurrence, $q_x \leq q_x^{\triangleright}$ et, aux points où la dérivée est définie (donc $q_x$ et $q_x^{\triangleright}$ sont localement constants)

$$(h - h^{\triangleright})'(x) = \Pr[S_{n+1} \leq s | S_n = q_x] - \Pr\left[S_{n+1}^{\triangleright} \leq s | S_n^{\triangleright} = q_x^{\triangleright}\right]$$

Nous appliquons la propriété 6.2,

$$(h - h^{\triangleright})'(x) \geq \Pr[S_{n+1} \leq s | S_n = q_x^{\triangleright}] - \Pr\left[S_{n+1}^{\triangleright} \leq s | S_n^{\triangleright} = q_x^{\triangleright}\right].$$

Par définition de $S^{\triangleright}$, $\Pr[S_{n+1} \leq s | S_n = q_x^{\triangleright}] = \Pr\left[S_{n+1}^{\triangleright} \leq s | S_n^{\triangleright} = q_x^{\triangleright}\right]$, d'où $(h - h^{\triangleright})'(x) \geq 0$. En intégrant sur $x \in [0,1]$, sachant que le nombre de points non définis est fini et que $h - h^{\triangleright}$ est continue par ailleurs, nous obtenons

$$(h - h^{\triangleright})(1) = \Pr_{s_0}[S_{n+1} \leq s] - \Pr_{s_0^{\triangleright}}\left[S_{n+1}^{\triangleright} \leq s\right] \geq 0$$

□

*Remarque* 6.7. Les propriétés données jusqu'ici servent à simplifier l'analyse du comportement des petits génomes, sachant que $s_{min}$ représente un pire cas. Le processus $(X_S^{\triangleright}, P_S^{\triangleright})$ nous servira quand il s'agira de montrer que les génomes plus petits que $s_{min}$ restent en partie plus petit que $s_{min}$. Si on arrive à le démontrer pour $s_{min}$, la propriété 6.6 assure que ce sera vrai également pour les états plus petits.

*Remarque* 6.8. Pour créer le pire cas, nous avons utilisé la propriété 6.2. On aurait pu choisir un représentant du pire cas d'une autre manière, par exemple en recensant toutes les transitions qui partent sous $s_{min}$ en cherchant celles qui autorisent de croître le plus haut au-dessus de $s_{min}$. On peut alors créer un état chimérique en sélectionnant les transitions les plus défavorables, celles qui autorisent la fraction la plus grande de la population à repasser au-dessus de $s_{min}$ (c'est possible car le nombre de transitions est fini). Un tel

cheminement permettrait de relaxer certaines hypothèses initiales, comme par exemple l'invariance par translation des indels en autorisant le détail des distributions à varier librement en fonction de $s_0$. Ceci étant dit, la propriété 6.2 semble plausible biologiquement.

Nous nous tournons maintenant vers les propriétés des grands génomes. Asymptotiquement, le processus est dominé par les grandes délétions et les duplications. Comme nous l'avons vu plusieurs fois, l'échelle adaptée pour analyser le processus est celle où l'impact des mutations tend vers une constante, en l'occurrence l'échelle logarithmique.

**Propriété 6.9.** Soit $\Delta(s) = \mathbb{E}\left[\log(S_{n+1})|S_n = s\right] - \mathbb{E}\left[\log(S_n)|S_n = s\right]$.

(a)
- si la (n+1)-ème mutation est une délétion

$$\forall s \geq 3, \quad \Delta(s) = -1 + \frac{1}{s}\left(\sum_{k=2}^{s-1} \log k - \int_0^s \log x dx\right) \underset{s\to+\infty}{\longrightarrow} -1$$

- si la (n+1)-ème mutation est une duplication

$$\forall s \geq 3, \quad \Delta(s) = 2\log 2 - 1 + \frac{1}{s}\left(\sum_{k=s+1}^{2s} \log k - \int_s^{2s} \log x dx\right) \underset{s\to+\infty}{\longrightarrow} 2\log 2 - 1$$

- si la (n+1)-ème mutation est une petite délétion ou une petite insertion (respectivement)

$$\Delta(s) \underset{s\to+\infty}{=} O\left(\frac{1}{s}\right) \underset{s\to+\infty}{\longrightarrow} 0$$

- si la (n+1)-ème mutation est une mutation ponctuelle, translocation ou inversion, $\Delta(s) = 0$

(b) Sans préjuger de la prochaine mutation, on a

$$\Delta(s) = \frac{(2\log 2 - 1)\mu_{dup} - \mu_{ldel}}{\mu} + \xi(s)$$

avec $\lim_{s\to+\infty} \xi(s) = 0$.

*Démonstration.* Si la (n+1)-ème mutation est une délétion,

$$\mathbb{E}\left[\log(S_{n+1})|S_n = s\right] = \sum_{k\geq 0} \log(k) \Pr\left[S_{n+1} = k|S_n = s\right] = \frac{1}{s}\sum_{k=2}^{s-1} \log k$$

$$= \frac{1}{s}\left(s(\log s - 1) + \sum_{k=2}^{s-1} \log k - \int_0^s \log x dx\right) = \log s - 1 + \left(\sum_{k=2}^{s-1} \log k - \int_0^s \log x dx\right)$$

On obtient alors facilement le résultat car $\mathbb{E}\left[\log(S_n)|S_n = s\right] = \log s$. La démonstration pour les duplications est similaire, elle est immédiate pour les autres types de mutation.

Le résulat (b) s'obtient un appliquant la loi des espérances totales en conditionnant sur chaque type de mutation. $\xi(s)$ contient tous les termes dus à la discrétisation, qui tendent vers 0 asymptotiquement. La limite s'obtient par des comparaisons sommes-intégrales et des développements en série de Taylor. En effet, pour les grandes délétions, on peut noter que

$$\forall k \geq 2, \quad \log k \leq \int_k^{k+1} \log(x)dx \leq \log(k+1)$$

(log étant croissante). On somme pour $k \in \{2, ..., s-1\}$ pour un $s \geq 3$ donné

$$\sum_{k=2}^{s-1} \log k \leq \int_2^s \log(x)dx \leq \sum_{k=2}^{s-1} \log(k+1) = \sum_{k=2}^{s-1} \log k - \log 2 + \log s$$

on réorganise chaque côté de l'inégalité séparément et on multiplie par $1/s$.

$$\frac{-\log s + \log 2}{s} \leq \frac{1}{s}\left(\sum_{k=2}^{s-1} \log k - \int_2^s \log(x)dx\right) \leq 0$$

qui tend bien vers 0 quand $s \to +\infty$. Nous avons ici omis le terme $\int_0^2 \log(x)dx$ présent dans l'expression originale qui tend également vers 0 une fois qu'il est divisé par $s$. En changeant les valeurs de sommation, on obtient le même résultat pour les duplications.

Pour les petites délétions, on pose $f_s(k) = \Pr\left[S_{n+1} = k|S_n = s\right]$ sachant que la (n+1)-ème mutation était une petite délétion. Pour $s \geq l_{sdel}$

$$\begin{aligned}\Delta(s) &= \sum_{k=-l_{sdel}}^{-1} f_s(s+k)(\log(s+k) - \log s) = \sum_{k=-l_{sdel}}^{-1} f_s(s+k)\log(1+k/s) \\ &\underset{s\to+\infty}{=} \sum_{k=-l_{sdel}}^{-1} f_s(s+k)O\left(\frac{1}{s}\right) = O\left(\frac{1}{s}\right)\end{aligned}$$

Le résultat est symétrique pour les petites insertions.

Pour revenir au cas général, on déconditionne en multipliant tous ces termes qui tendent vers 0 par des constantes, on a donc $\lim_{s\to+\infty} \xi(s) = 0$. □

Asymptotiquement le biais est déterminé par $(2\log 2 - 1)\mu_{dup} - \mu_{ldel}$. La propriété suivante s'applique au cas où le bias est négatif.

**Propriété 6.10.** Si $(2\log 2 - 1)\mu_{dup} - \mu_{ldel} < 0$, on pose $\delta = \frac{|(2\log 2-1)\mu_{dup}-\mu_{ldel}|}{2\mu} > 0$. Il existe une taille limite $\tilde{s} \in X_S$ telle que

$$\forall n \geq 0, \forall s \geq \tilde{s}, \quad \mathbb{E}\left[\log(S_{n+1})|S_n = s\right] \leq \mathbb{E}\left[\log(S_n)|S_n = s\right] - \delta$$

Cette propriété est simplement une implication de la propriété 6.9(b) en utilisant la définition de la limite.

*Remarque* 6.11. Cette propriété est en pratique assez forte puisqu'elle s'applique à n'importe quelle distribution dont le support est intégralement au-dessus de $\tilde{s}$. D'après le théorème d'espérance totale, on peut conditionner par l'une de ces distributions au lieu de la condition $S_n = s$.

Nous avons tous les outils nécessaires à la preuve du lemme 3.6. Les dernières propriétés indiquent que les grands génomes décroissent quand $(2\log 2 - 1)\mu_{dup} - \mu_{ldel} < 0$. Il reste à montrer à quelle vitesse ils vont sous le seuil $\tilde{s}$ et qu'ils restent sous le seuil asymptotiquement. D'après les propriétés sur les petits génomes, nous savons que pour analyser ce qui se passe sous le seuil $\tilde{s}$ nous pouvons utiliser le sous-processus $S^{\triangleright}$ (remarque 6.7). Nous allons adopter le point de vue des temps de premier retour : dire qu'un génome reste sous $\tilde{s}$ veut dire qu'il y revient infiniment souvent et que le temps entre deux passages a une espérance finie. Cela suffit pour prouver qu'une fraction de la population reste sous $\tilde{s}$ à chaque instant.

*Démonstration (du lemme 3.6).* a) (a) correspond exactement à la propriété 6.10.

b) Nous montrons (b) en quatre étapes. L'étape 1 montre qu'en attendant suffisamment longtemps, tout génome va sous $\tilde{s}$ avec une probabilité arbitrairement proche de 1, l'étape 2 que les petits génomes tendent à rester petits en analysant les temps de premier retour partant de $\tilde{s}$, l'étape 3 en conclut la relation pour les petits génomes et l'étape 4 utilise les étapes 1 et 3 pour déduire la relation pour les grands génomes.

1. Prenons $s_0 \geq \tilde{s}$. $\forall n \geq 0$, soit $A_n = \bigcap_{0\leq k\leq n}(S_k \geq \tilde{s})$, $p_n = \Pr_{s_0}[A_n]$ et $e_n = \mathbb{E}_{s_0}[\log(S_n)|A_n]$. En conditionnant sur la localisation du génome à l'étape $n+1$, on obtient

$$\begin{aligned}\mathbb{E}_{s_0}[\log(S_{n+1})|A_n] =& \mathbb{E}_{s_0}[\log(S_{n+1})|A_{n+1}]\,\frac{p_{n+1}}{p_n} \\ &+ \mathbb{E}_{s_0}[\log(S_{n+1})|A_n \cap (S_{n+1} < \tilde{s})]\,\frac{\Pr_{s_0}[A_n \cap (S_{n+1} < \tilde{s})]}{p_n}\end{aligned}$$

Comme $p_n > 0$ et $\mathbb{E}_{s_0}[\log(S_{n+1})|A_n \cap S_{n+1} < \tilde{s}] \geq 0$ (log est ici une fonction positive, voir remarque 6.1),

$$e_{n+1}p_{n+1} = \mathbb{E}_{s_0}[\log(S_{n+1})|A_{n+1}]\,p_{n+1} \leq \mathbb{E}_{s_0}[\log(S_{n+1})|A_n]\,p_n \leq (e_n - \delta)p_n$$

La dernière inégalité vient de la propriété 6.10 (voir remarque 6.11), combinée avec la propriété de Markov. Par récurrence immédiate, on obtient

$$e_n p_n \leq e_0 p_0 - \delta\sum_{k=0}^{n-1} p_k = \log s_0 - \delta\sum_{k=0}^{n-1} p_k \leq \log s_0 - n\delta p_n$$

La dernière inégalité vient du fait que la suite $(p_n)_{n\in\mathbb{N}}$ est décroissante. Finalement

$$p_n \leq \frac{\log s_0}{e_n + n\delta} \leq \frac{\log s_0}{\log\tilde{s} + n\delta}, \tag{II.6}$$

ce qui termine la démonstration de la première étape. De plus, si on revient à l'inégalité impliquant la somme des $p_k$, nous obtenons une borne sur la somme partielle

$$\delta \sum_{k=0}^{n} p_k \leq \log s_0 - e_{n+1} p_{n+1} \leq \log s_0$$

2. Nous utilisons la définition 6.3 avec $s_{min} = \tilde{s}$. Soit $T$ le temps de premier retour vers $\tilde{s}$ dans $(X_S^{\triangleright}, P_S^{\triangleright})$. $\forall n \geq 0$, soit $B_n = (S_0 \geq \tilde{s}) \cap \bigcap_{1 \leq k \leq n} (S_k > \tilde{s}) \subset A_n$. Notons qu'alors $\Pr_{\tilde{s}}[B_0] = 1$ et que pour $k \geq 1$, $\Pr_{\tilde{s}}[T = k] = \Pr_{\tilde{s}}[B_{k-1}] - \Pr_{\tilde{s}}[B_k]$. D'après la propriété 6.5, $\forall s_0 \geq \tilde{s}$, $\forall n \geq 1$,

$$\Pr_{s_0}\left[\bigcap_{k=1}^{n} (S_k^{\triangleright} > \tilde{s})\right] = \Pr_{s_0}\left[\bigcap_{k=1}^{n} (S_k > \tilde{s})\right] = \Pr_{s_0}[B_n] \leq \Pr_{s_0}[A_n] = p_n \qquad \text{(II.7)}$$

donc $\Pr_{\tilde{s}}[T < +\infty] = 1 - \Pr_{\tilde{s}}\left[\bigcap_{k \geq 1} (S_k^{\triangleright} > \tilde{s})\right] \geq 1 - \lim_n p_n = 1$. Cela équivaut à dire que $\tilde{s}$ est un état récurrent dans $(X_S^{\triangleright}, P_S^{\triangleright})$. De plus

$$\mathbb{E}_{\tilde{s}}[T] = \sum_{k \geq 1} k \Pr_{\tilde{s}}[T = k] = \sum_{k \geq 1} k \left(\Pr_{\tilde{s}}[B_{k-1}] - \Pr_{\tilde{s}}[B_k]\right)$$

Nous réorganisons les termes de la série via les sommes partielles

$$\sum_{k=1}^{n} k \Pr_{\tilde{s}}[B_{k-1}] - \sum_{k=1}^{n} k \Pr_{\tilde{s}}[B_k] = \sum_{k=0}^{n-1} (k+1) \Pr_{\tilde{s}}[B_k] - \sum_{k=1}^{n} k \Pr_{\tilde{s}}[B_k]$$

$$= \sum_{k=0}^{n-1} \Pr_{\tilde{s}}[B_k] - n \Pr_{\tilde{s}}[B_n] \leq \sum_{k=0}^{n-1} p_k \leq \frac{\log \tilde{s}}{\delta}$$

Les dernières inégalités proviennent de (II.7) et de la conclusion de l'étape 1. D'après la somme partielle, on voit que $\mathbb{E}_{\tilde{s}}[T] < +\infty$. Cela montre que $\tilde{s}$ est un état récurrent positif. La chaîne $(X_S^{\triangleright}, P_S^{\triangleright})$ est donc irréductible, apériodique (propiété 6.4) et récurrente positive. Le théorème de convergence pour les chaînes récurrentes positives montre qu'il existe une unique distribution stationnaire asymptotique et que $\lim_n \Pr[S_n^{\triangleright} = \tilde{s}] = 1/\mathbb{E}_{\tilde{s}}[T] > 0$ (voir par exemple Woess, 2009). De plus, $\forall n \geq 0$, $\Pr_{\tilde{s}}[S_n^{\triangleright} = \tilde{s}] > 0$ puisque $(P_S^{\triangleright})_{\tilde{s}\tilde{s}} > 0$. Ces deux remarques combinées impliquent que l'ensemble $\{\Pr_{\tilde{s}}[S_n^{\triangleright} = \tilde{s}], n \in \mathbb{N}\}$ est inférieurement borné par un réel $\varepsilon' > 0$.

3. La propriété 6.6 nous permet d'étendre la conclusion de l'étape précédente à $(X_S, P_S)$

$$\forall s_0 \leq \tilde{s}, \forall n \in \mathbb{N}, \quad \Pr_{s_0}[S_n \leq \tilde{s}] \geq \Pr_{\tilde{s}}[S_n^{\triangleright} = \tilde{s}] \geq \varepsilon'$$

Ceci prouve l'inégalité (b)i. du lemme. Comme la chaîne est homogène en temps, cette relation peut être généralisée :

$$\forall s \leq \tilde{s}, \forall k \geq 0, \forall n \geq 0, \quad \Pr[S_{n+k} \leq \tilde{s} | S_k = s] \geq \varepsilon' \qquad \text{(II.8)}$$

Il faut noter qu'à l'étape précédente, quand nous avons appliqué la propriété 6.4, nous avons omis le cas $\mu_{dup} = \mu_{ins} = 0$. Cependant, dans ce cas, les inégalités données ici sont évidentes même en prenant $\varepsilon' = 1$.

4. Soit $s_0 > \tilde{s}$. Nous calculons $\Pr_{s_0}[S_n \leq \tilde{s}]$ en partitionnant par le temps de premier passage sous $\tilde{s}$,

$$\begin{aligned}\Pr_{s_0}[S_n \leq \tilde{s}] = &\sum_{k\geq 1} \Pr_{s_0}[S_n \leq \tilde{s} | S_k \leq \tilde{s}, S_{k-1} > \tilde{s}, ..., S_0 > \tilde{s}] \\ &\times \Pr_{s_0}[S_k \leq \tilde{s}, S_{k-1} > \tilde{s}, ..., S_0 > \tilde{s}]\end{aligned}$$

Tous les termes $k > n$ sont nuls. Si on reprend la définition des $B_k$ et la relation (II.8) (en appliquant la remarque 6.11) pour $n \geq 1$

$$\Pr_{s_0}[S_n \leq \tilde{s}] \geq \varepsilon' \sum_{k=1}^{n} (\Pr_{s_0}[B_{k-1}] - \Pr_{s_0}[B_k]) = \varepsilon'(1 - \Pr_{s_0}[B_n])$$

Nous appliquons enfin les relations (II.7) and (II.6)

$$\Pr_{s_0}[S_n \leq \tilde{s}] \geq \varepsilon'(1 - p_n) \geq \varepsilon' \left(1 - \frac{\log s_0}{\log \tilde{s} + n\delta}\right)$$

Ceci prouve l'inégalité (b)ii. du lemme. La relation pour $n = 0$ est triviale car le côté droit de l'inégalité est négatif.

□

## 6.2 Détails des preuves de la section 4

*Démonstration (de la propriété 4.4).* L'élément le plus important est le changement des indices de sommation quand on somme en pondérant par la distribution de Poisson, comme dans

$$\begin{aligned}&\sum_{n\geq 0} n \frac{((\mu_{ldel} + \mu_{dup})s_0)^n}{n!} e^{-(\mu_{ldel}+\mu_{dup})s_0} \\ &= (\mu_{ldel} + \mu_{dup}) s_0 e^{-(\mu_{ldel}+\mu_{dup})s_0} \sum_{n\geq 1} \frac{((\mu_{ldel} + \mu_{dup})s_0)^{n-1}}{(n-1)!} \\ &= (\mu_{ldel} + \mu_{dup}) s_0\end{aligned}$$

Dans la suite, nous définissons $\mathrm{P}\Sigma[.]$ comme l'opérateur qui correspond à la sommation pondérée par les termes de la loi de Poisson (il s'agit en fait de l'espérance par rapport à la loi Poisson). Par exemple, la relation ci-dessus s'écrira simplement $\mathrm{P}\Sigma[n] = (\mu_{ldel} + \mu_{dup})s_0$. Similairement, on peut montrer que $\mathrm{P}\Sigma[n(n-1)] = (\mu_{ldel} + \mu_{dup})^2 s_0^2$. Nous nous autoriserons d'échanger librement les opérateurs $\mathrm{P}\Sigma[.]$ et $\mathbb{E}[.]$, nous le justifierons à la fin de la preuve.

$$\begin{aligned}\mathbb{E}_{s_0}\left[\log \hat{S}_f\right] = \mathrm{P}\Sigma\left[\mathbb{E}_{s_0}\left[\log \hat{S}_n\right]\right] &= \log s_0 + \mathbb{E}[J_n]\,\mathrm{P}\Sigma[n] \\ &= \log s_0 + s_0((2\log 2 - 1)\mu_{dup} - \mu_{ldel}).\end{aligned}$$

Pour simplifier les calculs du deuxième moment, nous introduisons les constantes $A = \mu_{ldel} - (2\log 2 - 1)\mu_{dup}$ et $B = \sqrt{2(\mu_{ldel} + (1-\log 2)^2\mu_{dup})}$. $\mathrm{P\Sigma}\left[n\mathbb{E}\left[J_n\right]\right] = -As_0$ et $\mathrm{P\Sigma}\left[n\mathbb{E}\left[J_n^2\right]\right] = B^2 s_0$. La variance de $\hat{S}_f$ est donnée par la formule

$$\mathbb{E}_{s_0}\left[\log^2 \hat{S}_f\right] - \left(\mathbb{E}_{s_0}\left[\log \hat{S}_f\right]\right)^2 = \mathrm{P\Sigma}\left[\mathbb{E}_{s_0}\left[\log^2 \hat{S}_n\right]\right] - (\log s_0 - As_0)^2$$

où

$$\begin{aligned}\mathbb{E}_{s_0}\left[\log^2 \hat{S}_n\right] = \sigma^2_{s_0}\left[\log \hat{S}_n\right] + \left(\mathbb{E}_{s_0}\left[\log \hat{S}_n\right]\right)^2 &= n\sigma^2\left[J_n\right] + (\log s_0 + n\mathbb{E}\left[J_n\right])^2 \\ &= \log^2 s_0 + 2\log s_0 n\mathbb{E}\left[J_n\right] + n(n-1)(\mathbb{E}\left[J_n\right])^2 + n\mathbb{E}\left[J_n^2\right]\end{aligned}$$

Nous appliquons la sommation de Poisson et utilisons les différentes définitions et relations établies jusqu'ici

$$\mathrm{P\Sigma}\left[\mathbb{E}_{s_0}\left[\log^2 \hat{S}_n\right]\right] = \log^2 s_0 - 2\log s_0 As_0 + A^2 s_0^2 + B^2 s_0 = (\log s_0 - As_0)^2 + B^2 s_0$$

On en déduit enfin

$$\sigma_{s_0}\left[\log \hat{S}_f\right] = \sqrt{B^2 s_0} = \sqrt{s_0}\sqrt{2(\mu_{ldel} + (1-\log 2)^2\mu_{dup})}$$

Interchanger $\mathrm{P\Sigma}\left[.\right]$ et $\mathbb{E}\left[.\right]$ est légitime d'après le théorème de Fubini-Tonelli. Comme $\mathrm{P\Sigma}\left[\mathbb{E}_{s_0}\left[\log^2 \hat{S}_n\right]\right] < +\infty$, les conditions du théorème sont remplies pour le deuxième moment, et comme, asymptotiquement, $|x| < x^2$, elles le sont aussi pour le premier moment. □

*Démonstration.* (du lemme 4.5) La dérivation de $Q_k(s_0)$ par rapport à $s_0$ donne

$$Q'_k(s_0) = Q_k(s_0)\left(\frac{1}{s_0} - A + \frac{kB}{2\sqrt{s_0}}\right) = Q_k(s_0)\frac{2 - 2As_0 + kB\sqrt{s_0}}{2s_0}$$

Le signe de la dérivée est donc déterminé par $2 - 2As_0 + kB\sqrt{s_0}$ qui s'annule pour une unique valeur

$$\sqrt{s_0} = \frac{kB + \sqrt{k^2B^2 + 16A}}{4A}$$

(car $A > 0$). En mettant au carré, on obtient $s_0^{max}$. Comme le coefficient principal $-2A$ est négatif, la dérivée est positive en dessous de $s_0^{max}$ et négative au-dessus.

Pour obtenir $s_{fixed}$, on pose $Q_k(s_0) = s_0$, ce qui donne $\sqrt{s_0} = kB/A$. Comme $A > 0$, on retrouve la valeur $s_{fixed}$ annoncée. De plus

$$Q'_k(s_{fixed}) = \frac{Q_k(s_{fixed})}{2s_{fixed}}\left(2 - 2k^2\frac{B^2}{A} + k^2\frac{B^2}{A}\right) = \frac{1}{2}\left(2 - k^2\frac{B^2}{A}\right).$$

On peut réécrire le ratio $B^2/A$ comme

$$\frac{B^2}{A} = 2 \times \frac{(\log 2 - 1)^2 \mu_{dup} + \mu_{ldel}}{\mu_{ldel} - (2\log 2 - 1)\mu_{dup}} = 2 \times \frac{1 + (\log 2 - 1)^2 (\mu_{dup}/\mu_{ldel})}{1 - (2\log 2 - 1)(\mu_{dup}/\mu_{ldel})} \geq 2.$$

Pour la dernière inégalité, on peut montrer que la fraction vaut 2 pour $\mu_{dup}/\mu_{ldel} = 0$ et croît strictement vers l'infini quand $\mu_{dup}/\mu_{ldel}$ tend vers $1/(2\log 2 - 1)$. Nous avons supposé $k \geq 1$, on obtient donc $Q'_k(s_{fixed}) \leq 0$. Comme le sous-ensemble où la dérivée est négative est $\{s \geq s_0^{max}\}$, on a nécessairement $s_{fixed} \geq s_0^{max}$. □

# Deuxième partie

# Évolution de la taille du génome en présence de sélection

# Chapitre III

# Implémentation des simulations

> Il y a lieu de s'arrêter une minute car cela va devenir noué [...] Arthur Eddington a donné le moyen de récupérer tous les lions [que le désert] contient ; il suffit de tamiser le sable et les lions restent sur la toile. Ceci comporte une phase – la plus intéressante – la phase d'agitation. À la fin, on a bien tous les lions sur la toile du tamis.
>
> Boris Vian, *L'Automne à Pékin*

Dans ce chapitre, nous proposons une implémentation numérique du modèle que nous avons étudié précédemment.

## 1 Présentation du modèle

### 1.1 Présentation théorique

Le modèle implémenté numériquement est dérivé du modèle général présenté dans l'introduction (section 3). Les hypothèses sur les mécanismes de mutation sont les mêmes que dans la partie analytique, mais nous détaillons davantage la structure du génome, en distinguant des parties fonctionnelles – qui vont contribuer à la fitness des individus – et des parties non fonctionnelles – qui n'auront pas d'effet sur la fitness. Ce que nous appelons

« gène » dans le modèle correspondrait dans un génome réel, soit à un segment d'ADN nécessaire à l'expression d'une protéine, avec sa séquence codante mais aussi son promoteur, son terminateur, ses parties 5'UTR et 3'UTR[1], les sites de fixation des protéines qui régulent sa transcription, soit à un segment d'ADN transcrit non traduit (non-coding RNA) ayant un rôle fonctionnel et donc un impact sur la fitness. Nous appellerons l'ensemble des « gènes » la partie « codante » du génome, ce qui constitue en toute rigueur un abus de langage, puisque ce terme désigne ici l'ensemble des segments d'ADN ayant un impact sur la fitness de l'organisme, qu'ils codent ou non pour une protéine.

**Structure des génomes** Nous considérons toujours une population infinie de génomes distribués sur un espace $X$, sauf que nous allons donner une structure à $X$ de manière à autoriser les variations de codant et de non-codant tout en maintenant un nombre de dimensions faible. Nous supposons que les génomes ont les propriétés suivantes :

- Un génome est représenté par le couple $(L, n) \in \mathbb{N}^2$, où $L$ est le nombre de bases non-codantes et $n$ est le nombre de gènes.
- Un génome est composé d'un seul chromosome circulaire. Cette propriété rapproche le modèle des structures bactériennes mais permet surtout d'éviter les effets de bord pour la longueur des réarrangement chromosomiques.
- Tous les gènes ont la même longueur $l_{gene}$.
- Les régions codantes sont régulièrement disposées le long du génome, c'est-à-dire que toutes les régions intergéniques ont la même longueur $l_{intergenic} = L/n$.

La taille totale du génome est donc $s = L + nl_{gene}$ et le pourcentage de codant $r = nl_{gene}/s$. Les hypothèses ci-dessus sont bien sûr extrêmement simplificatrices, mais nous verrons dans le chapitre suivant que ce modèle minimal a déjà une dynamique non triviale, qu'il convient de bien comprendre avant d'envisager de complexifier la représentation de la structure génomique. D'après la description, il est assez facile de voir que l'espace des génomes $X$ est en bijection avec $\mathbb{N}^2$. Cet espace est dénombrable, c'est-à-dire qu'il est également en bijection avec $\mathbb{N}$. Autrement dit, on peut lister tous les états de $X$ en leur donnant un identifiant entier unique. Dans le cas de l'espace à 2 dimensions $\mathbb{N}^2$, il s'agit d'un exercice classique. On peut commencer par l'état $(0, 0)$ qui porte l'identifiant 0, puis on s'éloigne en parcourant toutes les antidiagonales une à une. $(1, 0)$ porte l'identifiant 1, $(0, 1)$ le 2, $(2, 0)$ le 3, $(1, 1)$ le 4, $(0, 2)$ le 5, etc. (figure III.1). Cette remarque sera importante dans la suite car elle justifie le formalisme matriciel : même si l'espace initial porte naturellement 2 dimensions, tous ses états peuvent en pratique être listés dans un vecteur à une seule dimension. Ce raisonnement peut se généraliser pour un espace en bijection avec $\mathbb{N}^n$, où $n$ peut être arbitrairement grand, ce qui permettrait d'avoir une description du génome un peu moins caricaturale (pourvu que $X$ soit dénombrable).

[1] ADN transcrit non traduit, correspondant aux parties d'un ARN messager situées avant le codon START ou après le codon STOP.

Cependant, comme on va le voir dans ce chapitre, l'implémentation du système avec deux dimensions est déjà un peu délicate.

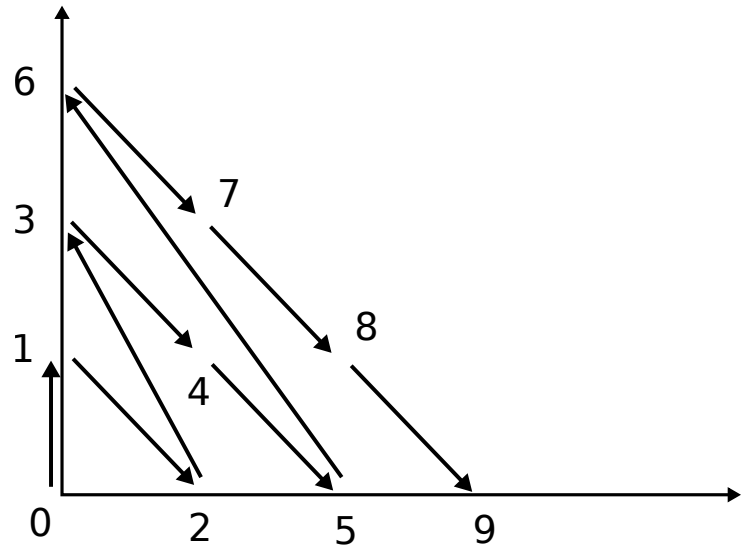


Figure III.1 – Exercice classique : comment attribuer un identifiant entier unique à tous les éléments de $\mathbb{N}^2$ ? Il s'agit de se convaincre qu'on ne peut pas le faire ligne par ligne ou colonne par colonne. La solution courante préconise les antidiagonales mais il existe une infinité d'autres possibilités.

Pour distinguer les notations du chapitre précédent et mettre en avant la structure particulière étudiée dans ce chapitre, nous introduisons la définition suivante, qui résume également le paragraphe précédent.

**Définition 1.1.** $X_{(L,n)}$ est l'espace de tous les états génomiques. Comme $X_{(L,n)} \simeq \mathbb{N}^2$ est dénombrable, nous pouvons associer un identifiant unique $i \in \mathbb{N}$ à chaque état $(L, n) \in X_{(L,n)}$ et définir une bijection $\varphi : X_{(L,n)} \to \mathbb{N}$ telle que $i = \varphi(L, n)$.

**Processus de mutation** Les mutations considérées sont les mêmes que dans la première partie, sauf qu'on va préciser des distributions pour les indels. On va considérer que les indels correspondent à des sauts de 1 à 6 paires de bases effectués par la polymérase qui réplique l'ADN. Il faut également préciser l'impact de chaque mutation sur la structure en codant et non-codant des génomes.

Les petites insertions et délétions (indels) sont définies comme suit :

- Les petites insertions ajoutent 1 à 6 paires de bases non-codantes (le nombre est tiré dans une distribution uniforme) sur une position aléatoire tirée uniformément le long du génome. Si l'insertion a lieu dans un gène, la région codante devient non-codante, un processus connu sous le nom de pseudogénisation. On néglige la probabilité que le gène continue à produire des protéines fonctionnelles.

- Les petites délétions retirent 1 à 6 paires de bases (le nombre est tiré dans une distribution uniforme) d'une position aléatoire tirée uniformément le long du génome. De même que pour les petits insertions, si la perte a lieu dans une région codante, le gène devient non-codant.

Nous prenons en compte les quatre types de réarrangement suivants :

- Pour chaque inversion, on tire deux positions aléatoires uniformément le long du génome, avec au moins une base entre les deux. Si un gène se trouve au niveau de l'une ou l'autre des positions, il est transformé en pseudogène, c'est-à-dire en non codant. L'inversion elle-même n'est pas effectuée, puisque la représentation des états génomiques ne contient pas l'ordre ni l'orientation des gènes. La taille totale du génome n'est pas affectée.
- Pour les translocations, on a un effet similaire aux inversions avec trois positions tirées uniformément le long du génome. Les deux premières donnent en théorie les points de cassure qui délimitent le segment du génome qui est extrait, la troisième indique où il est réinséré. En pratique, on s'intéresse uniquement à l'inactivation potentielle de gènes qui peut se produire à chacune de ces 3 positions.
- Pour les grandes délétions, on tire une position uniformément le long du génome ainsi qu'une longueur de délétion. Le segment délété a une longueur tirée uniformément entre 1 paire de bases et la totalité du génome. Si un gène est partiellement supprimé, la partie restante devient non codante.
- Pour les duplications, on tire une position uniformément le long du génome et une longueur de duplication. Le segment dupliqué a une taille tirée uniformément entre 1 paire de bases et la totalité du génome. On tire ensuite une position uniformément le long du génome pour insérer le segment dupliqué. Des gènes partiellement dupliqués deviennent des bases non codantes sur le segment dupliqué. Si le segment est inséré dans un gène, ce gène devient non codant.

Les mutations suivent à nouveau des processus de Poisson indépendants de taux $\mu_{ins}$, $\mu_{sdel}$, $\mu_{inv}$, $\mu_{trans}$, $\mu_{ldel}$ et $\mu_{dup}$ par paire de bases et par génération. Le taux de mutation total est $\mu = \mu_{ins} + \mu_{sdel} + \mu_{inv} + \mu_{trans} + \mu_{ldel} + \mu_{dup}$. Le nombre total de mutations par génération est donné par une loi de Poisson de paramètre $\mu s_0$, où $s_0$ est la taille initiale du génome considéré. Par l'indépendance des processus de Poisson, quand on connaît le nombre de mutations, on sait qu'une mutation donnée est du type petite insertion (par exemple) avec probabilité $\mu_{ins}/\mu$.

Nous déduisons de ces processus deux matrices de transition. La première décrit l'action d'une seule mutation.

**Définition 1.2.** $P_{(L,n)} = ((P_{(L,n)})_{ij})_{i,j \in X_{(L,n)}}$ représente l'action des petits indels et des réarrangements, sachant qu'exactement une mutation s'est produite. $(P_{(L,n)})_{ij}$ est la probabilité qu'un génome initialement situé dans l'état d'identifiant $i$ finisse dans l'état $j$

après *une mutation*. $P_{(L,n)}$ est une matrice stochastique. Les taux de transition de $i$ vers $j$ sont calculés d'après les définitions ci-dessus (voir section 2 pour le détail des probabilités).

Comme dans le chapitre précédent, $P_{(L,n)}$ dépend à la fois de la nature des mutations, mais aussi de leur taux respectifs. La seconde matrice donne les transitions pour une génération, quand les mutations ont été tirées selon les processus de Poisson indépendants.

**Définition 1.3.** $M_{(L,n)} = ((M_{(L,n)})_{ij})_{i,j\in X_{(L,n)}}$ représente l'action des petits indels et des réarrangements dans l'espace d'une génération. $(M_{(L,n)})_{ij}$ est la probabilité qu'un génome initialement situé dans l'état d'identifiant $i$ finisse dans l'état $j$ après *une génération*. $M_{(L,n)}$ est une matrice stochastique. Le calcul des transitions se fait à partir de $P_{(L,n)}$ et sera détaillé un peu plus bas (sous-section 1.2).

### Action de la sélection

**Définition 1.4.** Pour chaque état, nous définissons une fitness $f_i \geq 0$, où $i$ est l'identifiant de l'état considéré. Nous supposons que la fitness ne dépend que du nombre de gènes, il n'y a pas de coût pour les bases non codantes. Si on a $i, j$ tels que $i = \varphi(L, n)$ et $j = \varphi(L', n)$, alors $f_i = f_j$. Nous définissons une matrice de fitness $F = \mathrm{diag}((f_i)_{i\in\mathbb{N}})$.

Là encore, nous avons sciemment fait une hypothèse très simplificatrice, de sorte à obtenir un scénario qui semble facile à prédire par une simple « expérience de pensée ». Nous verrons pourtant que même avec une fonction de fitness aussi triviale, la dynamique du système, elle, ne l'est pas.

### Équation du modèle

**Définition 1.5.** $(n_t)_i$ est la densité de génomes dans l'état $i$ à la génération $t \in \mathbb{N}$. Le vecteur $n_t$ contient la densité pour tous les états génomiques. Pour chaque génération $t$, $\|n_t\| = \sum_{i\in\mathbb{N}}(n_t)_i = 1$.

Connaissant le vecteur de densité à la génération $t$, le vecteur de densité à la génération $t + 1$ est donné par

$$n_{t+1} = \frac{n_t F M_{(L,n)}}{\|n_t F\|} \tag{III.1}$$

À chaque génération, les génomes sont sélectionnés d'après leur fitness relative $F/\|n_t F\|$ puis mutés. On peut vérifier que $\|n_{t+1}\| = \|n_t\| = 1$, en accord avec la définition du vecteur de densité $n_t$.

On peut dériver le cas particulier où il n'y a pas de sélection ($F = I$) :

$$n_{t+1} = n_t M_{(L,n)} \tag{III.2}$$

On retrouve la description du chapitre précédent (équation (II.2), 68) sauf qu'ici, la matrice de transition est connue et s'applique au génomes structurés dans $X_{(L,n)}$ et non pas seulement dans l'espace des tailles $X_S$.

## 1.2 Lien entre $M_{(L,n)}$, $P_{(L,n)}$ et les lois de mutations du modèle

$M_{(L,n)}$ a été définie de façon formelle comme la matrice de mutation contenant toutes les transitions entre tous les couples d'états de $X_{(L,n)}$ d'après les contraintes imposées pour les mutations et les réarrangements dans notre système. Cependant, le calcul des transitions ne peut pas se faire de manière immédiate. Nous allons montrer ici comment on peut passer des lois de mutations individuelles à la matrice générationnelle, en passant par la matrice mutationnelle $P_{(L,n)}$.

Rappelons que les occurrences des mutations sont indépendantes et que le nombre total de mutations est donnée par une loi de Poisson de paramètre $s_0\mu$, où $s_0$ est la taille du génome au début de la génération pour l'état considéré et $\mu$ le taux total de mutation par base et par génération. Pour chaque type de mutation, on peut définir une matrice de transition correspondant à une mutation : $P^{ins}_{(L,n)}$ pour les petites insertions, $P^{sdel}_{(L,n)}$ pour les petites délétions, $P^{inv}_{(L,n)}$ pour les inversions, $P^{trans}_{(L,n)}$ pour les translocations, $P^{ldel}_{(L,n)}$ pour les délétions, $P^{dup}_{(L,n)}$ pour les duplications. Les coefficients de ces matrices sont obtenus directement à partir des définitions des processus de mutation. Nous reviendrons sur leur implémentation plus tard (section 2). En utilisant les propriétés des processus de Poisson indépendants on peut écrire

$$P_{(L,n)} = \frac{\mu_{ins}P^{ins}_{(L,n)} + \mu_{sdel}P^{sdel}_{(L,n)} + \mu_{inv}P^{inv}_{(L,n)} + \mu_{trans}P^{trans}_{(L,n)} + \mu_{ldel}P^{ldel}_{(L,n)} + \mu_{dup}P^{dup}_{(L,n)}}{\mu} \tag{III.3}$$

Il s'agit d'une application du théorème des probabilités totales, où l'on conditionne sur chaque type de mutation. $P_{(L,n)}$ s'obtient donc assez directement à partir des définitions : on sait ce qui se passe après une mutation. Pour accéder à $M_{(L,n)}$, il faut prendre en compte le fait qu'il peut y avoir plusieurs (ou aucune) mutations par génération en pondérant par les bonnes probabilités selon la taille initiale du génome. Intuitivement, il s'agit à nouveau de partitionner l'espace des possibles selon le nombre de mutations. On va donc appliquer le théorème des probabilités totales avec une partition qui contient une infinité d'éléments (le nombre de mutations). Comme les matrices sont elles-mêmes infinies, nous devons justifier que cette opération ne pose pas de problème.

Les matrices sont stochastiques : les coefficients sont compris entre 0 et 1, quand on somme les coefficients sur chaque ligne, on trouve 1. Dans ces conditions, on peut montrer que même si les matrices sont infinies, le produit matriciel entre deux matrices stochastiques est bien défini et associatif (en cas de produits multiples, le résultat ne dépend pas de l'ordre des opérations). En effet, chaque coefficient du produit est une somme infinie mais le résultat ne peut pas dépasser 1 et la convergence est absolue vu que les coefficients sont positifs. Plus techniquement, les matrices appartiennent à un sous-ensemble des matrices de Kojima. Les matrices de Kojima sont les matrices infinies pour lesquelles l'ensemble des sommes des valeurs absolues des coefficients de chaque ligne est borné supérieurement. La borne supérieure est une norme appelée norme de Kojima ou norme ligne. Les matrices de Kojima avec la norme de Kojima forment un espace de Banach (un espace vectoriel normé complet). Dans un espace de Banach, pour qu'une somme infinie converge, il suffit qu'elle converge en norme. L'idée est la suivante : grâce aux propriétés d'espace vectoriel, la sommation est bien définie, le fait que la norme converge indique qu'on converge vers un objet qui a du sens et, comme l'espace est complet, on garantit qu'il est bien compris dans l'espace considéré initialement. Ainsi, en assemblant les deux propriétés, on garantit l'existence des puissances de nos matrices infinies et l'existence de l'exponentielle, qui est simplement une somme pondérée de ces puissances qui ont toutes pour norme 1.

D'un point de vue biologique, $P^k_{(L,n)}$ contient les probabilités de transition d'état à état sachant qu'exactement $k$ mutations ont eu lieu. En effet, les occurrences des mutations sont indépendantes entre elles, le calcul en terme de probabilités se fait donc assez facilement par récurrence. On va ici faire le raisonnement intuitivement. Notons pour commencer que $P_{(L,n)}$ contient les transitions pour tous les points de départ. Partons d'un état $i_0$ et appliquons la première mutation. Pour connaître les transitions possibles, on va voir dans $P_{(L,n)}$ la ligne qui correspond à $i_0$ : chaque colonne correspond à un état d'arrivée et le coefficient donne la probabilité de la transition. Si on veut calculer la probabilité d'être en $i_2$ après 2 mutations, il suffit d'additionner les probabilités associées à tous les chemins possibles. Quand on effectue la multiplication matricielle, c'est bien ce qu'on fait, à condition toutefois que les occurrences des mutations soient indépendantes (figure III.2). En multipliant une nouvelle fois, on ajoute une tranche d'intermédiaires : on prend en compte tous les chemins à 3 mutations, et ainsi de suite.

Soit $(L_0, n_0) \in X_{(L,n)}$ et $i_0 = \varphi(L_0, n_0)$ son identifiant. On nomme la taille du génome correspondant à cet état $s_0 = L_0 + n_0 l_{gene}$. On appelle $\mathbf{1}_{i_0}$ le vecteur ligne infini qui contient le coefficient 1 sur la colonne $i_0$ et 0 partout ailleurs. Pour notre système, il indique que le génome se trouve sur l'état $(L_0, n_0)$ avec probabilité 1. D'après ce que nous avons vu, les vecteurs $\mathbf{1}_{i_0} P_{(L,n)}$, $\mathbf{1}_{i_0} P^k_{(L,n)}$, $\mathbf{1}_{i_0} M_{(L,n)}$ contiennent les probabilités de transition partant de $i_0$ vers tous les autres états en (respectivement) exactement une mutation, $k$ mutations et une génération. Si on applique le théorème des probabilités totales en partitionnant sur le nombre de mutations réalisées par le processus de Poisson, on obtient

$$\mathbf{1}_{i_0} M_{(L,n)} = \sum_{k=0}^{+\infty} \frac{e^{-\mu s_0} (\mu s_0)^k}{k!} \mathbf{1}_{i_0} P^k_{(L,n)} = e^{-\mu s_0} \mathbf{1}_{i_0} e^{\mu s_0 P_{(L,n)}} = \mathbf{1}_{i_0} e^{\mu s_0 (P_{(L,n)} - I)}$$

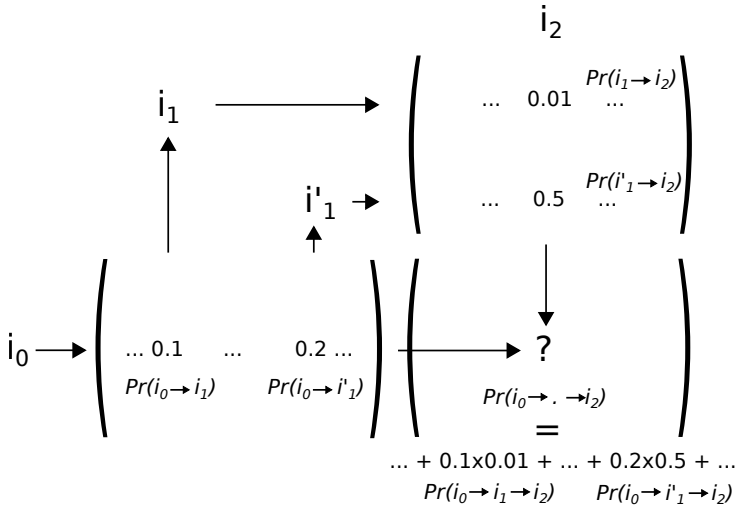


Figure III.2 – Quand on multiplie la matrice de mutation par elle-même, on explore toutes les combinaisons de mutations possibles. Ici on illustre le cas de 2 mutations : la multiplication combine les mutations de $i_0$ vers $i_2$ via tous les états intermédiaires possibles. Par récurrence, on peut continuer à multiplier et à explorer tous les cheminements correspondant à 3, 4, etc. mutations.

L'expression dépend de la taille initiale $s_0$. On ne peut donc pas définir $M_{(L,n)}$ simplement à partir de $P_{(L,n)}$. Cependant, la formule ci-dessus donne un moyen de définir $M_{(L,n)}$ par blocs : on extrait de $e^{\mu s_0(P_{(L,n)}-I)}$ toutes les lignes qui correspondent à des états qui ont effectivement la taille initiale $s_0$. On peut d'ailleurs noter qu'il y en a exactement $\lfloor s_0/l_{gene} \rfloor + 1$, selon la répartition en codant et en non-codant (le nombre de gène maximal est $\lfloor s_0/l_{gene} \rfloor$ et il faut compter le génome où il n'y a aucun gène). Plus $s_0$ est grand, plus on peut extraire de lignes à la fois.

## 1.3 Problèmes liés à l'implémentation

Nous avons défini jusqu'ici un modèle général avec des génomes qui évoluent dans un espace infini dénombrable à 2 dimensions : la longueur de non codant $L$ et le nombre de gènes $n$. Théoriquement, nous avons vu que toutes les grandeurs sont bien définies. Cependant, pour implémenter le modèle, la taille infinie des matrices pose un peu problème. En effet, $M_{(L,n)}$ étant obtenue à partir d'une somme infinie, il est difficile de calculer ses coefficients analytiquement pour les simulations. On va donc se contenter d'approximations en suivant trois étapes qui correspondent au cheminement donné ci-dessus pour aller des définitions des mutations vers $M_{(L,n)}$ en passant par $P_{(L,n)}$.

**Étape 1 : calcul des matrices atomiques** Nous avons vu que $P_{(L,n)}$ peut être subdivisée en une somme pondérée de matrices qui correspondent à chaque type de mutation (indels, inversion, translocation, délétion ou duplication). Il faut spécifier toutes les transitions possibles lorsqu'exactement un événement se produit. Cette étape est assez facile,

même si les choses se compliquent un peu lorsqu'on arrive aux délétions et aux duplications. Les calculs sont détaillés dans la section 2.

**Étape 2 : agrégations des transitions contenues dans les matrices atomiques** Le nombre de transitions atomiques est infini. Même si on se contente de simuler les génomes en dessous d'une certaine taille, le nombre de transitions croît très rapidement avec le nombre d'états considérés et on est rapidement bloqués par le nombre de possibilités. Par exemple, si on prend en compte les individus possédant jusqu'à 10 000 gènes et 1 Mb de non codant (ordres de grandeur pour une bactérie), il y a déjà $10^{10}$ états génomiques possibles. À cause des processus de Poisson, toutes les transitions d'état à état ont une probabilité non nulle (même si elle peut être infime). Cela fait donc $(10^{10})^2 = 10^{20}$ transitions possibles. Les probabilités sont enregistrées au format flottant de 8 octets. La matrice correspondant à ces transitions ferait donc $8.10^{11}$ Go. Actuellement, la mémoire vive d'un ordinateur fait au mieux quelques dizaines de Go, on est donc très loin du compte. Pour être plus proche des contraintes actuelles, on se limitera plutôt à $10^4$ états génomiques, qui donnent des matrices de transitions de 800 Mo, simulables sur un ordinateur classique.

Dans la section 3, nous allons donc agréger les états en des classes de génomes qui ont, autant que possible, des structures assez similaires. Ensuite, nous allons calculer des probabilités de transition pour ces méta-états en calculant les moyennes des transitions atomiques obtenues à la section 3. Nous étudierons deux types de découpages, en échelle linéaire et en échelle logarithmique. Dans les deux cas, l'agrégation naïve des transitions est beaucoup trop longue pour des paramètres réalistes, il va donc falloir les agréger analytiquement et recourir à des approximations supplémentaires. Cette étape est donc beaucoup plus sensible et technique que la précédente.

**Étape 3 : exponentiation approximative de $P_{(L,n)}$** Nous allons voir que $M_{(L,n)}$ est définie par blocs à partir d'exponentielles de $P_{(L,n)}$. Bien que $P_{(L,n)}$ soit une matrice creuse, $M_{(L,n)}$ ne l'est pas à cause des processus de Poisson : toutes les transitions sont possibles avec une probabilité non nulle. On ne pourra donc pas utiliser d'outils d'optimisations spécifiques aux matrices creuses. Le plus simple pour calculer l'exponentielle est de revenir au développement en série, qui permet un calcul efficace et rapide. Le défi pour ce genre de calculs est la stabilité numérique et la condition d'arrêt pour la sommation de la série. Nous présenterons l'algorithme retenu en section 4.

# 2 Calcul des transitions atomiques

Il s'agit dans cette section de déterminer quelles transitions sont contenues dans $P_{(L,n)}$. On a déjà vu que la matrice peut se décomposer selon chaque type de mutation selon l'équation (III.3). Nous allons donc étudier chaque type de mutation un à un, indépendamment du

taux. Lors des simulations, il suffit de charger les matrices correspondant à chaque type de mutations et de modifier les coefficients selon les taux utilisés, il n'est pas utile de calculer les transitions correspondant à une mutation à chaque fois.

Dans la suite, nous allons étudier les transitions issues du génome de départ $(L, n)$, où $L$ est la longueur totale du non codant et $n$ le nombre de gènes (dont la longueur est $l_{gene}$). Soit $i = \varphi(L, n)$ son identifiant entier. Les transitions calculées nous permettront donc de remplir la $i$-eme ligne de la matrice. En faisant varier $L$ et $n$ dans le génome de départ, on obtient bien toute la matrice. L'intérêt de cette section n'est pas l'impact des mutations sur la taille, que l'on connaît assez bien maintenant, mais l'impact sur la structure en codant et non codant, qui n'est pas si évident à déterminer pour tous les types de mutation.

## 2.1 Petites insertions et petites délétions

Le cas des petits indels est assez simple à décrire d'après leur définition. Ce qui importe le plus est le fait que l'indel tombe dans une zone codante ou non. S'il tombe dans une zone non codante, on dira qu'il est neutre, s'il tombe dans une région codante, un gène est perdu et transformé en non codant.

Le petit indel est neutre avec probabilité $p_{neutral} = L/(L+nl_{gene}))$. Dans ce cas, on ajoute ou on supprime 1 à 6 bases de non codant avec probabilité 1/6. En combinant les deux, on obtient alors les transitions

- Petites insertions : $\forall i \in \{1, ..., 6\}$, $\Pr[(L, n) \to (L + i, n)] = p_{neutral}/6$.
- Petites délétions :
  - $L = 0$ : la seule transition neutre dans ce cas est $\Pr[(0, 0) \to (0, 0)] = 1$.
  - $L \in \{1, ..., 5\}$ : $\forall i \in \{0, ..., L - 1\}$, $\Pr[(L, n) \to (i, n)] = p_{neutral}/L$.
  - $L \geq 6$ : $\forall i \in \{1, ..., 6\}$, $\Pr[(L, n) \to (L - i, n)] = p_{neutral}/6$.

Dans le cas où la mutation est une petite délétion et que $L < 6$, on répartit les probabilités uniformément entre les états disponibles. Ces probabilités ne sont pas exactes car elles ne prennent pas en compte les effets de bord pour les délétions. En théorie, la délétion pourrait commencer dans le non codant et déborder sur un gène, surtout si la taille de l'intergénique est très faible. D'ailleurs, la possibilité de déléter les dernières bases de non codant restantes alors qu'elles sont censées être réparties régulièrement le long du génome est incohérente avec la description de la structure. Nous avons choisi ici d'ignorer ces effets de bord (avec l'agrégation et les approximations qui vont suivre, leur effet est très marginal). Cela a tout de même un avantage : il est plus simple pour les individus d'éliminer tout leur non-codant s'ils le souhaitent.

Si l'indel n'est pas neutre, un gène est inactivé. On obtient alors les transitions suivantes

- Petites insertions : $\forall i \in \{1, ..., 6\}$, $\Pr[(L, n) \to (L + l_{gene} + i, n - 1)] = (1 - p_{neutral})/6$.
- Petites délétions : $\forall i \in \{1, ..., 6\}$, $\Pr[(L, n) \to (L + l_{gene} - i, n - 1)] = (1 - p_{neutral})/6$.

## 2.2 Inversions et translocations

Pour ces réarrangements, la taille du génome est constante, tout ce qui nous intéresse est la variation de codant et de non codant du fait de l'inactivation de gènes en pseudogènes. Comme pour les petits indels, il faut voir si les points des cassures et le point d'insertion de la translocation sont dans une région codante. La probabilité que l'un des points de cassure tombe dans un gène est $p = nl_{gene}/(L + nl_{gene})$. En théorie, il est possible que les deux points soient dans la même région codante mais, comme un peu plus haut, nous allons ignorer ce cas car il est assez marginal comparé aux approximations qui vont suivre. De même, la probabilité que le point d'insertion soit dans une région codante dépend du segment excisé mais en pratique elle est proche de $p = nl_{gene}/(L + nl_{gene})$. On supposera pour simplifier que le nombre de gènes perdus suit une distribution binomiale $\mathcal{B}(2, p)$ pour les inversions et $\mathcal{B}(3, p)$ pour les translocations.

Pour les inversions, on a donc

- Inactivation d'aucun gène : $\Pr[(L, n) \to (L, n)] = (1 - p)^2$.
- Inactivation d'un gène : $\Pr[(L, n) \to (L + l_{gene}, n - 1)] = 2p(1 - p)$.
- Inactivation de 2 gènes : $\Pr[(L, n) \to (L + 2l_{gene}, n - 2)] = p^2$.
- Exception : $\Pr[(L, 1) \to (L + l_{gene}, 0)] = 1 - (1 - p)^2$.

Pour les translocations

- Inactivation d'aucun gène : $\Pr[(L, n) \to (L, n)] = (1 - p)^3$.
- Inactivation d'un gène : $\Pr[(L, n) \to (L + l_{gene}, n - 1)] = 3p(1 - p)^2$.
- Inactivation de 2 gènes : $\Pr[(L, n) \to (L + 2l_{gene}, n - 2)] = 3p^2(1 - p)$.
- Inactivation de 3 gènes : $\Pr[(L, n) \to (L + 3l_{gene}, n - 3)] = p^3$.
- Exception : $\Pr[(L, 1) \to (L + l_{gene}, 0)] = 1 - (1 - p)^3$.
- Exception : $\Pr[(L, 2) \to (L + 2l_{gene}, 0)] = 1 - (1 - p)^2((1 - p) + 3p)$.

## 2.3 Grandes délétions et duplications

C'est le cas le plus compliqué, il nécessite quelques calculs. Avant de commencer, regardons l'impact de ces mutations sur le ratio de codant. Dès que la partie délétée ou copiée est assez grande, on copie une succession de gènes et de zones intergéniques dont le ratio est en première approximation proche de celui du génome global. En d'autres termes, l'impact des délétions et des duplications se ressent surtout sur la taille totale mais peu sur le ratio de codant. On pourrait donc se contenter de diminuer la taille du génome en gardant le ratio de codant le plus constant possible. Nous allons quand même effectuer les calculs pour justifier ce raisonnement intuitif et connaître les probabilités exactes.

On commence par chercher les états qu'il est possible d'atteindre pour le génome de départ $(L_0, n_0)$. Le génome est composé de gènes de longueur fixe $l_{gene}$ et de séquences intergéniques de longueur $l_{intergenic} = L_0/n_0$ qui sont disposées alternativement le long du génome. Lors d'une délétion, on ne peut donc pas supprimer plus de $l_{intergenic}$ bases non-codantes sans perdre de gènes. On peut donc se repérer selon le nombre de gènes qui ont été gagnés ou perdus. Si aucun gène n'est perdu, on perd au plus $l_{intergenic}$ bases. Si un gène est perdu, cela peut être dû au fait qu'une seule des bases du gène à été supprimée : on inactive le gène mais on gagne $l_{gene} - 1$ bases non codantes. À l'autre extrême, la délétion a pu toucher tout le gène et les 2 séquences intergéniques adjacentes. Si on perd $k \geq 2$ gènes, on perd au moins les intergéniques situés entre les deux, soit $(k-1)l_{intergenic}$ bases non codantes. Ensuite, si on perd juste une base de chaque gène aux extrémités, on regagne $2l_{gene}-2$ bases non-codantes, alors que si on supprime les gènes et les intergéniques adjacents, on perd $2l_{intergenic}$ bases non codantes supplémentaires. Le nombre de bases non codantes restantes est donc compris entre $L_0 - (k-1)l_{intergenic} + 2l_{gene} - 2$ et $L_0 - (k+1)l_{intergenic}$. On peut appliquer le même raisonnement aux duplications. Il reste maintenant à déterminer les probabilités de transition, qui ne sont pas uniformes.

Pour cela, nous allons commencer par nous intéresser aux grandes délétions et distinguer les cas selon la position des deux points qui délimitent la délétion. Le premier point de cassure sera noté BP1 et le second point BP2. Le segment délété est celui qui va de BP1 vers BP2, dans le sens horaire. On suppose que les deux points sont compris dans la délétion, il peuvent coïncider si une seule base est perdue.

Nous allons calculer les probabilités de transitions en regardant le nombre de combinaisons qui permettent d'aboutir à un état $(L_f, n_f)$ donné. Pour calculer le nombre de combinaisons qui permet d'aboutir à chaque état, il faut définir une feuille de route pour ne pas en manquer. Nous allons

1. Regarder toutes les transitions qui aboutissent à une ligne d'arrivée précise correspondant à la perte de $0 \leq k \leq n_0$ gènes. On regarde donc les transitions $(L_0, n_0) \to (., n_0 - k)$, où les valeurs de non codant restent à préciser.
2. Sur la ligne choisie, nous délimitons les valeurs extrêmes de non codant qui peuvent être atteintes, ainsi que le nombre de combinaisons qui permettent d'atteindre ces valeurs.

3. Nous calculons les combinaisons aboutissant à des valeurs de non codant intermédiaires.

Une fois que cette feuille de route aura été appliquée, on aura normalement associé un nombre de combinaisons de points de cassure menant à chaque état d'arrivée possible. Il suffira alors de diviser le nombre de combinaisons total pour obtenir les probabilités sous la forme $\Pr[(L_0, n_0) \to (L_f, n_f)]$.

On peut commencer par déterminer le nombre de combinaisons de points de cassure pour s'assurer qu'on n'aura pas oublié d'en compter à la fin. BP1 peut être sur n'importe quelle base du génome, soit $L_0 + n_0 l_{gene}$ positions. À chaque position de BP1 correspondent $L_0 + n_0 l_{gene}$ positions de BP2, qui peut être n'importe où sur le génome. Cela fait un total de $(L_0 + n_0 l_{gene})^2$ points de cassure. Nous allons éliminer les positions liées à la symétrie d'ordre $n_0$ du génome. Nous avons supposé que les gènes étaient régulièrement espacés le long du génome, donc que tous les régions intergéniques ont la même taille. Prenons une combinaison quelconque de BP1 et BP2. À cette combinaison est associée une délétion qui contient un certain nombre de bases codantes et non codantes. Si on décale cette combinaison de $l_{gene} + l_{intergenic}$, ce qui correspond à une rotation « jusqu'au prochain gène (ou région intergénique) », on se retrouve avec exactement la même délétion en termes de bases condantes et non codantes délétées. S'il y a $n_0$ gènes initialement, on peut trouver $n_0$ telles positions en décalant de gène en gène. Dans la suite, nous éviterons de compter ces positions : nous supposerons par exemple que BP1 tombe dans un gène précis et ne prendrons pas en compte le fait qu'il aurait aussi bien pu affecter les $n_0 - 1$ gènes restants. En faisant cela, nous divisons le nombre de combinaisons par $n_0$, ce qui réduit le total à $(L_0 + n_0 l_{gene})^2 / n_0$.

**Cas général : le nombre de gènes perdus $k$ est compris entre 2 et $n_0 - 1$** Si $2 \leq k \leq n_0 - 1$, la situation est simple car les points de cassure BP1 et BP2 ne peuvent pas se chevaucher. Le plus simple est ici de faire un dessin : BP1 et BP2 tombent nécessairement dans des régions intergéniques ou dans des gènes qui sont séparés par $k-2$ gènes et $k-1$ intergéniques sur le segment délété. Dans le segment conservé, il reste au moins un gène, ce qui garantit que BP1 et BP2 ne se croisent pas dans l'autre sens non plus. On peut donc facilement déterminer les positions possibles de BP1 et BP2 qui causent la perte (par délétion totale ou partielle) de $k$ gènes (figure III.3A). En plus de $k-2$ gènes et $k-1$ intergéniques de toute façon délétés, il faut déléter au moins une base des gènes de chaque côté. Il y a donc $l_{intergenic} + l_{gene}$ positions pour BP1 et autant pour BP2 de l'autre côté, en comptant toutes les positions de délétions dans le gène, puis celles dans l'intergénique qui suit.

Nous connaissons déjà le nombre de gènes perdus, nous nous intéressons ici au non codant perdu à cause de la délétion. Il y a trois configurations importantes. Au maximum, on supprime $(k+1) l_{intergenic}$ bases non codantes, si BP1 et BP2 sont aux bouts opposés des régions intergéniques, soit pour une seule combinaison (figure III.3A). Au minimum, si BP1 et BP2 sont au plus proche (sur la dernière base de chaque gène), on « supprime »

$(k-1)l_{intergenic} - 2(l_{gene} - 1)$ bases (en comptant en plus les bases devenues non codantes, figure III.3C). Entre les deux, si BP1 et BP2 sont le plus à gauche possible (l'un sur l'intergénique, l'autre sur le gène, figure III.3B), on « supprime » $kl_{intergenic} - (l_{gene} - 1)$ bases non codantes. Si on décale les deux points de cassure conjointement, on obtient le même résultat pour $l_{intergenic} + l_{gene}$ combinaisons. Pour le reste c'est assez simple, dès qu'on varie la taille de la délétion, on gagne ou on perd exactement une combinaison. Si on part de la plus longue délétion, on gagne une combinaison en diminuant la taille de la délétion jusqu'à arriver au cas de la figure III.3B. Ensuite, on perd une combinaison avec chaque diminution jusqu'à aboutir à la délétion la plus petite possible. On obtient donc un simple profil en triangle entre les trois positions caractéristiques, illustré sur la figure III.3.

**Cas particuliers** Quand on perd 1, 2 ou tous les gènes, on obtient des profils un peu plus compliqués dérivés du profil triangulaire. Les démonstrations suivent le même raisonnement que dans le cas général, avec quelques subtilités en plus. Comme ces cas sont des exceptions, nous n'en donnons pas la démonstration ici, elle est reportée dans l'annexe A, section 1.

**Profil complet des grandes délétions** Partant d'un nombre $n_0$ de gènes, si la délétion a conduit à la perte de $k$ gènes, alors la distribution de la quantité de non-codant restant après la délétion a la forme d'un triangle isocèle de même base pour presque toutes les valeurs de $k$, ce triangle étant simplement translaté de $l_{intergenic}$ bases vers la gauche quand on augmente $k$ de une unité. Le profil complet, composé de ces trinagles ainsi que des cas particuliers dont la démonstration a été différée en annexe, est illustré sur la figure III.4.

**Profil complet des duplications** Le cas des duplications est très similaire : on peut raisonner par complémentarité. Prenons une combinaison de points de cassure conduisant à la perte de $k$ gènes dans le génome $(L_0, n_0)$. La difficulté, c'est qu'on ne connaît pas le nombre de gènes *intacts* sur le segment délété : ça peut être $k$, $k-1$ ou $k-2$ si les gènes n'ont été délétés qu'en partie. Par contre, le segment non délété contient *exactement* $n_0 - k$ gènes intacts. À chaque combinaison de points de cassure correspond naturellement une délétion de $k$ gènes, mais on peut aussi associer une duplication de $n_0 - k$ gènes grâce au segment complémentaire. Il y a donc une bijection entre les délétions et les duplications. Par exemple, les combinaisons conduisant à la perte de tous les gènes ($k = n_0$) sont aussi celles qui conduisent aux duplications de $n_0 - k = 0$ gène. Les combinaisons conduisant à la perte de tous les gènes sauf 1 ($k = n_0 - 1$) sont celles qui correspondent aux duplications de 1 gène. Le même raisonnement peut s'appliquer au non codant, si bien que le profil des duplications est simplement le profil des délétions translaté (figure III.4).

Cependant, cette distribution est incomplète puisqu'on a ignoré la possibilité de perdre un gène à l'insertion. Pour prendre en compte ce phénomène, il suffit de calculer la probabilité d'insérer la séquence dans du non-codant qui vaut $p = L_0/(L_0 + n_0 l_{gene})$. On obtient alors le profil complet en séparant les deux cas : on pondère par $p$ les profils en triangle de

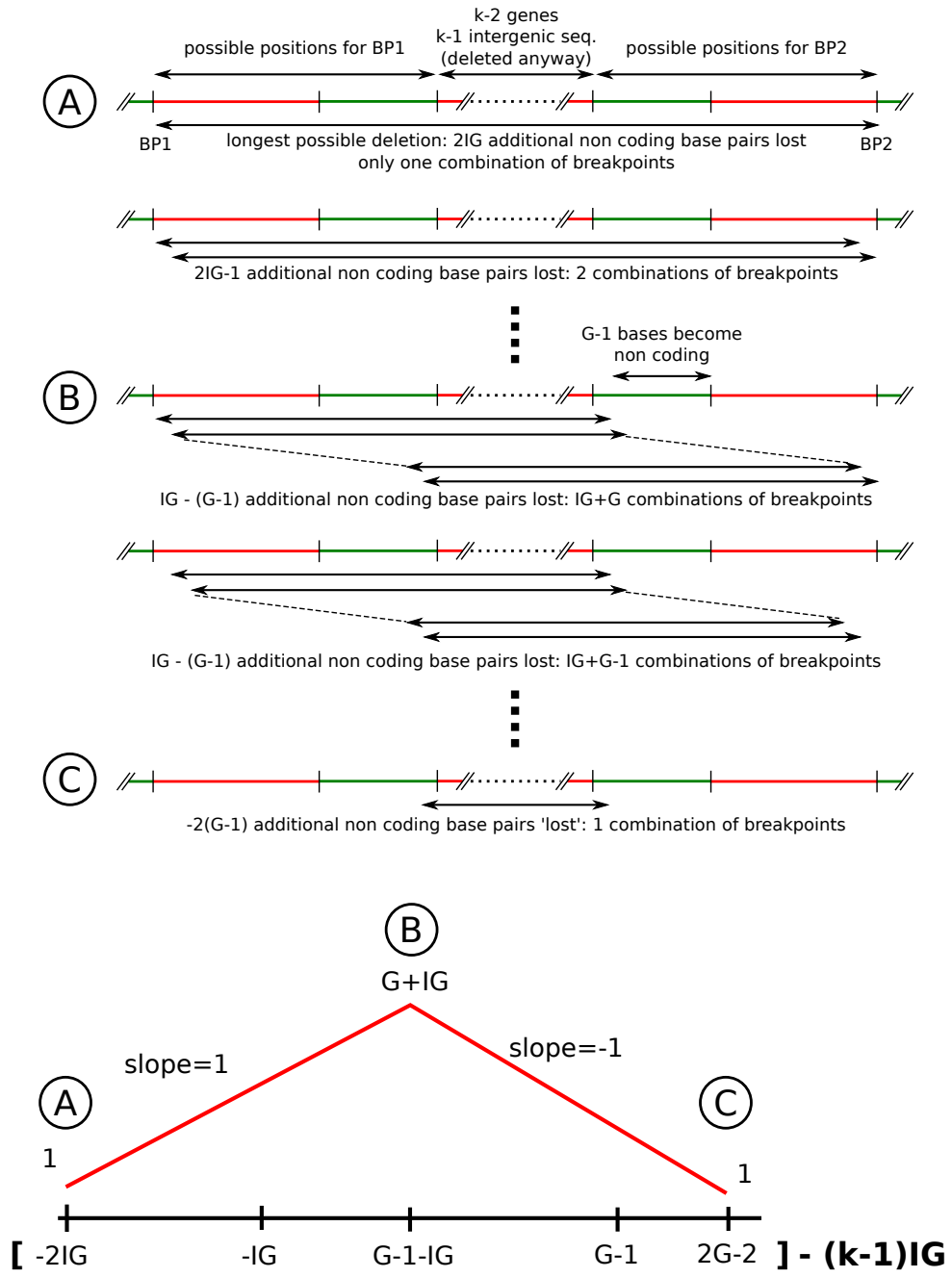


Figure III.3 – Distribution générale des transitions liées aux duplications et aux délétions, ici dans le cas de la perte de $k$ gènes. En variant la taille de la délétion de la plus longue possible (cas A) jusqu'à la plus faible (cas C), on se rend compte que le nombre de combinaisons de points de cassure qui permettent cette taille de délétion augmente dans un premier temps jusqu'au cas B, puis diminue. Pour raccourcir les notations sur le dessin, la longueur de l'intergénique est dénotée par IG ($l_{intergenic}$ dans le texte) et la longueur d'un gène G ($l_{gene}$ dans le texte).

la figure pour le cas où la séquence est insérée dans du non-codant, on les pondère pas $(1-p)$ et on les décale d'un gène vers le bas et vers la droite pour le cas où la séquence

casse un gène à l'insertion.

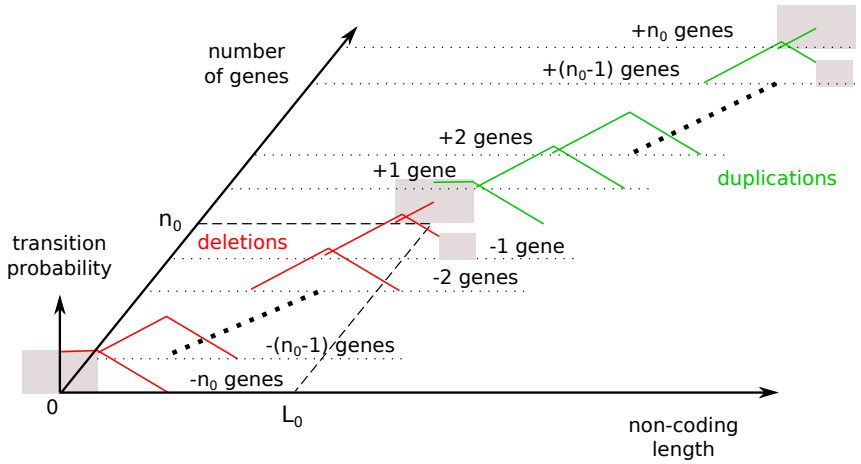


Figure III.4 – Distribution du nombre de gènes et de la quantité de non codant après une duplication (en ignorant la perte de gène à l'insertion) ou une délétion. On obtient le même profil en triangle sur toutes les lignes, à quelques exceptions près qui sont marquées d'un contour gris.

**Calcul des probabilités de transition** Les profils ont été donnés en nombre de combinaisons de points de cassure mais, en renormalisant, on obtient les mêmes profils compris entre 0 et 1 qui correspondent aux probabilités cherchées initialement. Le nombre de combinaisons pour un triangle complet vaut $(l_{gene} + l_{intergenic})^2$ et globalement, les parties tronquées se compensent, si bien que le nombre de combinaisons total vaut exactement $n_0(l_{gene} + l_{intergenic})^2 = n_0(l_{gene} + L_0/n_0)^2 = (n_0 l_{gene} + L_0)^2/n_0$, ce qui correspond à ce qui était attendu.

Sur la figure III.4, on peut interpréter l'axe z comme donnant les nombres de combinaisons ou les probabilités, puisque la renormalisation ne change pas la hauteur relative des triangles. On peut alors y lire les probabilités de transitions $\Pr[(L_0, n_0) \to (L_f, n_f)]$ : on se place au niveau de l'état $(L_f, n_f)$ dans le plan horizontal, la probabilité est donnée par la hauteur du profil.

Pour la suite, nous reviendrons sur les profils en triangle. Il est important de noter que

- Le triangle sur une ligne $n_f$ donne les probabilités de transitions qui s'écrivent sous la forme $\Pr[(L_0, n_0) \to (., n_f)]$.
- Après renormalisation, la somme des probabilités pour un triangle vaut exactement $1/n_0$ et sa hauteur $n_0/(l_{gene} + l_{intergenic})$.

Au passage (cela n'a pas d'importance pour la suite), on pourra noter que les pertes de gènes ne sont pas complètement équiprobables. Pour évaluer la probabilité de perdre un

nombre précis $k$ de gènes, il faut calculer

$$\sum_{i \geq 0} \Pr\left[(L_0, n_0) \to (i, n_0 - k)\right]$$

autrement dit, sommer toutes les probabilités associées à une ligne d'arrivée. Nous venons d'expliquer que cette somme vaut $1/n_0$ pour un triangle complet, mais un coup d'oeil à la figure III.4 devrait permettre de s'assurer que la somme ne sera pas la même pour $k \in \{0, 1, n_0\}$.

# 3 Agrégation des transitions

Comme expliqué dans l'introduction, générer toutes les transitions une à une serait extrêmement long et impossible à stocker, même sur un disque dur. Nous allons donc créer des classes de génomes en agrégeant des états qui ont une structure similaire et pour lesquelles les transitions sont similaires. Pour rappel, pour $10^4$ classes de génomes, on aurait déjà une matrice de 800 Mo.

L'agrégation doit se faire en partie analytiquement pour être réalisable en pratique. En effet, le nombre de transitions à additionner pour une classe de génome devient rapidement très grand à cause des duplications et des grandes délétions. Les mutations dont l'impact ne grossit pas avec la taille du génome ne posent pas de problème particulier puisque le nombre de transitions à agréger ne varie pas avec la structure. Au contraire, pour les duplications et les délétions, plus le génome est grand, plus il y a de transitions possibles. Par exemple, pour l'état de départ $(L, n)$, il y a $(2n+1) \times (2L/n + 2l_{gene}) > 4(L + nl_{gene})$ transitions possibles, soit plus de 4 fois le nombre de paire de bases du génome initial. Si on agrège les transitions naïvement, le nombre de sommes à effectuer ne dépend pas du découpage de l'espace choisi : chaque transition atomique sera sommée exactement une fois. En faisant quelques calculs, on peut déterminer que si l'espace de calcul contient $10^{10}$ états atomiques (ce qui correspond par exemple à une limite de 1000 gènes et de 10Mb de non codant), il faut agréger plus de $10^{12}$ transitions au total. De plus, l'augmentation asymptotique du nombre de transitions avec le nombre d'états atomiques est quadratique.

Nous allons commencer par proposer une subdivision de l'espace de simulation, puis nous proposerons des algorithmes qui permettent d'agréger les transitions atomiques de manière approximative, mais plus efficace.

## 3.1 Subdivision de l'espace

Pour commencer, il faut que le nombre de classes (ou subdivisions) de l'espace soit fini. Il y aura donc un certain nombre de classes aux bords qui contiendront une infinité d'états atomiques et pour lesquelles il y aura des conditions aux bords particulières. Il faudra

être attentif, lors de l'interprétation des résultats, à la signification des densités le long des bords. Nous allons présenter deux manières de subdiviser l'espace, l'une plus naïve, l'autre tirant partie de l'analyse du chapitre II. Les deux subdivisions se font simplement dans l'espace naturel (non codant, nombre de gènes).

### 3.1.1 Agrégation en échelle linéaire

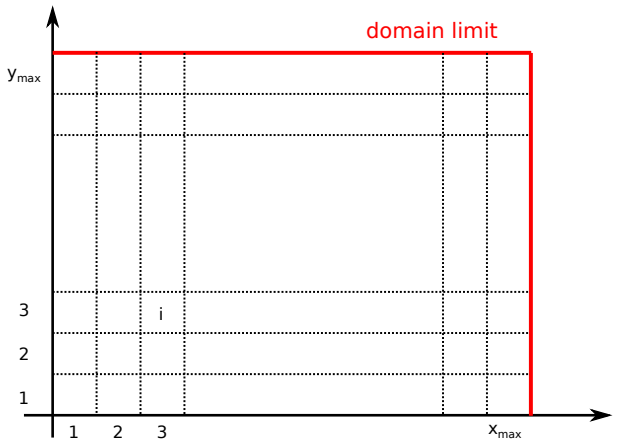


Figure III.5 – Agrégation en échelle linéaire.

La manière la plus simple de couper l'espace (non codant, nombre de gènes) est de le couper en rectangles réguliers (figure III.5). Pour que le nombre de classes soit fini, nous spécifions un nombre fini de subdivisions le long de chaque axe : $x_{max}$ pour le non codant, $y_{max}$ pour le codant. Chaque rectangle a une largeur de $\Delta L$ et une hauteur de $\Delta n$.

### 3.1.2 Agrégation en échelle logarithmique

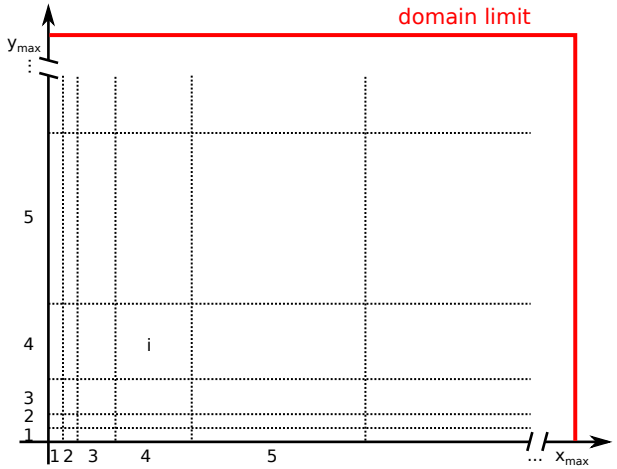


Figure III.6 – Agrégation en échelle logarithmique de base 2.

L'autre façon de couper l'espace vient du fait que les mutations qui dominent la dynamique globale sont de nature multiplicative : elle consiste à prendre les subdivisions en échelle logarithmique (figure III.6). Comme les duplications doublent au plus la taille du génome, le logarithme de base 2 est le plus naturel : on peut aller d'une subdivision à celle d'à côté en une mutation quelle que soit la position dans l'espace. À proximité de l'origine, l'échelle est mal définie, on spécifiera donc une plus petite subdivision à l'origine (voir la section 2.2 de l'annexe A pour plus de détails).

### 3.1.3 Répartition dans les subdivisions, transitions moyennes et conditions aux bords

Après avoir découpé l'espace en classes de génomes, il faut donner des transitions d'une classe vers une autre. Pour cela, il faut donner un *a priori* sur la position des individus au sein d'une classe et donc sur les transitions atomiques qui seront utilisées préférentiellement. La manière habituelle de gérer cette difficulté consiste à prendre une distribution uniforme, car une autre distribution introduit des biais vers des zones spécifiques de chaque subdivision qu'il est difficile de justifier *a priori*. C'est ce qui a été choisi ici : les transitions d'une classe vers une autre s'obtiennent simplement en faisant la moyenne de toutes les transitions issues de tous les états atomiques de la première classe et qui atterrissent sur n'importe quel état atomique de la classe d'arrivée.

Les subdivisions aux bords ont une géométrie particulière. Dès qu'une transition atomique sort de la limite du domaine, elle est réassignée à la subdivision la plus proche. Schématiquement, c'est comme si les traits des subdivisions se prolongeaient à l'infini sur les figures III.5 et III.6. Par contre, quand on calcule les transitions qui partent de ces subdivisions, on se contente de moyenner sur les états effectivement délimités dans les figures III.5 et III.6. Ce choix est un peu délicat, puisque cela revient à avoir un bord réflexif qui limite donc la croissance des génomes. Cependant, d'après l'analyse, le comportement de la population est assez binaire : soit la taille du génome a tendance à diminuer et on devrait être loin des bords, soit elle explose et la population devrait être collée aux bords. On verra dans les simulations comment ces prédictions se traduisent. Quand l'analyse prévoyait une convergence de la taille des génomes, nous avons pris des limites de domaine qui permettent une dynamique et une convergence de la population loin des bords. La réflexivité des bords a l'avantage de conserver la somme de la densité à 1.

### 3.1.4 Remarques et notations liées à la subdivision de l'espace

Les algorithmes présentés dans les sections suivantes sont indépendants du type de subdivision choisi, elles partent du principe que la subdivision est un rectangle, ce qui est le cas dans les deux échelles. Cela signifie qu'on pourrait varier encore le type de découpage, par exemple en prenant une échelle semi-logarithmique, linéaire en nombre de gènes et logarithmique en non codant.

Chaque subdivision est indexée par $(x, y)$, où $1 \leq x \leq x_{max}$ correspond à l'axe du nombre de paires de bases non codantes et $1 \leq y \leq y_{max}$ correspond à l'axe du nombre de gènes. Pour la subdivision $(x, y)$, on notera le plus petit nombre de bases non codantes pour un génome dans la subdivision $L_{min}(x)$, le plus grand $L_{max}(x)$ et la largeur $\Delta L(x)$. De même, on définit le plus petit nombre de gènes $n_{min}(y)$, le plus grand $n_{max}(y)$ et la largeur $\Delta n(y)$.

D'autres grandeurs sont introduites ponctuellement dans les algorithmes, comme le nombre moyen de gènes ou le ratio de codant moyen dans une subdivision. Ces valeurs sont spécifiées pour chaque échelle dans les sections 2.1 et 2.2 de l'annexe A. Elles font parfois intervenir des sommes qui ne peuvent pas être simplifiées analytiquement mais qui peuvent être calculées efficacement en utilisant un développement asymptotique. Cette technique ne sera pas abordée dans la présentation des algorithmes, il faudra se référer à la section 2.3 de l'annexe A pour une présentation de la technique utilisée et à la section 3 de l'annexe A pour voir comment on passe des sommes à calculer dans les algorithmes à des sommes classiques pour lesquelles le développement asymptotique s'applique.

Dans tous les algorithmes présentés, nous considérerons $(x_0, y_0)$ comme étant la subdivision de départ (celle pour laquelle on veut agréger les transitions atomiques sortantes). Pour simplifier les notations, les variables $x$ et $y$ seront omises quand il n'y a pas d'ambiguïté, par exemple on écrira $L_{max}$ au lieu de $L_{max}(x)$. Le nombre d'états atomiques dans la subdivision vaut $\Delta n \times \Delta L$ : il apparaîtra fréquemment dans les calculs de moyenne sans être justifié explicitement à chaque fois. De même, certains effets de bord dus à l'absence de gènes ou de bases non codantes ne sont pas explicitement cités si leur traitement ne pose pas de problème particulier.

## 3.2 Agrégation des indels neutres

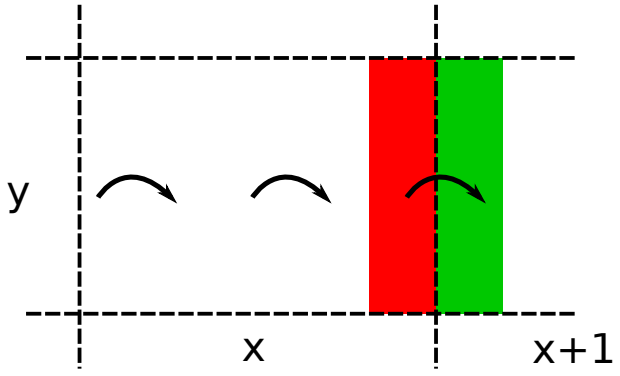


Figure III.7 – Agrégation de petites insertions neutres

Nous allons nous intéresser à l'agrégation des petites insertions neutres pour les individus qui partent de la subdivision $(x_0, y_0)$. Pour simplifier, nous supposons que $\Delta L(x_0) \geq 6$, de façon à ce qu'un indel puisse au maximum nous mener dans une subdivision voisine. Ceci sera toujours vrai en pratique. Parmi les petites insertions neutres partant de $(x_0, y_0)$

seules celles sur les 6 dernières colonnes peuvent quitter la subdivision de départ vers la subdivision voisine $(x_0 + 1, y_0)$ (figure III.7). En y regardant de plus près, toutes les transitions neutres de la dernière colonne vont vers la subdivision voisine, 5/6 de la colonne précédente, 4/6 de celle d'avant, etc. Rappelons que la probabilité qu'une insertion soit neutre est $L/(nl_{gene} + L)$. On a alors

$$\Pr\left[(x_0, y_0) \to (x_0 + 1, y_0)\right] = \frac{1}{\Delta n \Delta L} \sum_{k=1}^{6} \frac{7-k}{6} \sum_{n=n_{min}}^{n_{max}} \frac{L_{max} - (k-1)}{nl_{gene} + L_{max} - (k-1)}$$

Comme en pratique $L_{max} \gg k$, nous utilisons l'approximation suivante

$$\Pr\left[(x_0, y_0) \to (x_0 + 1, y_0)\right] \simeq \frac{3.5}{\Delta n \Delta L} \sum_{n_{min}}^{n_{max}} \frac{L_{max}}{nl_{gene} + L_{max}}$$

On peut appliquer le même raisonnement dans le cas des petites délétions : si $x_0 > 1$, les transitions partant des 6 colonnes tout à gauche peuvent aller vers la subdivisions $(x_0 - 1, y_0)$ et on a la même approximation :

$$\Pr\left[(x_0, y_0) \to (x_0 - 1, y_0)\right] \simeq \frac{3.5}{\Delta n \Delta L} \frac{L_{min}}{l_{gene}} \sum_{n_{min}}^{n_{max}} \frac{L_{min}}{nl_{gene} + L_{max}}$$

## 3.3 Agrégation de l'inactivation de gènes (indels non neutres, inversions, translocations)

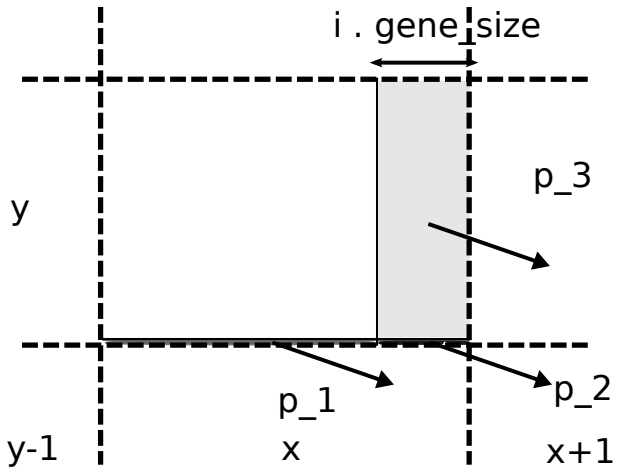


Figure III.8 – Agrégation de la perte de $i$ gènes

Pour le génome $(L, n)$, la probabilité qu'un indel ou un point de cassure tombe dans une partie codante est $p(L, n) = nl_{gene}/(nl_{gene} + L)$. Comme on l'a vu dans la section précédente, on fait l'approximation que le nombre de gènes perdus suit une loi binomiale

$\mathcal{B}(k, p(L,n))$, où $k = 1$, 2 ou 3, selon que l'on considère un indel, une inversion ou une translocation. Pour chaque cas, on se concentre spécifiquement sur les transitions sortantes : on ne donnera pas la valeur de la probabilité de rester sur place, elle s'obtient en retranchant de 1 la somme toutes les transitions sortantes (y compris les transitions neutres dans le cas des indels). Il y a 3 destinations possibles lors de l'inactivation de gènes : $(x_0+1, y_0)$, $(x_0+1, y_0-1)$ and $(x_0, y_0-1)$ (figure III.8). L'enjeu principal est de délimiter les zones sur lesquelles on prend les moyennes pour chaque destination. On donne ici le principe général, les sommes utilisées dans le programme, dérivées à partir de celles ci-dessous, figurent en section 3.2 de l'annexe A.

### 3.3.1 Indels non-neutres : inactivation d'un gène

Pour chaque destination possible $(x_f, y_f) \in \{(x_0+1, y_0), (x_0+1, y_0-1), (x_0, y_0-1)\}$, la moyenne des transitions qui s'y rendent s'écrit

$$\Pr\left[(x_0, y_0) \to (x_f, y_f)\right] = \frac{1}{\Delta L \Delta n} \sum_{L,n} p(L,n)$$

sauf que le domaine de sommation dépend de la destination. Pour obtenir la valeur de la transitions moyenne, il suffit donc de préciser ce domaine, ce que nous ferons dans la suite (on ne redonnera pas la somme à calculer pour obtenir la probabilité).

Les génomes qui ont initialement $n_{min}$ gènes (la ligne du bas du rectangle) vont quitter $(x_0, y_0)$ pour se rendre sur la subdivison d'ordonnée $y_0 - 1$. Les individus qui sont suffisamment proches du bord droite vont quitter la subdivision d'abscisse $x_0$ pour celle d'abscisse $x_0 + 1$ puisque le gène est converti en non codant. Plus précisément

- Les génomes $\{(L,n) : L_{min} \leq L \leq L_{max} - l_{gene} - 7, n = n_{min}\}$ vont en $(x, y-1)$.
- Les génomes $\{(L,n) : L_{max} - l_{gene} + 7 \leq L \leq L_{max}, n = n_{min}\}$ vont en $(x+1, y-1)$.
- Les génomes $\{(L,n) : L_{max} - l_{gene} + 7 \leq L \leq L_{max}, n > n_{min}\}$ vont en $(x+1, y)$.
- Pour les génomes tels que $L_{max} - l_{gene} - 6 \leq L \leq L_{max} - l_{gene} + 6\}$, il faut distinguer, pour chaque type d'indel, séparer les transitions qui restent dans la subdivision de coordonnée $x_0$ et celles qui vont dans en $x_0 + 1$. Le traitement de ce cas ressemble à l'agrégation des indels non-neutres et n'est pas détaillé ici.

### 3.3.2 Inversions : deux points de cassure

Le nombre de gènes perdus suit une distribution $\mathcal{B}(2, p(L,n))$.

Si un gène est perdu, les transitions sont moyennées en calculant la somme

$$\Pr\left[(x_0, y_0) \to (x_f, y_f)\right] = \frac{1}{\Delta L \Delta n} \sum_{L,n} 2p(L,n)(1 - p(L,n))$$

- Les génomes $\{(L,n) : L_{min} \leq L \leq L_{max} - l_{gene}, n = n_{min}\}$ vont en $(x_0, y_0 - 1)$.
- Les génomes $\{(L,n) : L_{max} - l_{gene} < L \leq L_{max}, n = n_{min}\}$ vont en $(x_0 + 1, y_0 - 1)$.
- Les génomes $\{(L,n) : L_{max} - l_{gene} < L \leq L_{max}, n > n_{min}\}$ vont en $(x_+1, y_0)$.

Si deux gènes sont perdus, les transitions sont moyennées en calculant

$$\Pr\left[(x_0, y_0) \to (x_f, y_f)\right] = \frac{1}{\Delta L \Delta n} \sum_{L,n} p(L,n)^2$$

- Les génomes $\{(L,n) : L_{min} \leq L \leq L_{max} - 2l_{gene}, n \in \{n_{min}, n_{min} + 1\}\}$ vont en $(x_0, y_0 - 1)$.
- Les génomes $\{(L,n) : L_{max} - 2l_{gene} < L \leq L_{max}, n \in \{n_{min}, n_{min} + 1\}\}$ vont en $(x_0 + 1, y_0 - 1)$.
- Les génomes $\{(L,n) : L_{max} - 2l_{gene} < L \leq L_{max}, n > n_{min} + 1\}$ vont en $(x_0 + 1, y_0)$.

Il y a une exception en échelle logarithmique pour les individus ayant 2 gènes (situées dans la subdivision $y_f = 3$). Quand un gène est perdu, la subdivison finale est celle d'ordonnée $y_f = 2$, quand deux sont perdus, la subdivision finale est celle d'ordonnée $y_f = 1$.

### 3.3.3 Translocations : 2 points de cassure et un point d'insertion

L'idée est la même que pour les inversions, sauf qu'on perd jusqu'à 3 gènes d'après une loi binomiale $\mathcal{B}(3, p(L,n))$.

Si un gène est perdu, les transitions sont moyennées en calculant la somme

$$\Pr\left[(x_0, y_0) \to (x_f, y_f)\right] = \frac{1}{\Delta L \Delta n} \sum_{L,n} 3p(L,n)(1 - p(L,n))^2$$

- Les génomes $\{(L,n) : L_{min} \leq L \leq L_{max} - l_{gene}, n = n_{min}\}$ vont en $(x_0, y_0 - 1)$.
- Les génomes $\{(L,n) : L_{max} - l_{gene} < L \leq L_{max}, n = n_{min}\}$ vont en $(x_0 + 1, y_0 - 1)$.
- Les génomes $\{(L,n) : L_{max} - l_{gene} < L \leq L_{max}, n > n_{min}\}$ vont en $(x_0 + 1, y_0)$.

Si deux gènes sont perdus, les transitions sont moyennées en calculant

$$\Pr\left[(x_0, y_0) \to (x_f, y_f)\right] = \frac{1}{\Delta L \Delta n} \sum_{L,n} 3p(L,n)^2(1 - p(L,n))$$

- Les génomes $\{(L,n) : L_{min} \leq L \leq L_{max} - 2l_{gene}, n \in \{n_{min}, n_{min}+1\}\}$ vont en $(x_0, y_0 - 1)$.
- Les génomes $\{(L,n) : L_{max} - 2l_{gene} < L \leq L_{max}, n \in \{n_{min}, n_{min}+1\}\}$ vont en $(x_0 + 1, y_0 - 1)$.
- Les génomes $\{(L,n) : L_{max} - 2l_{gene} < L \leq L_{max}, n > n_{min} + 1\}$ vont en $(x_0 + 1, y_0)$.

Si deux gènes sont perdus, les transitions sont moyennées en calculant

$$\Pr\left[(x_0, y_0) \to (x_f, y_f)\right] = \frac{1}{\Delta L \Delta n} \sum_{L,n} p(L,n)^3$$

- Les génomes $\{(L,n) : L_{min} \leq L \leq L_{max} - 3l_{gene}, n_{min} \leq n \leq n_{min} + 2\}$ vont en $(x_0, y_0 - 1)$.
- Les génomes $\{(L,n) : L_{max} - 3l_{gene} < L \leq L_{max}, n_{min} \leq n \leq n_{min} + 2\}$ vont en $(x_0 + 1, y_0 - 1)$.
- Les génomes $\{(L,n) : L_{max} - 3l_{gene} < L \leq L_{max}, n > n_{min} + 2\}$ vont en $(x_0 + 1, y_0)$.

On a le même type d'exception que pour les inversions en échelle logarithmique. Dans les toutes premières subdivisions, on peut éventuellement passer à $y_0 - 2$ ou $y_0 - 3$ au lieu de $y_0 - 1$.

## 3.4 Agrégation des grandes délétions

**Principe général** Mathématiquement, la transition entre subdivisons est reliée aux transitions atomiques par la relation

$$\boxed{\begin{array}{c}\Pr\left[(x_0, y_0) \to (x_f, y_f)\right] \stackrel{\text{def.}}{=} \\ \dfrac{1}{\Delta n(y_0)\Delta L(x_0)} \displaystyle\sum_{L_0=L_{min}(x_0)}^{L_{max}(x_0)} \sum_{n_0=n_{min}(y_0)}^{n_{max}(y_0)} \sum_{L_f=L_{min}(x_f)}^{L_{max}(x_f)} \sum_{n_f=n_{min}(y_f)}^{n_{max}(y_f)} \Pr\left[(L_0, n_0) \to (L_f, n_f)\right]\end{array}} \tag{III.4}$$

Dans l'idéal, il s'agirait de calculer analytiquement les quatre sommes : cela ne sera malheureusement pas possible. Nous allons considérer des versions où nous calculerons analytiquement deux ou trois de ces sommes, le reste se faisant par ordinateur.

**Rappels concernant les profils triangulaires et passage au continu** Dans la section 2.3, nous avons vu que les probabilités de transitions $\Pr[(L_0, n_0) \to (L_f, n_f)]$ s'obtenaient à partir d'une suite de profils triangulaires (ou presque). Quand nous cherchons à agréger ces transitions pour obtenir $\Pr[(x_0, y_0) \to (x_f, y_f)]$, nous sommes en train de moyenner ces profils triangulaires d'après la formule ci-dessus. Si on applique la division par $\Delta n(y_0)\Delta L(x_0)$ tout à la fin, il s'agit donc surtout de faire la somme des contributions données par les triangles. Pour cela, il faut commencer par repérer la position de tous les triangles dans l'espace (non codant, nombre de gènes).

Prenons un génome de taille initiale $(L_0, n_0)$. Si l'état génomique d'arrivée comporte $n$ gènes (et donc que $k = n_0 - n$ gènes ont été perdus), alors les quantités minimales et maximales pour la quantité d'ADN non codant dans l'état d'arrivée sont : (figure III.3, page 111) :

- $\min(L|\Pr[(L_0, n_0) \to (L, n)] \neq 0) = L_0 - ((n_0 - n) + 1)l_{intergenic} = L_0 - (n_0 - n + 1)L_0/n_0 = (n-1)L_0/n_0$
- $\max(L|\Pr[(L_0, n_0) \to (L, n)] \neq 0) = (n+1)L_0/n_0 + 2(l_{gene} - 1)$

Nous avons vu que les valeurs des probabilités pour une ligne donnée $n$ sont en général données par une densité en forme de triangle. De plus, nous avons établi que la somme des probabilités contenues dans un de ces triangles vaut $1/n_0$. Pour simplifier les calculs qui suivent, nous allons considérer que les profils triangulaires sont continus selon l'axe des $L$. Pour que la somme des probabilités soit conservée malgré cette approximation, nous introduisons un léger décalage dans les limites du profil :

- $\inf(L|\Pr[(L_0, n_0) \to (L, n)] \neq 0) = L_0 - ((n_0 - n) + 1)l_{intergenic} = L_0 - (n_0 - n + 1)L_0/n_0 = (n-1)L_0/n_0 - 1$
- $\sup(L|\Pr[(L_0, n_0) \to (L, n)] \neq 0) = (n+1)L_0/n_0 + 2(l_{gene} - 1) + 1$

La demi-largeur de chaque triangle est alors de $l_{intergenic} + l_{gene}$, la hauteur $1/[(l_{intergenic} + l_{gene})n_0]$ (figure III.9). Par simplicité, nous allons supposer que cette approximation est valable également pour la ligne $n = n_0 - 1$, où le triangle est théoriquement tronqué (figure III.4, page 112).

Le passage au continu nous permet de calculer plus simplement les expressions agrégées. Pour calculer les contributions des transitions se rendant entre deux points $a$ et $b$ (dans l'échelle discrète), il faudra intégrer les probabilités entre $a-0.5$ et $b+0.5$ pour compenser le passage au continu. Cette approximation génère un léger biais local comparé à l'expression discrète mais l'impact est très faible.

**Principe de l'agrégation « ligne à ligne »** L'élément central de l'agrégation repose sur l'aire des triangles, qui vaut $1/n_0$ : elle ne dépend pas de la taille du non codant, ni

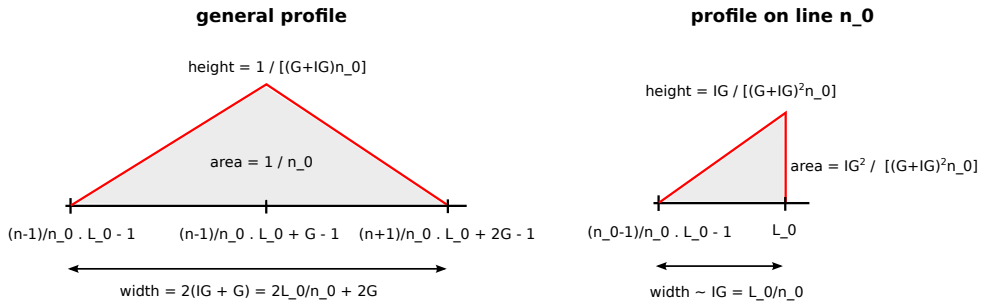


Figure III.9 – Profil atomique de délétion après renormalisation et passage au continu

de la ligne d'arrivée. Cela nous a permis de développer l'agrégation dite « ligne à ligne », qui correspond mathématiquement à

$$\Pr\left[(x_0,n_0)\to(x_f,n_f)\right] \overset{\text{def.}}{=} \sum_{L_f=L_{min}(x_f)}^{L_{max}(x_f)} \frac{1}{\Delta L(x_0)} \sum_{L_0=L_{min}(x_0)}^{L_{max}(x_0)} \Pr\left[(L_0,n_0)\to(L_f,n_f)\right]$$

La somme à l'intérieur contient les $\Delta L(x_0)$ triangles situés sur la ligne $n_f$ provenant des points $(L_{min}(x_0),n_0)$, ..., $(L_{min}(x_0),n_0)$. Dans la deuxième somme, on additionne les contributions entre les points $L_{min}(x_f)$ et $L_{max}(x_f)$. Cependant, on peut faire abstraction de cette deuxième somme dans un premier temps et se contenter d'obtenir la densité correspondant à la somme des $\Delta L(x_0)$ triangles sur toute la ligne $n_f$ pour toutes les valeurs de $L_f$ possibles. Si on y parvient, il suffira d'intégrer la densité entre $L_{min}(x_f)-0.5$ et $L_{max}(x_f)+0.5$ (dans l'approximation continue).

Quand on a déterminé ces densités pour toutes les lignes de départ et toutes les lignes d'arrivée, la probabilité de subdivision à subdivision s'obtient sommant par ordinateur via la formule

$$\Pr\left[(x_0,y_0)\to(x_f,y_f)\right] \overset{\text{def.}}{=} \frac{1}{\Delta n(y_0)} \sum_{n_0=n_{min}(y_0)}^{n_{max}(y_0)} \sum_{n_f=n_{min}(y_f)}^{n_{max}(y_f)} \Pr\left[(x_0,n_0)\to(x_f,n_f)\right]$$

**Principe de l'agrégation « rectangle à ligne »** Dans un deuxième temps, on verra que certaines densités données par l'agrégation « ligne à ligne » ont elles-mêmes des propriétés particulières qui peuvent être exploitées analytiquement, au prix d'une approximation supplémentaire. Cette deuxième approche utilise l'agrégation dite « rectangle à ligne », qui correspond mathématiquement à

$$\Pr\left[(x_0,y_0)\to(x_f,n_f)\right] \overset{\text{def.}}{=}$$
$$\sum_{L_f=L_{min}(x_f)}^{L_{max}(x_f)} \frac{1}{\Delta n(y_0)} \sum_{n_0=n_{min}(y_0)}^{n_{max}(y_0)} \frac{1}{\Delta L(x_0)} \sum_{L_0=L_{min}(x_0)}^{L_{max}(x_0)} \Pr\left[(L_0,n_0)\to(L_f,n_f)\right] \tag{III.5}$$

À l'intérieur ($\sum_{L_0=L_{min}(x_0)}^{L_{max}(x_0)} \Pr\left[(L_0, n_0) \to (L_f, n_f)\right] / \Delta L(x_0)$), on retrouve les densités obtenues via l'agrégation ligne à ligne. Ces densités sont sommées puis moyennées pour toutes les lignes de départ de $(x_0, y_0)$ vers la ligne d'arrivée $n_f$ : on obtient une nouvelle densité, exprimée en fonction du point d'arrivée $L_f$. On finit par calculer la contribution entre les points $L_{min}(x_f)$ et $L_{max}(x_f)$, ce qui correspond à une intégration dans l'approximation continue. Quand l'agrégation rectangle à ligne peut être utilisée, le calcul des probabilités de transition de subdivison à subdivison devient

$$\Pr\left[(x_0, y_0) \to (x_f, y_f)\right] = \sum_{n_f=n_{min}(y_f)}^{n_{max}(y_f)} \Pr\left[(x_0, y_0) \to (x_f, n_f)\right]$$

### 3.4.1 Agrégation ligne à ligne

Nous nous proposons d'agréger les transitions dues aux délétions issues de la ligne $n_0$, entre $L_{min}(x_0)$ et $L_{max}(x_0)$, à destination d'une ligne $n_f$. Nous avons vu avec la figure III.4 que si l'on part d'un état $(L_0, n_0)$, les états d'arrivée possibles comportent entre 0 et $n_0$ gènes et, que pour la majorité de ces nombres de gènes d'arrivée possibles, la distribution conditionnelle de la quantité d'ADN non-codant a une forme triangulaire. Ainsi, pour les nombres de gènes d'arrivée compris entre 1 et $n_0 - 1$, un état de départ $(L_0, n_0)$ « donne » $n_0 - 1$ triangles, chacun sur une ligne différente. Pour des nombres de gènes d'arrivée valant 0 ou $n_0$, chaque état de départ donne des triangles tronqués. Si l'on considère tous les points de départ de la forme $(L_0, n_0)$ avec $n_0$ fixé mais $L_0$ variant entre $L_{min}$ et $L_{max}$, on aura donc $\Delta L = L_{max} - L_{min} + 1$ triangles (de largeurs différentes mais de même aire) sur chaque ligne d'arrivée possible. En théorie, nous cherchons à calculer les contributions pour une subdivision d'arrivée précise mais, en pratique, il est plus simple de calculer les contributions sur toute une ligne d'arrivée et de découper en subdivisions ensuite.

Nous allons raisonner ligne d'arrivée par ligne d'arrivée, en regardant les transitions aboutissant à la ligne $n_f$. Nous allons commencer avec le cas $0 < n_f < n_0$, ce qui nous assure que les profils que nous agrégeons sont bien des triangles entiers. Cependant, la largeur des triangles dépend du point de départ car elle dépend de la taille de l'intergénique. Les triangles qu'on veut agréger ne sont donc pas identiques. De plus, les triangles qu'on agrège sont d'autant plus rapprochés les uns des autres que $n_f$ diminue. Globalement, l'aire du profil agrégé est constante car on fait la moyenne de triangles qui ont tous la même aire $1/n_0$. La hauteur du profil agrégé est donc de plus en plus haute (figure III.10).

Pour appliquer les algorithmes, il faut connaître la position des profils triangulaires. En fait, même si la largeur des triangles est variable pour une ligne d'arrivée donnée (puisque chaque triangle correspond à un point de départ différent avec une quantité spécifique d'ADN non-codant), les sommets des triangles sont régulièrement espacés. Si on connaît la position du triangle tout à gauche et de celui tout à droite, on peut reconstituer toute la suite. Les algorithmes n'utilisent donc que les 3 points caractéristiques des deux triangles aux extrémités. Nous nommons $L_1 = \inf(L | \Pr\left[(L_{min}, n_0) \to (L, n_f)\right] \neq 0)$ et

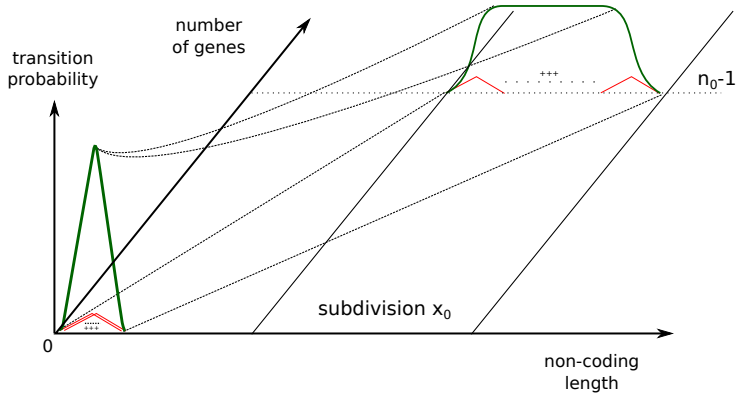


Figure III.10 – L'aire des profils agrégés vaut $1/n_0$, sauf pour $n_f = n_0$ et $n_f = 0$. Comme les triangles atomiques se rapprochent de plus en plus, la hauteur du profil agrégé augmente et sa forme devient plus complexe près de l'origine à cause du chevauchement : nous aurons besoin de plusieurs algorithmes d'agrégation pour prendre en compte ces phénomènes.

$L_3 = \sup(L|\Pr[(L_{min}, n_0) \to (L, n_f)] \neq 0)$ les points qui délimitent la base du triangle tout à gauche, et $L_2 = (L_1 + L_3)/2$ la position de son sommet. De même, nous nommons $L_4 = \inf(L|\Pr[(L_{max}, n_0) \to (L, n_f)] \neq 0)$ et $L_6 = \sup(L|\Pr[(L_{min}, n_0) \to (L, n_f)] \neq 0)$ les points qui délimitent la base du triangle tout à droite et $L_5 = (L_4 + L_6)/2$ la position de son sommet.

L'ordre de ces points caractéristiques va guider l'utilisation de différents algorithmes. En effet, on a toujours $L_1 < L_4$, $L_2 < L_5$ et $L_3 < L_6$, mais la relation entre les autres points n'est pas toujours claire. Quand peu de gènes sont perdus, on a en général $L_3 < L_4$ : les deux triangles extrêmes ne se chevauchent pas, ce qui est le cas le plus simple à agréger (algorithme LL pour ligne à ligne). Cependant, quand quasiment tous les gènes sont perdus, on commence à avoir un chevauchement prononcé entre les triangles extrêmes (et donc aussi avec tous les triangles entre les deux) et on aura $L_3 > L_4$ voire $L_3 > L_5$, qui nécessitent des algorithmes particuliers que nous allons lister dans la suite, accompagnés de la condition pour laquelle ils sont valables.

**Cas général (approximation LL, $n_0 > 1$, $0 < n_f < n_0$, $L_3 \leq L_5$)** Ce cas est le plus simple car il s'agit de sommer $\Delta L = L_{max} - L_{min} + 1$ triangles répartis régulièrement le long de l'axe des $L$. Les triangles se ressemblant beaucoup et ayant tous la même aire $1/n_0$, le processus d'agrégation ressemble à une intégration glissante. Supposons que $L_4 > L_3$, de telle manière à ce que les triangles extrêmes ne se chevauchent pas. Si les triangles additionnés étaient de même largeur, le profil agrégé consisterait approximativement en une première partie qui augmente de manière quadratique, une partie constante où les contributions des triangles qui entrent dans la sommation compensent ceux qui en sortent, puis une partie décroissante quadratique (figure III.11). Si $h$ est la hauteur de la partie constante, on peut montrer que la partie ascendante est approximativement (annexe A, section 3.3.1)

$$\begin{cases} f_1(L) &= C_1(L-L_1)^2 & \text{si } L \in [L_1, L_2] \\ f_2(L) &= h - C_1(L-L_3)^2 & \text{si } L \in [L_2, L_3] \end{cases}$$

Nous avons $f_1(L_2) = h/2$, d'où $C_1 = h/[2(L_2 - L_1)^2] = h/[2(L_{min}/n_0 + l_{gene})^2]$. Entre $L_3$ et $L_4$, nous avons $f_3(L) = h$. La partie décroissante est similaire au symétrique de la partie croissante

$$\begin{cases} f_4(L) &= h - C_2(L-L_4)^2 & \text{si } L \in [L_4, L_5] \\ f_5(L) &= C_2(L-L_6)^2 & \text{si } L \in [L_5, L_6] \end{cases}$$

où $C_2 = h/[2(L_5 - L_4)^2] = h/[2(L_{max}/n_0 + l_{gene})^2]$. Comme chaque profil atomique a une aire de $1/n_0$, le profil agrégé, qui est la moyenne des profils atomiques, a aussi une aire de $1/n_0$. Cette contrainte donne $h = 1/(n_0(L_5 - L_2)) = 1/(n_f(L_{max} - L_{min}))$, puis $C_1$ et $C_2$.

Cette approximation est bonne si la largeur des profils est constante, ce qui n'est pas le cas, puisqu'elle dépend de la taille $L_0$ de l'intergénique initial, $L_0$ variant entre $L_{min}$ et $L_{max}$. Cependant, comme l'aire des profils est constante, l'approximation est quand même très bonne en pratique : le plateau reste le même par compensation des aires des triangles ajoutés et retirés dans la somme, mais la partie descendante est plus large que la partie ascendante et n'est pas quadratique. En fait, comme $L_5 - L_4 > L_2 - L_1$, $C_1$ et $C_2$ ne seront pas égales, donc en utilisant les expressions ci-dessus, on reste dans une approximation quadratique mais on arrive à avoir une partie décroissante qui est plus large que la partie ascendante. Pour conclure, on obtient une bonne approximation du profil réel avec des expressions polynomiales de degré 2 et de classe $C^1$.

Si $L_4 < L_3$, le plateau du profil agrégé disparaît et les parties croissantes et décroissantes se chevauchent. Cependant, on peut ajuster les expressions pour prendre en compte ce chevauchement, qui correspond toujours à une forme d'intégration glissante. Par exemple, si $L_1 \leq L_2 \leq L_4 \leq L_3 \leq L_5 \leq L_6$ (figure III.12) :

$$\begin{cases} f_1(L) &= C_1(L-L_1)^2 & \text{si } L \in [L_1, L_2] \\ f_2(L) &= h - C_1(L-L_3)^2 & \text{si } L \in [L_2, L_4] \\ f_3(L) &= h - C_1(L-L_3)^2 - C_2(L-L_4)^2 & \text{si } L \in [L_4, L_3] \\ f_4(L) &= h - C_2(L-L_4)^2 & \text{si } L \in [L_3, L_5] \\ f_5(L) &= C_2(L-L_6)^2 & \text{si } L \in [L_5, L_6] \end{cases}$$

**Cas où la densité finale en gène est faible (LL_low_final_density, $n_0 > 1$, $0 < n_f < n_0$, $L_3 > L_5$)** Si $L_3 > L_5$, les profils en triangle sont très proches les uns des autres et l'approximation par intégration glissante fonctionne mal. Nous allons utiliser une autre approximation dans ce case.

Entre $L_4$ et $L_2$, tous les triangles sont dans leur partie strictement croissante. On en déduit que le profil agrégé doit lui aussi croître linéairement. Appelons $s$ la pente moyenne

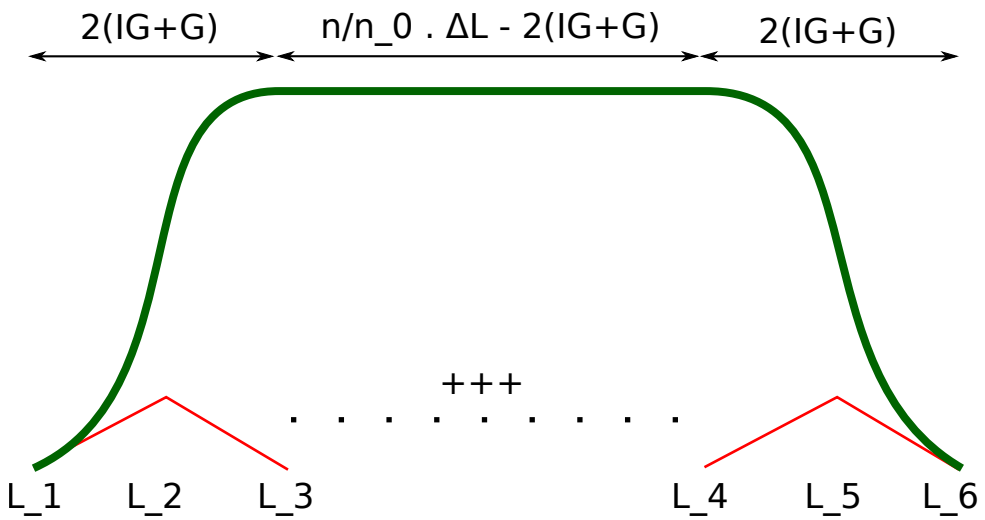


Figure III.11 – Profil agrégé pour les délétions sur la ligne $n_f$ pour des transitions de départ sur la ligne $n_0$ entre $L_{min}$ et $L_{max}$ quand les profils extrêmes ne se chevauchent pas.

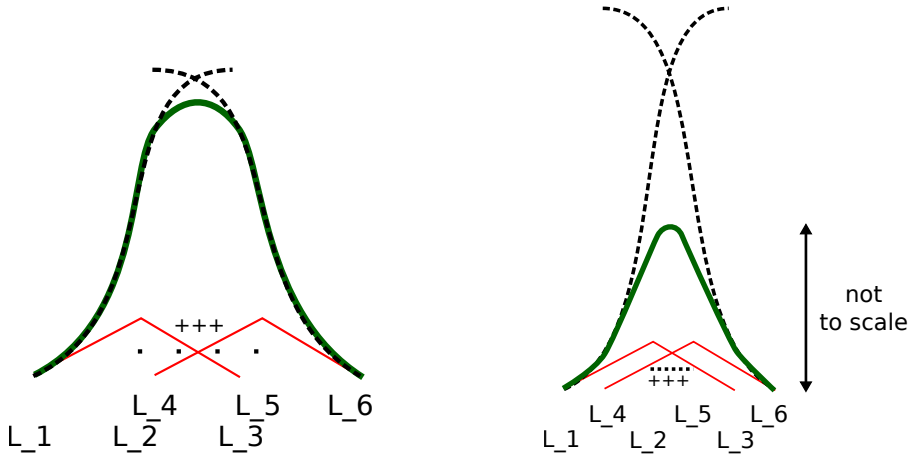


Figure III.12 – Profil agrégé pour les délétions sur la ligne $n_f$ pour des transitions de départ sur la ligne $n_0$ entre $L_{min}$ et $L_{max}$ quand les profils extrêmes commencent à se chevaucher.

de tous les profils agrégés. Entre $L_5$ et $L_3$, tous les profils atomiques sont dans leur partie strictement décroissante, donc le profil agrégé doit décroître linéairement, de pente exactement opposée $-s$ (III.13).

Comme pour le cas général, nous allons compléter les parties manquantes en cherchant une approximation en polynômes de degré 2 et globalement de classe $C^1$. Cette contrainte permet de résoudre toutes les inconnues sauf une (qui s'obtient en fait en renormalisant l'aire à $1/n_0$). Posons $N = s/2$, $IGG_{min} = L_{min}/n_0 + l_{gene}$ and $IGG_{max} = L_{max}/n_0 + l_{gene}$, nous obtenons

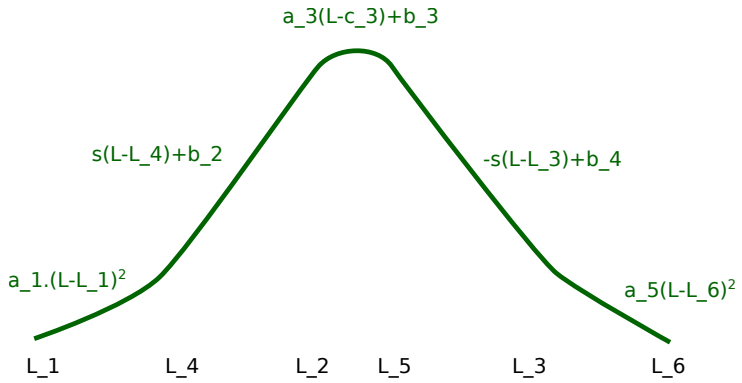


Figure III.13 – Deuxième technique d'approximation : quand la densité finale de gène est très faible, le chevauchement des triangles est tel que sur certaines portions, tous les triangles sont en train de croître ou décroître simultanément. Comme le profil agrégé représente la moyenne des triangles, il doit lui aussi croître ou décroître linéairement sur ces portions.

$$\begin{cases} f_1(L) &= N\frac{(L-L_1)^2}{|L_4-L_1|} & \text{si } L \in [L_1, L_4] \\ f_2(L) &= N(2L-(L_1+L_4)) & \text{si } L \in [L_4, L_2] \\ f_3(L) &= N\left(IGG_1 + IGG_2 - \frac{L_5-L_2}{2} - \frac{2\left(L-\frac{L_5+L_2}{2}\right)^2}{L_5-L_2}\right) & \text{si } L \in [L_2, L_5] \\ f_4(L) &= N(L_3+L_6-2L) & \text{si } L \in [L_5, L_3] \\ f_5(L) &= N\frac{(L-L_6)^2}{L_6-L_3} & \text{si } L \in [L_3, L_6] \end{cases}$$

La démonstration est donée dans l'annexe A, section 3.3.

**Cas particuliers** Quand les profils qu'il faut agréger ne sont pas des triangles, il faut développer d'autres algorithmes, selon le principe des deux algorithmes ci-dessus. Ces algorithmes sont expliqués dans l'annexe A, section 3.3. Comme ils concernent des cas exceptionnels, nous ne les traiterons pas ici et nous contentons de donner leur nom :

- LL_no_gene_loss : cas où aucun gène n'est perdu.
- LL_no_gene_loss_low_final_density : cas où aucun gène n'est perdu et la densité finale en gènes est faible.
- LL_empty : cas où le génome initial ne contient pas de gène.

**Calcul pratique des approximations : utilisation des cumulatives** En moyennant les triangles nous avons calculé la somme

$$\frac{1}{\Delta L(x_0)} \sum_{L_0=L_{min}(x_0)}^{L_{max}(x_0)} \Pr\left[(L_0, n_0) \to (L, n_f)\right]$$

qui exprimée en fonction de la ligne de départ $n_0$, la ligne d'arrivée $n_f$ et la quantité de non codant finale $L$. C'est cette quantité qui est donnée par les différentes fonctions $f_i(L)$ calculées ci-dessus. En pratique, quand on connaît la subdivision finale $(x_f, y_f)$, on voudra calculer

$$\Pr\left[(x_0, n_0) \to (x_f, n_f)\right] \stackrel{\text{def.}}{=} \sum_{L_f=L_{min}(x_f)}^{L_{max}(x_f)} \frac{1}{\Delta L(x_0)} \sum_{L_0=L_{min}(x_0)}^{L_{max}(x_0)} \Pr\left[(L_0, n_0) \to (L_f, n_f)\right]$$

Autrement dit, on aimerait sommer les $f_i(L)$ entre les points $L_{min}(x_f)$ et $L_{max}(x_f)$. Pour éviter de faire une telle somme, il est plus facile d'utiliser les cumulatives des profils moyens donnés ci-dessus (comme on est dans une approximation continue, c'est plus cohérent également). Par exemple, dans l'approximation LL, on peut cumuler les termes positifs du profil moyen :

$$\begin{cases} P(L) &= 0 & \text{si } L < L_1 \\ P(L) &= C_1 \frac{(L-L_1)^3}{3} & \text{si } L \in [L_1, L_2] \\ P(L) &= h(L-L_2) + C_1 \frac{(L_3-L)^3}{3} & \text{si } L \in [L_2, L_3] \\ P(L) &= h(L-L_2) & \text{si } L > L_3 \end{cases}$$

et les termes négatifs du profil moyen d'autre part

$$\begin{cases} N(L) &= 0 & \text{si } L < L_4 \\ N(L) &= C_2 \frac{(L-L_4)^3}{3} & \text{si } L \in [L_4, L_5] \\ N(L) &= h(L-L_5) + C_2 \frac{(L_6-L)^3}{3} & \text{si } L \in [L_5, L_6] \\ N(L) &= h(L-L_5) & \text{si } L > L_6 \end{cases}$$

la contribution moyenne des transitions arrivant entre $L_{min}(x_f)$ est $L_{max}(x_f)$ vaut alors simplement $[P(L_{min}(x_f)-0.5)-N(L_{min}(x_f)-0.5)]+[P(L_{max}(x_f)+0.5)-N(L_{max}(x_f)+0.5)]$. Les expressions des cumulatives pour les autres algorithmes sont données en annexe A, section 3.3).

### 3.4.2 Agrégation le long des colonnes de départ (approximation RL, rectangle vers ligne)

Dans les algorithmes présentés ci-dessus, il s'agissait de prendre une ligne de départ (c'est-à-dire des génomes ayant tous le même nombre de gène initialement) et d'agréger les transitions qui allaient vers la même ligne d'arrivée (c'est-à-dire des génomes ayant le même nombre de gènes après une délétion). Le but initial de notre travail est de moyenner les transitions issues de la subdivision $(x, y)$. On pourrait appliquer les algorithmes LL pour chaque ligne de la subdivision (et plus généralement de l'espace de calcul). On a alors une complexité quadratique avec le nombre maximal de gènes de l'espace : pour chaque ligne de l'espace $n_0$ il faut appliquer les algorithmes pour les $n_0 + 1$ lignes d'arrivées possibles, soit $1 + 2 + 3 + ... + n_{max} + 1 = (n_{max} + 1)(n_{max} + 2)/2$ fois. Cette complexité devient donc assez lourde quand on veut agrandir l'espace de manière conséquente.

Comme les profils de l'algorithme LL se ressemblent beaucoup, on est donc tentés de réduire la complexité en essayant d'agréger le plus de profils venant de la subdivision $(x, y)$. Cela implique de prendre les profils issus de lignes différentes, donc ayant des nombres de gènes initiaux différents, et de les additionner analytiquement. Nous proposons un algorithme (que nous appelons RL, rectangle vers ligne) qui fonctionne pour une partie des profils générés par l'algorithme LL général, car ces profils ont en commun d'avoir un plateau qui a une propriété particulière. Mathématiquement, nous pouvons relier l'algorithme RL aux densités obtenues via l'algorithme LL grâce à l'équation (III.5) donnée dans l'explication générale du principe de l'agrégation « rectangle à ligne », page 122.

Notons que cet algorithme RL, que nous allons détailler ci-dessous, ne pourra pas être utilisé pour les transitions pour lesquelles l'algorithme LL donne des profils moyens sans plateau ou celles qui ont été obtenues via des algorithmes autres que LL. Les transitions pour lesquelles l'algorithme RL ne s'applique devront être agrégées naïvement (c'est-à-dire sommées par ordinateur) ligne par ligne.

Pour commencer, nous nous débarrassons de la partie quadratique des profils de l'approximation LL de façon à obtenir une fonction constante par morceaux. Il existe plusieurs solutions, comme montré sur la figure III.14. Les deux simplifications utilisent les symétries du profil de manière à conserver l'aire totale. La simplification 1 a l'avantage de donner une fonction très simple puisqu'il n'y a plus qu'un seul plateau. Le problème, c'est que le support de la densité peut être assez largement réduit et qu'on perd l'asymétrie du profil initial : la partie décroissante est en réalité plus large que la partie décroissant. Nous avons préféré la simplification 2, qui permet une légère asymétrie et fonctionne très bien en pratique : elle bouge au maximum 1/10 de la distribution (quand le plateau du profil LL a une longueur nulle) et, à mesure que le plateau s'élargit, l'erreur commise décroît rapidement.

Le problème que nous avons à résoudre est le suivant : pour une ligne finale donnée $n_f$, nous devons agréger les $\Delta n$ profils issus des lignes $n_{min}$ à $n_{max}$. Dans le paragraphe explicitant l'algorithme LL, nous avons montré que la hauteur du plateau vaut $h = 1/(n_f(L_{max} - L_{min}))$ : *elle ne dépend pas de* $n_0$. Cela signifie que pour une valeur de $L$ donnée sur la ligne finale $n_f$, chacun des $\Delta n$ profils peut contribuer de trois façons : il peut valoir 0, $h/2$ et $h$. Il s'agit donc d'un problème de comptage : pour connaître la contribution moyenne, il suffit de compter combien de profils valent 0, combien valent $h/2$ et combien valent $h$.

Nous posons donc la question suivante : pour la valeur $L$ sur la ligne $n_f$, combien de profils sont actuellement sur leur premier plateau (entre $L'_1$ et $L'_2$), le plateau principal (entre $L'_2$ et $L'_3$) et le dernier plateau (entre $L'_3$ et $L'_4$) ?

Prenons le profil issu de la ligne $n_0$ entre $L_{min}$ et $L_{max}$ :

$$L'_1 = \frac{n_f}{n_0} L_{min} + (l_{gene} - 1) - \frac{1}{3}\left(\frac{L_{min}}{n_0} + (l_{gene} - 1)\right)$$

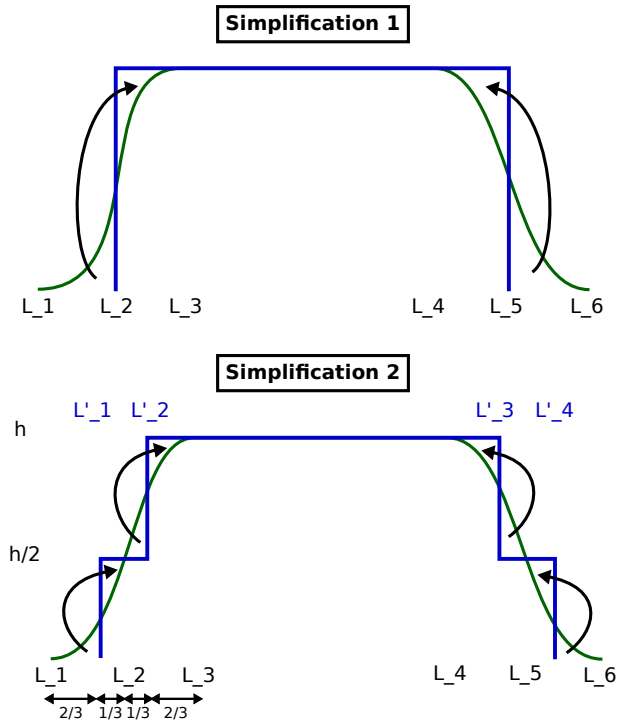


Figure III.14 – Simplifications possibles du profil donné par l'algorithme LL. À partir des points $L_1$, $L_2$, $L_3$, $L_4$, $L_5$, $L_6$ correspondant aux points caractéristique du profil, nous définissons un profil simplifié constant par morceaux. Dans la simplification 1, le support est donné par $[L_2, L_5]$. Dans la simplification 2, nous utilisons les propriétés de la croissance quadratique pour définir $L'_1$, $L'_2$, $L'_3$ et $L'_4$. Par exemple, $L'_1$ est le point tel que l'aire du profil original (en vert) entre $L_1$ et $L_2$ est égale à l'aire donnée par le rectangle de base $[L'_1, L_2$ et dont la hauteur est, en $L_2$ la même que celle du profil initial. On montre que c'est le cas la distance de $L'_1$ à $L_2$ vaut $1/3$ de la distance de $L_1$ à $L_2$.

d'où

$$L > L'_1 \quad \Leftrightarrow \quad L > \frac{n_f - 1/3}{n_0} L_{min} + \frac{2(l_{gene} - 1)}{3} \quad \Leftrightarrow \quad n_0 > \frac{n_f - 1/3}{L - \frac{2(l_{gene}-1)}{3}} L_{min}$$

Imaginons que $L$ est un curseur qui balaye le profil de gauche à droite. Si la condition ci-dessus est remplie, nous dirons que le point caractéristique $L'_1$ a été dépassé, sinon $L$ est toujours à gauche de $L'_1$ et l'atteindra plus tard. Tous les profils issus des lignes entre $n_{min}$ et $n_{max}$ qui vérifient cette condition ont donc dépassé le point caractéristique $L'_1$. Dans la suite, on nomme $N_{L'_1}(L) \in [0, \Delta n]$ le nombre de profils qui ont dépassé le point caractéristique $L'_1$ quand le curseur pointe sur $L$.

$$N_{L'_1}(L) = \begin{cases} 0 & \text{si } L < \frac{n_f - 1/3}{n_{max}} L_{min} + \frac{2(l_{gene}-1)}{3} \\ \Delta n & \text{si } L > \frac{n_f - 1/3}{n_{min}} L_{min} + \frac{2(l_{gene}-1)}{3} \\ n_{max} - \left\lceil \frac{n_f - 1/3}{L - \frac{2(l_{gene}-1)}{3}} L_{min} \right\rceil + 1 & \text{sinon} \end{cases}$$

De façon similaire, on définit $N_{L'_2}(L)$, $N_{L'_3}(L)$ and $N_{L'_4}(L)$ (voir annexe A, section 3.3.7).

Pour savoir combien de profils sont sur le premier plateau, il faut que le curseur pointe entre $L'_1$ et $L'_2$. Le nombre de profils sur le premier plateau (de hauteur $h/2$) est donc

$$N_{[L'_1,L'_2]}(L) = N_{L'_1}(L) - N_{L'_2}(L)$$

sur le plateau principal (de hauteur $h$)

$$N_{[L'_2,L'_3]}(L) = N_{L'_2}(L) - N_{L'_3}(L)$$

et sur le dernier plateau (de hauteur $h/2$)

$$N_{[L'_3,L'_4]}(L) = N_{L'_3}(L) - N_{L'_4}(L)$$

Comme nous sommes dans une approximation continue et que les contributions seront agrégées par intégration à la fin, l'appartenance des points caractéristiques au plateau de droite ou de gauche est indifférente, nous nous permettons donc d'utiliser des notations en intervalles fermés (juste pour le côté esthétique) qui peuvent paraître incohérentes à première vue. Le profil moyen au point $L$ vaut

$$\begin{aligned} C(L) &= \frac{1}{\Delta n}\left(N_{[L'_1,L'_2]}(L)\frac{h}{2} + N_{[L'_2,L'_3]}(L)h + N_{[L'_4,L'_3]}(L)\frac{h}{2}\right) \\ &= \frac{h}{2\Delta n}\left(N_{L'_1}(L) + N_{L'_2}(L) - N_{L'_3}(L) - N_{L'_4}(L)\right) \end{aligned}$$

En termes de probabilités, cette expression correspond à

$$C(L) = \frac{1}{\Delta n(y_0)} \sum_{n_0=n_{min}(y_0)}^{n_{max}(y_0)} \frac{1}{\Delta L(x_0)} \sum_{L_0=L_{min}(x_0)}^{L_{max}(x_0)} \Pr\left[(L_0,n_0) \to (L,n_f)\right]$$

Pour obtenir la contribution à la subdivision d'arrivée, il faudra intégrer cette expression pour $L$ variant entre $L_{min}(x_f) - 0.5$ et $L_{max}(x_f) + 0.5$. Comme on intègre une fonction en escalier, on peut convertir l'intégrale en une somme discrète, qui peut être calculée analytiquement par un développement asymptotique (annexe A, section 3.3.7). On peut donc agréger les probabilités provenant de tout le rectangle $(x, y)$ vers une ligne donnée $(x_f, n_f)$, tant que la conditions d'existence du plateau est remplie pour toutes les lignes de départ.

### 3.4.3 Bilan : calcul de l'agrégation des transitions en pratique

Finalement, nous disposons de six algorithmes pour agréger les transitions. Le plus efficace, l'algorithme RL, peut être utilisé pour calculer les contributions provenant d'un rectangle vers une ligne spécifique. Cependant, il ne fonctionne bien que quand les profils agrégés provenant de chaque ligne du rectangle de départ possèdent un plateau. Dans les

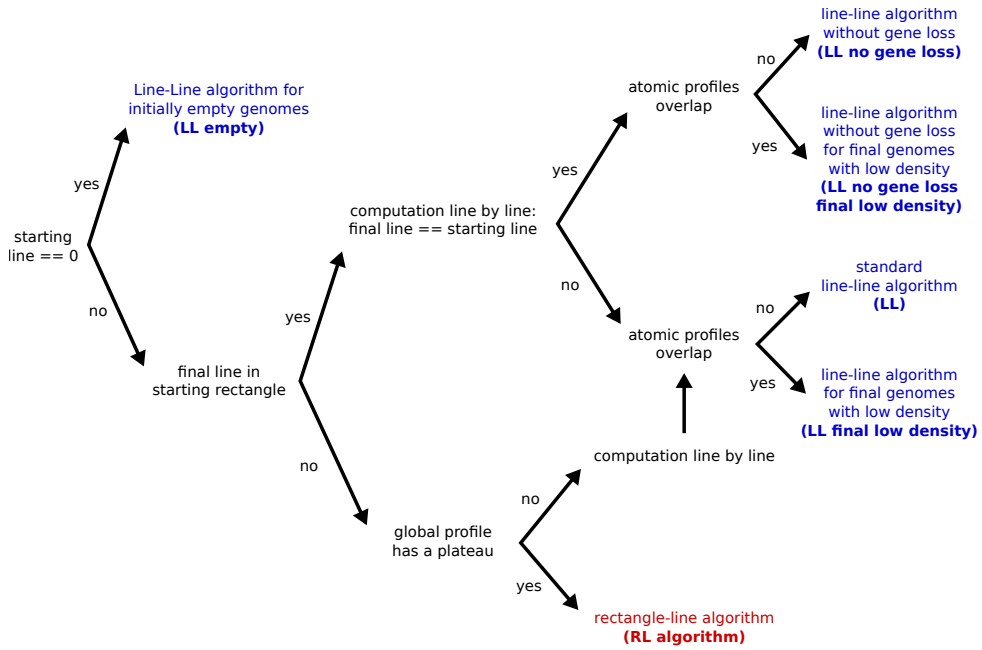


Figure III.15 – Vue schématique de l'utilisation des algorithmes pour l'agrégation des délétions.

autres cas, on est forcés de décomposer l'agrégation ligne par ligne, en utilisant l'un des algorithmes marqués LL. La figure III.15 résume comment en pratique on décide d'utiliser l'algorithme approprié.

Supposons que nous cherchions à agréger les transitions venant de $(x, y)$, délimité par $L_{min}$ et $L_{max}$ en largeur et $n_{min}$ et $n_{max}$ en hauteur. L'idée est simple : pour chaque ligne d'arrivée $n_f \in [0, n_{max}]$, nous répertorions les lignes qui vérifient la condition de plateau et nous utilisons l'algorithme RL pour ces lignes-là. Pour les autres, nous utilisons l'algorithme LL approprié tour à tour. La condition de plateau n'a jamais été donnée explicitement jusqu'ici. Elle est donnée dans la section 3.3.8 de l'annexe A, accompagnée d'une version un peu plus détaillée de l'utilisation de chaque algorithme.

## 3.5 Agrégation des duplications

Le cas des duplications est en fait très similaire à celui des délétions. Si on fait abstraction pour l'instant de l'inactivation d'un gène à l'insertion, le problème est quasiment toujours le même. Le profil des transitions venant de $(L_0, n_0)$ est une suite de triangles identiques à ceux des délétions (figure III.9). Si on raboute les cas particuliers des lignes $n_0$, $2n_0 - 1$ et $2n_0$, il y a également $n_0$ profils triangulaires en tout, donc chaque triangle a une aire $1/n_0$. On va donc pouvoir utiliser exactement les mêmes algorithmes que pour les délétions.

### 3.5.1 Algorithmes utilisés

Supposons d'abord qu'on fasse une agrégation ligne vers ligne de transitions issues de la ligne $n_0$, entre $L_{min}$ et $L_{max}$, à destination de la ligne $n_f$.

Comme les profils atomiques sont des triangles identiques à ceux des délétions, pour $n_0 < n_f < 2n_0$, nous reprendrons exactement les algorithmes LL utilisés dans la section 3.4.1, aux mêmes conditions. Sur la ligne $n_f = 2n_0$ pour les duplications, la troncature est similaire à celle qui existe pour la ligne $n_f = n_0$ dans le cas des grandes délétions. Nous pouvons donc appliquer les algorithmes LL_no_gene_loss et sa variante LL_no_gene_loss_low_final_density.

Il reste donc la ligne $n_f = n_0$ qui correspond au cas où aucun gène n'a été dupliqué. Pour éviter d'avoir à développer un algorithme spécifique à ce cas-là, nous utilisons le fait que les profils de délétion et de duplications sont presque complémentaires sur la ligne $n_f = n_0$. Pour obtenir le profil des duplications, on calcule le profil agrégé complet au moyen d'un algorithme $LL$ pour triangles complets et on retranche le profil obtenu via les algorithmes LL_no_gene_loss pour les délétions. Comme il s'agit d'approximations, on peut se retrouver avec des valeurs négatives dans quelques rares cas, notamment quand les deux profils se compensent exactement en théorie. Dans de tel cas, on force le profil à valoir 0 et on renormalise le reste pour que l'aire soit bien $1/n_0$.

Si $n_0 = 0$, l'algorithme LL_empty des délétions ne convient plus exactement. Il faut toujours sommer des profils uniformes mais l'ensemble de sommation n'est plus le même. Les détails sont donnés dans l'annexe A, section 3.4.1.

Enfin, il est possible d'utiliser l'algorithme RL, toujours à la condition que les profils issus de chaque ligne de départ aient un plateau.

### 3.5.2 Bilan : calcul de l'agrégation des transitions en pratique

On suit le même principe que pour les délétions : à cause de son efficacité, l'algorithme RL est utilisé en priorité, sinon on applique les algorithmes LL ligne à ligne (figure III.16). Par contre, une fois qu'on a appliqué ces algorithmes, on a en fait simplement déterminé le contenu du segment copié, mais pas encore pris en compte la possibilité d'inactiver un gène à l'insertion. Normalement, pour le génome $(L_0, n_0)$, quel que soit le segment copié, la probabilité d'insérer le segment dans un gène est $n_0 l_{gene}/(L_0 + n_0 l_{gene})$. Cette probabilité varie donc avec le point de départ, ce qui complique significativement l'agrégation.

Nous allons éviter la difficulté en utilisant une simplification. Nous prenons la probabilité moyenne $p$ d'inactiver un gène sur tout le rectangle de départ $(x, y)$. Comme expliqué plus haut, nous ignorons la possibilité d'inactiver un gène et appliquons les algorithmes RL et LL. Sans considérer pour l'instant l'inactivation de gène, on peut calculer la transition moyenne de la subdivision $(x_0, y_0)$ vers la ligne $n_f$ de la subdivision $(x_f, y_f)$. Soit $C$

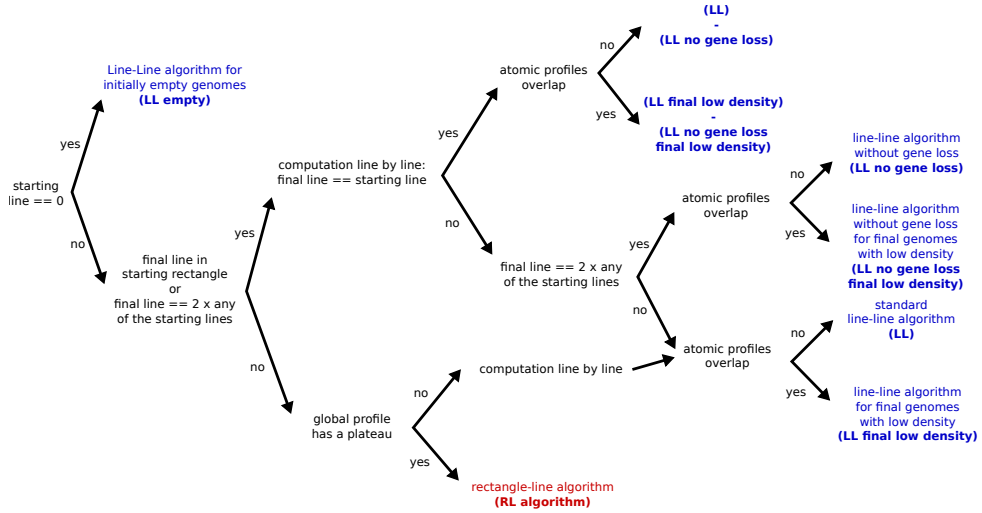


Figure III.16 – Vue schématique des algorithmes utilisés pour l'agrégation des duplications.

la valeur de la transition moyenne. Nous approximons l'inactivation de gène en disant que $(1-p)C$ individus vont effectivement vers la ligne $n_f$ entre $L_{min}(x_f)$ et $L_{max}(x_f)$, tandis que les $pC$ individus restants vont sur la ligne $n_f - 1$ entre $L_{min}(x_f) + l_{gene}$ et $L_{max}(x_f) + l_{gene}$. On a donc une partie des individus qui se retrouve dans la subdivision d'abscisse $x_f + 1$. En pratique pour que l'estimation soit plus précise, on calcule la part $C_1$ d'individus qui vont entre $L_{min}(x_f)$ et $L_{max}(x_f) - l_{gene}$, car on sait qu'ils restent sur la subdivision d'abscisse $x_f$, même après inactivation éventuelle d'un gène, et la part $C_2$ qui vont entre $L_{max}(x_f) - l_{gene} + 1$ et $L_{max}$ qui nous permet de déduire la quantité $pC_2$ d'individus qui vont dans la colonne $x_f + 1$. Il s'agit bien là d'une approximation. On pourrait légèrement l'améliorer, mais il serait malgré tout difficile d'obtenir une expression exacte et tout cela n'est pas d'une importance primordiale pour le comportement global du modèle.

## 4 Calcul de $M_{(L,n)}$ : approximation de l'exponentielle de $P_{(L,n)}$

Les sections précédentes nous permettent de calculer une version approximative des matrices $P^i_{(L,n)}ns$, $P^s_{(L,n)}del$, etc. (une par type de mutation). En prenant ensuite en compte les taux par base et par génération de chaque type de mutation, on peut en déduire la matrice de transition $P_{(L,n)}$ après exactement une mutation (équation (III.3), 102). Il reste à en déduire la matrice de transition $M_{(L,n)}$ après une *génération*. Nous avons vu à la section 1.2, que pour un état $i_0 = \varphi(L_0, n_0)$, la ligne correspondant dans $M_{(L,n)}$ peut être obtenue de $P_{(L,n)}$ via la formule

$$\mathbf{1}_{i_0} M_{(L,n)} = \mathbf{1}_{i_0} e^{\mu s_0 (P_{(L,n)} - I)}$$

où $s_0 = L_0 + n_0 l_{gene}$ est la taille du génome initial et $\mu$ le taux de mutation total. On peut donc en théorie calculer $M_{(L,n)}$ ligne par ligne en ajustant la valeur de $s_0$ pour chaque état retenu dans l'espace de calcul et en approximant l'exponentielle de matrice.

Cependant, l'espace de calcul a également été subdivisé en rectangles, on ne peut donc pas extraire à proprement parler une valeur de $s_0$ qui correspondra à tous les états d'une subdivision $(x, y)$, nous devrons travailler avec la valeur moyenne $\overline{s}$ des $s_0$ du rectangle. Rappelons que $s_0$ influence le nombre de mutations attendues : en prenant la valeur moyenne $\overline{s}$, nous allons surestimer le nombre de mutations pour un certain nombre de génomes et le sous-estimer pour d'autres.

Pour créer la matrice, nous devons attribuer à chaque subdivision $(x, y)$ un identifiant entier, comme on l'a fait pour les états génomiques atomiques. Pour faire simple, nous posons $R = x + (y-1)x_{max}$. La ligne numéro $R$ de $M_{(L,n)}$ correspond donc à toutes les mutations qui transitions du rectangle $(x, y)$. Nous nommons $N_{mut} = \overline{s}\mu$, l'espérance du nombre de mutation attendues pour un génome de taille $\overline{s}$. Nous cherchons à calculer (en notation Matlab)

$$\begin{aligned} M_{(L,n)}(R,:) &= \left(e^{N_{mut}(P_{(L,n)}-I)}\right)(R,:) = e^{-N_{mut}}\left(e^{N_{mut}P_{(L,n)}}\right)(R,:) \\ &= \left(\frac{I}{e^{N_{mut}}} + \frac{N_{mut}}{e^{N_{mut}}}P_{(L,n)} + \frac{N_{mut}^2}{2e^{N_{mut}}}P_{(L,n)}^2 + \frac{N_{mut}^3}{6e^{N_{mut}}}P_{(L,n)}^3 + ...\right)(R,:) \end{aligned}$$

Le développement en série permet de calculer toute la matrice assez rapidement, sans avoir à refaire tout le calcul pour chaque ligne de départ. Par contre, il faut se méfier de la stabilité numérique (prendre naïvement le développement de $\exp(P_{(L,n)} - I)$ marche en théorie mais pas en pratique) et choisir le bon moment pour arrêter la sommation.

Les seules valeurs qui dépendent de $R$ sont les coefficients $N_{mut}^k/[k!e^{N_{mut}}]$, les matrices du développement sont toujours des puissances de $P_{(L,n)}$, qu'il suffira donc de précalculer une seule fois. Comme la norme de $P_{(L,n)}$ est 1 et tous ses coefficients sont positifs ou nuls, le calcul de ses puissances est stable et peut être renormalisé efficacement à chaque étape. La norme de $P_{(L,n)}^k$ est 1 mais $N_{mut}^k/[k!e^{n_{mut}}]$ est toujours plus petit que 1 donc quand on avance dans la somme, on ajoute des contributions positives de plus en plus faibles. À chaque nouveau terme additionné, les coefficients de $M_{(L,n)}$ restent donc positifs et ses lignes somment à des valeurs qui se rapprochent de 1, leur valeur théorique. On peut donc arrêter la sommation quand on juge que chaque ligne a une valeur suffisamment proche de 1. Plus le nombre de mutations attendues $N_{mut}$ est grand, plus il faudra aller loin dans la somme pour remplir cette condition. Cependant, ces propriétés assurent un calcul stable et assez rapide de $M_{(L,n)}$.

L'algorithme schématique est le suivant :

**Données** : la matrice $P_{(L,n)}$, le nombre de subdivisions $R_{max}$, un vecteur contenant la taille de génome moyenne $\overline{s}(R)$ pour chaque subdivision.
**Résultat** : la matrice $M$, qui correspond à $M_{(L,n)}$ dans le texte.
on pose la matrice (de taille $R_{max} \times R_{max}$) $M = 0$ ;
on calcule le vecteur $N_{mut}(R) = \mu_{tot}\overline{s}(R)$ pour tout $R$;
on calcule $M(R,R) = \exp(-N_{mut}(R))$ pour tout $R$;
on pose la matrice $P_k = P_{(L,n)}$;
on pose l'entier $k = 1$ et le réel $logkfact = \log(k!) = 0$ (pour la stabilité numérique);
**tant que** *la somme sur une des lignes de la matrice $M$ n'est pas assez proche de 1* **faire**
  **si** $k \neq 1$ **alors**
    on met à jour $P_k = P_k \times P_{(L,n)}$ et $logkfact = logkfact + \log(k)$;
  **fin**
  **pour** *toute ligne $R$ de $M$ dont la somme n'est pas assez proche de 1* **faire**
    on calcule $coef = \exp(k \times \log(N_{mut}(R)) - logkfact - N_{mut}(R))$ ;
    on met à jour la ligne $M(R,:) = M(R,:) + coef \times P_k$;
  **fin**
  on met à jour k = k+1;
**fin**

Pour illustrer le problème de stabilité on peut comparer avec le développement suivant :

$$M_{(L,n)}(R,:) = \left( I + N_{mut}(P_{(L,n)} - I) + \frac{(N_{mut}(P_{(L,n)} - I))^2}{2} + \frac{(N_{mut}(P_{(L,n)} - I))^3}{6} + ... \right)(R,:)$$

En effet, les puissances des matrices sont de norme 0, car elles sont calculées à partir de $P_{(L,n)} - I$. Cela veut dire que chaque ligne somme à 0, donc est composée de coefficients positifs et négatifs qui se compensent exactement. Dans le développement utilisé en réalité, les coefficients sont forcément compris entre 0 et 1, alors qu'ici les coefficients peuvent être positifs ou négatifs et leur valeur peut devenir arbitrairement grande. La compensation des coefficients est en pratique de plus en plus précaire. On obtient assez rapidement des valeurs de coefficients qui sont absurdes (négatives ou supérieures à un), surtout quand la matrice $P_{(L,n)}$ est assez grosse.

# 5 Conclusion

La prise en compte d'une structure simple en ADN codant, modélisé par le nombre $n$ de gènes de longueur $l_{gene}$, et en ADN non codant, modélisé par le nombre $L$ de paire de bases non codantes, permet d'avoir un espace de calcul assez « réduit » avec seulement deux dimensions. Pour simuler le modèle, nous devons nous restreindre à un espace fini mais suffisamment grand pour que la population évolue loin des bords. On fixe donc en pratique un nombre de gènes maximal $n_M$ et un nombre de bases non codantes maximal $L_M$. Le nombre d'états génomiques possibles au sein de l'espace de calcul est alors $(L_M+1)(n_m+1)$, soit environ $L_M n_M$ (ce qui nous permettra d'omettre les « +1 » dans la suite).

Si on veut autoriser des fluctuations de nombre de gènes et de non codant assez grandes, on est déjà confronté au problème que le nombre d'états possibles est grand. Par exemple, pour $L_M = 1Mb$ et $n_M = 1000$, il y a déjà environ un milliard d'états possibles. De plus, si on veut simuler des structures proches des bactéries, ce choix ne remplit pas une condition importante : la population doit évoluer loin des bords. Si on multiplie $L_M$ et $n_M$ par 1000 pour « éloigner les bords », on passe déjà à $10^{15}$ (un million de milliards) d'états. Enfin, nous souhaitons calculer les transitions d'état à état pour chaque type de mutation. Chaque état peut, en théorie, aller vers n'importe lequel des $L_M n_M$ autres états de l'espace : il y a $(L_M n_M)^2$ transitions possibles. En pratique, pour un type de mutation donné, la plupart de ces transitions ont une probabilité nulle, notamment dans le cas des mutations locales. Pour les duplications et les grandes délétions, nous avons montré que le nombre de transitions atomiques partant d'un état atomique $(L, n)$ était supérieur à $4(L + nl_{gene})$. Si on somme sur tous les états possibles, cela fait de l'ordre de $O(L_M n_M(L_M + n_M l_{gene}))$ transitions. Si on reprend l'exemple de simulations de bactéries, cela signifie que le nombre de transitions est de l'ordre de $10^{24}$ transitions : on ne peut pas se permettre de les calculer une à une, il faut réduire le nombre d'états du système.

Nous avons introduit deux types de subdivisions, en échelle linéaire et en échelle logarithmique. Chaque subdivision comprend un ensemble d'états génomiques dits « atomiques ». Comme on ne peut pas savoir quels états atomiques sont occupées au sein d'une subdivision, nous faisons l'hypothèse que les génomes sont répartis uniformément. Plutôt que de considérer des transitions atomiques, nous considérons donc des transitions entre subdivisions calculées à partir des transitions atomiques. Néanmoins, nous ne pouvons pas envisager d'obtenir les transitions entre subdivisions à partir des transitions atomiques directement. En effet, on pourrait être tenté d'agréger les transitions atomiques une à une pour obtenir les transitions entre subdivisions, mais cela reviendrait à calculer puis sommer les $O(L_M n_M(L_M + n_M l_{gene}))$ transitions ($10^{24}$ pour simuler des bactéries). Il faut donc calculer analytiquement les probabilités de subdivision à subdivision.

Pour les mutations locales, ces calculs sont assez rapides. Pour les grandes délétions et les duplications, nous n'avons pas trouvé d'approximation analytique pour les probabilités de subdivision à subdivision. Nous avons donc proposé des solutions intermédiaires, avec des calculs d'une ligne d'une subdivision vers une ligne d'une autre subdivision (algorithmes LL) voire d'une subdivision entière vers une ligne d'une autre subdivision (algorithme RL). Quand on utilise l'approximation LL pour une ligne de départ $n_0$ et pour une subdivision d'abscisse $x_0$ donnés, il faut calculer les transitions vers les $2n_0+1$ lignes d'arrivée possibles. Les transitions sont ensuite réparties automatiquement dans les subdivisions d'arrivée : le nombre d'utilisations de l'algorithme ne dépend pas du nombre de subdivisions à l'arrivée, on calcule une fois la densité puis on la découpe. S'il y a $x_{max}$ subdivisions le long de l'axe du non codant, cela fait $x_{max}(2n_0 + 1)$ utilisations de l'algorithme LL pour la ligne $n_0$. En tout, on utilise l'algorithme $O(x_{max} n_M^2)$ fois.

- En échelle linéaire, $x_{max} = 100$ est un paramètre fixe qui est limité par le fait que les matrices doivent entrer dans la mémoire vive.
- En échelle logarithmique, $L_M$ est le nombre maximal de bases non codantes dans la

subdivision $x_{max}$, on a à peu près $1000 \times 2^{x}_{max-1} = L_M$, soit $x_{max} = O(\log L_M)$.

On a donc des complexités qui se réduisent à $O(x_{max} n_M^2)$ en échelle linéaire et $O(n_M^2 \log L_M)$. En reprenant l'exemple des bactéries ($L_M = 10^6 \times 1000$ et $n_M = 10^3 \times 1000$), cela fait de l'ordre de $10^{14}$ transitions à calculer en échelle linéaire et $10^{13}$ transitions à calculer en échelle logarithmique. Cela réduit déjà considérablement le nombre de transitions à agréger.

L'algorithme RL permet d'éliminer la complexité quadratique en $n_M$. Plaçons-nous dans un cadre idéal où l'algorithme RL peut être utilisé pour toutes les subdivisions de départ vers toutes les lignes d'arrivée. Considérons la subdivision de départ $(x_0, y_0)$. Le nombre de gène maximal au sein de cette subdivision vaut $n_{max}(y_0)$. Le nombre de lignes d'arrivée possibles pour les duplications et les délétions est donc $2 * n_{max}(y_0) + 1$. Pour la ligne d'ordonnée $y_0$, on appliquera l'algorithme RL $O(x_{max} n_{max}(y_0))$ fois. Si on somme sur tous les $y_0$, on a donc une complexité de l'ordre de $O(x_{max} \sum_{y_0} n_{max}(y_0))$.

- En échelle linéaire, $n_{max}(y_0) \simeq y_0 \Delta n$ (où $\Delta n$ est la hauteur d'une subdivision) donc $\sum_{y_0} n_{max}(y_0) = O(y_{max}^2 \Delta n) = O(n_M y_{max})$. $y_{max} = n_M / \Delta n$ est le nombre de subdivisions le long de l'axe du codant, en pratique, comme pour $x_{max}$, on prend $y_{max} = 100$ pour des raisons de mémoire.
- En échelle logarithmique, $n_{max}(y_0) \simeq 2^{y_0 - 1}$ donc $\sum_{y_0} n_{max}(y_0) = O(2^{y_{max}}) = O(2 n_M)$. En effet, $n_M$ étant le nombre de gènes maximal de la subdivision $y_{max}$, on a à peu près $2^{y_{max}-1} = n_M$.

On a donc des complexités de $O(x_{max} y_{max} n_M)$ en échelle linéaire et de $O(\log L_M n_M)$ en échelle logarithmique. Pour simuler nos bactéries ($L_M = 10^6 \times 1000$ et $n_M = 10^3 \times 1000$), cela fait de l'ordre de $10^{10}$ transitions à calculer en échelle linéaire et $10^7$ transitions à calculer en échelle logarithmique. On gagne un facteur supplémentaire en échelle logarithmique car les subdivisions le long de l'axe du codant deviennent de plus en plus grosses, ce qui est idéal pour l'utilisation de l'algorithme RL. Cependant, en réalité, il y a un facteur dans le grand $O()$ qui peut en pratique être comparable à $y_{max}$, on est vraisemblablement plus proche de $10^8$ en réalité.

Le passage de l'agrégation naïve (impossible à réaliser en pratique) à l'agrégation analytique via LL permet donc d'entrer dans des ordres de grandeurs plus réalistes mais quand même très longs en pratique. De plus, la dépendance quadratique en $n_M$, le nombre total de gènes dans l'espace, empêche d'étendre l'espace de calcul. En utilisant l'algorithme RL, on obtient des ordres de grandeurs qui permettent un calcul sur un ordinateur de bureau en quelques jours ou quelques semaines. La complexité devient linéaire en $n_M$, ce qui permet d'envisager des extensions, notamment pour simuler des génomes de l'ordre de grandeur des mammifères. En pratique, le programme utilise tout de même l'algorithme LL pour une partie des transitions, on a donc une complexité légèrement plus grande que linéaire.

# Chapitre IV

# Étude de l'évolution de la longueur du génome en fonction des taux de mutation et de la force de la sélection

> Nous estimons posséder la science d'une chose d'une manière absolue, quand nous croyons que nous connaissons la cause par laquelle la chose est, que nous savons que cette cause est celle de la chose, et qu'en outre il n'est pas possible que la chose soit autre qu'elle n'est.
>
> Aristote, *Seconds Analytiques I*, 2, 71b, 9-11.

Maintenant que les principes et les approximations qui régissent les implémentations ont été présentés, nous pouvons passer aux résultats des simulations et les comparer avec les prédictions théoriques. L'étude théorique, menée en l'absence de sélection darwinienne, a suggéré que la dynamique spontanée des duplications et des délétions empêchait le génome de croître infiniment même si l'on suppose que les duplications sont deux fois plus fréquentes que les délétions, et même si les petites insertions sont 10, 100 voire 1000 fois plus fréquentes que les petites délétions (par exemple du fait de l'activité d'éléments transposables). Les simulations vont nous permettre de tester si cette propriété de non-explosion reste valable en présence de sélection darwinienne, dans le cas le pire où la sélection favoriserait les génomes les plus grands. Nous allons implémenter dans les simulations une sélection hypothétique qui favorise directement les génomes ayant le plus de gènes. La fitness d'un état génomique sera donc une fonction strictement croissante et non bornée du nombre de gènes. Bien sûr, en réalité, dupliquer un gène n'est pas toujours

favorable, ou pas immédiatement. C'est tout l'intérêt de la modélisation que de permettre de telles « expériences » impossibles à réaliser à la paillasse et pourtant décisives pour comprendre les effets isolés puis combinés des différents facteurs. Dans le même ordre d'idées, nous simulons par défaut une taille de population infinie, permettant ainsi à tous les états d'être représentés, même avec une très faible densité. Il est en effet plus facile d'atteindre les tailles les plus longues avec une population infinie qu'avec une population finie, ce qui va aussi dans le sens de créer un « pire cas » de simulation pour la propriété de non-explosion. Cela dit, nous proposons également une variante du programme qui simule une population finie.

Après avoir présenté les paramètres par défaut et le déroulement d'une simulation, nous étudierons l'influence des différents taux de mutation sur la taille du génome à l'équilibre, dans le cas d'une population infinie et d'une fitness augmentant de façon logarithmique avec le nombre de gènes. Nous testerons ensuite des fonctions de fitness à croissance plus rapide que le logarithme, et enfin nous examinerons le comportement du modèle lorsque la population est de taille finie. Dans tous les cas, l'objectif n'est pas de simuler des conditions d'évolution les plus réalistes possibles, mais de tester dans des conditions extrêmes la « force de rappel » identifiée lors de l'étude analytique. Jusqu'à quel point peut-elle tenir ?

# 1 Paramètres par défaut et déroulement d'une simulation

Comme les parties techniques du chapitre précédent auront pu rebuter le lecteur souhaitant aller directement « en effet », nous allons résumer ici brièvement les paramètres qui contrôlent les simulations. Nous présentons également le résultat d'une simulation typique, ainsi que les mesures que nous avons retenues pour caractériser la taille du génome en fin de simulation.

## 1.1 Paramètres par défaut du modèle

Le modèle inclut deux variables : le nombre de gènes ($n$) et le nombre de paires de bases non codantes ($L$). Par ailleurs, il y a deux types de paramètres : ceux qui sont liés au modèle initial (mutations et sélection) et ceux qui sont liés au problème de l'implémentation (taille de l'espace et agrégation des transitions). Ils sont listés dans la table IV.1 avec les valeurs par défaut. Il faut noter que certains paramètres composites se déduisent à partir des paramètres donnés dans la table. On pensera notamment au taux de mutation total $\mu = \mu_{ins} + \mu_{sdel} + \mu_{inv} + \mu_{trans} + \mu_{dup} + \mu_{ldel}$ et aux tailles maximales de l'espace de calcul. Si on fait un découpage en échelle linéaire, le nombre de gènes maximal est $y_{max}\Delta n$ et le nombre de paires de bases non codantes maximal est $x_{max}\Delta L$. En échelle logarithmique,

on part d'un plus petit rectangle à l'origine correspondant aux génomes ayant 0 gène et moins de $l_{gene}$ bases de non codant, puis on obtient les subdivisions suivantes en multipliant par 2 la hauteur et/ou la largeur. Le nombre maximal de gènes est $2^{y_{max}-2}$ et le nombre de bases non codantes maximal est $l_{gene}2^{x_{max}-1}$. En échelle logarithmique, pour augmenter la taille maximale simulable, il suffit d'augmenter le nombre de subdivisions. En échelle linéaire, on contrôle à la fois le nombre et la largeur des subdivisions. Comme on l'a expliqué plus haut, on se limite à un maximum de $100 \times 100$ subdivisons. Pour changer la taille maximale simulable, on prend le nombre de subdivisions maximal et on change la largeur des subdivisions.

| Type | Nom du paramètre | Signification | Valeur par défaut |
|---|---|---|---|
| Taux de mutation | $\mu_{ins}$ | Taux de petites insertions (par base et par génération) | $10^{-7}$ |
| | $\mu_{sdel}$ | Taux de petites délétions (par base et par génération) | $10^{-7}$ |
| | $\mu_{inv}$ | Taux d'inversions (par base et par génération) | $10^{-7}$ |
| | $\mu_{trans}$ | Taux de transpositions (par base et par génération) | $10^{-7}$ |
| | $\mu_{dup}$ | Taux de duplications (par base et par génération) | $10^{-7}$ |
| | $\mu_{ldel}$ | Taux de grandes délétions (par base et par génération) | $10^{-7}$ |
| Sélection | $f(n)$ | Fonction de fitness (selon le nombre de gènes $n$) | $\log(n)$ |
| | $l_{gene}$ | Longueur d'une région codante (en paires de bases) | 1000 |
| | $N_{pop}$ | Taille de la population (en nombre d'individus) | $\infty$ |
| Découpage de l'espace : échelle linéaire | $x_{max}$ | Nombre de subdivisions le long de l'axe du non-codant | 100 |
| | $y_{max}$ | Nombre de subdivisions le long de l'axe du codant | 100 |
| | $\Delta L$ | Largeur d'une subdivision le long de l'axe du non-codant (en paires de bases) | $10^6$ |
| | $\Delta n$ | Largeur d'une subdivision le long de l'axe du codant (en nombre de gènes) | $10^3$ |
| Découpage de l'espace : échelle logarithmique | $x_{max}$ | Nombre de subdivisions le long de l'axe du non-codant | 20 |
| | $y_{max}$ | Nombre de subdivisions le long de l'axe du codant | 20 |

Table IV.1 – Paramètres par défaut du modèle.

Les indications données pour le découpage en échelle linéaire sont très théoriques. L'échelle linéaire a été implémentée mais, après quelques simulations typiques, il s'est avéré que, même pour les simulations, l'échelle linéaire n'est pas très bien adaptée au problème. Comme l'échelle logarithmique convient mieux, tous les résultats présentés dans la suite de ce manuscrit seront présentés dans cette échelle.

## 1.2 Préparation d'une simulation

Une simulation contient deux aspects fondamentalement différents : il faut générer la matrice $P_{(L,n)}$ des transitions pour une mutation, puis faire des opérations de calcul matriciel pour obtenir la matrice $M_{(L,n)}$ des transitions pour une génération, la matrice $F$ de fitness et le vecteur d'évolution en codant et non-codant de la population.

**Création d'une bibliothèque de matrices élémentaires** La première tâche – générer $P_{(L,n)}$ – implique beaucoup d'opérations élémentaires et de boucles de calcul. Le langage qui nous semblait le plus adapté dans ce cas est le C++, puisqu'il gère les boucles efficacement et que la structure en objets est très utile pour obtenir un code (presque) lisible et modulaire. Cette partie peut être extrêmement longue à calculer, selon le nombre d'états initiaux qu'il faut agréger. Cependant, nous avons fait la remarque plus haut que $P_{(L,n)}$ s'obtient à partir de matrices élémentaires, une par type de mutation. Nous générons donc une matrice par type de mutation *indépendante du taux de mutation*, qui sera prise en compte dans le calcul de $P_{(L,n)}$. Autrement dit, *pour un découpage de l'espace donné*, il suffit de générer les matrices élémentaires une seule fois : elles pourront être utilisées telles quelles pour calculer la matrice $P_{(L,n)}$ pour n'importe quels taux de mutation. Les matrices élémentaires sont stockées dans des fichiers listant tous les coefficients de la matrice selon le type de mutation et le découpage de l'espace représentés. Ces fichiers sont des fichiers textes, ils permettent de passer de C à Matlab facilement. Ils utilisent un format de matrice creuse, où on précise uniquement les coefficients non nuls avec leur état de départ et d'arrivée.

**Simulation d'un jeu de données** Quand les matrices élémentaires ont été générées pour un découpage de l'espace donné, on peut lancer une simulation en précisant les paramètres demandés dans la table IV.1. Les matrices élémentaires sont alors automatiquement lues par Matlab ou Octave (l'équivalent en logiciel libre) puis assemblées en $P_{(L,n)}$ en utilisant les taux de mutation. La matrice $M_{(L,n)}$ est alors calculée en utilisant la procédure donnée en section 4. La matrice $F$ est générée à partir de la fonction de fitness. En appliquant l'équation III.1, on peut calculer la densité pour n'importe quel point de départ. Le point de départ n'influence que transitoirement la dynamique de la simulation. En effet, le lemme 3.3 peut s'appliquer à la matrice $M_{(L,n)}$ des transitions entre états agrégés, donc il existe une distribution stationnaire unique, indépendante du point de départ. Ce lemme reste valide même en présence de sélection, car celle-ci module les valeurs des transitions sans les rendre nulles.

Toutes les opérations réalisées (calcul de $M_{(L,n)}$ à partir de $P_{(L,n)}$, déroulement d'une génération) étant du type matrice-matrice ou matrice-vecteur, l'utilisation de Matlab ou d'Octave convient parfaitement. Une fois que les matrices élémentaires ont été chargées, il y a donc deux étapes dans la simulation d'un jeu de paramètres : (1) la construction des matrices $M_{(L,n)}$ et $F$ et (2) l'itération des générations. En règle générale, les deux étapes sont assez rapides. Cependant, l'étape (1) peut être longue si l'espace est grand comparé aux taux de mutation : les génomes sur les bords subissent beaucoup de mutations, l'exponentiation prend alors plus de temps car il faut aller plus loin dans la série de l'exponentielle. L'étape (2) peut être longue si le ratio entre $\mu_{dup}$ et $\mu_{ldel}$ est proche de la condition d'explosion des génomes, à cause des très petits génomes. Sans sélection, il leur faut un temps très long pour croître à nouveau car ils rentrent dans des zones où le nombre de mutation par génération est très faible. Autrement dit, même si théoriquement on s'attend à ce que le génome croisse à nouveau et devienne arbitrairement grand, la convergence est délicate, d'autant que dans le cas où on prévoit une explosion de la taille, les génomes sont retenus par le bord et sont surreprésentés dans la zone lente par rapport aux prévisions théoriques.

## 1.3 Comportement typique et condition d'arrêt de la simulation

On obtient pour chaque génération une densité de population qui tend progressivement vers une distribution stationnaire. On peut représenter la répartition de la population dans l'espace (non codant, codant) à chaque génération (figure IV.1). La distribution stationnaire indépendante du point de départ. La dynamique transitoire se fait en général en deux phases. D'abord, le ratio de codant des génomes à tendance à être conservé : on observe surtout des variations de taille via les délétions et les duplications (dans ce modèle ces deux types de mutations conservent à peu près le ratio de codant). Ensuite, quand une taille maximale semble avoir été atteinte, celle-ci ne varie plus beaucoup, mais le contenu en codant évolue rapidement via une élimination du non codant en surplus. S'il n'y avait pas de non codant en surplus, les deux phases sont mélangées : du non codant est acquis en même temps que l'ajustement de la longueur. Ultimement, la population se stabilise autour d'une valeur d'équilibre en s'étalant très largement autour. Cette variabilité, aussi bien en taille qu'en pourcentage de codant, est très loin de la réalité : elle reflète notamment le choix de la fonction de fitness, comme nous le verrons au cours de ce chapitre et dans le chapitre V.

Comme le processus ne converge qu'asymptotiquement, il faut un critère pour arrêter la simulation. La convergence étant assez rapide, nous avons décidé de le faire de manière très simple. Nous regardons à chaque étape de combien la distribution change en absolu dans chaque subdivision. Quand la somme de ces changements est inférieure à $10^{-6}$ ou $10^{-7}$, nous considérons que la distribution stationnaire est atteinte et nous arrêtons la simulation.

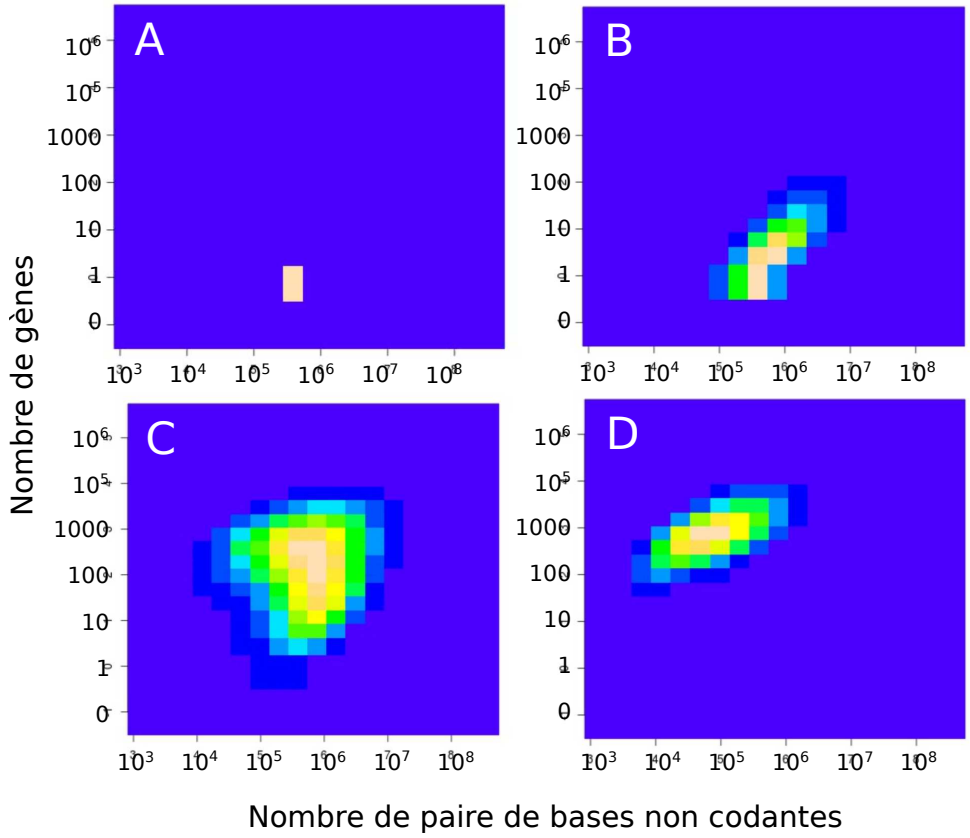


Figure IV.1 – Exemple d'une simulation dans le plan (non codant, codant) en échelle logarithmique avec les paramètres par défaut. Initialement, on a donné aux génomes un gène et environ 500kb de bases non codant (A). Assez rapidement, les individus qui arrivent à dupliquer leur gène sont sélectionnés (B), même s'ils acquièrent du non codant au passage, puisque cela ne change pas leur fitness. En milieu de simulation, les individus augmentent progressivement leur nombre de gènes sans que le non codant soit fortement affecté (C). En fin de simulation, les individus restent autour d'une certaine taille limite mais échangent progressivement le non codant contre du codant, jusqu'à une certaine valeur limite, puis se stabilisent autour d'une distribution stationnaire (D).

## 1.4 Mesures réalisées lors des simulations

À chaque génération, nous sauvegardons la totalité de la distribution, ce qui permet de réaliser des mesures dans le temps et dans l'espace. Les mesures dans l'espace se heurtent à un problème important : comme on s'intéresse à une population infinie, la variabilité peut être assez grande et il faut être sûr que les mesures retenues soient représentatives. En général, la distribution stationnaire est très régulière. Comme les processus de duplications et délétions sont prédominants, ils donnent à la distribution un aspect quasiment gaussien en échelle logarithmique. Dans ce cas, la moyenne et l'écart-type sont des mesures qui suffisent à caractériser toute la distribution. Il est également possible de prendre en compte d'autres types de mesures, comme le mode (avec des algorithmes dédiés à la recherche du mode pour une distribution quasi gaussienne) ou la médiane. Il faut noter que pour une

gaussienne parfaite, mode, médiane et moyenne coïncident. En pratique, la moyenne en échelle logarithmique s'avère être la mesure la plus simple et la plus robuste. On pourrait alors donner l'évolution de la moyenne et de l'écart-type en fonction du temps ou d'un paramètre, comme par exemple sur la figure IV.2A.

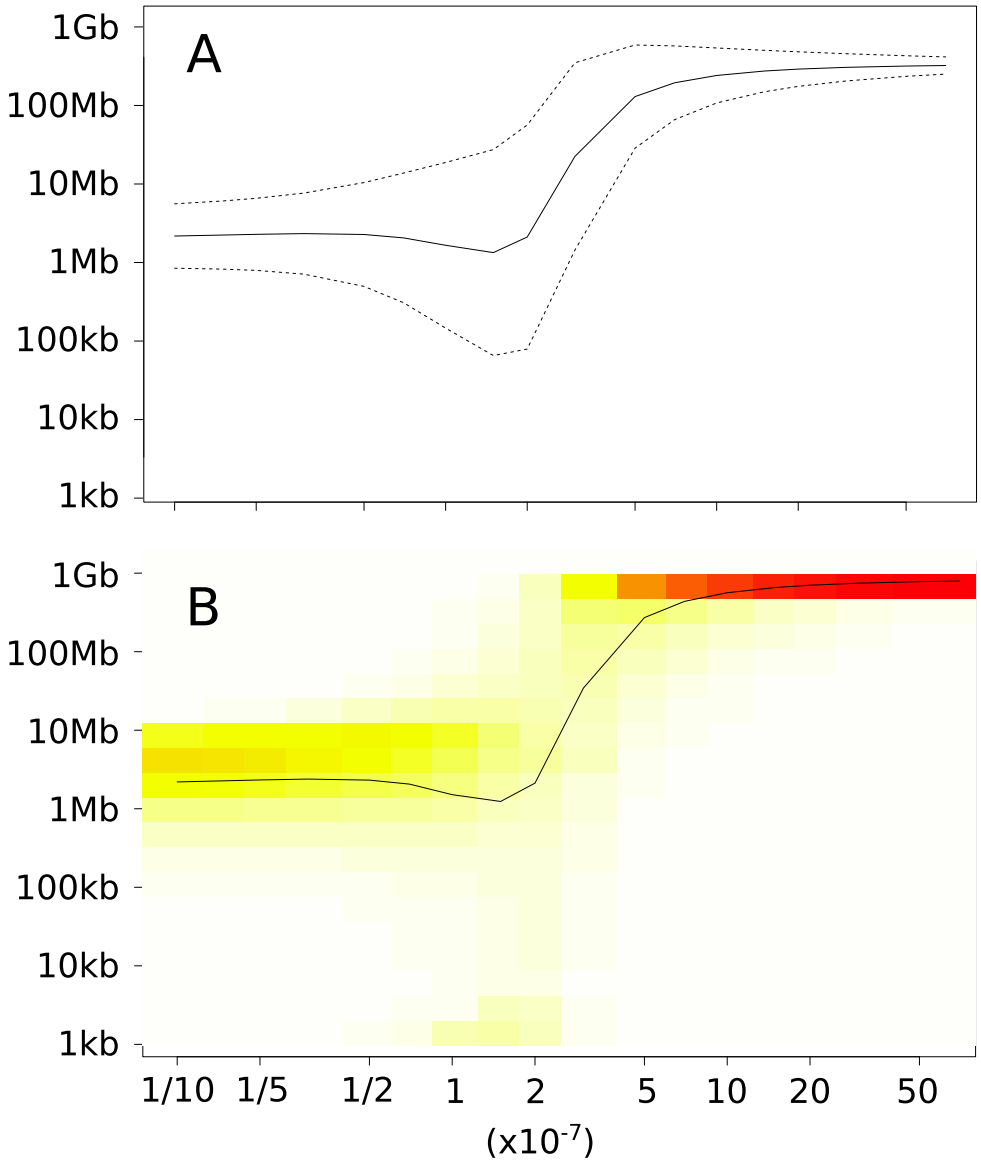


Figure IV.2 – Exploitation typique du modèle : on étudie ici l'évolution de la taille du génome en fonction du taux de duplications, les autres taux étant fixés à la valeur par défaut, $10^{-7}$. La représentation classique consiste à dessiner en trait plein la moyenne et en pointillés l'écart-type (A). On peut aussi remplacer les écarts-types par la densité sous forme de niveaux de couleurs (B). On voit alors mieux les détails de la répartition de la population : on se rend compte qu'au niveau de la condition d'explosion, la moyenne et l'écart-type ne capturent que très partiellement la structure de la population.

Cependant, quand le ratio entre le taux de duplications et de grandes délétions est proche de la condition d'explosion, les génomes se trouvent dans des positions plus extrêmes. Une partie de la population se trouve près des bords du domaine de simulation correspondant

aux grandes tailles. Parallèlement, certains deviennent très petits et vont donc se coller aux bords opposés. On a donc une déstructuration de la population qui n'est pas correctement capturée par la description donnée par la moyenne et l'écart-type. Pour cela, nous utilisons une description en niveaux de couleur de la densité de la taille de génome (figure IV.2B). En comparant les deux représentations, on voit que l'utilisation de l'écart-type n'était pas complètement adaptée au cas illustré sur la figure. Nous avons donc préféré la deuxième représentation : elle permet de détecter rapidement si les distributions stationnaires sont piquées autour de la moyenne ou si au contraire la population est plus déstructurée.

## 1.5 Une variation : simulation de populations finies

Dans cette variation, il s'agit d'adapter le plus simplement possible le modèle pour qu'il prenne en compte les effets de taille finie liés aux petites populations. En faisant cela, on diminue le nombre de transitions effectivement réalisées à chaque génération et on s'attend à ce que les transitions rares, notamment celles qui font grandir le génome soient difficiles à réaliser et à maintenir sur le long terme. On sort donc du scénario pire cas pour la croissance et on se dirige vers des hypothèses un peu plus réalistes. Cette variation a pour but d'illustrer que le scénario pire cas en population infini est bien un « pire cas ». En se plaçant dans des conditions très légèrement plus réalistes, on verra que les tailles de génome atteintes en population infinie sont difficilement tenables dans un certain nombre de cas si on bascule en population finie.

L'idée la plus naturelle serait ici d'implémenter un modèle individu-centré car il permettrait, au prix d'un temps de calcul certes plus long, de réaliser les mutations sans aucune approximation. Néanmoins, nous avons voulu rester proche du modèle général, pour mieux illustrer comment il peut être modifié et adapté. Rappelons que dans le modèle général, la population subit deux étapes : la sélection et la mutation. Nous appliquons ces deux étapes comme pour le cas de la population infinie :

$$n_{t+1}^{inf} = \frac{MFn_t}{\|Fn_t\|}$$

La densité dans $n_{t+1}^{inf}$ nous donne toutes les transitions qui peuvent être réalisées, avec leur probabilité, à partir de la population dans $n_t$. Dans le cas de la population infinie, nous supposons qu'elles le sont toutes et on a $n_{t+1} = n_{t+1}^{inf}$. Dans le cas d'une population finie de $N_{pop}$ individus, nous réalisons un tirage multinomial basé sur les probabilités données dans $n_{t+1}^{inf}$. La répartition des individus dans l'espace d'états est alors inscrite dans le vecteur $n_{t+1}$, qui contient donc uniquement des nombres rationnels multiples de $1/N_{pop}$.

Comparé au modèle à population infinie, on ajoute une étape de nature probabiliste. La population ne converge plus vers une distribution stationnaire, même si les tendances sont les mêmes qu'en populations infinie. Cela ajoute de la variabilité et une difficulté supplémentaire pour représenter son comportement en fonction des différents paramètres. Pour

rester simple, nous avons choisi de moyenner la densité sur les 500 dernières générations de 5 populations indépendantes. Avant ces 500 générations qui servent pour la mesure, 1500 générations initiales sont réalisées pour limiter l'impact de la condition initiale et garantir des variations autour de « valeurs représentatives ». À titre de comparaison, en population infinie, la convergence se fait en quelques dizaines de générations pour les valeurs de paramètres et les conditions initiales choisies dans ce chapitre.

# 2 Étude de l'effet des taux de mutation sur la taille du génome

En faisant varier les taux de mutation associés à chaque mutation, nous pouvons étudier leur impact sur la taille du génome et vérifier nos prédictions analytiques. Par défaut, les valeurs des taux de mutation sont prises égales pour tous les types de mutation, la fonction de fitness est la fonction logarithme et le découpage de l'espace est fait en échelle logarithmique. D'après nos prévisions, la population devrait converger vers une taille finie avec les paramètres par défaut et la taille limite devrait dépendre essentiellement des grandes délétions et des duplications. Nous allons ici étudier les effets des différents types de mutation en allant du moins intéressant vers le plus intéressant : inversions et translocations, indels puis grandes délétions et duplications.

## 2.1 Effet des taux d'inversions, de translocations et d'indels sur la taille du génome

Les taux d'inversions ou de translocations influencent la probabilité de perdre des gènes par pseudogénisation, mais ces mutations ne changent pas la taille totale du génome. Ainsi, lorsqu'on fait varier le taux d'inversions ou le taux de translocations, la taille moyenne du génome à l'équilibre ne change quasiment pas (figure IV.3A et B). Nous verrons cependant au chapitre suivant que ces réarrangements ont une influence sur le ratio de codant.

Les taux de petites insertions et de petites délétions influencent eux aussi la probabilité d'inactiver un gène, mais ces mutations modifient également la taille des génomes. Cependant, les simulations montrent que leur impact sur la taille moyenne du génome à l'équilibre est en fait très faible (figure IV.3C et D), surtout si cette taille d'équilibre est assez grande. Ainsi, même si le taux de petites insertions est 50 fois plus grand que le taux de petites délétions, ou inversement 50 fois plus petit, la taille d'équilibre est quasiment la même que si les taux étaient égaux. Cela est dû au fait que, comme prévu analytiquement, les duplications et les grandes délétions ont un impact beaucoup plus fort que les petits indels sur la taille à l'équilibre. Comme les inversions et les translocations, les petits indels influencent surtout le contenu en codant ou en non codant, ce que nous détaillerons au chapitre suivant.

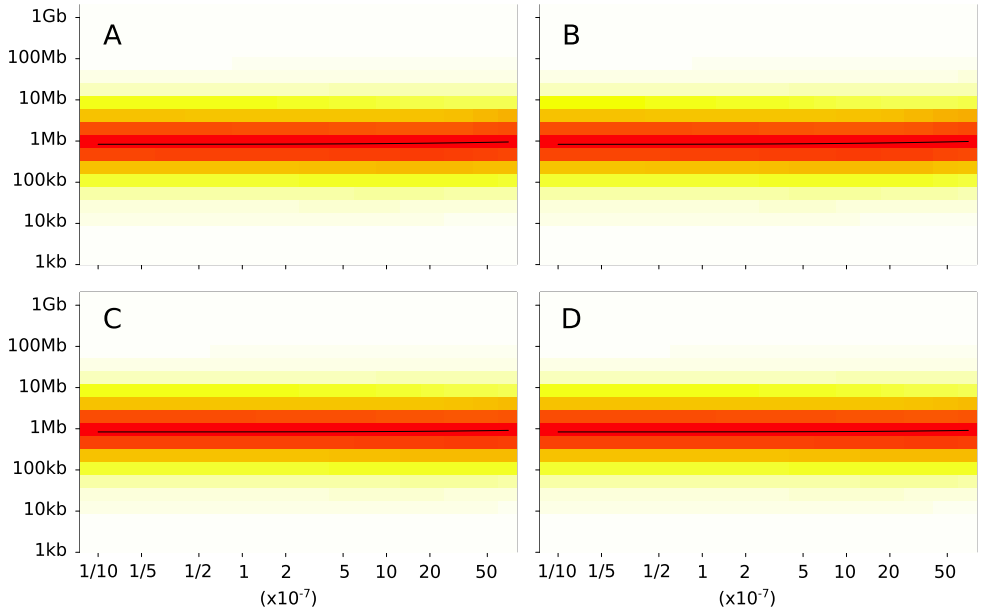


Figure IV.3 – Évolution de la taille du génome en fonction du taux d'inversions (A), de translocations (B), de petites insertions (C) et de petites délétions (D).

## 2.2 Effet des taux de duplications et de grandes délétions sur la taille du génome

Les taux de grandes délétions (figure IV.4) et de duplications (figure IV.5) ont un impact beaucoup plus important que les autres mutations sur la taille moyenne du génome à l'équilibre. Le déterminant principal de la taille du génome est donc clairement le processus de grandes délétions et duplications.

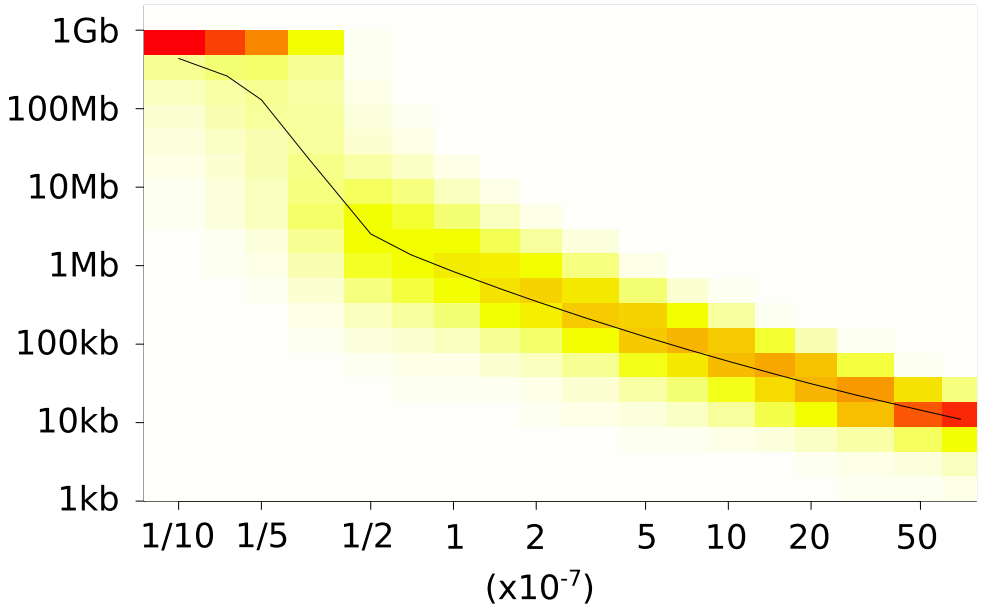


Figure IV.4 – Évolution de la taille du génome en fonction du taux de grandes délétions.

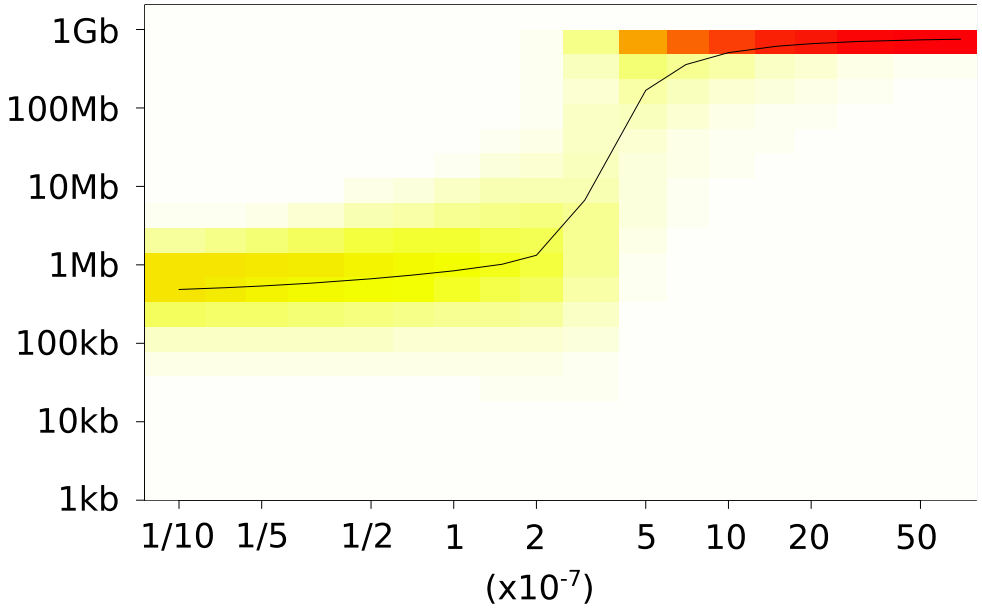


Figure IV.5 – Évolution de la taille du génome en fonction du taux de duplications.

Notons que les valeurs de paramètres pour lesquels la taille du génome « explose » sont en fait celles pour lesquelles elle plafonne à 1 Gb, la limite du domaine de simulation. On observe que, comme prévu analytiquement, le seuil d'explosion du génome ne se trouve pas à une duplication pour une délétion, mais à environ 2.5 duplications pour une délétion, et ce malgré une pression de sélection qui favorise directement les génomes contenant le plus de gènes. Cet effet s'observe aussi bien en augmentant le taux de duplication (figure IV.5) qu'en diminuant le taux de grandes délétions (figure IV.4).

Ainsi, dans le scénario où les duplications sont spontanément deux fois plus fréquentes que les délétions et où les duplications de gènes sont systématiquement favorables, le génome ne croît pas infiniment mais se stabilise sur une taille d'équilibre (éloignée des bords du domaine de simulation).

Ce résultat confirme ce que nous avions prédit analytiquement : la sélection, même caricaturale, ne peut pas à elle seule faire croître infiniment les génomes si la dynamique spontanée des duplications et des délétions les rappelle vers des tailles finies, ce rappel s'exerçant au-delà du ratio intuitif de 1 duplication pour 1 délétion, jusqu'à environ 2.5 duplications pour une délétion.

Lorsque le ratio excède 2.6, la dynamique spontanée des duplications et des délétions fait au contraire croître les génomes, une croissance accélérée par la sélection. Comme l'espace de calcul est fini, cela se traduit par des effets de bord plus ou moins prononcés, qui rendent la transition vers l'explosion moins nette que si l'on avait pu garder un nombre infini d'états. Comme les courbes après explosion plafonnent artificiellement à cause de la finitude du domaine de simulation, nous allons examiner les tendances de la taille moyenne quand la taille de génome reste finie. Nous avions donné au chapitre II des bornes pour

la taille du génome à l'équilibre. Grâce à la simulation, nous allons pouvoir aborder plus précisément cette question : où exactement sous cette borne la taille se stabilise-t-elle en présence de sélection ? Est-ce que les taux de duplication et de délétion ont une influence sur la valeur précise de la taille ? Nous allons voir qu'une fois de plus, les tendances pour les duplications et les grandes délétions ne sont pas symétriques.

Il semble y avoir un lien très simple entre le taux de délétion et la taille à l'équilibre lorsque celle-ci est finie : la taille à l'équilibre décroît quasiment linéairement (en échelle logarithmique) avec le taux de délétion (figure IV.4). Ainsi, les délétions délimitent une taille au-delà de laquelle un génome est trop instable pour survivre à long terme. Cette limite est une limite probabiliste : elle n'implique pas que les génomes en dessous de la limite vont tous conserver leur structure et ceux au dessus la perdre. Elle n'a donc pas de positionnement précis. Comme la sélection favorise les plus gros génomes et uniquement ça, tout pas vers un génome plus gros est directement sélectionné, même si l'individu n'a aucune chance de se reproduire à l'identique, ce qui crée un large étalement autour de la valeur moyenne. Cela ne serait pas le cas dans une population d'organismes réels, où toutes les duplications de gènes ne sont pas nécessairement favorables.

Lorsqu'on fait varier le taux de duplications, la taille à l'équilibre change assez peu tant qu'on reste sous le seuil de 2.5 duplications pour une délétion. Contrairement aux grandes délétions, les duplications ne contrôlent pas directement le comportement moyen (on va voir à la prochaine section qu'en moyenne, la taille peut même diminuer quand on augmente le taux de duplications). Le taux de duplication a cependant une influence plus prononcée sur la dispersion de la population (ce qui se caractérise par un évasement de la densité autour de la moyenne). Ainsi, quand on augmente le taux de duplication, on augmente la pression à l'augmentation de la taille des génomes, ce qui conduit une partie de la population dans une zone instable, où les individus sont incapables de se reproduire à l'identique, notamment à cause des grandes délétions. Les descendants à chaque génération sont ainsi très éparpillés en taille.

Pour résumer, le taux relatif des duplications comparativement aux délétions permet de contrôler la variance, tandis que le taux des délétions limite le comportement moyen tant que la taille moyenne des génomes de la population n'explose pas. On peut complètement contrôler le comportement moyen en faisant varier conjointement les taux de duplications et des grandes délétions, ce qui garantit que la population n'explose pas. Analytiquement, dans le cas où les taux de duplications et de délétions sont égaux, nous avions donné une borne supérieure de la médiane de la population sous la forme $\overline{s} = C/\mu_{dupdel}$, où $\mu_{dupdel} = \mu_{dup} = \mu_{ldel}$ (équation (II.4), page 79). Quand on fait varier les taux de duplications et de grandes délétions conjointement, on a effectivement une variation de la taille moyenne qui semble inversement proportionnelle à ces taux (figure IV.6). Nous avons déjà souligné que cette relation rappelle fortement celle de Drake (1991), qui reliait la taille de génomes de différents microbes à l'inverse de leurs taux de mutation « global ». (Rappelons que la ressemblance est très visuelle : pour que la comparaison soit pertinente, il faudrait montrer que l'estimation des taux de mutation par Drake corrèle avec les taux spontanés de réarrangements). Nous pouvons déterminer notre constante $C$ par simple régression linéaire en échelle logarithmique : on trouve $C \simeq 0.1$. Cela signifie qu'au niveau du mode de

la population, les individus subissent en moyenne une grande délétion ou une duplication toutes les 10 générations. Nous n'avions pas pu déterminer la valeur de cette constante analytiquement. Il y a une bonne raison à cela : on s'attend à ce que la constante varie en fonction de la force de la sélection. En effet, les prédictions analytiques donnent des bornes supérieures qui ne peuvent pas être dépassées, mais la sélection détermine à quel point on s'en rapproche. Cette question est explorée plus en détails dans la prochaine section.

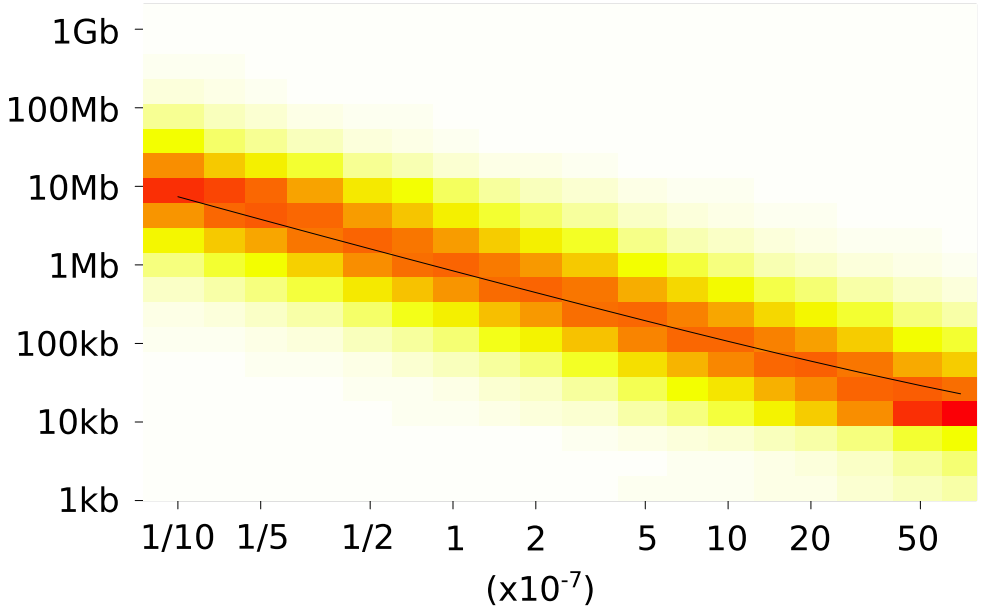


Figure IV.6 – Évolution de la taille du génome quand le taux de duplications et de grandes délétions varient simultanément.

# 3 Étude de l'impact de la force de sélection sur la taille du génome

Dans cette section, nous ne ferons plus varier les taux des indels, des inversions et des translocations, étant donné que leur impact sur la taille est négligeable. D'après les prévisions analytiques, la limite de viabilité imposée par les réarrangements ne peut pas être surmontée par la sélection. Jusqu'à maintenant, la fitness n'augmentait « que » logarithmiquement avec le nombre de gènes. Nous allons donc répéter les expériences précédentes sous des conditions de sélection plus stringentes.

Pour illustrer l'évolution de la population quand elle est soumise à une pression de sélection qui l'oriente vers des tailles plus grandes, nous avons choisi la famille des fonctions $f(n) = n^{alpha}$ qui croissent asymptotiquement plus rapidement que la fonction $\log(n)$ et d'autant plus rapidement que $\alpha$ est grand. Pour illustrer notre propos, il suffit de comparer les résultats des simulations pour $\alpha \in \{0.2, 0.4, 0.6, 0.8, 1\}$. Les figures IV.7 et IV.8

illustrent l'évolution de la taille du génome quand on fait varier respectivement le taux de duplications et le taux de grandes délétions, sous ces différentes pressions de sélection.

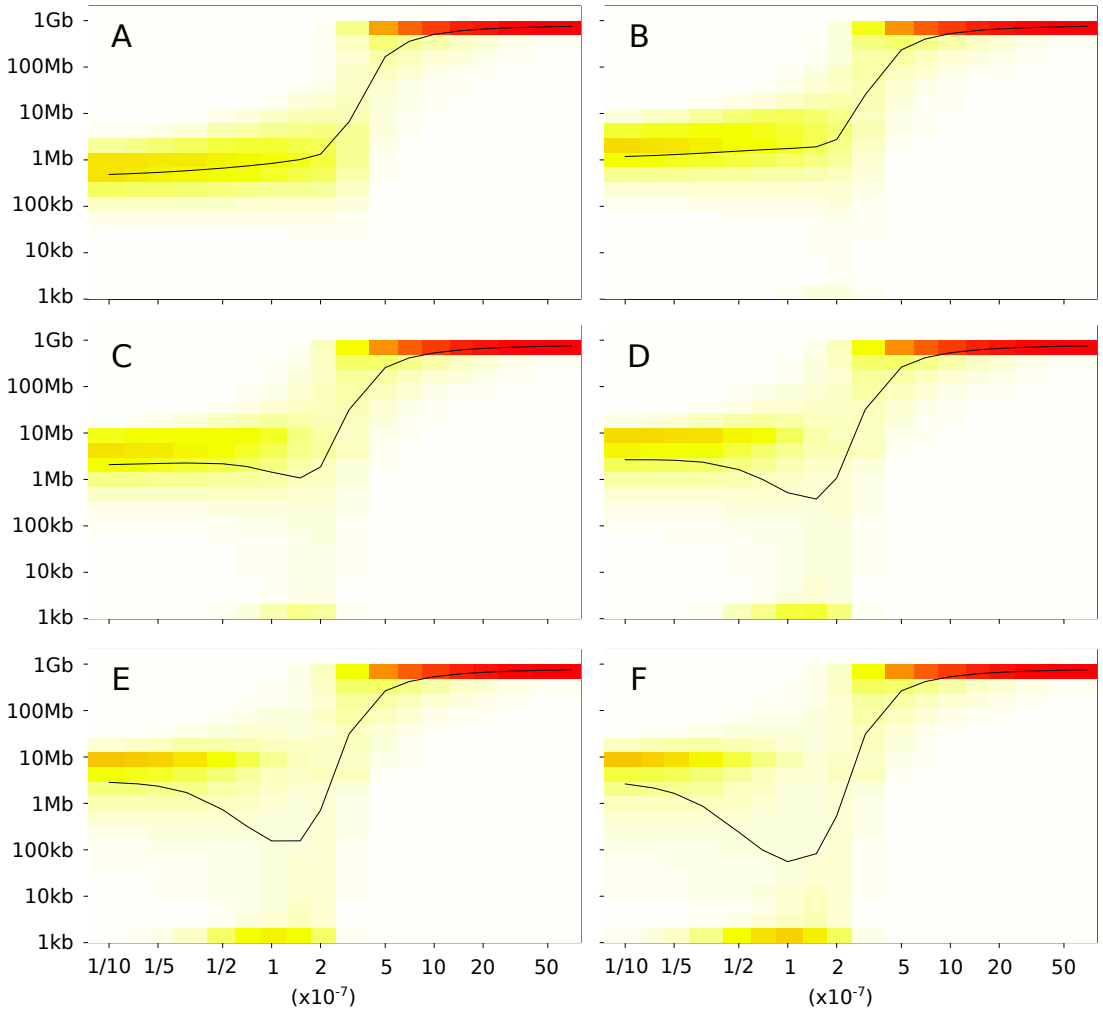


Figure IV.7 – Évolution de la taille du génome en fonction du taux de duplications pour différentes formes de sélection : $\log(n)$ (A), $n^{0.2}$ (B), $n^{0.4}$ (C), $n^{0.6}$ (D), $n^{0.8}$ (E), $n^1$ (F).

Quand on augmente la force de la sélection, on obtient des effets contradictoires : pour des taux de duplications faibles, la taille moyenne du génome à l'équilibre est plus grande si la pression de sélection est plus forte, mais pour des taux de duplications compris entre (environ) 0.5 et 2 fois le taux de délétions, la taille moyenne à l'équilibre est plus basse si la pression de sélection est plus forte. En effet, quand on augmente la force de sélection, la taille moyenne à l'équilibre part de plus haut (tout en restant bornée) pour de faibles taux de duplications mais diminue plus vite ensuite quand le taux de duplications augmente.

Ainsi, si l'on donne un très grand avantage compétitif aux longs génomes, passer de 1 à 2 duplications pour une délétion conduit paradoxalement à une *diminution* de la taille moyenne du génome.

Dans l'intervalle où le taux de duplications est compris entre 0.5 et 2 fois le taux de délé-

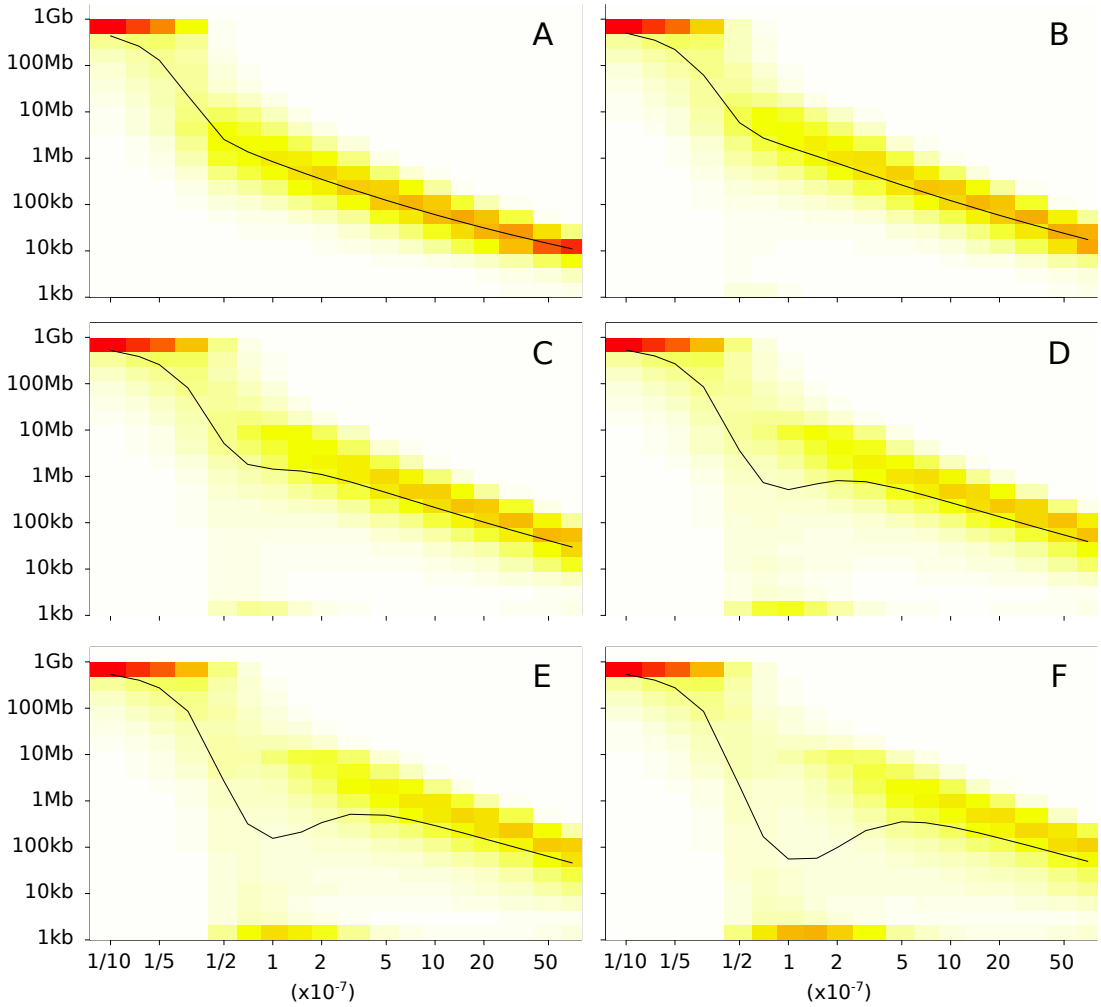


Figure IV.8 – Évolution de la taille du génome en fonction du taux de grandes délétions pour différentes formes de sélection : $\log(n)$ (A), $n^{0.2}$ (B), $n^{0.4}$ (C), $n^{0.6}$ (D), $n^{0.8}$ (E), $n^1$ (F).

tions, la diminution de la taille moyenne est en fait due à un éclatement de la population en deux sous-populations, dont l'une avec des génomes de taille quasiment nulle, inférieure à la taille d'un gène. Cela est dû au fait que, comme nous l'avions vu analytiquement, la relation entre la taille avant réplication et la médiane de la taille après réplication n'est pas monotone (voir figure II.6 page 79) : au-delà d'une certaine taille de départ, cette relation est décroissante. Donc plus un génome est grand au départ, plus les génomes de ses descendants tendent à être courts. En donnant un très grand avantage compétitif aux grands génomes, on sur-représente leurs descendants à la génération suivante, et ces descendants tendent à avoir des génomes très courts. Ces petites génomes n'ont quasiment aucune chance de survivre à moyen terme : il leur faut longtemps pour reconstruire ce qui a été perdu car ils se trouvent dans une zone où le nombre de mutations par génération est très faible. On a donc une sélection des quelques génomes assez grands qui ont subsisté, mais qui restent instables.

Pour des taux de duplications et de délétions égaux, la taille moyenne à l'équilibre varie

toujours comme l'inverse des taux pour les pressions de sélection les plus faibles (figures IV.9A et B). Quand la sélection devient très forte, on entre progressivement dans la zone instable, la population éclate de plus en plus, d'où un abaissement du niveau auquel la moyenne de la population se stabilise (figures IV.9C à F). Ainsi, si le taux de duplications et de délétions détermine une borne supérieure pour la taille du génome, la force de la sélection détermine (de façon non triviale) l'endroit effectif où la taille va effectivement se stabiliser : très près de la borne supérieure si la sélection en faveur des grands génomes est modérée, plus bas si cette sélection est forte.

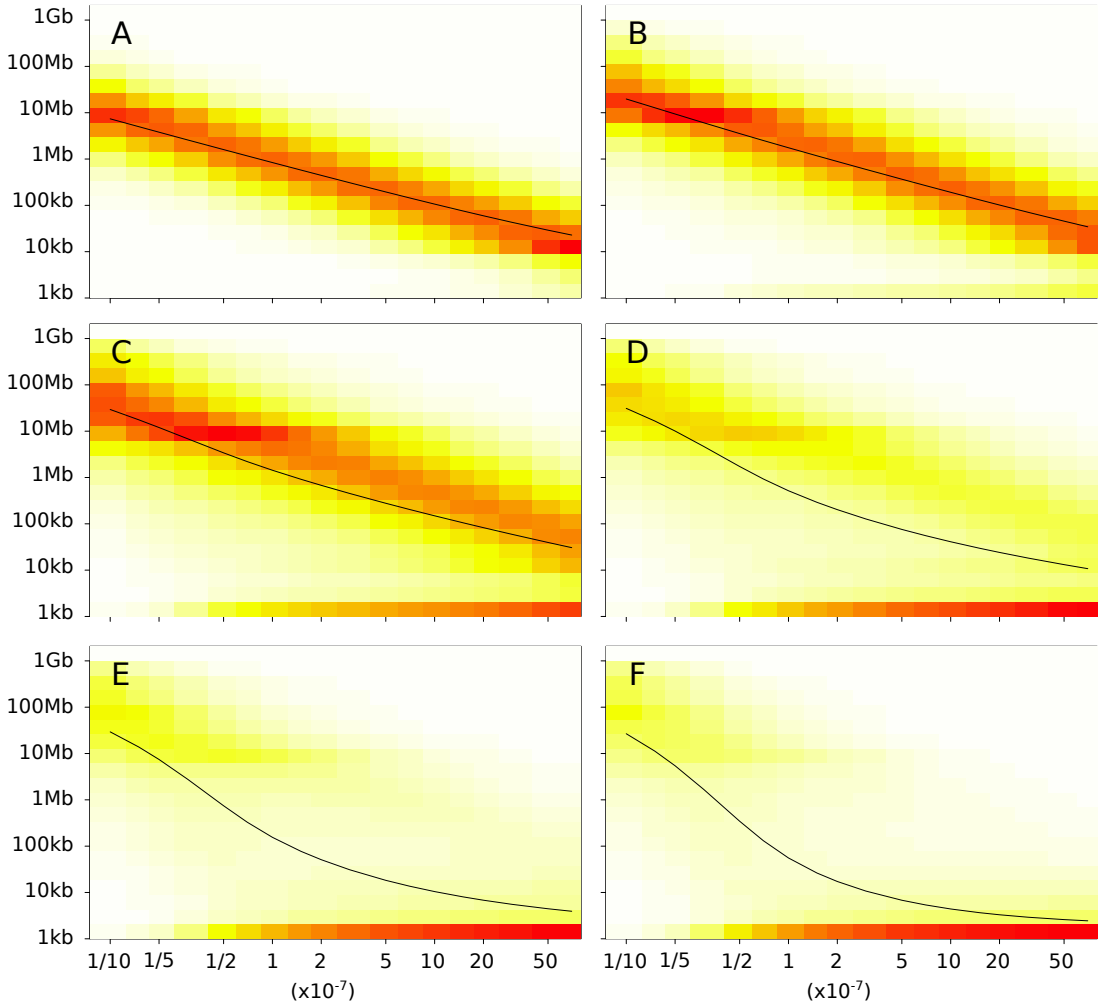


Figure IV.9 – Évolution de la taille du génome quand on varie conjointement le taux de duplications et de délétions : $\log(n)$ (A), $n^{0.2}$ (B), $n^{0.4}$ (C), $n^{0.6}$ (D), $n^{0.8}$ (E), $n^{1}$ (F).

Il est peu probable que les populations réelles aient une dynamique aussi caricaturale, car il n'y a pas de relation simple entre le nombre de gènes et la fitness, en tout cas pas aussi simple que celle que nous avons choisie pour pousser le système dans ses retranchements. On peut envisager, d'un point de vue purement combinatoire, que plus de gènes puissent être exploités pour obtenir un avantage sélectif, mais pas de manière aussi simple et immédiate que dans le modèle. D'autre part, dans le modèle, l'absence de gènes essentiels joue un rôle important. Dans une population réelle, l'entrée en zone instable ne devrait donc pas se faire aussi facilement. Elle est facilitée dans le modèle car il est possible

de regagner une bonne fitness après une délétion en dupliquant tout ou partie des gènes restants, sans tenir compte de *l'information* perdue. Dans des conditions plus réalistes, les génomes devraient se stabiliser dans une zone où la probabilité de perdre un gène essentiel est suffisamment faible. On pourrait alors envisager que la variation soit plus resserrée, la population plus stable et qu'elle se stabilise à des tailles plus faibles que celles réalisables en théorie. Enfin, les duplications sont très avantageuses dans le modèle alors qu'elles posent en réalité des problèmes de dosage car elles peuvent augmenter l'expression de certains gènes. Elles peuvent donc être directement délétères. Cet argument plaide également pour une stabilisation autour de tailles plus faibles quand les taux de duplications augmentent. Pour toutes ces raisons, on pourrait donc observer une baisse de la taille moyenne quand on augmente le taux de duplication dans des conditions plus réalistes biologiquement, mais pour des raisons en partie différentes que celles qui conduisent à une baisse dans ce modèle.

Pour tester cette hypothèse, il faudrait utiliser un modèle plus complexe, comme par exemple aevol (Knibbe *et al.*, 2007) ou le modèle de réseau de gènes de Crombach et Hogeweg (2008), où les gènes sont plus ou moins importants pour la survie et où leur dosage est pris en compte dans la fitness de l'individu. L'avantage serait d'avoir une sélection beaucoup plus subtile, qui n'agirait pas directement sur le nombre de gènes, mais sur l'information contenue dans le génome. Mais le revers méthodologique de cet avantage (sans parler des nombreux paramètres *ad hoc* de la chimie artificielle qu'il faut définir pour décoder le génome et lui assigner une fitness) est que si l'on ne contrôle pas la façon dont la sélection agit sur le nombre de gènes, on peut toujours l'accuser de participer à la stabilisation de la taille du génome. On ne peut alors pas identifier précisément les causes de la stabilisation des génomes. Or à quoi sert un modèle d'un système biologique s'il est presque aussi compliqué à analyser que le système lui-même ? Le modèle minimaliste que nous avons développé ici se veut complémentaire de ces approches de vie artificielle : en sacrifiant la notion d'information, nous avons pu réduire drastiquement le nombre d'états génomiques et le nombre de paramètres, et nous avons pu choisir une sélection qui, avec certitude, n'aide pas le génome à se stabiliser. Cela nous a permis de mettre en évidence la puissance de la « force de rappel » liée à la dynamique spontanée des duplications et des délétions. Maintenant que ceci est bien établi et compris, il sera possible de retourner à des approches plus « riches », permettant d'obtenir des dynamiques moins caricaturales et donc plus réalistes.

# 4 Étude de l'évolution de la taille du génome en population finie

En suivant la procédure donnée en introduction du chapitre, on peut ajouter de la dérive en diminuant la taille de la population. *A priori*, la diminution de la taille de la population rend la sélection moins efficace et a donc un effet contraire à celui de l'augmentation de la force de la sélection. La figure IV.10 montre que l'allure générale des courbes en population finie est similaire à celles obtenues en population infinie, tout en étant plus bruitées.

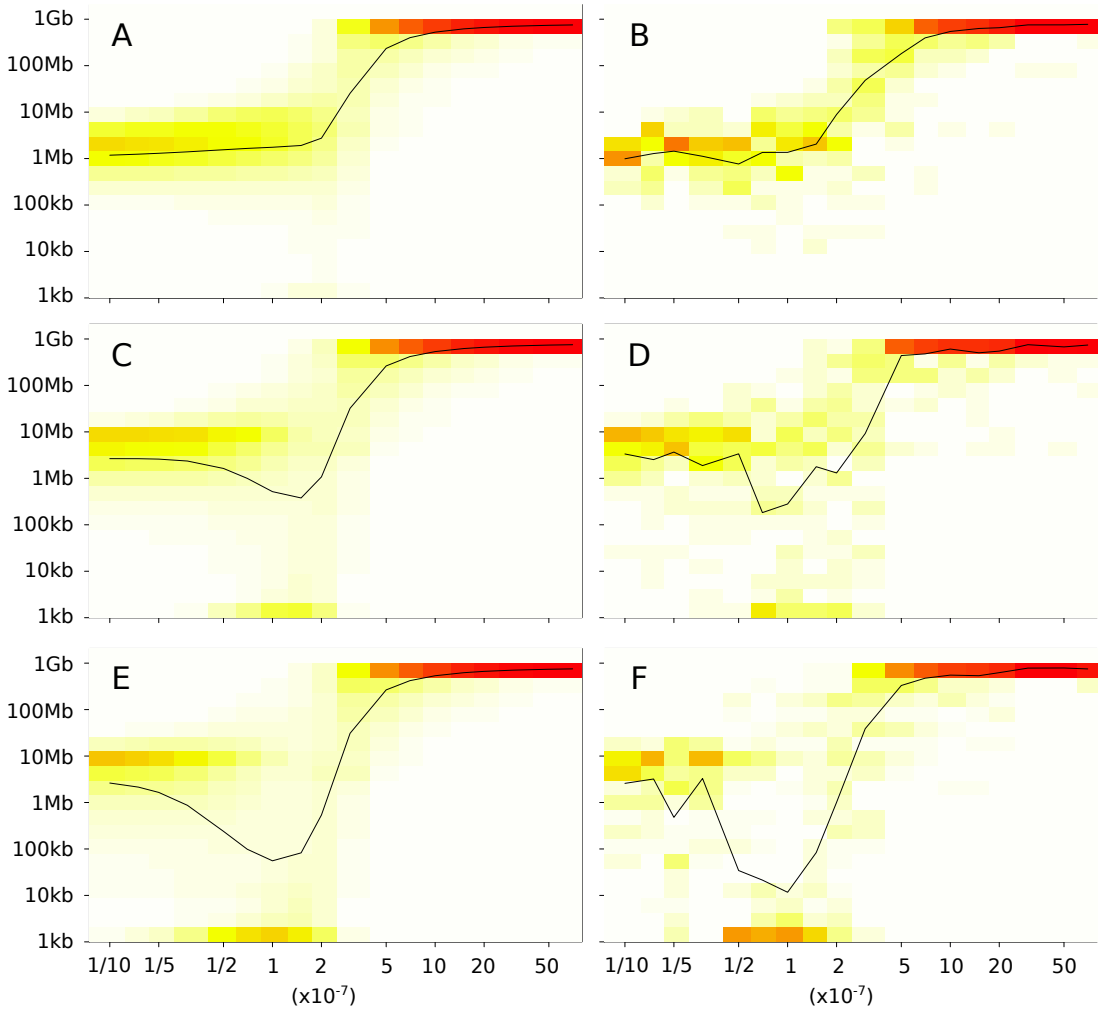


Figure IV.10 – Comparaison de l'évolution de la taille du génome en fonction du taux de duplications pour différentes formes de sélection en population finies et infinies. Fitness $n^{0.2}$ : population infinie (A), $N = 50$ (B). Fitness $n^{0.6}$ : population infinie (C), $N = 50$ (D). Fitness $n^1$ : population infinie (E), $N = 50$ (F). Pour les simulations en population finie (B, D,F), les densités représentées sont des densités moyennées sur les 500 dernières générations de 5 populations indépendantes.

Le fait que la population soit finie ne change donc pas fondamentalement la dynamique du système. Cependant, cela empêche la distribution de converger aussi bien que dans le cas déterministe. En effet, en population infinie, une faible fraction de la population maintient une taille assez grande en permanence mais ses descendants sont beaucoup plus petits. À l'équilibre, seule une petite partie, correspondant à la fraction initiale, parvient à se maintenir, ce qui permet la convergence de la distribution. Au contraire, en population finie, il est plus difficile de maintenir la même taille, ce qui peut entraîner par moments un effondrement de la taille du génome pour la totalité des individus. Si un grand génome est fortement sélectionné mais qu'il est trop instable, il y a une probabilité non négligeable que tous ses descendants aient subi des pertes importantes. Pour des sélections très fortes, les simulations en population finie sont donc très instables, il y a peu de continuité visuelle d'une génération à l'autre, on a l'impression que la structure des génomes change en

permanence, ce qui explique les courbes très bruitées.

# 5 Conclusion

Les résultats des simulations menées ici sont totalement conformes aux prédictions analytiques présentés dans la première partie. Ils donnent cependant des informations quantitatives sur la stabilisation de la population auxquelles nous n'avions pas accès dans le cadre analytique.

L'explosion des génomes est bien contrôlée par les grandes délétions et les duplications avec un seuil aux alentours de 2.59 pour le biais en faveur des duplications, les taux de petites insertions et délétions n'influençant quasiment pas la taille d'équilibre même s'ils ne sont pas égaux. La valeur précise est difficile à obtenir à cause des effets de bord mais les simulations semblent en cohérence avec la théorie. Les deux prédictions annexes sont également confirmées. Une sélection même caricaturale ne semble pas pouvoir permettre de surmonter les instabilités chromosomiques. La seule façon d'obtenir un génome plus grand est donc de baisser les taux de délétions et de duplications, avec une taille moyenne proportionnelle à l'inverse du taux de mutation quand la sélection n'est pas trop forte.

Cependant, les simulations auront apporté plus que la simple confirmation des prévisions analytiques. Elles illustrent la difficulté à prédire par la pensée l'effet de l'augmentation du taux de duplications ou de la sélection. Alors qu'on peut simplement s'attendre à une augmentation de la taille moyenne, on introduit surtout plus d'instabilité. Certains génomes vont donc pour grossir mais, globalement, on assiste à un éclatement progressif de la population. Grâce aux simplifications du modèle, une fraction de la population parvient tout de même à maintenir des génomes de grande taille mais il est très probable que cela ne soit pas possible dans des conditions un peu plus réalistes. Le modèle ne permet donc pas de prévoir de manière définitive comment les génomes évolueraient si les taux de duplications augmentaient, mais au vu des résultats, la prédiction intuitive d'augmentation de la taille paraît trop simple : la taille des génomes pourrait très bien diminuer pour des raisons de robustesse.

# Chapitre V

# Étude de l'évolution du pourcentage de codant en fonction des taux de mutation

La démarche utilisée dans ce chapitre et les paramètres par défaut sont strictement les mêmes que pour le chapitre précédent. Ce chapitre est cependant plus exploratoire, nous allons présenter les résultats des simulations et un certain nombre d'éléments d'interprétation. Nous avons vu dans les chapitres précédents que la borne sur la taille du génome s'applique sur la somme des parties codantes et non codantes, nous essayons ici de comprendre la dynamique du non codant. Dans un déroulement classique de simulation, les gènes sont dupliqués, initialement sans que le non codant soit particulièrement éliminé. Quand la limite de taille est atteinte, le non codant est progressivement éliminé au profit de séquences codantes. Cependant, il semble que la population à l'équilibre maintient une certaine quantité de non codant qui dépend des paramètres du modèle. Nous présentons dans ce chapitre comment le pourcentage de codant varie en fonction des paramètres et explorons deux types d'interprétation : biais du modèle ou phénomène de robustesse ?

## 1 Avertissement : biais lié à l'implémentation du modèle

Avant d'explorer les résultats, il nous semble judicieux de mettre en garde le lecteur sur un biais d'implémentation qui affecte directement le pourcentage de codant, en tout cas dans un certain nombre de situations. Le modèle a été conçu pour permettre à la taille de varier librement et autoriser les génomes à contrôler leur pourcentage de codant pour illustrer le fait que le processus de duplications et de délétions limite la taille du génome, codant et non codant compris. Le découpage de l'espace que nous avons retenu est pratique du point de vue de la fitness : à chaque ligne correspond une valeur de fitness.

Cependant, cela entraîne d'autres effets indésirables qui joueront un rôle important dans l'interprétation des résultats. On va voir que le pourcentage de codant est essentiellement contrôlé par les mutations qui occasionnent des inactivations de gène. Or, il se trouve que ces mutations donnent des résultats un peu surprenants, *en l'absence du processus de délétion/duplication*, à cause justement des inactivations de gènes. Pour mieux comprendre ce phénomène, reprenons la figure qui explique l'agrégation des pertes de gènes par pseudogénisation (reproduite sur la figure V.1).

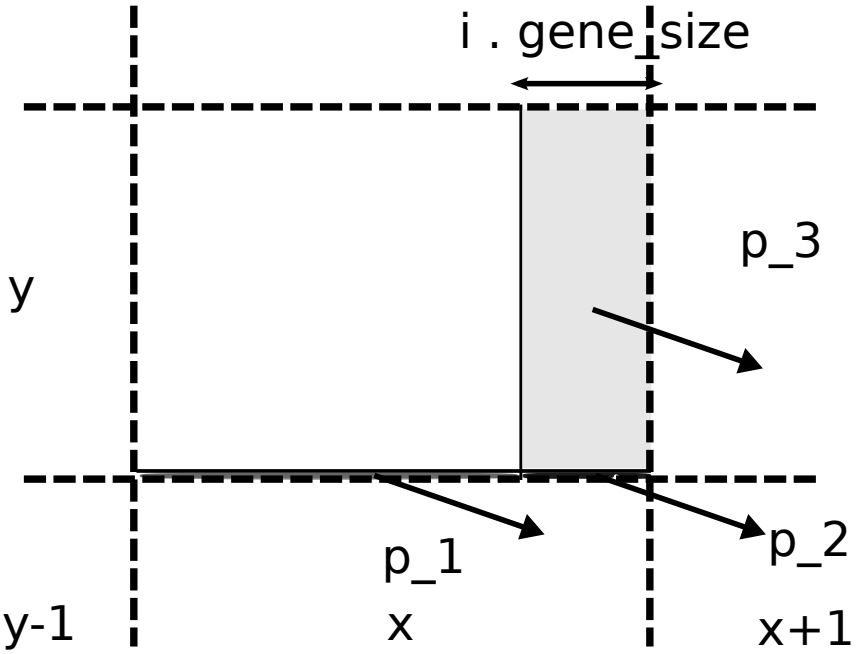


Figure V.1 – Peu d'individus changent de case en cas de pseudogénisation d'un gène (suite à une petite délétion dans du codant par exemple), mais la majorité de ceux qui se déplacent vont dans la subdivision directement à droite. Comme on homogénéise les populations à l'intérieur des cases à chaque génération, les choses se passent comme si les individus qui sont allés en $(x_0+1, y_0)$ avaient gardé le même nombre de gènes et augmenté leur fraction de non-codant.

On peut s'attendre à ce que la plupart des individus restent dans la subdivision initiale, seuls les individus sur les bords de la subdivision peuvent en changer. C'est le cas, notamment, des individus situés à moins de 1 kb (la taille par défaut d'un gène) du bord droit. Ceux-ci vont se déplacer d'une subdivision vers la droite quand ils perdent un gène. Parmi ces individus, la quasi-totalité va dans la subdivision directement à droite. Comme on homogénéise la population, on ne peut pas savoir quel est leur nombre de gènes exact, tout se passe donc comme s'ils conservaient leur nombre de gènes initial et qu'ils acquéraient simplement du non codant. Cette explication étant un peu abstraite, il vaut mieux prendre des exemples plus précis pour illustrer les problèmes que l'agrégation peut engendrer.

**Illustration 1 : mettre tous les taux à 0 sauf les petites délétions.** Si on ne conserve que les petites délétions, les génomes ne peuvent pas acquérir de gènes ni, en

théorie, augmenter la taille de leur génome. On prend une population qui a initialement un nombre arbitraire de gènes avec une sélection assez forte. On choisit un taux de petites délétions tel que les pertes de gènes ne soient pas rares mais pas systématiques non plus. La population est initialement étalée uniformément dans la subdivision qui contient le nombre de gènes choisi. Quand les individus qui se trouvent tout en bas de la subdivision perdent un gène, ils passent à la subdivision en dessous et sont éliminés par la sélection, si bien que la population est confinée sur la ligne de l'espace qui contient le nombre de gène initial. Ceci n'est pas très surprenant, puisque la population est infinie et que les individus ne peuvent que perdre des gènes, la sélection maintient l'avantage initial.

Le phénomène artefactuel se produit sur les bords à gauche et à droite. On s'attend à ce que le non codant soit érodé : c'est le contraire qui se produit. En effet, spontanément, les individus situées sur les 6 bp tout à gauche d'une subdivision peuvent perdre une partie du non codant et se diriger dans la subdivision sur la gauche. Le problème, c'est que les individus situés sur la droite à moins de 1kb du bord se dirigent vers la subdivision à droite en cas de perte de gènes, sans que la sélection ne les désavantage, comme l'agrégation masque le fait qu'un gène vient d'être perdu. On a donc deux flux inverses : d'un côté, l'érosion légitime du non codant qui pousse la population vers la gauche, de l'autre, le biais d'implémentation de la perte de gène qui pousse la population vers la droite. La force de ces courants dépend du nombre d'individus concernés (une bande de 6 bp d'un côté, de 1kb de l'autre) et de la probabilité de perdre un gène. Très grossièrement, on pourrait dire que le flux vers la gauche compense le flux qui va vers la droite quand

$$
\begin{aligned}
6\Pr\left[\text{petite délétion neutre}\right] &= 1000\Pr\left[\text{perte de gène}\right] \\
6\frac{L}{L+n} &= 1000\frac{n}{L+n} \\
6L + 6n &= 1000n + 6n \\
\frac{n}{L+n} &= \frac{6}{1006}
\end{aligned}
$$

Ce calcul est très approximatif mais il indique que la population se stabilise quand elle a un pourcentage de codant d'à peu près 0.6 %. On est donc très loin de la valeur théorique de 100 %. En pratique cependant, ce déplacement pathologique vers la droite (observé en l'absence de duplications et de délétions) est beaucoup plus lent que la convergence vers un état densément codant observée dans les simulations normales.

**Illustration 2 : calcul du déplacement moyen de la population subissant une petite délétion.** Pour quantifier plus précisément ce biais vers un gain de codant illégitime, on peut aussi calculer le déplacement moyen d'une population qui part d'une subdivision donnée $(x_0, y_0)$ et qui subit exactement une petite délétion. On peut faire le bilan des déplacements de la population, à l'aide de la figure V.1, en prenant en compte en plus les déplacements liés aux petites délétions neutres (pour simplifier, on suppose

que les délétions font exactement 6 bp). Les individus situés à moins de 6 bp du bord gauche vont en $(x_0 - 1, y_0)$ si la délétion est neutre. Si la délétion n'est pas neutre, ceux situés à moins de 1 kb du bord droite vont en $(x_0 + 1, y_0 - 1)$ s'ils sont sur la ligne tout en bas et en $(x_0 + 1, y_0)$ sinon. Ceux sur la ligne du bas à plus de 1 kb du bord gauche vont en $(x_0, y_0 - 1)$ en cas d'inactivation de gène. Dans tous les autres cas, les individus restent en $(x_0, y_0)$. En prenant en compte l'homogénéisation des cases, on peut calculer approximativement la variation moyenne de la taille totale du génome, qui doit en théorie valoir -6 bp. Les calculs étant élémentaires nous n'en donnons pas le détail ici. On trouve

$$\Delta s \simeq \frac{-3(-l_{gene}^2 \Delta n(y_0) + 6\Delta L(x_0))}{4(l_{gene}\Delta n(y_0) + \Delta L(x_0))}$$

Par exemple, considérons la subdivision de départ de coordonnées $(x_0 = 8, y_0 = 11)$ en échelle logarithmique, centrée autour d'un génome comportant 768 gènes et environ 96000 bases non codantes (pourcentage de codant de 89 %,). Le déplacement moyen vaut alors à peu près +667 bp, soit une erreur aux alentours de 0.08 % de la taille initiale. Si l'on considère à présent la subdivision de départ de coordonnées $(x_0 = 14, y_0 = 11)$, centrée autour du génome de 768 gènes mais avec 6144000 bases non codantes (pourcentage de codant de 11 %) le déplacement moyen vaut environ +85 bp, soit 0.001 % de la taille initiale. Ces valeurs sont cohérentes avec la valeur donnée dans la première illustration, mais l'approximation est moins grossière. On retrouve le fait que le biais à l'augmentation de la taille peut être assez fort comparé à la valeur attendue et ce d'autant plus que la fraction de codant est importante.

**Conclusion** L'homogénéisation des cases conduit donc à un biais en faveur de l'accumulation de séquences non codantes. Ce biais est cependant limité et, même si nous avons pris pour l'illustrer et le quantifier des exemples caricaturaux, ce biais sera, en pratique, filtré par la sélection. En effet, il faut faire attention à ne pas confondre spontané et fixé, moyenne et distribution. Même si le déplacement de $(x_0, y_0)$ vers $(x_0, y_0 + 1)$ est avantagé, le déplacement de $(x_0, y_0)$ vers $(x_0 - 1, y_0)$ existe bel et bien et peut, en théorie, être sélectionné. Nous avons vu que ce n'était pas le cas quand on conserve uniquement les petites délétions, mais l'enjeu n'est plus le même en présence de duplications et délétions puisqu'il existe une taille limite et un processus de création de gènes. Un génome fortement codant sera avantagé par la sélection, ce qui rend une convergence vers 0.6 % de codant assez improbable. Les petites délétions offrent alors une possibilité de diminuer la quantité de non codant qui peut être sélectionnée (même si cela sera plus difficile que cela ne devrait l'être).

# 2 Évolution du taux de codant en fonction des paramètres dans les simulations

Nous utilisons les mêmes représentations que dans le chapitre précédent, en indiquant la moyenne et la densité en couleurs pour donner une idée de la répartition de la population autour de la moyenne.

L'impact le plus simple à étudier est celui des mutations responsables d'inactivation de gènes (en comptant dans ces mutations les indels). Rappelons que l'impact de ces mutations sur la taille est très faible et remarquablement identique d'une mutation à l'autre. Le fait que la taille varie peu simplifie l'interprétation : on voit directement sur quel compartiment chaque mutation agit, une baisse du pourcentage de codant signifie forcément que le codant a diminué tandis que le non codant a augmenté.

Dans nos simulations, les taux d'inactivations de gènes semblent déterminer en grande partie le taux de codant. Pour le taux de petites insertions par exemple, le ratio moyen de codant à l'équilibre vaut 90% pour $\mu_{ins} = 10^{-8}$ et descend jusqu'à 60% pour $\mu_{ins} = 5.10^{-6}$. L'effet est plus prononcé quand le nombre de gènes potentiellement inactivés par événement est plus grand. Les petites insertions et les petites délétions (figure V.2A et B) peuvent causer l'inactivation d'un gène à chaque événement avec la même probabilité, les courbes semblent à première vue identiques. Les inversions (figure V.2C) générent deux points de cassure : on remarque que la pente de la courbe est plus accentuée que pour les indels. Le pourcentage de codant moyen part d'une valeur légèrement plus élevée que lors de la variation des indels et descend jusqu'à une valeur plus basse. Quand on modifie le taux de translocations (figure V.2D), cet effet est encore plus prononcé. En effet, une translocation peut causer jusqu'à l'inactivation de 3 gènes : 2 pour les points de cassure et un à l'insertion. On a donc une relation simple entre les taux d'inactivations de gènes et la proportion de codant.

Le cas des grandes délétions et des duplications est plus compliqué à analyser puisque la taille du génome à l'équilibre varie. Par exemple, une baisse de codant peut être due à une baisse du nombre de gènes, mais aussi à une augmentation plus rapide du non codant comparé au codant. De plus, les duplications sont elles-mêmes responsables d'inactivations de gènes à cause de la réinsertion du segment dupliqué.

Avant d'analyser les figures, il faut donc garder en tête l'évolution de la taille, quitte à se référer rapidement au chapitre précédent. Pour les duplications le comportement avant explosion est à première vue assez proche de celui des indels (figure V.3A). Ceci semble cohérent avec la faible variation de taille que nous avons observée au chapitre précédent, couplée au fait que l'inactivation d'un gène se fait avec la même probabilité que pour les indels. Cependant, il n'est en réalité pas directement comparable aux indels dans la mesure où la taille augmente légèrement. En moyenne, aussi bien la quantité de codant que la quantité de non codant augmentent quand le taux de duplications augmente, tant qu'on reste sous le seuil d'explosion. Il est donc possible que la quantité de non codant

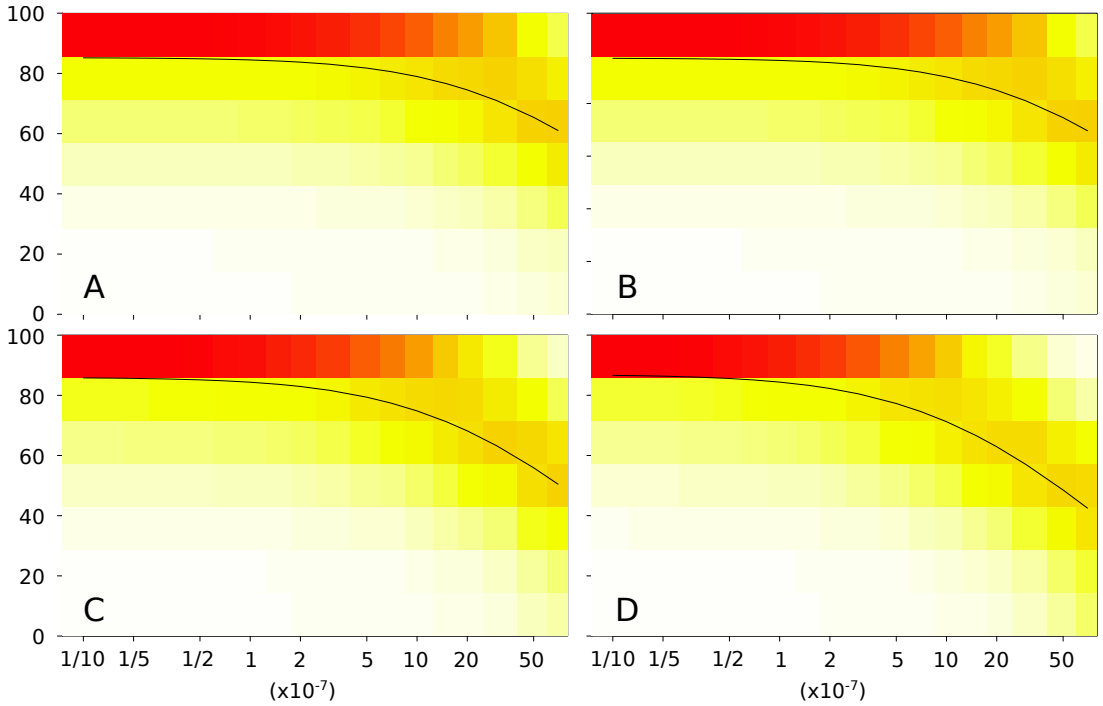


Figure V.2 – Évolution du pourcentage de codant en fonction des taux des mutations essentiellement responsables d'inactivations de gènes : petites insertions (A), petites délétions (B), inversions (C), translocations (D). Les taux de grandes délétions et de duplications sont fixés à $10^{-7}$, leur valeur par défaut. La baisse du pourcentage de codant est d'autant plus prononcée que le nombre de gènes qui peuvent être inactivés en un événement est grand (1 pour les indels, 2 pour les inversions, 3 pour les translocations).

soit expliquée par l'inactivation de gènes plus fréquente, mais la baisse du ratio de codant ne correspond pas à une baisse du nombre de gènes. On voit d'ailleurs qu'il semble y avoir une quantité optimale de duplications pour le ratio de codant : quand le taux de duplications devient très faible, le pourcentage de codant baisse aussi. Il y a donc bien une dynamique propre des duplications qu'on pourra chercher à mettre en rapport avec le fait que les duplications sont les seules mutations qui permettent la création de nouveaux gènes dans le modèle. Après explosion, les génomes tous vers le bord de l'espace simulé, d'où le taux de codant de 50 % qui n'a absolument aucune signification. En effet, après explosion, il n'y a plus de limite à la taille de génome, donc les génomes se stabilisent sur la ligne avec le plus de gènes (tout en haut) et l'élimination du non codant n'apporte pas d'avantage sélectif. Comme nous allons le voir un peu plus tard, l'acquisition de non codant est plus simple que son élimination, c'est pourquoi la population se stabilise sur la colonne tout à droite, avec la plus grande quantité de non codant.

L'évolution du pourcentage de codant avec le taux de grandes délétions correspond au cas où la taille change fortement (elle diminue quand on augmente le taux de grandes délétions), mais le taux d'événements responsables d'inactivation de gènes est constant. On observe une variation importante du pourcentage de codant (figure V.3B). On ignore à nouveau la partie qui correspond à l'explosion des génomes où le ratio est à 50 %. Quand on s'éloigne de la zone aux abords de l'explosion, le ratio baisse avec l'augmentation du

taux de délétions. Cette baisse de la proportion codante accompagne donc la baisse de la taille du génome observée au chapitre précédent : quand le taux de délétions augmente, on a donc des génomes de plus en plus courts et de moins en moins denses en gènes. Dans le détail, on observe une baisse du nombre de gènes et du non codant, mais la baisse du nombre de gènes est plus rapide. Il semble donc y avoir un effet dû à la taille du génome difficile à analyser *a priori*. En théorie, les inactivations de gènes par indels, inversions ou translocations sont plus rares dans un génome court puisque les taux de mutations sont exprimés par paire de bases. Leur impact relatif ne devrait donc pas varier avec la taille du génome, elles ne peuvent pas être tenues pour responsables de la diminution de la fraction de codant. On peut illustrer cette absence d'impact relatif en faisant varier les taux des mutations responsables d'inactivations de gènes de manière à avoir un nombre d'événements identique quelle que soit la taille du génome. D'après le chapitre précédent, la taille des génomes est inversement proportionnelle aux taux de délétions et de duplications. Si on varie tous les taux conjointement, les variations de tailles sont exactement compensées par le fait que les inactivations se font plus fréquentes ou plus rares, donc le nombre d'inactivations attendues est le même quelle que soit la taille du génome. Dans ce cas, la baisse du pourcentage de codant est encore plus dramatique (figure V.4), car on additionne l'effet des délétions et des inactivations de gènes. Il semble donc que, pour le cas initial de la figure V.3B, les grandes délétions soit seules responsables de la baisse de pourcentage de codant. Les grandes délétions peuvent également être responsables d'inactivations simples de gènes, il est donc plausible qu'elles agissent aussi bien sur la taille que sur le pourcentage de codant.

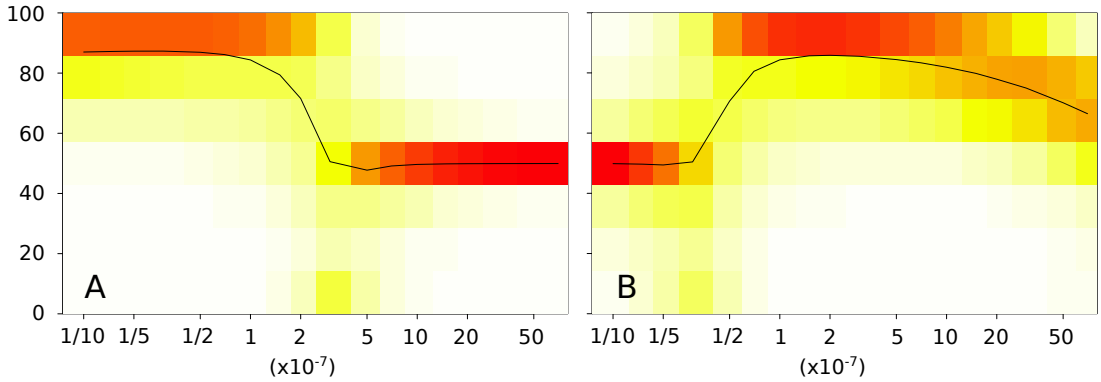


Figure V.3 – Évolution du pourcentage de codant en fonction des taux de duplications (A) et de grandes délétions (B).

## 3 Nouveau paradoxe : peut-on, en théorie, converger vers un pourcentage de codant qui ne soit pas 100% ?

Dans le modèle, les séquences non codantes n'apportent aucun avantage sélectif mais peuvent être présentes en assez grande quantité à l'équilibre. Intuitivement, on pourrait s'attendre à ce qu'elles soient éliminées, de façon à atteindre le plus grand nombre de gènes possibles. Le fait que la population ne converge pas vers 100% de codant peut

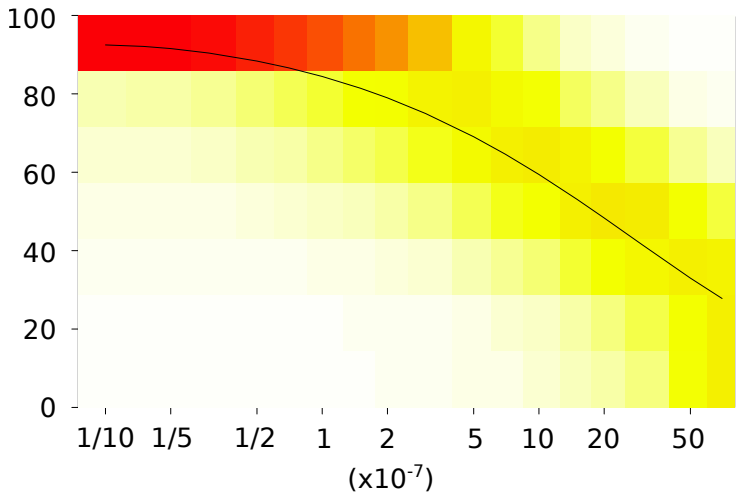


Figure V.4 – Évolution du pourcentage de codant lorsqu'on fait varier conjointement tous les taux de mutation.

donc paraître surprenant. Nous disposons pour l'instant de deux éléments importants. Premièrement, l'implémentation semble biaiser vers l'acquisition de non codant. Pourtant, avec les paramètres par défaut, les duplications et les délétions prennent globalement le dessus et permettent théoriquement à la population de converger vers une proportion de non codant relativement faible. Deuxièmement, dans les simulations, le pourcentage de non codant semble dépendre des inactivations de gènes (par les indels, les inversions et les translocations notamment), sans toutefois atteindre une proportion extravagante du génome pour les gammes de paramètres testés. Nous devons donc reposer la question initiale : ces résultats sont-il dus au biais d'implémentation ou s'agit-il d'un phénomène réaliste ?

## 3.1 Arguments en faveur du biais d'implémentation

Prenons d'abord le parti du biais d'implémentation. Cette suggestion fait sens puisque quand on augmente les inactivations de gènes via les taux d'indels, d'inversions ou de translocations (via les paramètres $\mu_{ins}$, $\mu_{sdel}$, $\mu_{inv}$ ou $\mu_{ldel}$), on peut estimer que le biais vers le gain de non codant est renforcé. D'autre part, comme dans le cas où on a éliminé toutes les mutations sauf les petites délétions, on peut se demander s'il y a une quantité théorique de non codant qui devrait pouvoir être atteinte. Une telle quantité peut être proposée : du point de vue de la robustesse, on pourrait estimer que les individus qui n'ont pas de non codant du tout sont avantagés à long terme comparé à des individus qui ont le même nombre de gènes et des séquences non codantes en plus. Dans ce cas, la convergence devrait se faire vers des états à 100 % de codant, ce qui indiquerait que nos résultats sont dû au biais d'implémentation. On peut comprendre ce raisonnement assez simplement en prenant « l'individu final typique », disons celui qui est situé au niveau du mode de la distribution, et en le comparant avec un clone duquel on a retiré toutes les séquences non codantes (figure V.5).

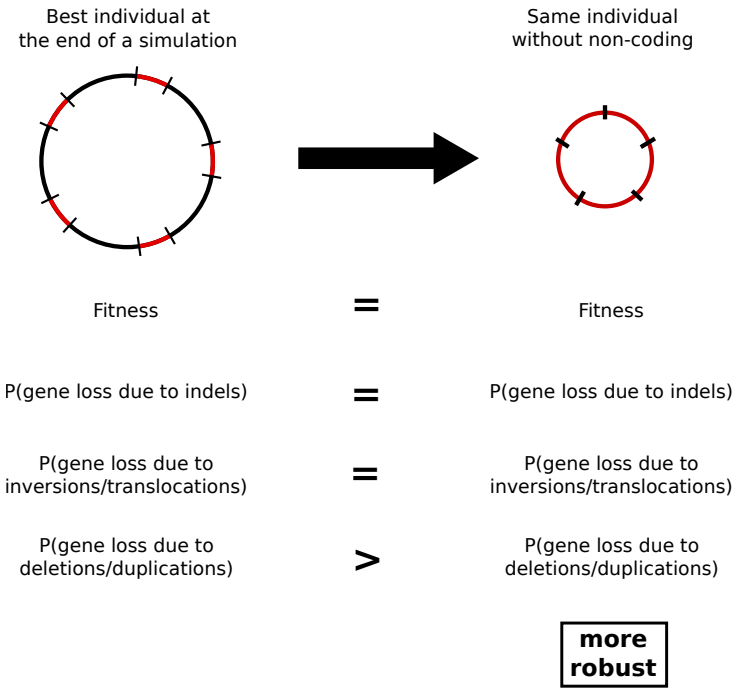


Figure V.5 – Nouveau paradoxe : si on compare le meilleur individu final avec une copie dont on a retiré les séquences codantes, on se rend compte que la copie est plus robuste.

Les deux génomes ont le même nombre de bases codantes. Comme les taux sont exprimés par paire de bases, la distribution de probabilité du nombre de gènes inactivés en une génération à cause d'un indel, d'une mutation ponctuelle ou d'un point de cassure sont les mêmes. Les génomes ont exactement la même robustesse vis-à-vis de la pseudogénisation. Par contre, les grandes délétions sont plus délétères pour le grand génome. En effet, une délétion donnée aura à peu près le même effet en nombre de gènes perdus, à cause de sa nature multiplicative. Par contre, le non codant offre plus de possibilités de points de cassures, le grand génome subira globalement plus de délétions en une génération, ce qui le rend moins robuste. Le cas des duplications est plus compliqué. Premièrement, à cause de la nature multiplicative, lors d'une duplication donnée, le nombre de gènes copiés est le même pour les deux génomes mais la probabilité d'inactiver un gène à l'insertion pour cette duplication est plus grande pour le petit génome : en moyenne, le petit génome gagne moins de gènes que le grand à chaque duplication. Deuxièmement, comme une duplication est dans ce modèle favorable, le grand génome en subira plus mais cela lui apporte en général un avantage sélectif. Cependant, on rappelle que l'individu choisi initialement est celui qui était l'individu typique, donc proche de la limite de taille. On peut donc estimer que les duplications sont délétères car les gains effectués ne sont que transitoires. En adoptant ce point de vue, le petit génome est à nouveau plus robuste puisqu'il subit moins de duplications : il a plus de chances de se reproduire à l'identique.

On pourrait donc argumenter que le petit génome devrait logiquement remporter la compétition face au grand génome. Dans cette hypothèse, la population devrait se stabiliser autour de 100% de codant. Si elle s'en éloigne, c'est donc à cause du biais d'implémentation.

## 3.2 Arguments alternatifs

Même si le biais d'implémentation existe, l'explication ci-dessus ne paraît pas complètement satisfaisante. Le point faible vient des duplications. En effet, même si le comportement dû aux duplications est transitoire, il a une influence sur la dynamique de la population. L'argument de robustesse pour le génome sans non codant tient dans la mesure où la sélection va le favoriser face à des descendants qui ont perdu des gènes. Pour comprendre vraiment ce qui se passe, il faut regarder le comportement à moyen terme en adoptant le même raisonnement qu'au début du manuscrit : quelle trajectoire un génome suit-il spontanément, les sauts effectués dans l'espace des structures sont-ils symétriques ?

Si on fait pour l'instant exception de la sélection, on peut étudier les trajectoires qui partent du génome entièrement codant. On obtient les déplacements les plus importants via les inactivations de gènes, les grandes délétions et les duplications. Dans tous les cas, on peut s'attendre à un gain de non codant par inactivation de gènes. On peut maintenant se poser la question : comment revenir à l'état initial ? Si un gène a été perdu, il faut, dans le modèle, en redupliquer un, ce qui s'accompagnera à nouveau par un gain de non codant car il y a peu de chances que le segment dupliqué contienne exactement un gène sans ADN non codant de part et d'autre. Sinon, il faut éliminer les gènes en surplus et le non codant restant par délétions (petites ou grandes). Ce processus d'élimination du non codant est extrêmement long puisqu'il est difficile, d'un point de vue probabiliste, d'éliminer le non codant dans chaque région intergénique sans affecter les gènes alentours. Moins il reste de non codant à éliminer, moins il est probable de réussir à l'éliminer, puisque le nombre de combinaisons qui permettent de le faire devient de plus en plus petit. S'il reste une seule base à éliminer, il est beaucoup plus probable d'inactiver un gène et regagner du non codant que de réussir effectivement à la supprimer. Résumons cela en une phrase. Les états à 100% de codant sont hautement inaccessibles via les processus de mutations : il est très facile d'en sortir et extrêmement difficile d'y retourner. Par comparaison, un état à 80 % de codant a une robustesse structurelle plus forte : on peut facilement (du point de vue probabiliste) gagner/perdre du non codant et gagner/perdre des gènes. On peut donc argumenter que les génomes pourront plus facilement se stabiliser dans une telle zone. Ce raisonnement est représenté schématiquement en figure V.6. Dans la suite, nous allons nous référer à ce phénomène sous le nom de « robustesse structurelle » des génomes non entièrement codants.

Maintenant qu'on connaît le comportement spontané des génomes, on peut essayer d'extrapoler des cheminements possibles via le prisme de la sélection. Commençons par signaler qu'on peut adopter deux points de départ. Si la population possède initialement du non codant, il faut déjà qu'elle réussisse à l'éliminer pour parvenir à 100% de codant, ce qui peut être long mais pas impossible, par exemple avec une sélection hautement purificatrice et un biais vers les petites délétions. On peut également supposer que la population se trouve déjà dans cet état et voir à quelles conditions elle peut en sortir. Si on suppose que la sélection est plutôt faible et que l'inactivation de certains gènes n'est pas létale, alors la population va pouvoir emprunter en partie la dynamique spontanée qui l'entraîne vers les zones plus robustes du point de vue de la structure. Il est également possibles que

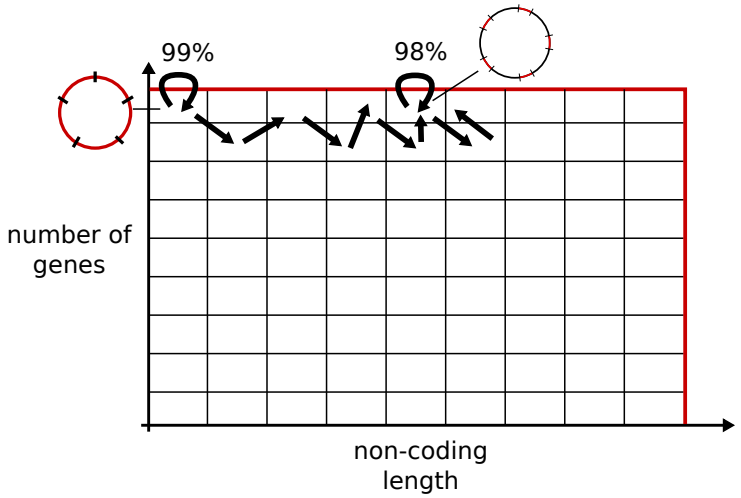


Figure V.6 – Idée de robustesse structurelle : spontanément, il est hautement improbable de maintenir 100% de codant. Il faut donc que la sélection y maintienne les individus.

certaines duplications favorisent un individu. La limite de taille est une borne supérieure, on a vu que la sélection influence en grande partie la position finale des génomes sous cette borne. Dans ce cas, si l'évolution s'accompagne de gains de gènes occasionnels, ils seront accompagnés du gain de séquences non codantes (qui jouxtent le gène dupliqué ou transféré) qui n'auront pas le temps d'être érodés (la probabilité de les supprimer en totalité avant le gain du prochain gène étant faible, même s'il y a un biais vers les petites délétions).

Il est difficile de prévoir l'effet exact de la fitness. Dans notre modèle, on peut estimer que l'état à 100 % de codant n'est pas stable à cause des duplications qui sont toujours favorables. Si on part de l'état entièrement codant, un individu qui duplique un gène (et gagne au passage du non codant) sera sélectionné, ce qui déplace la population vers la zone avec plus de non codant. Il est alors plausible que la population se stabilise dans une zone où la structure est plus robuste, avec un nombre de gènes aussi élevé que possible. Le biais d'implémentation rend l'érosion du non codant plus difficile, mais on se stabilise malgré tout à des valeurs de codant sont assez élevées, ce qui indique qu'il n'influence que très partiellement la dynamique. Les valeurs atteintes ne semblent donc pas avoir un fondement biologique profond, mais elles donnent cependant une indication de tendance qui pourrait avoir un sens biologiquement.

L'idée de robustesse structurelle *spontanée* peut, à notre avis, s'appliquer en biologie, mais il s'agit d'un phénomène qui sera très probablement contré par la sélection. Si on se place dans un cadre où le non codant est neutre, il semble malgré tout qu'il faille une forte érosion spontanée du non codant et une sélection fortement purificatrice pour pouvoir observer une disparition du non codant. Si une partie du répertoire génique est quasi neutre et fréquemment renouvelée, par duplication ou transfert horizontal, on peut envisager que les gains de gènes s'accompagnent de parties non codantes qui ne pourront pas être entièrement érodées. Dans cette interprétation, proche du *Junk DNA*, le non codant pourrait résulter d'un équilibre entre la fréquence du renouvellement du répertoire

et la vitesse d'érosion du non codant, qui sont eux-mêmes éventuellement liés à la sélection. On remarquera au passage que dans les simulations, on balaye les pourcentages de codant des bactéries, mais qu'on reste loin des eucaryotes.

Nous concluons en insistant sur le fait qu'il s'agit d'éléments d'interprétation et non d'une recherche d'explication du non codant observé dans les organismes réels. Nous pensons que les valeurs observées en simulation n'ont pas de signification profonde mais qu'elles permettent d'identifier un mécanisme qui peut éventuellement influencer le non codant, s'il n'est pas noyé dans une pléthore d'autres mécanismes qui ne sont pas pris en compte dans notre modèle. À ce stade, rien ne permet de totalement conclure et d'autres modèles sont requis pour avancer sur cette question.

# 4 Conclusion

Les variations de la fraction de codant dans le modèle varient de 20% à 90% pour les valeurs de paramètres illustrés. Cette fraction est, en apparence, contrôlée par les taux d'inactivations de gènes. On est *a priori* assez loin de pouvoir représenter la diversité observée dans la nature. Si on met en lien ces résultats avec ceux du chapitre précédent, cela confirme que les limites de taille s'appliquent bien au codant et au non codant, quel que soit le pourcentage de codant. Comme le modèle ne prend pas en compte un certain nombre d'événements qui affectent le non codant (comme la prolifération d'éléments transposables – qu'on pourrait intégrer théoriquement dans la distribution des insertions), que celui-ci est considéré comme complètement neutre (contrairement aux théories nucléotypiques ou nucléosquelettiques) et que la sélection soit fortement biaisée en faveur du codant, le fait que les valeurs soient loin des valeurs de références pour de nombreux eucaryotes n'est pas complètement surprenant. On peut en effet s'attendre à ce que les valeurs soient biaisées par le modèle et son implémentation.

Malgré cela, les tendances affichées dans le modèle peuvent être attribuées à un phénomène qui peut affecter de manière passive les « vrais » organismes, si son action n'est pas masquée par la sélection. En effet, certaines structures génomiques sont plus robustes que d'autres du point de vue des processus de mutation. Si la sélection autorise des changements de structure, comme des variations dans le répertoire génique, certaines structures sont faciles à quitter via les processus de mutation, mais difficiles à atteindre. Il est donc possible qu'elles ne surviennent que dans des cas très rares. Cet argument est en fait assez intuitif : si un génome subit fréquemment des réarrangements ou des transferts de gènes, un peu de non codant est nécessaire pour que ces opérations puissent se faire spontanément assez facilement. S'il n'y a pas de non codant initialement, les premiers transferts ou réarrangements ont de grandes chances d'en faire apparaître.

On en reste à ce stade au niveau de l'expérience de pensée, que nous avons tant décriée en introduction. Comme pour la taille, il semble qu'on atteigne ici les limites du modèle. Pour confirmer cette interprétation et lui donner une interprétation biologique plus profonde,

l'étude de modèles prenant en compte d'autres pressions sur le non codant s'impose.

# Chapitre VI

# Discussion

Dans cette thèse, nous avons utilisé un modèle de dynamique des génomes soumis à des petites mutations et des mutations dépendantes de la taille, avec ou sans sélection, pour explorer les mécanismes qui régissent la taille du génome. Les principaux résultats de notre modèle sont : (i) la condition d'existence pour une distribution stationnaire de la taille des génomes en l'absence de sélection, (ii) les bornes supérieures pour les quantiles de la distribution de la taille des génomes à chaque génération et (iii) le rôle de la fitness dans la stabilisation de la structure du génome.

## 1 Condition d'existence d'une distribution stationnaire pour la taille des génomes

Le cadre de modélisation et les résultats présentés ici permettent de déterminer la dynamique spontanée de la taille des génomes, dans de nombreuses conditions, puisque les généralisations de la fin du chapitre II autorisent un choix assez large pour les distributions spontanées de taille des petits indels, des duplications et des délétions. La condition d'existence d'une distribution asymptotique, c'est-à-dire la condition garantissant que la population ne s'échappe pas vers des tailles de génome infinies, ne dépend alors que de la moyenne de ces distributions après un changement d'échelle adapté à la distribution qui varie le plus amplement avec la taille du génome (vraisemblablement les grandes délétions et/ou les duplications). La partie la plus difficile peut être d'établir l'existence et la forme de la mise à l'échelle mais, une fois qu'elle est déterminée, la dynamique globale est donnée par une condition très simple sur les taux de mutations énoncée dans le corollaire 5.1. Pour de nombreux types de distributions où les indels ont un effet local alors que les duplications et les délétions ont un effet dépendant de la taille du génome, cette condition ne prendra en compte que le ratio des taux de duplication et de délétions. Ainsi, dans le cas particulier où les petits indels ont des effets additifs et que la taille des duplications et délétions est tirée selon une loi uniforme bornée par la taille courante du génome, la

condition d'existence d'une distribution stationnaire ne fait pas intervenir les taux de petits indels, et elle n'est *pas* trivialement $\mu_{dup} < \mu_{ldel}$ comme on pourrait le penser à première vue. La condition dans ce cas est $\mu_{dup} < 2.59\mu_{ldel}$, ce qui signifie que même si les duplications sont spontanément deux fois plus fréquentes que les délétions, la population ne s'échappe pas vers des tailles de génome infinies mais converge vers une distribution stationnaire. En effet, la dynamique spontanée des duplications et des délétions exerçant une « force de rappel » des très grands génomes vers de plus petites tailles.

Plus généralement, le corollaire 5.1 montre que quand des processus multiplicatifs (ou au moins plus qu'additifs) et additifs sont mélangés, il est possible d'obtenir des distributions stationnaires sans avoir à ajouter de sélection. Il s'applique à de nombreux cas non intuitifs, dans lesquels les taux relatifs favorisent les gains tandis que le déplacement moyen suite à une mutation après mise à l'échelle prédit des pertes. Même si mathématiquement le sens du déplacement après mise à l'échelle est évident, le biais apparent vers les gains peut être perturbant, parce que l'espérance de la taille du génome augmente à chaque génération. Par exemple, s'il y a deux fois plus de duplications que de délétions dans le cas uniforme, il y a création de « matière » et augmentation du nombre de paires de bases totales à l'échelle de la population. Pourtant la taille de la majorité des génomes aura tendance à être très faible, car la médiane se déplace dans le sens opposé de la moyenne.

Dans l'introduction, nous avons montré comment Falush et Iwasa (1999) ont obtenu par simulation une distribution de ce type (figure I.4, page 45) dans le cas des microsatellites avec un mode aux alentours de 10 éléments. À titre de comparaison, de nombreux modèles de microsatellites (par exemple Krüger et Vogel, 1975; Walsh, 1987) ont dû incorporer de la sélection pour que la distribution ne soit pas piquée en 1 ou divergente. Le fait d'ajouter un processus additif (comme les indels) à un processus multiplicatif ne change pas l'existence d'une distribution stationnaire mais change certaines propriétés de cette distribution.

Cette interaction entre processus multiplicatifs et additifs peut également être soulignée dans les études d'évolution réductive de la taille du génome. Mira *et al.* (2001) ont argumenté que les biais vers les petites délétions et les grandes délétions étaient des bons candidats pour expliquer l'évolution réductive observée chez certaines espèces de bactéries. En comparaison, Petrov base ses modèles exclusivement sur les petits indels pour prévoir la taille à l'équilibre du génome (Petrov, 2000, 2002). D'après notre modèle, s'il existe un processus de grandes délétions qui varie plus amplement que les indels avec la taille du génome, alors le biais vers les petites délétions ne contrôle que la dynamique spontanée des petits génomes, même si elles peuvent participer à l'érosion globale. Pour revenir à la situation étudiée par Mira *et al.* (2001), le cas des endosymbiotes, notre modèle montre aussi qu'un biais vers les petites délétions n'est pas nécessaire à une évolution réductive. Dans nos simulations, toute augmentation des taux des mutations « multiplicatives » (grandes délétions ou duplications) est susceptible de conduire à une évolution réductive. Plus précisément, les grandes délétions et les duplications imposent une limite de viabilité structurelle, mais dans le domaine de viabilité, l'effet drastique de ces mutations s'estompe en grande partie. Les effets de sélection, de dérive génétique et un biais vers les petites délétions peuvent alors contrôler en partie la convergence de la population

vers une structure génomique précise.

L'hypothèse la plus importante pour l'existence de la distribution stationnaire est la mise à l'échelle de la distribution des pertes et des gains dus aux grandes délétions et duplications. Nous avons initialement proposé des distributions uniformes qui permettent de copier ou de supprimer un segment de n'importe quelle taille du génome. Ce choix peut soulever de nombreuses objections, notamment si on se réfère à la taille de réarrangements fixés, qui est vraisemblablement plus proche d'une exponentielle, voire de lois un peu plus complexes (Sankoff *et al.*, 2005; Darling *et al.*, 2008). Cependant, ces réarrangements fixés, même s'il s'agit d'inversions, sont probablement filtrés par la sélection et peuvent donc être très éloignés de la distribution spontanée. Comme l'effet des duplications et des délétions est d'autant plus important qu'elles sont longues, on peut même s'attendre à ce que les distributions fixées soient fortement biaisées vers les petits événements. La létalité des grands événements rend une mesure de la distribution spontanée quasiment impossible en pratique. Cependant, il existe de nombreux éléments indiquant que la distribution uniforme n'est pas si caricaturale.

Les réarrangements se produisent physiquement par ruptures du chromosome et réassemblage. Elles peuvent se produire entre des zones homologues ou non-homologues du chromosome. On peut donc imaginer que la taille des réarrangements dépend surtout de la proximité physique dans la cellule des différentes zones du chromosome. Par exemple, si un bout de chromosome est physiquement voisin, dans la cellule, d'une séquence diamétralement opposée sur le chromosome, mais très éloigné d'une autre séquence située à 1 kb, il pourrait y avoir plus de chances de provoquer un réarrangement qui implique la moitié du chromosome que le petit segment de 1 kb. En plaçant des bactéries dans des conditions de laboratoire, on relâche en partie des pressions sélectives, ce qui peut conduire à des tailles de délétions et duplications importantes (Porwollik *et al.*, 2004). Nilsson *et al.* (2005) ont pu observer des délétions de plus de 200kb, même entre des séquences non homologues, indiquant que les points de cassure peuvent survenir n'importe où. L'apparition de la même délétion dans plusieurs populations évoluant indépendamment suggère que les délétions sont fréquentes et généralement filtrées par la sélection (Nakatsu *et al.*, 1998) et qu'elles peuvent être pilotées par les éléments transposables (Schneider *et al.*, 2000; Gaffé *et al.*, 2011). Cooper *et al.* (2001) suggèrent même que la délétion peut se produire via transposition d'un élément transposable à une position aléatoire du génome puis recombinaison immédiate entre les deux copies. En effet, dans leurs expériences, ils obtiennent des délétions dans 12 populations indépendantes qui délètent toutes le même gène à partir d'un point de cassure correspondant à un élément transposable bien identifié, mais le deuxième point de cassure survient à une position aléatoire à la sortie du gène. Par ailleurs, dans le génome humain, de grandes duplications et délétions ont été identifiées parce qu'elles étaient à l'origine de nombreux cas sporadiques de maladies génétiques. Par exemple, pour la moitié des patients atteints d'une maladie de Charcot-Marie-Tooth[1], la maladie est causée par une duplication de 1.4 Mb, et ce réarrangement se produit suffisamment souvent pour qu'à la fois des cas hérités et des cas *de novo* soient observés

[1]Les maladies de Charcot-Marie-Tooth regroupent un ensemble de maladies neurologiques génétiques parmi les plus fréquentes (1 naissance sur 2 500 en France). La CMT entraîne des troubles de la marche et une déformation fréquente des pieds.

dans la même famille (Lupski, 2007). De même, 90% des patients atteints du syndrome de Smith-Magenis[1] ont une délétion d'une partie du chromosome 17, allant de 650kb à 9Mb selon les patients (Lupski, 2007). Pour donner un ordre de grandeur, une délétion de 9Mb correspond environ à deux fois la taille du génome complet de la bactérie *E. coli* K12 (4.6 Mb).

Même dans les données citées ci-dessus, la létalité limite directement la taille des délétions observables. Les expériences menées par Lynch actuellement sur plusieurs espèces de bactéries visent à étudier le mieux possible leur dynamique de mutation via des expériences d'accumulations de mutations, qui augmentent au maximum la dérive pour permettre la fixation d'événements délétères (mais pas létaux). L'équipe de Lynch aurait été surprise par l'impact des grandes délétions dans ces expériences, y compris en ce qui concerne le lien avec la taille du génome (données non publiées, communication personnelle de Way Sung, post-doctorant qui travaille sur le projet).

Même si la distribution uniforme n'est probablement pas la distribution réelle des réarrangements, il est difficile avec les données dont nous disposons d'argumenter en faveur d'une autre distribution. Prendre la distribution fixée serait une erreur dans la mesure où nous montrons que ce sont les caractéristiques de la distribution spontanée qui donne les conditions de convergence. Comme la distribution est inconnue, nous avons tenu à proposer la formalisation plus générale du corollaire 5.1. Quelles que soient les distributions, tant qu'il existe une mise à l'échelle non linéaire, on retrouve les phénomènes intéressants décrits dans le manuscrit. Du point de vue d'un génome en tant que marcheur le long d'une chaîne de Markov, la symétrie est celle de la réversibilité des sauts : les pertes et les gains que je subis quand je bouge le long de la chaîne se compensent-ils ? La mise à l'échelle permet de répondre à cette question si la taille moyenne des sauts tend vers une constante pour chaque point de départ. Dès qu'il existe une mise à l'échelle, la symétrie apparente de gains et de pertes pourra alors être brisée et la dynamique spontanée des génomes (croissance ou décroissance) ne pourra pas se calculer intuitivement.

Si les processus de duplications et de délétions sont de nature quasi-multiplicative, la condition pour la dynamique spontanée s'obtient en échelle logarithmique, comme illustré dans le chapitre II. On peut alors remplacer la distribution par une distribution uniforme tronquée, exponentielle tronquée ou même multimodale, tant que la distribution s'exprime en fonction de la taille relative des pertes et des gains au lieu de la perte absolue. Par contre, si on prend des familles de distributions intermédiaires (plus qu'additives, moins que multiplicatives), par exemple avec des tailles moyennes de gains et de pertes qui n'évoluent pas linéairement avec le génome, l'établissement même de l'existence de la mise à l'échelle peut être compliqué. Il serait intéressant de déterminer des conditions suffisantes pour l'existence d'une mise à l'échelle appropriée des distributions, voire une façon de déterminer explicitement la mise à l'échelle quand elle existe. Cela permettrait de donner un contour plus net aux familles de distributions pour lesquels notre théorème s'applique, ce qui renforcerait très certainement sa plausibilité biologique.

[1] Le syndrome de Smith-Magenis est l'association d'un visage caractéristique, d'un retard de développement, de troubles cognitifs et des anomalies du comportement.

L'unicité de la fonction de mise à l'échelle est certainement la plus simple à établir, à condition de prendre en compte les équivalences asymptotiques pour les fonctions, vu que le théorème s'applique pour des tailles de génomes qui tendent vers l'infini. Dans ces conditions, il est assez clair que si $f$ est une fonction de mise à l'échelle répondant au problème, les fonctions $\{\lambda f, \lambda \in \mathbb{R}^*\}$ conviennent également. On peut s'attendre à une unicité à $O(f)$ près. Le problème de l'existence revient surtout à assurer une compatibilité (toujours asymptotique) de la mise à l'échelle pour les distributions des différents points de départ. En effet, si on prend en compte un point de départ isolé, on peut trouver une mise à l'échelle qui envoie sa distribution de pertes et de gains vers une distribution souhaitée (aux effets discrets près). Intuitivement, si on imagine une distribution cible sous forme d'histogramme, cela revient à déplacer tous les points d'arrivée (en maintenant leur ordre) dans les barres de l'histogramme pour les remplir successivement : ce n'est pas un exercice très difficile. Le problème, c'est que la manière de disposer les points d'arrivée doit aussi correspondre aux points de départs plus petits et plus grands. C'est dans ce sens là qu'il doit exister une forme de compatibilité entre les différentes distributions.

Pour finir, le modèle fait l'hypothèse que l'impact et le taux des mutations est homogène le long du génome, ce qui n'est pas le cas dans des organismes réels. Certaines parties du génome subissent spontanément plus de mutations locales ou de réarrangements (Aguilera et Gómez-González, 2008). Proposer un modèle qui prenne en compte ces inhomogénéités nous semble hors de portée, il paraît plus raisonnable de commencer par un modèle homogène. Cependant, on pourrait restreindre le modèle pour l'appliquer aux plasmides ou à des zones particulières du génome, voire découper le génome en plusieurs zones pour prendre en compte une partie des inhomogénéités, tout en conservant le résultat principal.

## 2 Bornes supérieures des quantiles de la distribution de la taille des génomes

En utilisant une approximation continue des processus de duplication et délétions, nous avons quantifié la fragilité des grands génomes en calculant des bornes supérieures pour les quantiles de la distribution de la taille des génomes après une génération. Pour la médiane de la distribution par exemple, la borne supérieure obtenue ne varie pas de façon monotone avec la taille de génome initiale : au-delà d'une certaine taille initiale $s_0^{max}$, la taille médiane après une génération *décroît* lorsque la taille initiale $s_0$ augmente, ce qui signifie que les descendants d'un génome de départ A plus grand qu'un génome B tendront à avoir des génomes *plus courts* que ceux des descendants de B.

L'hypothèse que plusieurs mutations puissent se produire à chaque réplication est importante pour le calcul de ces bornes (alors qu'elle n'influençait pas la condition de dynamique spontanée). En effet, contrairement à la chaîne de Markov mutationnelle pour laquelle chaque saut correspond à une mutation, la chaîne de Markov générationnelle montre qu'il y a une probabilité non négligeable que les génomes situés au dessus d'un certain seuil, aussi grande que puisse être leur taille, s'effondrent. Ainsi, plus la taille initiale est grande,

plus la probabilité d'effondrement est grande, car les pertes dues au nombre croissant de mutations augmentent plus vite que la taille du génome. Par exemple, l'ensemble des génomes qui parviennent à conserver au moins leur taille initiale, et *a fortiori* leur structure, avec une probabilité plus grande que 0.5 est confiné à un domaine fini.

Le fait que le nombre de mutations soit donné par une loi de Poisson qui dépend uniquement de la taille initiale simplifie les calculs mais peut être discuté. En effet, cela signifie que, même si le génome a subi une importante délétion, le nombre d'événements n'est pas revu à la baisse : il continuera à accumuler des indels, des délétions et des duplications dont les effets dépendront de la taille actuelle, mais dont le nombre a été calculé grâce à la taille initiale. Cette hypothèse permet de réaliser les simulations plus facilement, mais ne semble pas nécessaire dans l'analyse mathématique. On peut envisager un autre type de processus de Poisson, dit non homogène, qui permettrait de corriger cet effet. Dans un processus de Poisson non homogène, les occurrences des mutations restent indépendantes, l'effet des mutations continue de dépendre de la taille actuelle, mais on peut ajuster les fréquences des mutations de manière à prendre en compte la taille actuelle. Plus précisément, dans le processus homogène, on peut obtenir le nombre et les temps d'occurrences d'une mutation de taux par base et par génération $\mu_i$ en simulant un processus ponctuel de Poisson de paramètre $\lambda = \mu_i L$ sur l'intervalle $[0, 1]$. Dans le processus non homogène, le paramètre dépend de la taille actuelle du génome et devient $\lambda(t) = \mu_i L(t)$. Par exemple, si le génome devient petit, les événements deviennent plus rares. Nous avons déjà utilisé cette idée dans des simulations individu-centrées pour un problème similaire. Il serait intéressant de refaire l'analyse du modèle avec ce processus et de voir si le résultat sur les quantiles reste valide. Dans ce cas, on s'attend à ce que les gros génomes mutent toujours autant mais cessent de muter quand ils reviennent dans la zone où les génomes sont stables, puisque les mutations y sont rares. À première vue, cela ne semble pas remettre en cause le fait que les génomes puissent revenir sous la borne de stabilité, mais ils ne deviendraient pas aussi petits.

L'existence de ces bornes a deux implications importantes. D'abord, ce ne sont pas uniquement des bornes pour la distribution stationnaire mais pour l'ensemble du processus. Cela signifie que même en appliquant de la sélection, elles restent valables pour les descendants des individus qui ont pu se reproduire. En d'autres termes, la sélection ne peut pas permettre de dépasser ces bornes, même si la sélection favorise les plus gros génomes. En pratique, les individus ayant une probabilité importante de maintenir leur taille sont ceux qui ne subissent que rarement des réarrangements. Dans ce sens, notre modèle prédit que les génomes se stabilisent spontanément dans un régime de faible taux de mutation (au moins pour les duplications et grandes délétions), ce qui est pris comme hypothèse par de nombreux autres modèles. Plus spécifiquement, la plupart des individus vont subir au plus une duplication ou grande délétion (ou même une mutation si les taux sont similaires) par génération. Le fait d'autoriser plusieurs mutations en une génération entraîne une pression indirecte qui limite la taille des génomes mais qui est en fait assez rarement observée en pratique.

L'impossibilité d'une replication fidèle à long terme pour les grands génomes rappelle l'*error threshold* d'Eigen (Eigen, 1971; Eigen *et al.*, 1988). Dans son modèle, si un po-

lymère dépasse la taille critique, le nombre de mutations par réplication est si élevé que l'information contenue dans la séquence du polymère est progressivement perdue dans les générations qui suivent. Notre étude montre que la nature des mutations prises en compte dans le modèle est importante quand il s'agit d'étudier l'évolution globale de la structure du génome, incluant codant et non codant. Si seules les mutations locales sont considérées, l'*error threshold* s'applique à la partie codante du génome (Eigen *et al.*, 1988). Nous avons montré que si les réarrangements sont pris en compte, la partie non codante du génome est également bornée, car l'ADN non codant est mutagène pour les gènes environnants dans la mesure où il augmente le nombre de points de cassure potentiels pour les délétions et les duplications. Ce phénomène avait déjà été observé dans aevol (Knibbe *et al.*, 2007). En ajoutant des réarrangements dans le modèle original d'Eigen, nous prédisons donc un *error threshold* généralisé qui s'applique au génome entier.

Il est important de rappeler que l'existence des bornes sur la taille du génome repose sur des taux de réarrangements chromosomiques exprimés par paire de bases pour déterminer le nombre d'événements par génération. Il est plausible que les taux de réarrangements ne doivent pas être exprimé par paire de bases mais en termes du nombre d'éléments responsables des réarrangements. C'est certainement le cas pour la recombinaison homologue, qui nécessite au niveau des deux points de rupture des séquences fortement similaires et assez longues, comme des éléments transposables par exemple. Pour la recombinaison NHEJ par contre, l'hypothèse d'un taux par paire de base est plus défendable. Pour étudier la pertinence de notre analyse, il faudrait mieux caractériser la façon dont la taille et les taux spontanés des réarrangements varient avec la taille du génome. Certains réarrangements sont dus à des séquences particulières au sein du génomes, comme les éléments transposables ou les séquences répétées en tandem. Certaines données indiquent que le nombre d'éléments transposables augmente avec la taille du génome (Oliver *et al.*, 2007), mais leur impact sur le taux spontané de réarrangements reste à établir.

Le lien entre réarrangements et taille du génome est souvent mis en avant chez les plantes, en réponse à l'article de Bennetzen et Kellogg (1997) sur la nécessité d'existence d'un mécanisme de délétion d'ADN. Leur analyse a été reprise par d'autres groupes chez d'autres espèces de plantes qui arrivent à la même conclusion : les phases de réduction seraient même plus fréquentes que les phases de grossissement (Wendel *et al.*, 2002). Dans ce débat, les éléments transposables ont un effet contradictoire : ils seraient éventuellement responsables de la croissance et de la diminution. Chia *et al.* (2012) ont comparé l'évolution de 28 espèces de maïs sauvages et domestiques et calculé les corrélations entre abondance de familles d'éléments transposables et taille du génome. Certaines familles, localisées dans l'hétérochromatine dans des structures très denses appelées *knobs*, corrèlent positivement avec la taille du génome (plus exactement la taille corrèle avec le nombre de *knobs*), tandis que la majorité des familles corrèle négativement avec la taille du génome. On observe un effet similaire chez *Arabidopsis* : la réduction de la taille du génome pourrait être liée à la capacité d'une famille spécifique d'éléments à produire des réarrangements (Devos *et al.*, 2002).

Ces résultats sont un peu compliqués à interpréter tels quels et sont controversés pour plusieurs raisons. D'abord, il existe deux types d'analyse distincts. En considérant la

recombinaison au sens de recombinaison homologue allélique, on ne trouve pas spécifiquement de lien avec les éléments transposables (Rees et Durrant, 1986; Ross-Ibarra, 2007). A l'opposé, on peut se focaliser sur la recombinaison ectopique (non allélique) et donc rechercher plutôt les délétions intrachromosomiques. Pour estimer leur taux, on compte les délétions d'éléments transposables à séquences LTR (*Long Terminal Repeat*). En l'absence de délétion, on s'attend à trouver deux LTR pour un gène codant une transposase, donc un rapport 2 :1. Si le rapport est plus grand que 2 :1, on déduit que de nombreux éléments ont été délétés par recombinaison des séquences LTR (Vicient *et al.*, 1999; Rabinowicz, 2000; Devos *et al.*, 2002). De telles méthodes commencent à être utilisées pour d'autres organismes à gros génome, comme les salamandres. Les salamandres à plus petit génome auraient un excès de LTR « solos » plus prononcé que celles à grands génomes, indiquant des délétions (fixées) plus fréquentes (Rachel Lockridge Mueller, présentation SMBE 2013 et communication personnelle).

L'autre point de la controverse porte sur le lien entre taille du génome et le nombre de rétrotransposons. En prenant en compte plusieurs espèces de plantes, on observe une corrélation faible mais positive entre la taille du génome et le nombre de rétrotransposons (Vicient *et al.*, 1999). Cela ne contredit pas pour autant le lien entre recombinaison et éléments transposables. Il y aurait plus de chances que la recombinaison soit liée à la densité et pas au nombre absolu : dans un génome avec deux éléments transposables, ceux-ci ont plus de chances de « se rencontrer » lors du repliement si le chromosome est petit. Par ailleurs, nous venons de voir que l'activité recombinante dépend des familles d'éléments transposables et de leur position dans le génome. Enfin, des études comparatives entre *Arabidopsis* et le coton (proche phylogénétiquement mais avec un génome 20x plus grand) suggèrent que l'activité de recombinaison plus intense pourrait être reliée à des défauts de la réparation de rupture double brin de l'ADN (Kirik *et al.*, 2000; Orel et Puchta, 2003). S'il existe vraiment des différences de ce type, *Arabidopsis* pourrait recombiner plus fréquemment, même sans avoir une densité d'éléments transposables plus élevée.

Sloan *et al.* (2012) suggèrent que la taille des génomes mitochondriaux de certaines plantes pourrait être due à la recombinaison via les éléments répétés. Ils étudient 4 génomes mitochondriaux : 2 petits et 2 grands, avec des nombres et des densités très variables en éléments répétés qui ne sont pas reliés directement à la taille de génome (le plus grand génome est celui qui a la plus forte densité d'éléments répétés). Ici, l'activité de recombinaison serait plutôt corrélée avec la taille des séquences répétées, mais de toute façon beaucoup plus forte pour les petits génomes, même à taille de séquence répétée égale. Paradoxalement, les gros génomes seraient donc ceux qui recombinent spontanément le moins et évoluent pourtant le plus vite car ils fixent plus d'événements. L'analyse du système de Cairns-Foster chez *E. coli* (Andersson, 1998; Kugelberg *et al.*, 2006) dans lequel un gène essentiel est inactivé par frameshift semble suggérer une conclusion similaire. Apparemment, les individus arrivent mieux à survivre s'ils parviennent à dupliquer le gène déficient (cela augmente la probabilité de produire une protéine fonctionnelle par « erreur ») mais ne semblent pas à dépasser un certain nombre de copies. Il existe 2 sous-populations : la première atteint 10 copies délimitées par des séquences répétées de 1kb, l'autre atteint 100 copies délimitées par des séquences répétées de quelques bases seulement. Les auteurs concluent à des « instabilités » empêchant d'acquérir plus de

copies, sans rentrer plus avant dans les détails, mais on peut l'interpréter comme une impossibilité d'acquérir plus de copies à cause de la probabilité de délétion. Chez les bactéries comme chez les plantes, le lien entre éléments transposables et recombinaison reste difficile à élucider (Touchon et Rocha, 2007), les taux de recombinaisons pourraient plutôt refléter les problèmes de réparation de rupture d'ADN, notamment l'activité du système MMR (*MisMatch Repair*) qui semble assurer la stabilité du chromosome (Nilsson *et al.*, 2005).

## 3 Rôle de la sélection

Le chapitre II prévoyait une limite supérieure de stabilité pour les génomes qui ne pouvait pas être surmontée par la sélection. Pourtant, cela ne permet absolument pas de prévoir où exactement la population va se positionner sous cette limite : la sélection joue un rôle crucial dans la détermination de la structure des génomes. En implémentant numériquement le modèle, nous avons pu confirmer les prédictions analytiques, mais aussi étudier comment la sélection influence la convergence de la population.

La sélection des génomes les plus grands permet aux individus d'acquérir des fractions de codant importantes, mais on observe toujours une convergence vers une distribution stationnaire (et non une croissance infinie), y compris dans les cas où les gains sont favorisés (dans la limite de 2.58 duplications pour une délétion). Ce résultat contraste avec d'autres modèles pour l'évolution avec sélection incorporant une fitness non bornée, qui montrent que la vitesse de croissance du premier moment de la distribution de fitness de la population converge vers une constante positive, même si un cut-off empêche les individus les plus adaptés de se reproduire (Tsimring *et al.*, 1996; Brunet et Derrida, 1997). Ici, nous n'avons pas besoin d'introduire un cut-off pour empêcher la croissance infinie des génomes, même si la sélection favorise les grands génomes.

De plus, l'impact d'une pression à l'augmentation des génomes a des effets qui semblent délétères, du moins à l'échelle de la population. Dans le cas d'une population infinie, cela favorise la sélection de quelques individus fortement instables qui parviennent tout juste à se maintenir mais produisent une très large quantité de mutants. On obtient donc une population éclatée, et ce d'autant plus que la sélection est forte et que la quantité de duplications est élevée. Les duplications et les délétions n'ont donc pas du tout la même influence. Les délétions sont les premières responsables de la limitation de la taille du génome, tandis que les duplications sont essentielles pour l'évolution du répertoire génique, mais pas absolument nécessaires pour maintenir la taille du génome. Quand on diminue le taux de duplication, les génomes parviennent à maintenir à peu près leur taille alors qu'une augmentation du taux de délétion a des répercussions drastiques.

Ces résultats suggèrent que la compréhension des processus de délétions peut être plus importante que celle des processus de duplications pour comprendre l'origine évolutive de la taille du génome. Par contre, pour un taux de délétions donné, il semble exister une

valeur optimale de taux de duplications en terme de pourcentage de codant. Cette valeur est difficile à analyser, mais laisse penser qu'il y a une valeur de duplication intermédiaire qui permet une création suffisamment fréquente de nouveaux gènes mais suffisamment rare pour ne pas provoquer une croissance du génome qui conduise, à terme, à déstabiliser la structure. Cela impliquerait que la taille des génomes puisse effectivement décroître avec le taux de duplications, la population convergeant suffisamment loin de la zone d'instabilité (où loin dépend du taux de duplication).

L'interprétation des pourcentages de codant est très délicate, à cause des hypothèses du modèle et du biais d'implémentation induit par l'agrégation des états élémentaires en méta-états. Elle laisse cependant entrevoir un nouveau conflit entre dynamique spontanée des génomes et action de la sélection. D'un point de vue probabiliste, il semble difficile de converger vers des états où les génomes ne contiennent que du non codant. Il y a donc une sorte de dérive vers une acquisition de non codant qui devrait logiquement être contrecarrée par la sélection. Cependant, cette dérive ne semble pas pouvoir être totalement contrecarrée par la sélection à cause de la création de nouveaux gènes, qui s'accompagne quasiment systématiquement d'un apport de non codant. Nous n'avons à ce stade pas de prédiction analytique ni de modèle adapté à une compréhension exacte de cette dynamique, d'autant plus qu'il faut prendre en compte le fait que les génomes sont à la limite de l'instabilité avec la sélection que nous utilisons. On peut tout de même conclure que la limite de la taille des génomes s'applique bien indépendamment du pourcentage de codant.

À partir de ces observations, on pourrait essayer de concevoir un modèle indépendant de celui étudié ici, qui prenne en compte le renouvellement du répertoire génique (par duplication ou transfert horizontal). En effet, chez les bactéries, ce renouvellement peut être assez rapide et important pour l'évolution (Ochman et Moran, 2001; Touchon *et al.*, 2009), il pourrait donc avoir un effet colatéral sur la structure des génomes. De plus, d'après Ochman et Davalos (2006), la quantité de non codant chez les bactéries a longtemps été sous-évaluée, notamment celle due aux pseudogènes. Le modèle devrait prendre en compte l'afflux de séquences non codantes et son élimination (par biais mutationnels ou sélection) en prenant en compte la difficulté croissante de supprimer le non codant. On peut néanmoins s'attendre à ce que ce mécanisme soit responsable au mieux d'une fraction assez faible de non codant.

Le modèle, que nous avons voulu aussi simple que possible, permet donc de nombreuses prédictions mais montre simultanément ses limites. L'absence d'éléments transposables explicites ne semble pas remettre en cause le raisonnement fondamental, puisque leur effet direct serait de même nature que les indels, donc asymptotiquement plus faible que les délétions. Par contre, ils pourraient avoir un effet majeur sur le pourcentage de codant. Pour prendre en compte des effets plus fins de robustesse, il existe plusieurs solutions. Il est envisageable de modifier la simulation pour qu'elle prenne en compte une dimension supplémentaire qui quantifie la redondance génétique. Cette modification rentre toujours dans le cadre du modèle général, mais la dimension supplémentaire alourdirait considérablement l'implémentation et nécessiterait l'introduction d'un nouveau modèle indépendant pour quantifier l'évolution de la redondance génétique.

Il serait intéressant de prendre en compte certaines formes de sélection analytiquement. Pour cela, il faudra vraisemblablement changer de formalisme, tout en restant dans un cadre stochastique. Si on veut que la fitness dépende de la fréquence des autres types d'individus présents dans la population ou que le nombre d'individus puisse varier, il faut ajouter un processus qui caractérise la population. Il existe des exemples de processus de Markov sur l'espace des mesures appliquées à l'ensemble des génomes, comme dans les modèles proposés par Fleming et Viot (1979) et Champagnat *et al.* (2006), mais l'incorporation de la sélection complique toujours l'analyse du processus. D'autres processus prennent en compte une fitness de population dépendant de la fréquence d'individus présents, mais utilisent un espace de structure de génomes simplifiés par rapport à notre modèle (Lambert, 2005, 2006). Nous ne disposons actuellement pas des outils pour faire le lien entre notre modèle et les processus cités ci-dessus mais, à première vue, il semble compliqué d'articuler l'espace des structures avec l'espace des fitness.

La dernière solution consiste à utiliser la modélisation individus-centrée, qui permet de prendre en compte tous les aspects manquants. Cela est un peu contraire à notre démarche, qui visait au développement d'un modèle simple et analysable. Pour prendre en compte les effets dus à la dérive ou à la sélection indirecte pour la robustesse dans un modèle individu-centré, aevol offre un très bon compromis. En effet, il est soumis à la limitation de la taille du génome par les processus de délétions, mais la convergence exacte de la taille du génome dépend en réalité des autres paramètres du génome. Un étude récente montre ainsi qu'une sélection plus faible, augmentant les effets liés à la dérive pour les gènes non-essentiels, mène à une érosion conséquente du génome, notamment du non codant (Batut *et al.*, 2013).

Même si nos résultats révèlent une contrainte sur la taille des génomes basée sur la dynamique spontanée des réarrangements chromosomiques, ils n'invalident pas pour autant les explications incluant une taille optimale ou donnant un rôle prépondérant aux petits indels. D'après nos résultats, le processus de délétion et de duplication impose une limite au-delà de laquelle les génomes deviennent instables. Réciproquement, cela signifie qu'il existe toute une zone où les génomes sont stables. Les conclusions de cette thèse ne contredisent donc formellement aucune de ces théories, même si elles peuvent en modifier l'interprétation ou le cadre. Dans la zone où les génomes sont stables, les réarrangements sont suffisamment rares pour qu'ils n'entravent ni les adaptations ni la dérive. De nombreuses pressions s'exercent alors simultanément sur les génomes réels mais les effets spontanés peuvent être masqués par la sélection directe. L'avantage d'études telle que celle présentée ici est que les pressions directes peuvent être retirées ou contrôlées pour évaluer la force des effets spontanés.

Du point de vue des théories en faveur d'une quantité d'ADN optimal, cela signifie surtout que les taux de délétions et de duplications soient suffisamment faibles pour que la taille optimale soit dans la zone de stabilité. Une plus grande quantité d'ADN requerrait donc des mécanismes de stabilisation des chromosomes plus importants. Les effets de dérive pourraient également orienter la convergence de la distribution, que ce soit vers une augmentation via les éléments transposables (comme suggéré par Lynch et Conery (2003), en sachant qu'il existe alors une limite supérieure de stabilité) ou une diminution par érosion

de gènes non essentiels et de non codant (comme suggéré par Kuo et Ochman (2009)). On peut noter au passage que, d'après ces deux théories, la dérive influence aussi bien la taille totale que le pourcentage de codant et joue donc de toute façon un rôle dans le cadre des théories de taille d'ADN optimale. Il y a donc une complémentarité possible entre toutes les théories et avec nos résultats.

# Conclusions et perspectives

Dans ce manuscrit, nous avons proposé un modèle mathématique de l'évolution structurelle des génomes en mettant l'accent sur l'influence des différents mécanismes de mutation. Notre étude montre que le simple fait que la distribution de taille des réarrangements intrachromosomiques s'élargisse avec la taille du génome entraîne des effets difficiles à prévoir *a priori*. Pour des distributions de gains par duplications et de pertes par grandes délétions symétriques, une duplication ne compense pas, en moyenne, une grande délétion. En l'absence de sélection, les génomes ont donc tendance à rétrécir quand les taux de duplications et de grandes délétions sont égaux, la croissance vers une taille infinie n'intervient que pour des taux de duplications plus élevés (2.59 duplications pour une délétion pour des pertes et des gains distribués uniformément). Si le nombre de réarrangements spontanés par génération croît assez rapidement avec la taille du génome, l'incorporation de la sélection ne suffit pas pour surmonter ces instabilités chromosomiques. Même en sélectionnant les individus portant le plus de gènes, les génomes se stabilisent dans une zone où les mutations sont rares. Si on augmente le taux de duplications ou la force de la sélection, on force les génomes à quitter la zone stable, ce qui conduit à un éclatement de la population et, paradoxalement, à une baisse de la taille moyenne des génomes.

Biologiquement, la dynamique spontanée des réarrangements imposeraient donc une limite de stabilité au génome, un peu à la manière de l'*error threshold* d'Eigen, à ceci près que le mécanisme s'applique à la totalité du génome, séquences codantes et non codantes. Nos résultats sont complémentaires aux théories classiques de détermination de la taille du génome : les réarrangements délimiteraient simplement une zone dans laquelle ces théories peuvent s'exercer. Dans cette zone stable où les réarrangements sont rares, la sélection, la dérive et les mutations additives (comme les petits indels dus au glissement de l'ADN polymérase ou la transposition réplicative d'un élément transposable) agissent selon les mécanismes prévus par chacune de ces théories.

Le modèle que nous avons développé est volontairement minimaliste : la longueur et le nombre de réarrangements sont définis de manière phénoménologique, et la direction de la sélection – favorisant systématiquement le plus grand nombre de gènes – a été choisie pour tester si l'effet d'*error threshold* tenait dans des conditions certainement plus extrêmes que celles rencontrées par les populations réelles. Cette simplicité nous a permis d'aboutir aux conclusions ci-dessus et il serait intéressant de voir si ces conclusions peuvent s'étendre à des cas légèrement plus réalistes. Deux grands axes peuvent ainsi être exploités. Dans le premier axe, il s'agit d'étendre l'analyse mathématique en explicitant des familles de

distributions spontanées de réarrangements pour lequel notre théorème s'applique, en proposant des mécanismes biologiques qui permettent d'aboutir à ces distributions spontanées. On pourrait également déterminer à quelle vitesse le nombre de réarrangements par génération doit croître avec la taille du génome pour que la sélection ne permette pas de surmonter les instabilités chromosomiques. Le deuxième axe se rapporte aux simulations : il s'agirait de prendre en compte les notions de redondance génétique et de divergence pour éviter que les duplications soient immédiatement et systématiquement favorables.

Dans ce manuscrit, nous nous sommes attachés à identifier de nombreux paradoxes, biologiques ou mathématiques, qui se rapportent à l'« expérience de pensée ». Dans ce type d'expériences, les résultats s'obtiennent souvent sous un éclairage précis – un *a priori* – qui n'est pas toujours explicité ou vérifié strictement. Quand on l'applique à un processus complexe comme l'évolution darwinienne, cela revient souvent à entretenir une certaine ambiguïté sur les différents mécanismes de mutation et de sélection en jeu, qui peuvent tous avoir un rôle important à jouer. En développant un modèle minimaliste, nous nous plaçons volontairement dans un cadre que ne prétend pas reproduire ce qui passe dans la réalité. Cependant, ce type de modèle permet d'étudier l'effet de chaque paramètre indépendamment puis en interaction, et d'arriver à des conclusions parfois contre-intuitives. Ces conclusions pourront être interprétées au sein du processus réel comme une nouvelle donnée, une nouvelle pression qu'il aurait été difficile de prédire uniquement par l'expérience de pensée. Ceci étant dit, l'expérience de pensée reste une étape importante dans le développement de la connaissance. Notre approche plaide pour un va-et-vient entre l'expérience de pensée et la modélisation formelle : l'expérience de pensée engendre des idées qui peuvent être soumises à modélisation qui, par ses conclusions, confirme ou infirme mais, dans tous les cas, enrichit l'expérience de pensée originale.

> Les déterminations numériques de la science repassent sur le pointillé d'une constitution du monde déjà faite avant elles.
>
> Maurice Merleau-Ponty, *Phénoménologie de la perception*, 1945.

# Bibliographie


Adami, C. (2002). What is complexity ? *Bioessays*, 24(12):1085–1094.

Aguilera, A. et Gómez-González, B. (2008). Genome instability : A mechanistic view of its causes and consequences. *Nat Rev Genet*, 9(3):204–217.

Alexander, R. P., Fang, G., Rozowsky, J., Snyder, M. et Gerstein, M. B. (2010). Annotating non-coding regions of the genome. *Nat Rev Genet*, 11(8):559–571.

Ancel Meyers, L., Ancel, F. D. et Lachmann, M. (2005). Evolution of genetic potential. *PLoS Comput Biol*, 1(3):e32.

Andersson, D. I. (1998). Evidence that gene amplification underlies adaptive mutability of the bacterial lac operon. *Science*, 282(5391):1133–1135.

Barbosa, V. C., Donangelo, R. et Souza, S. R. (2012). Quasispecies dynamics with network constraints. *J Theor Biol*, 312:114–119.

Basten, C. J. et Moody, M. E. (1991). A branching-process model for the evolution of transposable elements incorporating selection. *J Math Biol*, 29(8):743–761.

Batut, B., Parsons, D., Fischer, S., Beslon, G. et Knibbe, C. (2013). *In silico* experimental evolution : A tool to test evolutionary. *BMC Bioinfo.*

Beaton, M. J. et Cavalier-Smith, T. (1999). Eukaryotic non-coding DNA is functional : Evidence from the differential scaling of cryptomonad genomes. *P Roy Soc Lond B Bio*, 266(1433):2053–2059.

Bennett, M. D. (1972). Nuclear DNA content and minimum generation time in herbaceous plants. *P Roy Soc Lond B Bio*, 181(1063):109–135. PMID : 4403285.

Bennetzen, J. L. et Kellogg, E. A. (1997). Do plants have a one-way ticket to genomic obesity ? *Plant Cell*, 9(9):1509–1514. PMID : 12237393.

Bensasson, D., Petrov, D. A., Zhang, D.-X., Hartl, D. L. et Hewitt, G. M. (2001). Genomic gigantism : DNA loss is slow in mountain grasshoppers. *Mol Biol Evol*, 18(2): 246–253. PMID : 11158383.

Bhargava, A. et Fuentes, F. F. (2009). Mutational dynamics of microsatellites. *Mol Biotechnol*, 44(3):250–266.

Bird, A. P. (1995). Gene number, noise reduction and biological complexity. *Trends Genet*, 11(3):94–100.

Brunet, E. et Derrida, B. (1997). Shift in the velocity of a front due to a cutoff. *Phys Rev E*, 56(3):2597–2604.

Cavalier-Smith, T. (1978). Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA c value paradox. *J Cell Sci*, 34(1):247–278.

Cavalier-Smith, T. (1985). *The Evolution of genome size*. John Wiley & Sons Ltd, Chichester, UK.

Cavalier-Smith, T. et Beaton, M. J. (1999). The skeletal function of non-genic nuclear DNA : new evidence from ancient cell chimaeras. *Genetica*, 106(1-2):3–13.

Champagnat, N., Ferrière, R. et Méléard, S. (2006). Unifying evolutionary dynamics : From individual stochastic processes to macroscopic models. *Theor Popul Biol*, 69(3):297–321.

Champagnat, N. et Lambert, A. (2007). Evolution of discrete populations and the canonical diffusion of adaptive dynamics. *Ann Appl Probab*, 17(1):102–155.

Charlesworth, B. (2009). Fundamental concepts in genetics : Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet*, 10(3):195–205.

Charlesworth, B., Sniegowski, P. et Stephan, W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature*, 371(6494):215–220.

Chen, J.-M. (2011). Genomic rearrangements : Mutational mechanisms. *In Encyclopedia of Life Sciences*. John Wiley & Sons, Ltd, Chichester, UK.

Chia, J.-M., Song, C., Bradbury, P. J., Costich, D., de Leon, N., Doebley, J., Elshire, R. J., Gaut, B., Geller, L., Glaubitz, J. C., Gore, M., Guill, K. E., Holland, J., Hufford, M. B., Lai, J., Li, M., Liu, X., Lu, Y., McCombie, R., Nelson, R., Poland, J., Prasanna, B. M., Pyhäjärvi, T., Rong, T., Sekhon, R. S., Sun, Q., Tenaillon, M. I., Tian, F., Wang, J., Xu, X., Zhang, Z., Kaeppler, S. M., Ross-Ibarra, J., McMullen, M. D., Buckler, E. S., Zhang, G., Xu, Y. et Ware, D. (2012). Maize HapMap2 identifies extant variation from a genome in flux. *Nat Genet*, 44(7):803–807.

Cooper, V. S., Schneider, D., Blot, M. et Lenski, R. E. (2001). Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli* b. *J Bacteriol*, 183(9):2834–2841.

Crombach, A. et Hogeweg, P. (2008). Evolution of evolvability in gene regulatory networks. *PLoS Comput Biol*, 4(7):e1000112.

Darling, A. E., Miklós, I. et Ragan, M. A. (2008). Dynamics of genome rearrangement in bacterial populations. *PLoS Genet*, 4(7):e1000128.

Devos, K. M., Brown, J. K. M. et Bennetzen, J. L. (2002). Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res*, 12(7):1075–1079.

Dieckmann, U. et Law, R. (1996). The dynamical theory of coevolution : A derivation from stochastic ecological processes. *J Math Biol*, 34(5-6):579–612.

Diekmann, O., Jabin, P.-E., Mischler, S. et Perthame, B. (2005). The dynamics of adaptation : An illuminating example and a Hamilton–Jacobi approach. *Theor Popul Biol*, 67(4):257–271.

Doolittle, W. F. et Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284(5757):601–603.

Draghi, J. A., Parsons, T. L., Wagner, G. P. et Plotkin, J. B. (2010). Mutational robustness can facilitate adaptation. *Nature*, 463(7279):353–355.

Drake, J. W. (1991). A constant rate of spontaneous mutation in DNA-based microbes. *P Natl Acad Sci USA*, 88(16):7160–7164.

Eigen, M. (1971). Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58(10):465–523.

Eigen, M., McCaskill, J. et Schuster, P. (1988). Molecular quasi-species. *J Phys Chem-US*, 92(24):6881–6891.

Eigen, M. et Schuster, P. (1979). *The hypercycle, a principle of natural self-organization*. Springer-Verlag, Berlin Heidelberg.

Falush, D. et Iwasa, Y. (1999). Size-dependent mutability and microsatellite constraints. *Mol Biol Evol*, 16(7):960.

Fleming, W. et Viot, M. (1979). Some measure-valued markov processes in population-genetics theory. *Indiana U Math J*, 28(5):817–843.

Gaffé, J., McKenzie, C., Maharjan, R. P., Coursange, E., Ferenci, T. et Schneider, D. (2011). Insertion sequence-driven evolution of *Escherichia coli* in chemostats. *J Mol Evol*, 72(4):398–412.

Geritz, S. a. H., Kisdi, E., Meszéna, G. et Metz, J. a. J. (1998). Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree. *Evol Ecol*, 12(1):35–57.

Gregory, T. (2003). Is small indel bias a determinant of genome size ? *Trends Genet*, 19(9):485–488.

Gregory, T. (2004). Insertion–deletion biases and the evolution of genome size. *Gene*, 324:15–34.

Gregory, T. R. (2001). Coincidence, coevolution, or causation ? DNA content, cellsize, and the c value enigma. *Biol Rev*, 76(1):65–101.

Gu, W., Zhang, F. et Lupski, J. R. (2008). Mechanisms for human genomic rearrangements. *PathoGenetics*, 1(1):4.

Hahn, M. W. et Wray, G. A. (2002). The g value paradox. *Evol Dev*, 4(2):73–75.

Jaillon, O., Aury, J.-M. et Wincker, P. (2009). "Changing by doubling", the impact of whole genome duplications in the evolution of eukaryotes. *C R Biol*, 332(2-3):241–253.

Kimura, M. (1964). Diffusion models in population genetics. *J Appl Probab*, 1(2):177–232.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626.

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge [Cambridgeshire] ; New York.

Kirik, A., Salomon, S. et Puchta, H. (2000). Species-specific double-strand break repair and genome evolution in plants. *EMBO J*, 19(20):5562–5566.

Knibbe, C. (2006). *Structuration des génomes par sélection indirecte de la variabilité mutationnelle : une approche de modélisation et de simulation*. Thèse de doctorat, INSA de Lyon.

Knibbe, C., Coulon, A., Mazet, O., Fayard, J.-M. et Beslon, G. (2007). A long-term evolutionary pressure on the amount of noncoding DNA. *Mol Biol Evol*, 24(10):2344–2353.

Krüger, J. et Vogel, F. (1975). Population genetics of unequal crossing over. *J Mol Evol*, 4(3):201–247.

Kugelberg, E., Kofoid, E., Reams, A. B., Andersson, D. I. et Roth, J. R. (2006). Multiple pathways of selected gene amplification during adaptive mutation. *P Natl Acad Sci USA*, 103(46):17319–17324.

Kuo, C.-H., Moran, N. A. et Ochman, H. (2009). The consequences of genetic drift for bacterial genome complexity. *Genome Res*, 19(8):1450–1454.

Kuo, C.-H. et Ochman, H. (2009). Deletional bias across the three domains of life. *Genome Biol Evol*, 1:145–152.

Lambert, A. (2005). The branching process with logistic growth. *Ann Appl Probab*, 15(2):1506–1535.

Lambert, A. (2006). Probability of fixation under weak selection : A branching process unifying approach. *Theor Popul Biol*, 69(4):419–441.

Langley, C. H., Montgomery, E., Hudson, R., Kaplan, N. et Charlesworth, B. (1988). On the role of unequal exchange in the containment of transposable element copy number. *Genet Res*, 52(3):223–235.

Leibler, S. et Kussell, E. (2010). Individual histories and selection in heterogeneous populations. *P Natl Acad Sci USA*, 107(29):13183–13188.

Leushkin, E. V., Bazykin, G. A. et Kondrashov, A. S. (2013). Strong mutational bias toward deletions in the *Drosophila melanogaster* genome is compensated by selection. *Genome Biol Evol*, 5(3):514–524.

Lisch, D. (2009). Epigenetic regulation of transposable elements in plants. *Annu Rev Plant Biol*, 60(1):43–66.

Lupski, J. R. (2007). Genomic rearrangements and sporadic disease. *Nat Genet*, 39(7s): S43–S47.

Lynch, M. (2006). Streamlining and simplification of microbial genome architecture. *Ann Rev Microbiol*, 60(1):327–349.

Lynch, M. (2010). Evolution of the mutation rate. *Trends Genet*, 26(8):345–352.

Lynch, M. et Conery, J. S. (2003). The origins of genome complexity. *Science*, 302(5649):1401–1404.

Mira, A., Ochman, H. et Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends Genet*, 17(10):589–596.

Moody, M. E. (1988). A branching process model for the evolution of transposable elements. *J Math Biol*, 26(3):347–357.

Nakatsu, C. H., Korona, R., Lenski, R. E., de Bruijn, F. J., Marsh, T. L. et Forney, L. J. (1998). Parallel and divergent genotypic evolution in experimental populations of *Ralstonia sp.* *J Bacteriol*, 180(17):4325–4331.

Nilsson, A. I., Koskiniemi, S., Eriksson, S., Kugelberg, E., Hinton, J. C. D. et Andersson, D. I. (2005). Bacterial genome size reduction by experimental evolution. *P Natl Acad Sci USA*, 102(34):12112–12116.

Nowak, M. et Schuster, P. (1989). Error thresholds of replication in finite populations mutation frequencies and the onset of muller's ratchet. *J Theor Biol*, 137(4):375–395.

Nowak, M. A. (1992). What is a quasispecies ? *Trends Ecol Evol*, 7(4):118–121.

Ochman, H. et Davalos, L. M. (2006). The nature and dynamics of bacterial genomes. *Science*, 311(5768):1730–1733.

Ochman, H. et Moran, N. A. (2001). Genes lost and genes found : Evolution of bacterial pathogenesis and symbiosis. *Science*, 292(5519):1096–1099.

Ohno, S. (1970). *Evolution by gene duplication.* Springer-Verlag, London ; New York.

Ohno, S. (1972). So much "junk" DNA in our genome. *Brookhaven Sym Biol*, 23. MEDLINE :5065367.

Ohta, T. et Kimura, M. (1981). Some calculations on the amount of selfish DNA. *P Natl Acad Sci USA*, 78(2):1129–1132.

Oliver, M. J., Petrov, D., Ackerly, D., Falkowski, P. et Schofield, O. M. (2007). The mode and tempo of genome size evolution in eukaryotes. *Genome Res*, 17(5):594–601.

Orel, N. et Puchta, H. (2003). Differences in the processing of DNA ends in *Arabidopsis thaliana* and tobacco : possible implications for genome evolution. *Plant Mol Biol*, 51(4):523–531.

Orgel, L., Crick, F. et Sapienza, C. (1980). Selfish DNA. *Nature*, 288(5792):645–646.

Orgel, L. E. et Crick, F. H. C. (1980). Selfish DNA : the ultimate parasite. *Nature*, 284(5757):604–607.

Petrov, D. A. (2000). Evidence for DNA loss as a determinant of genome size. *Science*, 287(5455):1060–1062.

Petrov, D. A. (2001). Evolution of genome size : new approaches to an old problem. *Trends Genet*, 17(1):23–28.

Petrov, D. A. (2002). Mutational equilibrium model of genome size evolution. *Theor Popul Biol*, 61(4):531–544.

Petrov, D. A. et Hartl, D. L. (1998). High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. *Mol Biol Evol*, 15(3):293–302.

Petrov, D. A., Lozovskaya, E. R. et Hartl, D. L. (1996). High intrinsic rate of DNA loss in *Drosophila*. *Nature*, 384(6607):346–349.

Poole, A. M., Phillips, M. J. et Penny, D. (2003). Prokaryote and eukaryote evolvability. *Biosystems*, 69(2–3):163–185.

Porwollik, S., Wong, R. M.-Y., Helm, R. A., Edwards, K. K., Calcutt, M., Eisenstark, A. et McClelland, M. (2004). DNA amplification and rearrangements in archival *Salmonella enterica* serovar typhimurium LT2 cultures. *J Bacteriol*, 186(6):1678–1682.

Pál, C. et Hurst, L. D. (2000). The evolution of gene number : are heritable and non-heritable errors equally important ? *Heredity*, 84(4):393–400.

Rabinowicz, P. D. (2000). Are obese plant genomes on a diet ? *Genome Res*, 10(7):893–894.

Rees, H. et Durrant, A. (1986). Recombination and genome size. *Theor Appl Genet*, 73(1):72–76.

Ross-Ibarra, J. (2007). Genome size and recombination in angiosperms : a second look. *J Evol Biol*, 20(2):800–806.

Rouzic, A. L. et Deceliere, G. (2005). Models of the population genetics of transposable elements. *Genet Res*, 85(03):171.

Sankoff, D., Lefebvre, J.-F., Tillier, E., Maler, A. et El-Mabrouk, N. (2005). The distribution of inversion lengths in bacteria. *In* Lagergren, J., éditeur : *Comparative Genomics*, numéro 3388 de Lecture Notes in Computer Science, pages 97–108. Springer, Berlin Heidelberg.

Schad, E., Tompa, P. et Hegyi, H. (2011). The relationship between proteome size, structural disorder and organism complexity. *Genome Biol*, 12(12):1–13.

Schneider, D., Duperchy, E., Coursange, E., Lenski, R. E. et Blot, M. (2000). Long-term experimental evolution in *Escherichia coli*. IX. characterization of insertion sequence-mediated mutations and rearrangements. *Genetics*, 156(2):477–488.

Sloan, D. B., Alverson, A. J., Chuckalovcak, J. P., Wu, M., McCauley, D. E., Palmer, J. D. et Taylor, D. R. (2012). Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS Biol*, 10(1):e1001241.

Smith, G. P. (1976). Evolution of repeated DNA sequences by unequal crossover. *Science*, 191(4227):528–535.

Stephan, W. (1987). Quantitative variation and chromosomal location of satellite DNAs. *Genet Res*, 50(1):41–52.

Stroock, D. W. (2005). *An Introduction to Markov Processes*. Springer.

Tenaillon, O., Silander, O. K., Uzan, J.-P. et Chao, L. (2007). Quantifying organismal complexity using a population genetic approach. *PloS ONE*, 2(2):e217.

Thomas, C. A. (1971). The genetic organization of chromosomes. *Annu Rev Genet*, 5(1):237–256.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., Calteau, A., Chiapello, H., Clermont, O., Cruveiller, S., Danchin, A., Diard, M., Dossat, C., Karoui, M. E., Frapy, E., Garry, L., Ghigo, J. M., Gilles, A. M., Johnson, J., Le Bouguénec, C., Lescat, M., Mangenot, S., Martinez-Jéhanne, V., Matic, I., Nassif, X., Oztas, S., Petit, M. A., Pichon, C., Rouy, Z., Ruf, C. S., Schneider, D., Tourret, J., Vacherie, B., Vallenet, D., Médigue, C., Rocha, E. P. C. et Denamur, E. (2009). Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet*, 5(1):e1000344.

Touchon, M. et Rocha, E. P. C. (2007). Causes of insertion sequences abundance in prokaryotic genomes. *Mol Biol Evol*, 24(4):969–981.

Touzain, F., Petit, M.-A., Schbath, S. et Karoui, M. E. (2011). DNA motifs that sculpt the bacterial chromosome. *Nat Rev Micro*, 9(1):15–26.

Tsimring, L. S., Levine, H. et Kessler, D. A. (1996). RNA virus evolution via a fitness-space model. *Phys Rev L*, 76(23):4440–4443.

VAN NIMWEGEN, E., CRUTCHFIELD, J. P. et HUYNEN, M. (1999). Neutral evolution of mutational robustness. *P Natl Acad Sci USA*, 96(17):9716–9720.

VICIENT, C. M., SUONIEMI, A., ANAMTHAWAT-JÓNSSON, K., TANSKANEN, J., BEHARAV, A., NEVO, E. et SCHULMAN, A. H. (1999). Retrotransposon BARE-1 and its role in genome evolution in the genus hordeum. *Plant Cell*, 11(9):1769–1784.

WAGNER, A. (2008). Robustness and evolvability : a paradox resolved. *P Roy Soc Lond B Bio*, 275(1630):91–100.

WALSH, J. B. (1987). Persistence of tandem arrays : Implications for satellite and simple-sequence DNAs. *Genetics*, 115(3):553–567.

WENDEL, J. F., CRONN, R. C., JOHNSTON, J. S. et PRICE, H. J. (2002). Feast and famine in plant genomes. *Genetica*, 115(1):37–47.

WHITNEY, K. D. et GARLAND, T. (2010). Did genetic drift drive increases in genome complexity ? *PLoS Genet*, 6(8):e1001080.

WILKE, C. O. (2003). Probability of fixation of an advantageous mutant in a viral quasispecies. *Genetics*, 163(2):467–474.

WOESS, W. (2009). *Denumerable Markov Chains : Generating Functions, Boundary Theory, Random Walks on Trees*. European Mathematical Society, Zurich.

ZELDOVICH, K. B., CHEN, P. et SHAKHNOVICH, E. I. (2007). Protein stability imposes limits on organism complexity and speed of molecular evolution. *P Natl Acad Sci USA*, 104(41):16152–16157.

# Annexe A

# Informations complémentaires pour l'implémentation des simulations des structures de génome

## 1 Informations supplémentaires concernant le calcul des transitions atomiques : duplications et grandes délétions

Cette section donne la démonstration des profils correspondant aux cas particuliers des calculs des transitions de grandes délétions et duplications. Le détail de la procédure est donné dans le texte principal (chapitre III, section 2.3, page 108). Les démonstrations sont toutes dérivées de celle pour le cas général, que nous avons reproduite ici. Les démonstrations des cas particuliers ne sont pas aussi détaillées : on ne réexplique pas comment compter les combinaisons une fois que la distance entre les deux points de cassure est fixée, on s'attache surtout à mettre en avant les similitudes et les différences avec le cas général.

**Cas général : le nombre de gènes perdus $k$ est compris entre 2 et $n_0 - 1$** Si $2 \leq k \leq n_0 - 1$, la situation est simple car les points de cassure BP1 et BP2 ne peuvent pas se chevaucher. Le plus simple est ici de faire un dessin : BP1 et BP2 tombent nécessairement dans des intergéniques ou des gènes qui sont séparés par $k-2$ gènes et $k-1$ intergéniques sur le segment délété. Dans le segment conservé, il reste au moins un gène, ce qui garantit que BP1 et BP2 ne se croisent pas dans l'autre sens non plus. On peut donc facilement délimiter les positions accessibles pour BP1 d'un côté et pour BP2 de l'autre (figure A.1A). Elles sont symétriques : en plus de $k-2$ gènes et $k-1$ intergéniques de toute façon délétés, il faut déléter au moins une base des gènes de chaque côté. Il y a donc $l_{intergenic} + l_{gene} - 1$ positions accessibles de chaque côté, en comptant toutes les positions de délétions dans le

gène à part la première base, puis celles dans l'intergénique qui suit.

Nous connaissons déjà le nombre de gènes perdus, nous nous intéressons ici au non codant perdu à cause de la délétion. Il y a trois configurations importantes. Au maximum, on supprime $(k+1)l_{intergenic}$ bases non codantes, si BP1 et BP2 sont aux bouts opposés des intergéniques, soit pour une seule combinaison (figure A.1A). Au minimum, si BP1 et BP2 sont au plus proche (sur la dernière base de chaque gène), on « supprime » $(k-1)l_{intergenic} - 2(l_{gene}-1)$ bases (en comptant en plus les bases devenues non codantes, figure A.1C). Entre les deux, si BP1 et BP2 sont le plus à gauche possible (l'un sur l'intergénique, l'autre sur le gène, figure A.1B), on « supprime » $kl_{intergenic} - (l_{gene}-1)$ bases non codantes. Si on décale les deux points de cassure conjointement, on obtient le même résultat pour $l_{intergenic} + l_{gene}$ combinaisons. Pour le reste c'est assez simple, dès qu'on varie la taille de la délétion, on gagne ou on perd exactement une combinaison. Si on part de la plus longue délétion, on gagne une combinaison en diminuant la taille de la délétion jusqu'à arriver au cas de la figure III.3B. Ensuite, on perd une combinaison avec chaque diminution jusqu'à aboutir à la délétion la plus petite possible. On obtient donc un simple profil en triangle entre les trois positions caractéristiques illustré sur la figure III.3.

**Perte d'un seul gène avec $n_0 \geq 2$** Ce cas est similaire au cas précédent sauf que les points de cassure peuvent se croiser à l'intérieur du gène. Pour être exact, le profil est le même, mais il faut retirer toutes les combinaisons où BP2 précède BP1. Comparons en nous concentrant sur BP1 : on place BP1 tout à gauche et BP2 tout à droite pour avoir la plus grande délétion possible (comme sur la figure A.1A, $2l_{intergenic}$ bases non codantes perdues). On garde BP1 à gauche et on décale BP2 vers la gauche : on diminue la taille de la délétion en gagnant une combinaison à chaque décalage, comme dans le cas général, jusqu'à ce que BP2 bute sur la fin du gène délété : on atteint le sommet du triangle comme dans le cas général (A.1B, $l_{intergenic} - (l_{gene}-1)$ bases non codantes perdues). Maintenant on laisse BP2 tout à gauche du gène et on décale BP1 vers la droite : on continue à diminuer la taille de la délétion en perdant une combinaison à chaque fois, ce qui correspond toujours au cas général. Quand BP1 arrive au même niveau que BP2, la délétion atteint sa taille minimale de 1 base, qui correspond à un gain de $l_{gene}-1$ bases non codantes par pseudogénisation, ce qui est possible pour $l_{gene}$ combinaisons. On a alors épuisé toutes les combinaisons : on a parcouru le triangle du cas général, mais la fin du triangle est tronquée (figure A.2). C'est relativement logique : la fin du triangle correspondrait à la pseudogénisation de deux gènes différents, ce qui n'est pas possible ici puisqu'on s'est placé dans le cas où un seul gène est perdu.

**Perte de tous les gènes avec $n_0 \geq 2$** Si tous les gènes sont perdus, cela implique que BP2 se trouve avant BP1 (ce qui est possible étant donné la circularité du génome). On va donc utiliser le raisonnement ci-dessus à l'envers, en partant de la plus petite délétion possible et en comparant avec le cas général. Plaçons BP1 tout à droite de son gène et BP2 tout à gauche du sien (perte de $(n_0-1)l_{intergenic} - 2(l_{gene}-1)$ bases non codantes, une combinaison). On décale BP1 vers la gauche, on gagne une taille de

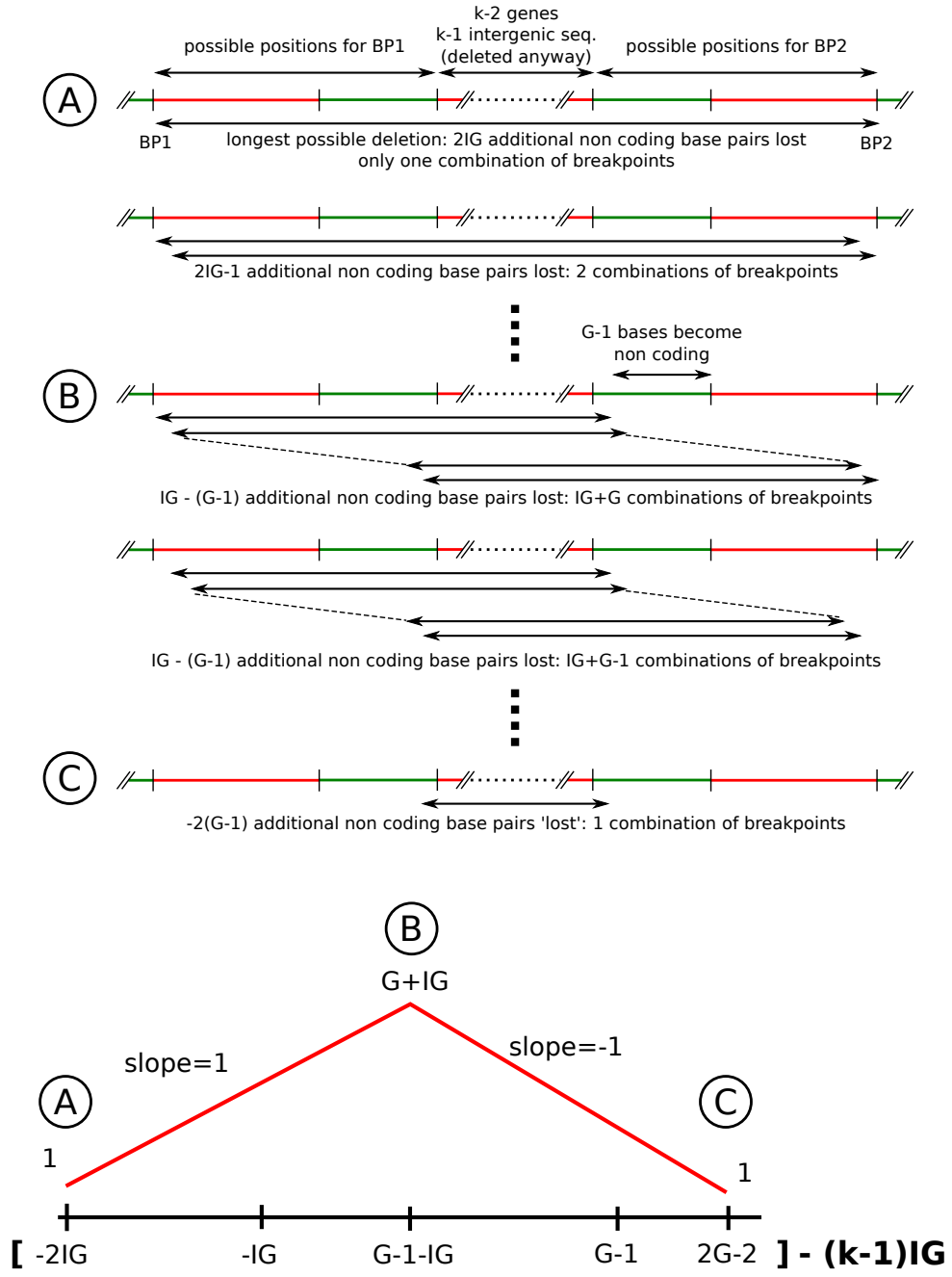


Figure A.1 – Distribution générale des transitions liées aux duplications et aux délétions, ici dans le cas de la perte de $k$ gènes. En variant la taille de la délétion de la plus longue possible (cas A) jusqu'à la plus courte (cas C), on se rend compte que le nombre de combinaisons de points de cassure qui permettent cette taille de délétion augmente dans un premier temps jusqu'au cas B, puis diminue. Pour raccourcir les notations sur le dessin, la longueur de l'intergénique est dénotée par IG ($l_{intergenic}$ dans le texte) et la longueur d'un gène G ($l_{gene}$ dans le texte).

délétion et une combinaison jusqu'à ce que BP1 bute sur la fin de l'intergénique (perte de $n_0 l_{intergenic} - (l_{gene} - 1)$ bases non codantes, $l_{intergenic} + l_{gene}$ combinaisons). À ce moment, on

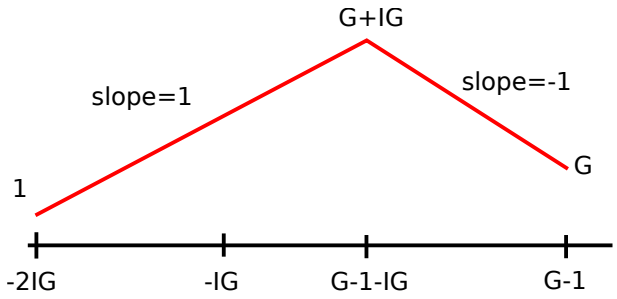


Figure A.2 – Nombre de combinaisons de points de cassure qui permettent le gain ou la perte de $i$ bases non-codantes quand exactement un gène est perdu.

peut continuer à bouger BP1 vers la gauche jusqu'à le mettre à côté de BP2, ce qui n'était pas possible dans le cas général pour ne pas augmenter le nombre de gènes perdus. La symétrie devient un peu complexe à prendre en compte ici et le nombre de combinaisons valides difficile à calculer rigoureusement... Prenons le cas où BP1 et BP2 partent du gène de BP2, aux extrêmes opposés. En les décalant conjointement jusqu'à ce que BP1 bute sur la fin de son gène, on obtient $l_{intergenic} + l_{gene} + 1$ combinaisons. Néanmoins, dans la combinaison finale, BP1 et BP2 occupent les extrêmes opposés du gène de BP1, qui est une des positions symétriques d'ordre $n_0$ de la position initiale : on n'a pas le droit de la compter. Il y a donc en réalité $l_{intergenic} + l_{gene}$ combinaisons. En fait, si on continue à décaler BP1 vers BP2, le nombre de combinaisons reste $l_{intergenic} + l_{gene}$ à cause de la symétrie d'ordre $n_0$ (c'est d'ailleurs le maximum possible compte tenu de la symétrie). On s'arrête quand BP1 est juste à côté de BP2, puisqu'on supprime alors tout le génome. Pour résumer, la partie de droite du profil est la même que dans le cas général, mais ensuite, on a un plateau jusqu'à la suppression totale du génome (figure A.3).

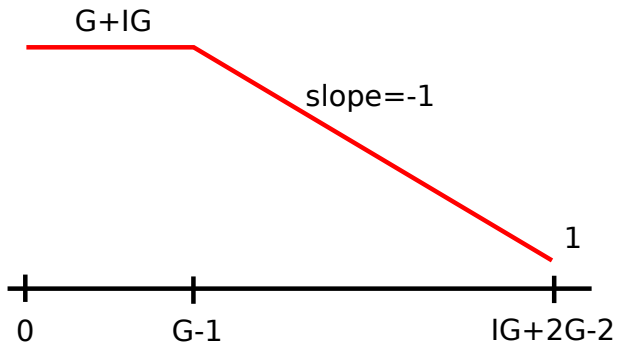


Figure A.3 – Nombre de combinaisons de points de cassure qui permettent de conserver $i$ bases non-codantes quand tous les gènes sont perdus.

**Perte de tous les gènes (sic) avec $n_0 = 1$** Là, on cumule les tares des deux cas précédents : les points de cassure peuvent se chevaucher à la fois dans le gène à déléter et en faisant le tour du génome. Comparé au cas général, on a la troncature à droite (impossible de convertir 2 gènes) et le plateau qui remplace la partie ascendante à gauche. Nous omettons les détails pour ce cas.

**Pas de perte de gène avec $n_0 \geq 1$** Si aucun gène n'est perdu, les deux points de cassure sont dans le même intergénique. Partons de la plus grosse délétion : BP1 est à gauche et BP2 à droite, soit une combinaison pour une perte de $l_{intergenic}$ bases non codantes. La présence du gène assure que les notions « à gauche » et « à droite » ont un sens. On rapproche BP2 de BP1 : on perd une taille de délétion mais on gagne une combinaison. Quand BP2 coïncide avec BP1, on arrive à une délétion de 1 base et $l_{gene}$ combinaisons. On obtient donc un début de triangle, soit la partie tout à gauche du cas général (figure A.4).

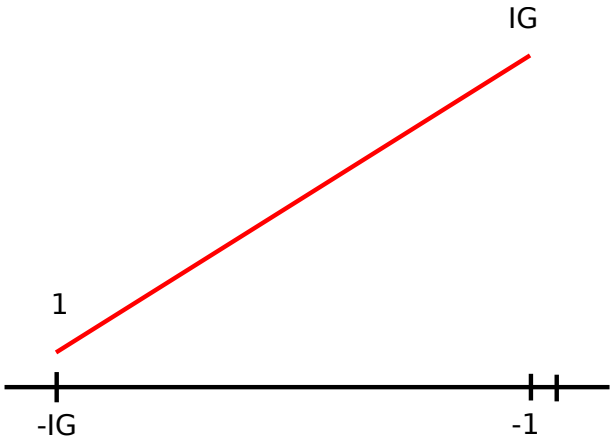


Figure A.4 – Nombre de combinaisons de points de cassure qui permettent la perte de $i$ bases non-codantes quand aucun gène n'est perdu.

**Pas de perte de gène avec $n_0 = 0$** Dans ce cas, il n'y a plus de symétrie d'ordre $n_0$. On peut raisonner de beaucoup de manières différentes, mais le résultat est très intuitif. Quelle que soit la taille de la délétion, on peut la placer n'importe où le long du génome, soit $L_0$ combinaisons. On obtient donc une distribution uniforme avec 0 à $L_0 - 1$ bases non codantes restantes.

**Bilan - généralisation** En général, on obtient un profil en triangle, qu'on décale d'un intergénique pour chaque perte de gène supplémentaire. Il faut faire cependant attention aux cas extrêmes (perte de 0 ou 1 gène et perte de tous les gènes). Le cas des duplications est très similaire : on peut raisonner par complémentarité. À chaque paire de points de cassure correspond une perte de $k$ gènes : le segment complémentaire contient donc $n_0 - k$ gènes *intacts*. À chaque délétion de $k$ gènes on peut donc associer, par complémentarité, une duplication avec un segment qui contient $n_0 - k$ gènes. Pour avoir l'équivalent du profil des duplications sans gain de gènes, on copie le profil de perte de tous les gènes, et ainsi de suite. On obtient le profil en figure A.5. Cependant, cette distribution est incomplète puisqu'on a ignoré la possibilité de perdre un gène à l'insertion. Pour prendre en compte ce phénomène, il suffit de calculer la probabilité d'insérer la séquence dans du non-codant qui vaut $p = L_0/(L_0 + n_0 l_{gene})$. On obtient alors le profil complet en séparant les deux cas : on pondère les profils en triangle de la figure pour le cas où la séquence est insérée

dans du non-codant, on les pondère et on les décale d'un gène vers le bas et vers la droite pour le cas où la séquence casse un gène à l'insertion.

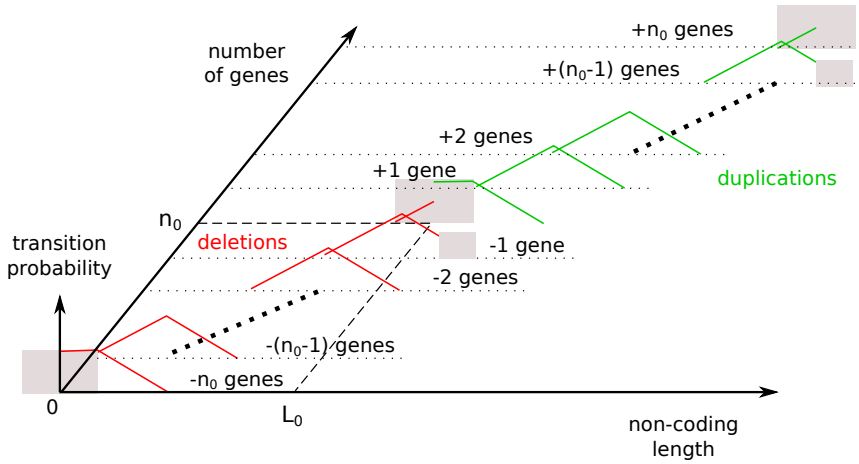


Figure A.5 – Distribution de la taille du génome après une duplication (en ignorant la perte de gène à l'insertion) ou une délétion. On obtient le même profil en triangle sur toutes les lignes, à quelques exceptions près qui sont marquées d'un contour gris.

Les profils sont donnés en nombre de combinaisons de points de cassure mais, en renormalisant, on obtient les mêmes profils compris entre 0 et 1 qui correspondent aux probabilités cherchées initialement. Le nombre de combinaisons pour un triangle complet vaut $(l_{gene} + l_{intergenic})^2$ et globalement, les parties tronquées se compensent, si bien que le nombre de combinaisons total vaut exactement $n_0(l_{gene} + l_{intergenic})^2$ pour chaque type de mutation. En divisant, la somme des probabilités pour un triangle vaut donc exactement $1/n_0$ et sa hauteur $n_0/(l_{gene} + l_{intergenic})$.

# 2 Calcul des valeurs moyennes et approximation des sommes

Cette section donnent les valeurs moyennes des mesures usuelles en échelles linéaire et logarithmique, ainsi qu'une façon classique d'accélérer le calcul de bon nombre de sommes.

## 2.1 Propriétés des subdivisions en échelle linéaire

La largeur de chaque subdivision est constante. On a $\forall x \in \{1, ..., x_{max}\}$,

- $\Delta L(x) = \Delta L$.
- $L_{min}(x) = (x-1)\Delta L$.

- $L_{max}(x) = x\Delta L - 1$.
- $\overline{L(x)} = x\Delta L - (\Delta L + 1)/2$.

De même, la hauteur de chaque subdivision est identique et $\forall y \in \{1, y_{max}\}$,

- $\Delta n(y) = \Delta n$.
- $n_{min}(y) = (y-1)\Delta n$.
- $n_{max}(y) = y\Delta n - 1$.
- $\overline{n(y)} = y\Delta n - (\Delta n + 1)/2$.

Ces équations peuvent être inversées : on cherche dans quelle subdivision se trouve le génome $(L, n)$. Schématiquement, cela revient à résoudre $L = (x-1)\Delta L$ et $n = (y-1)\Delta n$. Comme en réalité on a des valeurs entières, la solution exacte est

- $x = \lfloor \frac{L}{\Delta L} + 1 \rfloor$
- $y = \lfloor \frac{n}{\Delta n} + 1 \rfloor$

## 2.2 Propriétés des subdivisions en échelle logarithmique

Dans le cas de l'échelle logarithmique, les subdivisions sont données par des puissances de 2, sauf celles près de l'origine, dont la hauteur et la largeur sont fixées arbitrairement à un gène. Le long de l'axe des $x$, on obtient alors

- Pour $x = 1$ :
  - $L_{min}(x) = 0$
  - $L_{max}(x) = l_{gene} - 1$
  - $\Delta L(x) = l_{gene}$
  - $\overline{L(x)} = (l_{gene} - 1)/2$
- $\forall x \in \{2, x_{max}\}$ :
  - $L_{min}(x) = 2^{x-2} l_{gene}$
  - $L_{max}(x) = 2^{x-1} l_{gene} - 1$
  - $\Delta L(x) = 2^{x-2} l_{gene}$
  - $\overline{L(x)} = (3 \times 2^{x-2} l_{gene} - 1)/2$

Le long de l'axe des $y$, on a

- Pour $y = 1$ :
  - $n_{min}(y) = 0$
  - $n_{max}(y) = 0$
  - $\Delta n(y) = 1$
  - $\overline{n(y)} = 0$
- $\forall y \in \{2, y_{max}\}$ :
  - $n_{min}(y) = 2^{y-2}$
  - $n_{max}(x) = 2^{y-1} - 1$
  - $\Delta n(y) = 2^{y-2}$
  - $\overline{n(y)} = (3 \times 2^{y-2} - 1)/2$

On peut inverser ces relations : on cherche à quelle subdivision appartient le génome $(L, n)$. Schématiquement, on résout les équations $L = 2^{x-2} l_{gene}$ et $n = 2^{y-2}$. En prenant en compte la particularité de la subdivision à l'origine et le fait que la réponse doit être entière, on obtient

- Si $L < l_{gene}$, $x = 1$, sinon $x = \lfloor \log_2 (L/l_{gene}) + 2 \rfloor$
- Si $n < 1$, $y = 1$, sinon $y = \lfloor \log_2 (n) + 2 \rfloor$

## 2.3 Approximation des sommes

Dans les calculs présentés dans cette section, nous avons beaucoup de sommes à calculer qui sont du type $\sum_{n_{min} \leq k \leq n_{max}} \frac{1}{k}$ ou $\sum_{n_{min} \leq k \leq n_{max}} \frac{1}{k^2}$ (voir par exemple les équations apparaissant dans les transitions correspondant aux inactivations de gènes, section 3.2, page 204). Comme dans un certain nombre de cas, ces sommes peuvent comprendre un très grand nombre de termes, additionner les termes un à un peut être inutilement long. Pour accélérer le processus, nous utilisons les comparaisons sommes-intégrales pour obtenir une expression approximative des sommes. De plus, nous utilisons la formule d'Euler-Maclaurin, qui permet de calculer l'erreur qui est commise en passant de la somme à l'intégrale. Grâce à cette formule, on peut affiner le résultat jusqu'à obtenir une précision arbitrairement grande.

$$\sum_{n=a}^{b} f(n) = \int_a^b f(x)dx + \frac{f(a) + f(b)}{2} + \sum_{k=1}^{+\infty} \frac{B_{2k}}{(2k)!} (f^{(2k-1)}(b) - f^{(2k-1)}(a))$$

Les $B_n$ sont les nombres de Bernoulli et les $f^(k)$ sont les dérivées $k$-ièmes de $f$. La formule d'Euler-Maclaurin fait le lien entre l'intégrale et la somme via la formule des trapèzes pour une subdivision régulière. Appliquons la formule pour $a = 1$ et $b = n$ (nous généraliserons après).

$$\sum_{k=1}^{n} f(n) = \int_1^n f(x)dx + \frac{f(1)+f(n)}{2} + \sum_{k=1}^{+\infty} \frac{B_{2k}}{(2k)!}(f^{(2k-1)}(n) - f^{(2k-1)}(1))$$

Dans les formules qui nous intéressent, dans le membre de droite, la série de termes qui ne dépendent pas de $n$ tend vers une constante $C$ quand $n \to \infty$. On a donc la formule générale

$$\sum_{k=1}^{n} f(n) = \int_1^n f(x)dx + C + \frac{f(n)}{2} + \sum_{k=1}^{+\infty} \frac{B_{2k}}{(2k)!} f^{(2k-1)}(n)$$

Appliquons cette formule aux sommes rencontrées dans nos calculs

- $f(x) = 1/x$, avec $f^{(2k-1)}(x) = -\frac{(2k-1)!}{x^{2k}}$

  $$\sum_{k=1}^{n} \frac{1}{k} = \ln n + \gamma + \frac{1}{2n} - \sum_{k=1}^{+\infty} \frac{B_{2k}}{2k} \frac{1}{x^{2k}} = \ln n + \gamma + \frac{1}{2n} - \frac{1}{12n^2} + \frac{1}{120n^4} - \frac{1}{252n^6} + \frac{1}{240n^8} + ...$$

  ($\gamma \simeq 0,5772156649015328$ est la constante d'Euler-Mascheroni)

- $f(x) = 1/x^2$, avec $f^{(2k-1)}(x) = -\frac{2k!}{x^{2k+1}}$

  $$\sum_{k=1}^{n} \frac{1}{k^2} = \frac{-1}{n} + \frac{\pi^2}{6} + \frac{1}{2n^2} - \sum_{k=1}^{+\infty} B_{2k} \frac{1}{x^{2k+1}} = \frac{\pi^2}{6} - \frac{1}{n} + \frac{1}{2n^2} - \frac{1}{6n^3} + \frac{1}{30n^5} - \frac{1}{42n^7} + \frac{1}{30n^9} + ...$$

  (montrer que $\zeta(2) = \sum_{k=1}^{+\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$ est un problème classique connu comme « le problème de Bâle »)

- $f(x) = 1/x^3$, avec $f^{(2k-1)}(x) = -\frac{(2k+1)!}{2x^{2k+2}}$

  $$\sum_{k=1}^{n} \frac{1}{k^3} = \frac{-1}{2n^2} + \zeta(3) + \frac{1}{2n^3} - \sum_{k=1}^{+\infty} B_{2k} \frac{2k+1}{2x^{2k+2}} = \zeta(3) - \frac{1}{2n^2} + \frac{1}{2n^3} - \frac{1}{4n^4} + \frac{1}{12n^6} - \frac{1}{12n^8} + \frac{3}{20n^{10}} + ...$$

  ($\zeta(3) \simeq 1.2020569031595942$ est la constante d'Apéry)

L'avantage de ces développements est qu'ils convergent beaucoup plus vite que la somme initiale. Si on prend $n = 10$, on peut vérifier qu'en sommant 3 termes des développements on a déjà un résultat avec 3 décimales correctes.

Dans nos études, les sommes ne commencent pas à 1 en général, on a plutôt des sommes de la forme $\sum_{n+1 \leq k \leq n+\Delta} \frac{1}{k}$ ou $\sum_{n+1 \leq k \leq 2n} \frac{1}{k}$. Dans ce cas, il suffit de couper la somme en

deux : $\sum_{n+1\leq k\leq n+\Delta} = \sum_{1\leq k\leq n+\Delta} - \sum_{1\leq k\leq n}$. Appliquons les développements vus précédemment :

$$\sum_{k=n+1}^{n+\Delta} \frac{1}{k} = \frac{\Delta}{n}\left(1 - \frac{\Delta+1}{2n} + \frac{2\Delta^2+3\Delta+1}{6n^2} - \frac{\Delta^3+2\Delta^2+\Delta}{4n^3} + ...\right)$$

L'erreur relative en ne sommant que les $j$ premiers termes est de l'ordre de $\left(\frac{\Delta}{n}\right)^j$, ce qui permet de choisir une condition d'arrêt acceptable. Par exemple, une erreur relative de $10^{-15}$ signifie que les 15 premières décimales *à partir de la première décimale non nulle* du résultat sont correctes (l'erreur absolue est une mauvaise indication pour trouver le bon moment pour arrêter la sommation).

$$\sum_{k=n+1}^{2n} \frac{1}{k} = \ln(2) + \frac{1}{4n} - \frac{1}{16n^2} + \frac{1}{96n^4} - \frac{1}{256n^6} + ...$$

Dans cette somme, la confusion entre erreur absolue et erreur relative est moins dangereuse vu que le terme principal vaut $\ln(2) \simeq 0.693$, les deux sont donc du même ordre de grandeur. En pratique, la condition d'arrêt s'obtient en comparant le dernier terme qui a été additionné au résultat actuel : quand on juge que ce qui est ajouté est négligeable comparé au résultat, on s'arrête.

# 3 Informations supplémentaires concernant l'agrégation des transitions

## 3.1 Agrégation de petits indels neutres

Pas d'information supplémentaire.

## 3.2 Agrégation de l'inactivation de gènes

Nous réécrivons ici les sommes trouvées en section 3.3 du chapitre III sous la forme sous laquelle elles sont utilisées dans le programme pour que le développement asymptotique présenté en section 2.3 s'applique facilement.

### 3.2.1 Petits indels : inactivation d'un gène

Transition $(x, y) \to (x, y-1)$ :

$$\frac{1}{\Delta n \Delta L} \sum_{L=L_{min}}^{L_{max}-l_{gene}} \frac{n_{min} l_{gene}}{n_{min} l_{gene} + L} = \frac{n_{min} l_{gene}}{\Delta n \Delta L} \sum_{k=L_{min}+n_{min} l_{gene}}^{L_{max}-l_{gene}+n_{min} l_{gene}} \frac{1}{k}$$

Transition $(x, y) \to (x+1, y-1)$ :

$$\frac{n_{min} l_{gene}}{\Delta n \Delta L} \sum_{k=L_{max}-l_{gene}+1+n_{min}l_{gene}}^{L_{max}+n_{min}l_{gene}} \frac{1}{k}$$

Transition $(x, y) \to (x+1, y)$ :

$$\frac{l_{gene}}{\Delta n \Delta L} \sum_{n=n_{min}+1}^{n_{max}} n \sum_{k=L_{max}-l_{gene}+1+nl_{gene}}^{L_{max}+nl_{gene}} \frac{1}{k}$$

En fonction du type de la mutation (petite délétion ou petite insertion), il faut ajouter ou retirer 6 au domaine de sommation du non codant.

### 3.2.2 Inversions : inactivation potentielle de deux gènes

**Inactivation d'un gène** Transition $(x, y) \to (x, y-1)$ :

$$\begin{aligned}
&\frac{2}{\Delta n \Delta L} \sum_{L=L_{min}}^{L_{max}-l_{gene}} \frac{n_{min} l_{gene}}{n_{min} l_{gene} + L} \left(1 - \frac{n_{min} l_{gene}}{n_{min} l_{gene} + L}\right) \\
&= \frac{2 n_{min} l_{gene}}{\Delta n \Delta L} \left( \sum_{L=L_{min}+n_{min}l_{gene}}^{L_{max}-l_{gene}+n_{min}l_{gene}} \frac{1}{k} - n_{min} l_{gene} \sum_{k=L_{min}+n_{min}l_{gene}}^{L_{max}-l_{gene}+n_{min}l_{gene}} \frac{1}{k^2} \right)
\end{aligned}$$

Transition $(x, y) \to (x+1, y-1)$ :

$$\frac{2 n_{min} l_{gene}}{\Delta n \Delta L} \left( \sum_{k=L_{max}-l_{gene}+n_{min}l_{gene}+1}^{L_{max}+n_{min}l_{gene}} \frac{1}{k} - n_{min} l_{gene} \sum_{k=L_{max}-l_{gene}+n_{min}l_{gene}+1}^{L_{max}+n_{min}l_{gene}} \frac{1}{k^2} \right)$$

Transition $(x, y) \to (x+1, y)$ :

$$\frac{2 l_{gene}}{\Delta n \Delta L} \sum_{n=n_{min}+1}^{n_{max}} \left( n \sum_{k=L_{max}-l_{gene}+nl_{gene}+1}^{L_{max}+nl_{gene}} \frac{1}{k} - n^2 l_{gene} \sum_{k=L_{max}-l_{gene}+nl_{gene}+1}^{L_{max}+nl_{gene}} \frac{1}{k^2} \right)$$

**Inactivation de deux gènes** Transition $(x, y) \to (x, y-1)$ :

$$\frac{l_{gene}^2}{\Delta n \Delta L} \left( n_{min}^2 \sum_{k=L_{min}+n_{min}l_{gene}}^{L_{max}-2l_{gene}+n_{min}l_{gene}} \frac{1}{k^2} + (n_{min}+1)^2 \sum_{k=L_{min}+(n_{min}+1)l_{gene}}^{L_{max}-2l_{gene}+(n_{min}+1)l_{gene}} \frac{1}{k^2} \right)$$

Transition $(x, y) \to (x+1, y-1)$ :

$$\frac{l_{gene}^2}{\Delta n \Delta L}\left(n_{min}^2 \sum_{k=L_{max}-2l_{gene}+n_{min}l_{gene}+1}^{L_{max}+n_{min}l_{gene}} \frac{1}{k^2} + (n_{min}+1)^2 \sum_{k=L_{max}-2l_{gene}+(n_{min}+1)l_{gene}+1}^{L_{max}+(n_{min}+1)l_{gene}} \frac{1}{k^2}\right)$$

Transition $(x, y) \to (x+1, y)$ :

$$\frac{l_{gene}^2}{\Delta n \Delta L} \sum_{n=n_{min}+2}^{n_{max}} n^2 \sum_{k=L_{max}-2l_{gene}+nl_{gene}+1}^{L_{max}+nl_{gene}} \frac{1}{k^2}$$

### 3.2.3 Translocations : inactivation potentielle de trois gènes

**Inactivation d'un gène** Transition $(x, y) \to (x, y-1)$ :

$$\frac{3}{\Delta n \Delta L} \sum_{L=L_{min}}^{L_{max}-l_{gene}} \frac{n_{min}l_{gene}}{n_{min}l_{gene}+L}\left(1 - \frac{n_{min}l_{gene}}{n_{min}l_{gene}+L}\right)^2$$

$$= \frac{3n_{min}l_{gene}}{\Delta n \Delta L}\left(\sum_{k=L_{min}+n_{min}l_{gene}}^{L_{max}+n_{min}l_{gene}-l_{gene}} \frac{1}{k} - 2n_{min}l_{gene} \sum_{k=L_{min}+n_{min}l_{gene}}^{L_{max}+n_{min}l_{gene}-l_{gene}} \frac{1}{k^2} + (n_{min}l_{gene})^2 \sum_{k=L_{min}+n_{min}l_{gene}}^{L_{max}+n_{min}l_{gene}-l_{gene}} \frac{1}{k^3}\right)$$

Transition $(x, y) \to (x+1, y-1)$ :

$$\frac{3n_{min}l_{gene}}{\Delta n \Delta L}\left(\sum_{k=L_{max}+n_{min}l_{gene}-l_{gene}+1}^{L_{max}+n_{min}l_{gene}} \frac{1}{k} - 2n_{min}l_{gene} \sum_{k=L_{max}+n_{min}l_{gene}-l_{gene}+1}^{L_{max}+n_{min}l_{gene}} \frac{1}{k^2} + (n_{min}l_{gene})^2 \sum_{k=L_{max}+n_{min}l_{gene}-l_{gene}+1}^{L_{max}+n_{min}l_{gene}} \frac{1}{k^3}\right)$$

Transition $(x, y) \to (x+1, y)$ :

$$\frac{3l_{gene}}{\Delta n \Delta L} \sum_{n=n_{min}+1}^{n_{max}}\left(n \sum_{k=L_{max}+(n-1)l_{gene}+1}^{L_{max}+nl_{gene}} \frac{1}{k} - 2n^2 l_{gene} \sum_{k=L_{max}+(n-1)l_{gene}+1}^{L_{max}+nl_{gene}} \frac{1}{k^2} + n^3 l_{gene}^2 \sum_{k=L_{max}+(n-1)l_{gene}+1}^{L_{max}+nl_{gene}} \frac{1}{k^3}\right)$$

**Inactivation de deux gènes** Transition $(x, y) \to (x, y-1)$ :

$$\frac{3l_{gene}^2}{\Delta n \Delta L}\left(n_{min}^2\left(\sum_{k=L_{min}+n_{min}l_{gene}}^{L_{max}+n_{min}l_{gene}-2l_{gene}}\frac{1}{k^2}-n_{min}l_{gene}\sum_{k=L_{min}+n_{min}l_{gene}}^{L_{max}+n_{min}l_{gene}-2l_{gene}}\frac{1}{k^3}\right)\right.$$
$$\left.+(n_{min}+1)^2\left(\sum_{k=L_{min}+n_{min}l_{gene}+l_{gene}}^{L_{max}+n_{min}l_{gene}-l_{gene}}\frac{1}{k^2}-(n_{min}+1)l_{gene}\sum_{k=L_{min}+n_{min}l_{gene}+l_{gene}}^{L_{max}+n_{min}l_{gene}-l_{gene}}\frac{1}{k^3}\right)\right)$$

Transition $(x, y) \to (x+1, y-1)$ :

$$\frac{3l_{gene}^2}{\Delta n \Delta L}\left(n_{min}^2\left(\sum_{k=L_{max}+n_{min}l_{gene}-2l_{gene}+1}^{L_{max}+n_{min}l_{gene}}\frac{1}{k^2}-n_{min}l_{gene}\sum_{k=L_{max}+n_{min}l_{gene}-2l_{gene}+1}^{L_{max}+n_{min}l_{gene}}\frac{1}{k^3}\right)\right.$$
$$\left.+(n_{min}+1)^2\left(\sum_{k=L_{max}+n_{min}l_{gene}-l_{gene}+1}^{L_{max}+n_{min}l_{gene}+l_{gene}}\frac{1}{k^2}-(n_{min}+1)l_{gene}\sum_{k=L_{max}+n_{min}l_{gene}-l_{gene}+1}^{L_{max}+n_{min}l_{gene}+l_{gene}}\frac{1}{k^3}\right)\right)$$

Transition $(x, y) \to (x+1, y)$ :

$$\frac{3l_{gene}^2}{\Delta n \Delta L}\sum_{n=n_{min}+2}^{n_{max}}\left(n^2\sum_{k=L_{max}+(n-2)l_{gene}+1}^{L_{max}+nl_{gene}}\frac{1}{k^2}-n^3l_{gene}\sum_{k=L_{max}+(n-2)l_{gene}+1}^{L_{max}+nl_{gene}}\frac{1}{k^3}\right)$$

**Inactivation de trois gènes** Transition $(x, y) \to (x, y-1)$ :

$$\frac{l_{gene}^3}{\Delta n \Delta L}\left(n_{min}^3\sum_{k=L_{min}+n_{min}l_{gene}}^{L_{max}+n_{min}l_{gene}-3l_{gene}}\frac{1}{k^3}+(n_{min}+1)^3\sum_{k=L_{min}+n_{min}l_{gene}+l_{gene}}^{L_{max}+n_{min}l_{gene}-2l_{gene}}\frac{1}{k^3}\right.$$
$$\left.+(n_{min}+2)^3\sum_{k=L_{min}+n_{min}l_{gene}+2l_{gene}}^{L_{max}+n_{min}l_{gene}-l_{gene}}\frac{1}{k^3}\right)$$

Transition $(x, y) \to (x+1, y-1)$ :

$$\frac{l_{gene}^3}{\Delta n \Delta L}\left(n_{min}^3\sum_{k=L_{max}+n_{min}l_{gene}-3l_{gene}+1}^{L_{max}+n_{min}l_{gene}}\frac{1}{k^3}+(n_{min}+1)^3\sum_{k=L_{max}+n_{min}l_{gene}-2l_{gene}+1}^{L_{max}+n_{min}l_{gene}+l_{gene}}\frac{1}{k^3}\right.$$
$$\left.+(n_{min}+2)^3\sum_{k=L_{max}+n_{min}l_{gene}-l_{gene}+1}^{L_{max}+n_{min}l_{gene}+2l_{gene}}\frac{1}{k^3}\right)$$

Transition $(x, y) \to (x+1, y)$ :

$$\frac{l_{gene}^3}{\Delta n \Delta L}\sum_{n=n_{min}+3}^{n_{max}}\left(n^3\sum_{k=L_{max}+(n-3)l_{gene}+1}^{L_{max}+nl_{gene}}\frac{1}{k^3}\right)$$

## 3.3 Agrégation des délétions

Dans cette section, nous donnons les démonstrations des densités annoncées dans le chapitre III, à la section 3.4 (page 120). Nous donnons également les algorithmes que servent à agréger les différents cas particuliers. Pour une explication complète des grandeurs recherchées et des termes utilisés, on se reportera au texte principal.

### 3.3.1 Cas général (approximation LL, $n_0 > 1$, $0 < n_f < n_0$, $L_3 \leq L_5$)

Nous allons ici expliciter ce que nous appelions l'intégration glissante dans la section 3.4.1 du chapitre III. Pour cela nous allons supposer que nous sommes en train d'agréger des profils triangulaires identiques régulièrement espacés le long d'un axe 1D de variable $L$. Nous supposons que la hauteur des triangles à agréger est $h$, leur largeur $w$ et que $d$ est la distance qui sépare les sommets de 2 triangles adjacents. Nous allons montrer que sommer les profils triangulaires revient approximativement à faire une intégration glissante pondérée. Pour cela, nous allons nous concentrer sur la contribution apportée par chacun des triangles quand on fait glisser un curseur le long de l'axe $L$.

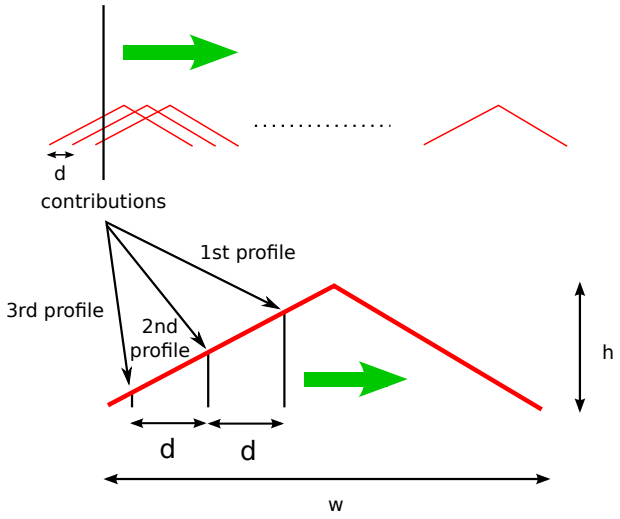


Figure A.6 – Contribution de chaque triangle au profil agrégé. On peut repérer la contribution d'un triangle par rapport à son sommet et mettre côte-à-côte les contributions de tous les triangles sur un profil-type. Quand on avance le long de l'axe des $x$, la contribution de chaque triangle bouge le long du profil-type mais la distance entre deux contributions est toujours $d$.

Plaçons le curseur sur une position $L$. Nous pouvons recenser quels triangles participent au profil global et à quelle hauteur (figure A.6). On peut placer chaque contribution le long d'un seul triangle que nous appelerons « carte des contributions ». Appelons $s = h/(w/2) = 2h/w$ la pente ascendante du triangle. Dans l'exemple, si $C_i$ est la contribution du $i$-eme profil, on voit que $C_2 = C_3 + ds$ et $C_1 = C_2 + ds$ donc la contribution totale est $C = C_1 + C_2 + C_3 = 3C_3 + ds + 2ds$.

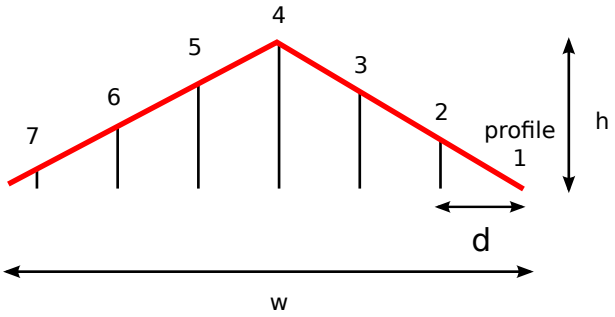


Figure A.7 – Contribution de chaque triangle au profil agrégé quand $L$ pointe sur la fin du premier profil atomique

Si on pousse le curseur assez loin, jusqu'à la fin du premier triangle, et qu'on répertorie les contributions de chaque triangle sur un profil-type, les contributions sont réparties tout le long du profil-type (figure A.7). On peut déjà intuiter qu'en faisant la somme, on a quelque chose qui ressemble au calcul de l'aire du profil-type. Faisons le calcul : la contribution des premiers triangles (ceux avec le plus petit identifiant, à droit sur la figure A.7) est 0, $sd$, $2sd$, etc. En tout, il y a $N = \lfloor w/(2d) \rfloor$ contributions *non nulles* le long de la partie descendante. La contribution du dernier triangle sur la partie descendante est donc celle du triangle $N+1$, elle vaut $Nds$. La distance entre cette contribution et le sommet du profil est $r = w/2 - dN$. Le triangle $N+2$ est le premier triangle en partant de la droite sur la partie ascendante, sa contribution est $(w-(N+1)d)s = (N.2d+2r-(N+1)d)s = (N.d+2r-d)s = N.ds + (2r-d)s$. À partir de là, les contributions décroissent de $ds$ avec chaque nouveau triangle jusqu'à atteindre 0. La contribution totale de la partie descendante est donc

$$C_d = 0 + ds + 2ds + ... + Nds = \frac{N(R+1)}{2}ds$$

et sur la partie ascendante (les crochets indiquent une contribution optionnelle, selon le signe de $2r-d$)

$$C_a = (Nds+(2r-d)s)+...+(ds+(2r-d)s)+[(2r-d)s] = \frac{N(N+1)}{2}ds+(N[+1])(2r-d)s$$

La contribution totale est donc

$$C = C_d + C_a = N(N+1)ds + (N[+1])(2r-d)s$$

En substituant $N$ par $N = \lfloor w/(2d) \rfloor = w/(2d) - r/d$ et en réarrangeant, on trouve

$$C = \left[ \left(\frac{w}{2d}\right)^2 - \left(\frac{r}{d}[-1]\right)^2 \right] ds$$

En pratique $d \leq 2$ et $w = 2(l_{intergenic} + l_{gene}) \geq 2000$ donc $w \gg d$. De plus, $0 \leq r < d$, donc le deuxième terme entre crochets peut être négligé.

$$C \simeq \left(\frac{w}{2d}\right)^2 ds = \frac{w^2}{4d}s = \frac{w^2}{4d}\frac{2h}{w} = \frac{hw/2}{d} = \frac{\text{aire du profil atomique}}{d}$$

De fait, la somme des profils triangulaires au niveau de la fin du premier triangle revient donc à intégrer l'aire sous un seul de ces triangles avec un facteur $1/d$ : plus $d$ est petit, plus on introduit de redondance dans le calcul de l'aire.

Revenons maintenant au cas montré dans la figure A.7. La somme des contributions au niveau de la fin du premier triangle vaut à peu près $hw/(2d)$. Si on continue à avancer le long de l'axe des $x$, on a un jeu de chaises musicales : les contributions des premiers profils décroissent vers 0 mais de nouveaux profils arrivent de l'autre côté et compensent ces pertes. Graphiquement, on voit bien que la somme n'est pas parfaitement constante mais les fluctuations sont négligeables. Tant que de nouveaux profils arrivent, on est donc sur un plateau de hauteur $H \simeq hw/(2d)$.

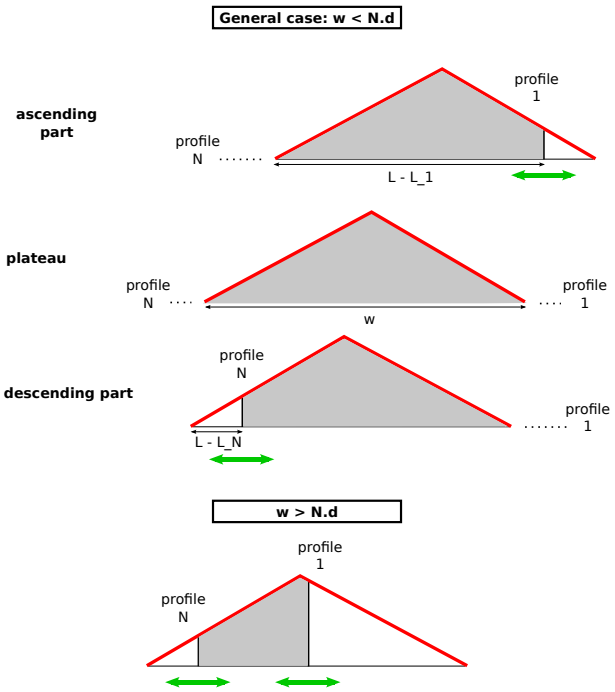


Figure A.8 – Les contributions ne couvrent pas toujours le profil-type de la carte des contributions. Cette vue schématique résume tous les cas particuliers pouvant survenir.

Il reste à déterminer le comportement avant et après le plateau. La réponse est plutôt simple si on utilise la carte des contributions. Sur la figure A.8), nous avons recensé tous les cas particuliers qui peuvent survenir. On quitte le plateau s'il manque des contributions sur la droite ou sur la gauche. Au début, il manque des contributions entre le premier triangle et le côté droit du profil-type. Appelons $L_1$ le point de départ du premier triangle. L'espace du profil-type qui n'est occupé par aucune contribution est $w - (L - L_1)$. L'aire non occupée est donc

$$\begin{cases} \frac{hw}{2} - \int_0^{L-L_1} \frac{h}{w/2} u du = \frac{hw}{2} - \frac{2h}{w}(L-L_1)^2 & \text{si } (L-L_1) \in [0, w/2] \\ \int_0^{w-(L-L_1)} \frac{h}{w/2} u du = \frac{2h}{w}(w-(L-L_1))^2 & \text{si } (L-L_1) \in [w/2, w] \end{cases}$$

Pour passer à la valeur en termes de contributions, on multiplie par $1/d$. Soit $C = 2h/(wd)$.

L'aire qui manque pour arriver au plateau vaut

$$MC_1(L) = \begin{cases} H & \text{si } L < L_1 \\ H - C(L - L_1)^2 & \text{si } L \in [L_1, L_1 + w/2] \\ C((L_1 + w) - L)^2 & \text{si } L \in [L_1 + w/2, L_1 + w] \\ 0 & \text{si } L > L_1 + w \end{cases}$$

À la fin, il y a des contributions manquantes quand on atteint le dernier triangle à agréger. S'il y a $T$ triangles à additionner et que $L_T$ est la position qui marque le début du dernier triangle, l'espace non occupé sur le profil-type est $L - L_T$. La situations est parfaitement symétrique au cas précédent. Les contributions qui manquent par rapport au plateau sont

$$MC_2(L) = \begin{cases} 0 & \text{si } L < L_T \\ C(L - L_T)^2 & \text{si } L \in [L_T, L_T + w/2] \\ H - C((L_T + w) - L)^2 & \text{si } L \in [L_T + w/2, L_T + w] \\ H & \text{si } L > L_T + w \end{cases}$$

Pour une position quelconque $L$, la contribution totale est donc donnée par

$$f(L) = H - MC_1(L) - MC_2(L)$$

Pour obtenir la contribution moyenne, il suffit de diviser par $T$.

En pratique, cela signifie que le plateau n'existe pas forcément. Il existe quand $w < Td$ car au moins une des deux fonctions qui répertorient les contributions manquantes est nulle. Au début, $MC_2(L) = 0$, d'où

$$f(L) = H - MC_1(L) = \begin{cases} 0 & \text{si } L < L_1 \\ C(L - L_1)^2 & \text{si } L \in [L_1, L_1 + w/2] \\ H - C((L_1 + w) - L)^2 & \text{si } L \in [L_1 + w/2, L_1 + w] \\ H & \text{si } L > L_1 + w \end{cases}$$

À la fin, quand le dernier triangle est atteint, $MC_1(L) = 0$ et nous obtenons une fonction décroissante donnée par

$$f(L) = H - MC_2(L) = \begin{cases} H & \text{si } L < L_T \\ H - C.(L - L_T)^2 & \text{si } L \in [L_T, L_T + w/2] \\ C.((L_T + w) - L)^2 & \text{si } L \in [L_T + w/2, L_T + w] \\ 0 & \text{si } L > L_T + w \end{cases}$$

On a donc un profil avec une croissance quadratique, un plateau et une décroissance quadratique (figure A.9). Il y a un point de symétrie centrale au milieu de la partie croissante, idem pour la partie décroissante.

Si $w > T.d$, la partie croissante et décroissante se chevauchent : il n'y a plus de plateau. Ce cas est illustré dans la section principale dédiée à l'agrégation des délétions (chapitre III, section 3.4.1, page 123).

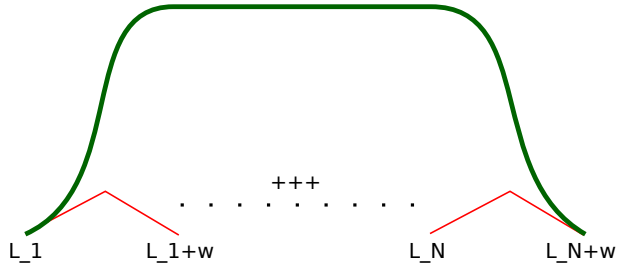


Figure A.9 – Profil agrégé quand il y a un plateau $w < T.d$.

Finalement, pour intégrer les contributions, nous nous servons de la fonction cumulative $F(L) = h(L-L_1) - CMC_1(L) - CMC_2(L)$, où $CMC_{1/2}$ sont les cumulatives des contributions manquantes par rapport au plateau. Dans le programme, nous calculons (en notant $L_1 = L_1$, $L_2 = L_1 + w/2$, $L_3 = L_1 + w$)

$$h(L-L_1) - CMC_1(L) = \begin{cases} 0 & \text{si } L < L_1 \\ C_1 \frac{(L-L_1)^3}{3} & \text{si } L \in [L_1, L_2] \\ h(L-L_2) + C_1 \frac{(L_3-L)^3}{3} & \text{si } L \in [L_2, L_3] \\ h(L-L_2) & \text{si } L > L_3 \end{cases}$$

puis retranchons (en notant $L_4 = L_T$, $L_5 = L_T + w/2$, $L_6 = L_T + w$)

$$CMC_2(L) = \begin{cases} 0 & \text{si } L < L_4 \\ C_2 \frac{(L-L_4)^3}{3} & \text{si } L \in [L_4, L_5] \\ h(L-L_5) + C_2 \frac{(L_6-L)^3}{3} & \text{si } L \in [L_5, L_6] \\ h(L-L_5) & \text{si } L > L_6 \end{cases}$$

### 3.3.2 Agrégation ligne vers ligne, cas où la densité finale en gène est faible (LL_low_final_density, $n_0 > 1$, $0 < n_f < n_0$, $L_3 > L_5$)

Comme expliqué dans la section principale, nous utilisons une approximation basée sur 2 parties linéaires de pentes opposées sur $[L_4, L_2]$ et $[L_5, L_3]$, on complète avec 3 parties quadratiques en imposant que le tout soit $C^1$. Mathématiquement, cela s'écrit (dans certains cas extrêmes, il faut échanger la position de $L_1$ et $L_4$)

$$f(L) = \begin{cases} a_1(L-L_1)^2 & \text{si } L \in [L_1, L_4] \\ sL + b_2 & \text{si } L \in [L_4, L_2] \\ a_3(L-c_3)^2 + b_3 & \text{si } L \in [L_2, L_5] \\ -sL + b_4 & \text{si } L \in [L_5, L_3] \\ a_5(L-L_6)^2 + b_5 & \text{si } L \in [L_3, L_6] \end{cases}$$

La dérivée vaut

$$f'(L) = \begin{cases} 2a_1(L - L_1) & \text{si } L \in [L_1, L_4] \\ s & \text{si } L \in [L_4, L_2] \\ 2a_3(L - c_3) & \text{si } L \in [L_2, L_5] \\ -s & \text{si } L \in [L_5, L_3] \\ 2a_5(L - L_6) & \text{si } L \in [L_3, L_6] \end{cases}$$

En $L_1$ et $L_6$, la fonction est bien $C^1$. En $L_4$, il faut

- $s = 2a_1(L_4 - L_1)$ d'où $a_1 = s/[2(L_4 - L_1)]$.
- $sL_4 + b_2 = a_1(L_4 - L_1)^2 = s(L_4 - L_1)/2$, d'où $b_2 = -s(L_1 + L_4)/2$.

Symétriquement, en $L_3$,

- $-s = 2a_5(L_3 - L_6)$ d'où $a_5 = s/[2(L_6 - L_3)]$.
- $-sL_2 + b_4 = a_5(L_3 - L_6)^2 = s(L_6 - L_3)/2$, d'où $b_4 = s(L_6 + L_3)/2$.

En $L_2$ et $L_5$,

- $s = 2a_3(L_2 - c_3)$ d'où $a_3 = s/[2(L_2 - c_3)]$.
- $s(2L_2 - (L_1 + L_4))/2 = a_3(L_2 - c_3)^2 + b_3 = s(L_2 - c_3)/2 + b_3$.
- $-s = 2a_3(L_5 - c_3)$ d'où $a_3 = -s/[2(L_5 - c_3)]$.
- $s((L_2 + L_6) - 2L_5)/2 = a_3(L_5 - c_3)^2 + b_3 = s(L_5 - c_3)/2 + b_3$.

Des deux expressions concernant $a_3$, on obtient $L_2 - c_3 = -(L_5 - c_3)$, d'où $c_3 = (L_2 + L_5)/2$ et $a_3 = -s/[L_5 - L_2]$. De la deuxième expression, on déduit $b_3 = s(2L_2 - (L_1 + L_4))/2 - s(L_2 - c_3)/2$. Pour simplifier, exprimons tout en fonction de $L_2$ et $L_5$. On a $L_1 = L_2 - L_{min}/n_0 - l_{gene}$ et $L_4 = L_5 - L_{max}/n_0 - l_{gene}$. On pose $IGG1 = L_{min}/n_0 + l_{gene}$ et $IGG2 = L_{max}/n_0 + l_{gene}$, alors

$$\begin{aligned} b_3 &= \frac{s}{2}\left((2L_2 - L_2 - L_5 + IGG1 + IGG2) - (L_2 - (L_2 + L_5)/2)\right) \\ &= \frac{s}{2}\left(IGG1 + IGG2 + \frac{L_2 - L_5}{2}\right) \end{aligned}$$

Cette expression est cohérente avec l'autre contrainte donnée sur $b_3$ est $c_3$. Finalement, en posant $N = s/2$, il nous reste

$$f(L) = \begin{cases} N\frac{(L-L_1)^2}{L_4-L_1} & \text{si } L \in [L_1, L_4] \\ N(2L - (L_1 + L_4)) & \text{si } L \in [L_4, L_2] \\ N\left((IGG1 + IGG2 - \frac{L_5-L_2}{2} - \frac{2(L-(L_2+L_5)/2)^2}{L_5-L_2}\right) & \text{si } L \in [L_2, L_5] \\ N(L_3 + L_6 - 2L) & \text{si } L \in [L_5, L_3] \\ N\frac{(L-L_6)^2}{L_6-L_3} & \text{si } L \in [L_3, L_6] \end{cases}$$

La fonction cumulative est :

$$F(L) = \begin{cases} N\frac{(L-L_1)^3}{3(L_4-L_1)} & \text{si } L \in [L_1, L_4] \\ N\frac{(L_4-L_1)^2}{3} & \text{si } L = L_4 \\ F(L_4) + N(L - L_1)(L - L_4) & \text{si } L \in [L_4, L_2] \\ F(L_4) + N(L_2 - L_1)(L_2 - L_4) & \text{si } L = L_2 \\ F(L_2) + N\left(\left((IGG1 + IGG2 - \frac{L_5-L_2}{2}\right)(L - L_2)\right. & \\ \quad \left. - \frac{2}{3(L_5-L_2)}\left(\left(L - \frac{L_2+L_5}{2}\right)^3 - \left(\frac{L_5-L_2}{2}\right)^3\right)\right) & \text{si } L \in [L_2, L_5] \\ F(L_2) + N(L_5 - L_2)\left(IGG1 + IGG2 - \frac{2}{3}(L_5 - L_2)\right) & \text{si } L = L_5 \\ F(L_5) + N(L - L_5)(L_3 + L_6 - (L + L_5)) & \text{si } L \in [L_5, L_3] \\ F(L_5) + N(L_3 - L_5)(L_6 - L_5) & \text{si } L = L_3 \\ F(L_3) + N\frac{(L_6-L_3)^3-(L_6-L)^3}{3(L_6-L_3)} & \text{si } L \in [L_3, L_6] \end{cases}$$

### 3.3.3 Agrégation ligne vers ligne, cas où tous les gènes sont perdus ($n_0 > 1$, $n_f = 0$)

Pour la ligne $n_f = 0$, les triangles à agréger sont tronqués et la partie ascendante remplacée par un plateau. La partie qui s'étend en dessous de $L = 0$ doit (logiquement) être retirée. Pour éviter de faire un algorithme supplémentaire, nous utilisons l'approximation LL_low_final_density en tronquant la partie entre $L_1 < 0$ et 0.

### 3.3.4 Agrégation ligne vers ligne, cas où aucun gène n'est perdu (LL_no_gene_loss, $n_0 > 1$, $n_f = n_0$, $\tilde{L}_2 < \tilde{L}_3$)

Dans le cas où aucun gène n'est perdu, les profils à agréger ne sont pas des triangles complets, il faut en tronquer la plus grande partie. On va appliquer le même type de raisonnement, mais il va falloir changer le nom des points caractéristiques des deux profils extrêmes. Cette fois, les profils ne sont composés que d'une pente ascendante, ils sont simplement caractérisés par leur support. On nomme $\tilde{L}_1$ et $\tilde{L}_2$ les extrémités du support du profil tronqué tout à gauche, et $\tilde{L}_3$ et $\tilde{L}_4$ les extrémités du support du profil tronqué tout à droite.

Ce cas est proche de l'approximation LL, sauf que les profils à agréger ne sont pas des triangles complets. Si $\tilde{L}_2 < \tilde{L}_3$, les profils atomiques extrêmes ne se chevauchent pas, on peut refaire l'approximation de l'intégration glissante. On obtient alors pour le profil agrégé une partie ascendante quasi-quadratique, puis une partie linéaire qui croît lentement au lieu du plateau de l'approximation LL, enfin une partie descendante quasi-quadratique. Cette fois-ci, il n'y a pas de plateau car la troncature ne donne pas la même aire à chaque profil atomique : l'aire augmente avec la quantité de non-codant initial.

Dans cette section, on pose $IG_1 = L_{min}/n_0$ la longueur de l'intergénique du profil tout à gauche et $IG_2 = L_{max}/n_0$ celui du profil tout à droite. On pose aussi $IGG_1 = IG_1 + l_{gene}$ et $IGG_2 = IG_2 + l_{gene}$.

L'idée est la même que pour l'approximation LL, sauf que les triangles agrégés sont tronqués. On réutilisera ici les constantes $C_1 = h/[2(L_{min}/n_0 + l_{gene})^2] = h/[2IGG_1^2]$ et $C_2 = h/[2IGG_2^2]$ de l'approximation LL. Par contre, le plateau de hauteur $h$ ne sera jamais atteint à cause de la troncature.

Quand $\tilde{L}_3 > \tilde{L}_2$, le profil agrégé est le même que celui de l'approximation LL pour $L < L_2$, soit $f(L) = C_1(L - \tilde{L}_1)^2$. Posons $h_1 = f(\tilde{L}_2) = C_1(\tilde{L}_2 - \tilde{L}_1)^2 = h(IG_1 + 1)^2/[2IGG_1^2]$. De l'autre côté, posons $h_2 = f(\tilde{L}_3)$. Nous savons que $f(\tilde{L}_4) = 0$, et par intégration glissante, $f(L) = h_2 - C_2(L - \tilde{L}_3)^2$ pour $L \in [\tilde{L}_3, \tilde{L}_4]$. On a donc $h_2 = C_2(\tilde{L}_4 - \tilde{L}_3)^2 = h(IG_2 + 1)^2/[2IGG_2^2]$. Comme $IG_2 > IG_1$, $h_2 > h_1$.

Entre $\tilde{L}_2$ et $\tilde{L}_3$, l'analogie avec l'intégration glissante nous dit qu'on est en train d'intégrer des profils dont l'aire est progressivement croissante. L'aire d'un profil tronqué est $l_{intergenic}^2/[2(l_{intergenic} + l_{gene})^2 n_0]$, qui tend vers une constante quand $L_0$ tend vers l'infini. À part près de l'origine, où la taille de l'intergénique varie fortement au sein de la même subdivision, cette aire varie suffisament lentement pour qu'une approximation linéaire entre $\tilde{L}_2$ et $\tilde{L}_3$ soit bonne. L'approximation est donc moins bonne quand $x = 1$ mais ce n'est pas très grave puisque dans ce cas-là, toutes les transitions restent forcément dans la subdivision de départ $(x_0, y_0)$...

Finalement, on a donc (figure A.10) :

$$f(L) = \begin{cases} C_1.(L - \tilde{L}_1)^2 & \text{si } L \in [\tilde{L}_1, \tilde{L}_2] \\ h_1 + (L - \tilde{L}_2)\frac{h_2 - h_1}{\tilde{L}_3 - \tilde{L}_2} & \text{si } L \in [\tilde{L}_2, \tilde{L}_3] \\ h_2 - C_2(L - \tilde{L}_3)^2 & \text{si } L \in [\tilde{L}_3, \tilde{L}_4] \end{cases}$$

Toutes les constantes s'obtiennent par renormalisation. L'aire d'un profil tronqué est

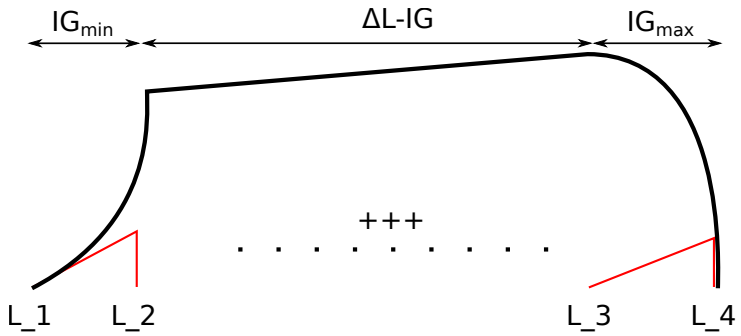


Figure A.10 – Profil agrégé pour des délétions conservant $n_0$ gènes et partant avec $n_0$ gènes et entre $L_{min}$ et $L_{max}$ bases non-codantes.

$l^2_{intergenic}/[2(l_{intergenic}+l_{gene})^2 n_0]$, l'aire totale vaut alors

$$\begin{aligned}\mathcal{A} &= \frac{1}{\Delta L}\sum_{L=L_{min}}^{L_{max}} \frac{l^2_{intergenic}}{2(l_{intergenic}+l_{gene})^2 n_0} = \frac{1}{2n_0\Delta L}\sum_{L=L_{min}}^{L_{max}} \frac{(L/n_0)^2}{(L/n_0+l_{gene})^2} \\ &= \frac{1}{2n_0\Delta L}\sum_{L=L_{min}}^{L_{max}} \left(1-\frac{n_0 l_{gene}}{L+n_0 l_{gene}}\right)^2 \\ &= \frac{1}{2n_0} - \frac{l_{gene}}{\Delta L}\sum_{L=L_{min}+n_0 l_{gene}}^{L_{max}+n_0 l_{gene}} \frac{1}{k} + \frac{n_0 l^2_{gene}}{2\Delta L}\sum_{L=L_{min}+n_0 l_{gene}}^{L_{max}+n_0 l_{gene}} \frac{1}{k^2}\end{aligned}$$

Comme pour l'approximation LL, on calcule les cumulatives en séparant contributions positives et négatives.

$$F(L) = \begin{cases} C_1\frac{(L-\tilde{L}_1)^3}{3} & \text{si } L\in[\tilde{L}_1,\tilde{L}_2] \\ F(\tilde{L}_2)+h_1(L-\tilde{L}_2)+(L-\tilde{L}_2)^2\frac{h_2-h_1}{2(\tilde{L}_3-\tilde{L}_2)} & \text{si } L\in[\tilde{L}_2,\tilde{L}_3] \\ F(\tilde{L}_2)+(\tilde{L}_3-\tilde{L}_2)\frac{h_1+h_2}{2} & \text{si } L=\tilde{L}_3 \\ F(\tilde{L}_3)+h_2(L-\tilde{L}_3)-C_2\frac{(L-\tilde{L}_3)^3}{3} & \text{si } L\in[\tilde{L}_3,\tilde{L}_4] \end{cases}$$

### 3.3.5 Agrégation ligne vers ligne, cas où aucun gène n'est perdu, densité finale en gène faible (LL_no_gene_loss_low_final_density, $n_0 > 1$, $n_f = n_0$, $\tilde{L}_2 > \tilde{L}_3$)

Quand $\tilde{L}_3 < \tilde{L}_2$, tous les profils tronqués se chevauchent et l'approximation LL_no_gene_loss ne fonctionne pas très bien. Entre $\tilde{L}_3$ et $\tilde{L}_2$, tous les triangles contribuent au profil moyen et sont linéairement croissants donc le profil moyen croît linéairement aussi. Nous ajoutons les contraintes suivantes : (i) les autres parties du profil moyen varient en $L^2$, (ii) une condition de régularité : le profil doit être $C^1$ jusqu'en $\tilde{L}_2$ puis continu (figure A.11) et (iii) la partie quadratique initiale doit correspondre à celle obtenue dans le cas précédent, autrement dit elle varie en $(L-\tilde{L}_3)^2$. Mathématiquement,

$$f(L) = \begin{cases} a_1(L-\tilde{L}_1)^2 & \text{si } L \in [\tilde{L}_1, \tilde{L}_3] \\ sL + b_2 & \text{si } L \in [\tilde{L}_3, \tilde{L}_2] \\ a_3(L-\tilde{L}_3)^2 + b_3 & \text{si } L \in [\tilde{L}_2, L_5] \end{cases}$$

Les contraintes de régularité en $\tilde{L}_1$ donnent $a_1 = s/[2(\tilde{L}_3 - \tilde{L}_1)]$ et $b_2 = -s(\tilde{L}_1 + \tilde{L}_3)/2$. En $\tilde{L}_4$, elles donnent $b_3 = -a_3(\tilde{L}_4 - \tilde{L}_3)^2$ et en $\tilde{L}_2$, elles donnent $a_3 = s/2.(2\tilde{L}_2 - \tilde{L}_1 - \tilde{L}_3)/[(\tilde{L}_2 - \tilde{L}_4)(\tilde{L}_2 + \tilde{L}_4 - 2\tilde{L}_3)]$. Soit $N = s/2$, on obtient (figure A.11)

$$f(L) = \begin{cases} N\frac{(L-\tilde{L}_1)^2}{|\tilde{L}_3-\tilde{L}_1|} & \text{si } L \in [\tilde{L}_1, \tilde{L}_3] \\ N(2L - \tilde{L}_1 - \tilde{L}_3) & \text{si } L \in [\tilde{L}_3, \tilde{L}_2] \\ N\left(2\tilde{L}_2 - \tilde{L}_1 - \tilde{L}_3 \times \frac{(L-\tilde{L}_4)(L-2\tilde{L}_3+\tilde{L}_4)}{(\tilde{L}_2-\tilde{L}_4)(\tilde{L}_2-2\tilde{L}_3+\tilde{L}_4)}\right) & \text{si } L \in [\tilde{L}_2, \tilde{L}_4] \end{cases}$$

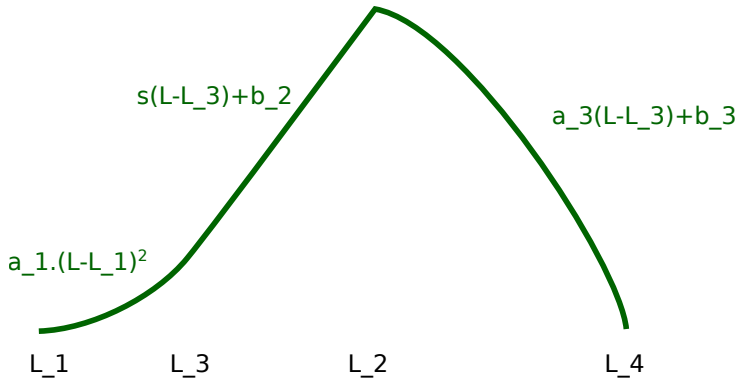


Figure A.11 – Profil agrégé pour des délétions conservant $n_0$ gènes et partant avec $n_0$ gènes et entre $L_{min}$ and $L_{max}$ bases non codantes dans le cas où $L_3 < L_2$.

La fonction cumulative $F(L)$ vaut

$$\begin{cases} N\frac{(L-\tilde{L}_1)^3}{3|\tilde{L}_3-\tilde{L}_1|} & \text{si } L \in [\tilde{L}_1, \tilde{L}_3] \\ F(\tilde{L}_3) + N(L-\tilde{L}_3)(L-\tilde{L}_1) & \text{si } L \in [\tilde{L}_3, \tilde{L}_2] \\ F(\tilde{L}_2) + N\left(\frac{2\tilde{L}_2-\tilde{L}_1-\tilde{L}_3}{(\tilde{L}_2-\tilde{L}_4)(\tilde{L}_2-2\tilde{L}_3+\tilde{L}_4)}\left(\frac{(\tilde{L}_2-\tilde{L}_3)^3-(L-\tilde{L}_3)^3}{3} + (L-\tilde{L}_2)(\tilde{L}_4-\tilde{L}_3)^2\right)\right) & \text{si } L \in [\tilde{L}_2, \tilde{L}_4] \end{cases}$$

### 3.3.6 Agrégation ligne vers ligne, cas où le génome initial ne contient pas de gène (LL_empty, $n_0 = 0$)

Dans ce cas, pour chaque point de départ $(L_0, 0)$, la taille finale est distribuée uniformément entre 0 et $L_0 - 1$. On a dont un profil plat de hauteur $1/L_0$. La ligne finale est forcément $n_f = 0$. Ici les calculs se font en discret, on obtient immédiatement

- $\forall L \in [0, L_{min} - 1]$, $f(L) = 1/\Delta L. \sum_{L_{min} \leq k \leq L_{max}} 1/k$.

- $\forall L \in [L_{min}, L_{max}-1]$, $f(L) = 1/\Delta L . \sum_{L+1 \leq k \leq L_{max}} 1/k$.

Il reste à déterminer les contributions totales pour une subdivision finale $(x_f, 1)$ donnée. Comme les expressions ne dépendent pas de $L$, nous obtenons immédiatement

- $\forall x_f \in [0, x_0 - 1]$, $p(x_f) = \Delta L(x_f) \times f(L) = \Delta L(x_f)/\Delta L(x_0) \sum_{L_{min} \leq k \leq L_{max}} 1/k$.
- Pour $x_f = x_0$, $p(x_0) = 1 - \sum_{x=1}^{x_0-1} p(x) = 1 - L_{min}/\Delta L(x_0) \sum_{L_{min} \leq k \leq L_{max}} 1/k$.

### 3.3.7 Agrégation le long des colonnes de départ (approximation RL, rectangle vers ligne)

Si on reprend les limites du profil simplifié vu dans le chapitre III (section 3.4.2, page 128), on a

- $L'_1 = L_2 - 1/3(l_{intergenic} + l_{gene}) = (n_f - 1/3)L_{min}/n_0 + 2(l_{gene} - 1)/3$
- $L'_2 = L_2 + 1/3(l_{intergenic} + l_{gene}) = (n_f + 1/3)L_{min}/n_0 + 4(l_{gene} - 1)/3$
- $L'_3 = L_4 - 1/3(l_{intergenic} + l_{gene}) = (n_f - 1/3)L_{max}/n_0 + 2(l_{gene} - 1)/3$
- $L'_4 = L_4 + 1/3(l_{intergenic} + l_{gene}) = (n_f + 1/3)L_{max}/n_0 + 4(l_{gene} - 1)/3$

On peut rappeler et généraliser les expressions qui donnent le nombre de profils qui ont dépassé les points caractéristiques $L'_1$, $L'_2$, $L'_3$ et $L'_4$ quand le curseur pointe sur $L$.

$$N_{L'_1}(L) = \begin{cases} 0 & \text{si } L < \frac{n_f - 1/3}{n_{max}} L_{min} + \frac{2(l_{gene}-1)}{3} \\ \Delta n & \text{si } L > \frac{n_f - 1/3}{n_{min}} L_{min} + \frac{2(l_{gene}-1)}{3} \\ n_{max} - \left\lceil \frac{n_f - 1/3}{L - \frac{2(l_{gene}-1)}{3}} L_{min} \right\rceil + 1 & \text{sinon} \end{cases}$$

$$N_{L'_2}(L) = \begin{cases} 0 & \text{si } L < \frac{n_f + 1/3}{n_{max}} L_{min} + \frac{4(l_{gene}-1)}{3} \\ \Delta n & \text{si } L > \frac{n_f + 1/3}{n_{min}} L_{min} + \frac{4(l_{gene}-1)}{3} \\ n_{max} - \left\lceil \frac{n_f + 1/3}{L - \frac{4(l_{gene}-1)}{3}} L_{min} \right\rceil + 1 & \text{sinon} \end{cases}$$

$$N_{L'_3}(L) = \begin{cases} 0 & \text{si } L < \frac{n_f - 1/3}{n_{max}} L_{max} + \frac{2(l_{gene}-1)}{3} \\ \Delta n & \text{si } L > \frac{n_f - 1/3}{n_{min}} L_{max} + \frac{2(l_{gene}-1)}{3} \\ n_{max} - \left\lceil \frac{n_f - 1/3}{L - \frac{2(l_{gene}-1)}{3}} L_{max} \right\rceil + 1 & \text{sinon} \end{cases}$$

$$N_{L'_4}(L) = \begin{cases} 0 & \text{si } L < \frac{n_f+1/3}{n_{max}} L_{max} + \frac{4(l_{gene}-1)}{3} \\ \Delta n & \text{si } L > \frac{n_f+1/3}{n_{min}} L_{max} + \frac{4(l_{gene}-1)}{3} \\ n_{max} - \left\lceil \frac{n_f+1/3}{L - \frac{4(l_{gene}-1)}{3}} L_{max} \right\rceil + 1 & \text{sinon} \end{cases}$$

Pour calculer les probabilités agrégées, il faut intégrer ces fonctions entre deux points $a$ et $b$ qui dépendent de la subdivision finale visée. En effet, si on intègre le profil moyen donné en section 3.4.2 du chapitre III, on obtient

$$\int_a^b C(L)dL = \frac{h}{2\Delta n}\left(\int_a^b N_{L'_1}(L)dL + \int_a^b N_{L'_2}(L)dL - \int_a^b N_{L'_3}(L)dL - \int_a^b N_{L'_4}(L)dL\right)$$

Sur la figure A.12, nous illustrons ce que vaut l'une des intégrales à calculer. L'aire se calcule facilement en découpant des tranches le long de l'axe des $y$, puisque les fonctions sont des fonctions de comptage : elles sont en escalier avec des marches de hauteur 1.

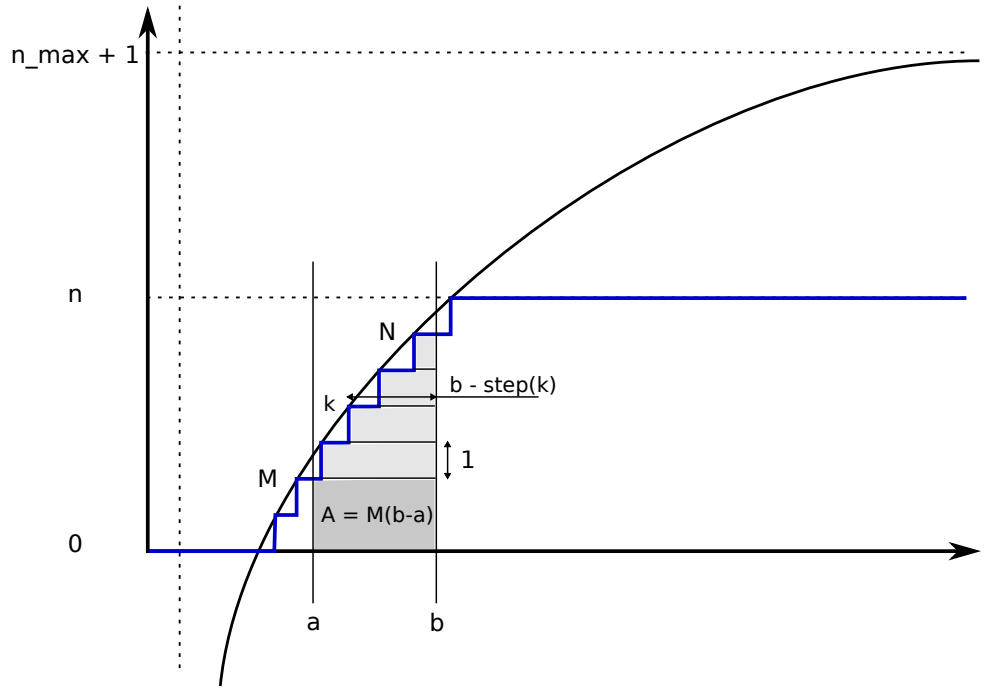


Figure A.12 – Intégration du nombre de profils qui ont dépassé l'un des points caractéristiques : le calcul de l'aire se fait facilement en découpant des tranches le long de l'axe $y$.

Il ne nous reste donc qu'à trouver la largeur des marches en déterminant la position où débute chaque marche. Par exemple, pour le point $L'_1$, la marche de hauteur $k$ commence quand

$$k = n_{max} - \frac{n_f - 1/3}{L - \frac{2(l_{gene}-1)}{3}} L_{min} + 1$$

$$\frac{n_f - 1/3}{L - \frac{2(l_{gene}-1)}{3}} L_{min} = n_{max} + 1 - k$$

$$L = \frac{n_f - 1/3}{n_{max} + 1 - k} L_{min} + \frac{2(l_{gene} - 1)}{3} = \text{step}_1(k)$$

On peut obtenir des expressions similaires pour les autres points caractéristiques.

Soit $M_1 = N_{L'_1}(a)$ and $N_1 = N_{L'_1}(b)$. Grâce à la figure A.12, on s'assure facilement que

$$\begin{aligned}
\int_a^b N_{L'_1}(L)dL &= M_1(b-a) + \sum_{k=M_1+1}^{N_1} (b - \text{step}_1(k)) \\
&= M_1(b-a) + (N_1 - M_1)b - \sum_{k=M_1+1}^{N_1} \text{step}_1(k) \\
&= N_1 b - M_1 a - (N_1 - M_1)\frac{2(l_{gene}-1)}{3} - (n_f - 1/3)L_{min} \sum_{k=M_1+1}^{N_1} \frac{1}{n_{max}+1-k} \\
&= N_1 b - M_1 a - \frac{2(N_1 - M_1)}{3}(l_{gene} - 1) - (n_f - 1/3)L_{min} \sum_{K=n_{max}+1-N_1}^{n_{max}-M_1} \frac{1}{K}
\end{aligned}$$

De même, on montre que

$$\int_a^b N_{L'_2}(L)dL = N_2 b - M_2 a - \frac{4(N_2 - M_2)}{3}(l_{gene} - 1) - (n_f + 1/3)L_{min} \sum_{K=n_{max}+1-N_2}^{n_{max}-M_2} \frac{1}{K}$$

$$\int_a^b N_{L'_3} dL = N_3 b - M_3 a - \frac{2(N_3 - M_3)}{3}(l_{gene} - 1) - (n_f - 1/3)L_{max} \sum_{K=n_{max}+1-N_3}^{n_{max}-M_3} \frac{1}{K}$$

$$\int_a^b N_{L'_4} dL = N_4 b - M_4 a - \frac{2(N_4 - M_4)}{3}(l_{gene} - 1) - (n_f + 1/3)L_{max} \sum_{K=n_{max}+1-N_4}^{n_{max}-M_4} \frac{1}{K}$$

Finalement, la contribution entre $a$ et $b$ vaut

$$\begin{aligned}
\int_a^b C(L)dL = &\frac{h}{2\Delta n}\Bigg[(N_1 + N_2 - N_3 - N_4)b - (M_1 + M_2 - M_3 - M_4)a \\
&- ((N_1 - M_1) + 2(N_2 - M_2) - (N_3 - M_3) - 2(N_4 - M_4))\frac{2(l_{gene}-1)}{3} \\
&- (n_f - 1/3)\left(L_{min} \sum_{K=n_{max}+1-N_1}^{n_{max}-M_1} \frac{1}{K} - L_{max} \sum_{K=n_{max}+1-N_3}^{n_{max}-M_3} \frac{1}{K}\right) \\
&- (n_f + 1/3)\left(L_{min} \sum_{K=n_{max}+1-N_2}^{n_{max}-M_2} \frac{1}{K} - L_{max} \sum_{K=n_{max}+1-N_4}^{n_{max}-M_4} \frac{1}{K}\right)\Bigg]
\end{aligned}$$

Si on fait le lien avec la section principale (voir équation (III.5), page 122 et conclusion de la section 3.4.2, 128), cette expression nous permet de calculer

$$\begin{aligned}
&\Pr\left[(x_0, y_0) \to (x_f, n_f)\right] \\
&= \sum_{L_f=L_{min}(x_f)}^{L_{max}(x_f)} \frac{1}{\Delta n(y_0)} \sum_{n_0=n_{min}(y_0)}^{n_{max}(y_0)} \frac{1}{\Delta L(x_0)} \sum_{L_0=L_{min}(x_0)}^{L_{max}(x_0)} \Pr\left[(L_0, n_0) \to (L_f, n_f)\right] \\
&= \int_{L_{min}(x_f)}^{L_{max}(x_f} C(L) dL
\end{aligned}$$

Au passage, on peut calculer l'aire totale du profil donné par $C(L)$ sur la ligne $n_f$. Dans le calcul de $C(L)$, on somme $\Delta L(x_0)$ densités d'aire $1/n_0$. En prenant en compte les termes de renormalisations, l'aire vaut

$$\frac{1}{\Delta L(x_0)\Delta n(y_0)} \sum_{n_0=n_{min}}^{n_{max}} \frac{\Delta L(x_0)}{n_0} = \frac{1}{\Delta n(y_0)} \sum_{n_0=n_{min}}^{n_{max}} \frac{1}{n_0}$$

### 3.3.8 Calcul de l'agrégation des transitions en pratique

Dans cette section, nous regardons comment utiliser l'algorithme RL en pratique. L'accent est mis sur la subdivision de départ $(x_0, y_0)$ et sur la ligne finale $n_f$. Comme précédemment on fait ici abstraction de la subdivision d'arrivée : quand on connaît la densité pour $n_f$ en entier, il suffit d'intégrer entre $L_{min}(x_f)$ et $L_{max}(x_f)$.

**Condition pour utiliser l'algorithme RL** Pour utiliser l'algorithme RL, il faut que les deux triangles extrêmes provenant de la ligne $n_0$ ne se chevauchent pas, soit $L_3 > L_4$. Il faut donc déterminer pour quelle ligne finale la ligne $n_0$ participe à l'algorithme RL ou, réciproquement, quelles lignes de départ peuvent être utilisées pour calculer la contribution vers une ligne finale $n_f$. Rappelons que

$$L_3 = \frac{n_f + 1}{n_0} L_{min} + 2l_{gene} - 1$$

et

$$L_4 = \frac{n_f - 1}{n_0} L_{max} - 1$$

Ainsi, la densité de la quantité de non-codant finale sur la ligne $n_f$ (sachant qu'on est parti de $n_0$ gènes avec entre $L_{min}$ et $L_{max}$ bp non codantes) comporte un plateau si et seulement si

$$L_3 \leq L_4 \qquad \Leftrightarrow \qquad \frac{n_f + 1}{n_0} L_{min} + 2l_{gene} - 1 \leq \frac{n_f - 1}{n_0} L_{max} - 1$$

$$\Leftrightarrow \qquad \frac{L_{max}}{n_0} + \frac{L_{max}}{n_0} + 2l_{gene} \leq n_f \left( \frac{L_{max}}{n_0} - \frac{L_{max}}{n_0} \right)$$

$$\Leftrightarrow \qquad n_f \geq \frac{L_{max} + L_{min} + 2n_0 l_{gene}}{L_{max} - L_{min}}$$

Soit $a = (L_{max} - L_{min})/(2l_{gene})$ et $b = (L_{max} + L_{min})/(2l_{gene})$. Nous pouvons utiliser l'algorithme RL pour toutes les lignes de départ $n_0$ vérifiant

$$\boxed{n_f \geq \frac{b + n_0}{a} \qquad \Leftrightarrow \qquad n_0 \leq a n_f - b}$$

**Utilisation optimale de l'algorithme RL pour agréger les délétions** Pour utiliser l'algorithme RL de manière optimale, il faut voir combien de lignes de $(x_0, y_0)$ peuvent y participer. Disons que pour que l'algorithme RL soit préféré à une sommation naïve via les algorithmes LL, il faut qu'il y ait au moins 100 lignes qui participent à l'algorithme RL. Si on connaît les conditions *a priori*, on gagne du temps à l'implémentation.

**Condition pour que l'algorithme RL s'applique avec toutes les lignes d'une subdivision de départ** $(x_0, y_0)$ Si $n_f \geq (b + n_{max}(y_0))/a$, toutes les lignes $n_0$ vérifient $n_f \geq (b + n_0)/a$. Il existe alors une plus petite ligne finale pour laquelle l'algorithme RL peut être appliqué, qu'on nomme $n_f^{whole} = \lceil (b + n_{max}(y_0))/a \rceil$. L'algorithme RL peut donc être utilisé pour tout $(x_0, y_0)$ pour $n_f^{whole} \leq n_f \leq n_{min}(y_0) - 1$ ($n_{min}$ n'est pas éligible car le profil venant de $n_{min}$ est tronqué). Si $n_f^{whole} \geq n_{min}(y_0)$, l'algorithme RL ne peut pas être utilisé avec le rectangle de départ en entier.

**Condition pour utiliser RL avec une partie du rectangle de départ quand la ligne finale est en dehors de** $(x_0, y_0)$ D'après la condition pour le plateau, on voit que si $n_0$ ne remplit pas la condition pour un certain $n_f$ alors toutes les lignes situées au-dessus de $n_0$ ne la remplissent pas non plus. Si on veut qu'au moins 100 lignes remplissent la condition, il faut donc que $n_f \geq (b + n_{min} + 100 - 1)/a$.

**Condition pour utiliser RL avec une partie du rectangle de départ quand la ligne finale est à l'intérieur de** $(x, y)$ Dans ce cas, aucune des lignes $n_0 \leq n_f$ ne participe à l'agrégation. Pour les autres, il faut que $n_0 \leq a n_f - b$. La ligne la plus haute éligible est donc $\lfloor a n_f - b \rfloor$ et celles en dessous remplissent la condition de plateau. Au plus, il y a alors $\lfloor a n_f - b \rfloor - (n_f + 1) + 1 = \lfloor (a-1) n_f - b \rfloor$ lignes éligibles. Si $a < 1$ on ne pourra jamais utiliser l'algorithme RL, sinon on peut l'utiliser à condition que

$$n_f \geq \frac{100 + b}{a - 1} \qquad \text{et} \qquad n_f \leq n_{max} - 100 + 1$$

La première condition garantit qu'il y a au moins 100 lignes au-dessus de $n_f$ qui peuvent participer. Comme on l'a dit plus haut, toutes les lignes sous $\lfloor a n_f - b \rfloor$ sont éligibles, à

commencer par $n_f + 1$. La deuxième condition vérifie que parmi ces lignes au moins 100 sont à l'intérieur du rectangle de départ. Plus simplement, si on veut qu'au moins 100 lignes participent à l'agrégation, cela ne sert à rien de tenter sa chance avec un $n_f$ pour lequel il y a moins de 100 lignes de départ au-dessus.

**Bilan : calcul des transitions entre la subdivision $(x_0, y_0)$ et l'ensemble des lignes d'arrivées possibles $n_f$** La partie difficile avec toutes ces conditions est de répertorier correctement toutes les contributions pour lesquels il faut utiliser les algorithmes ligne à ligne. Dans le programme, on boucle sur la ligne finale. Pour chaque nombre de gènes final, on va considérer les lignes de la subdivision (méta-état) de départ $(x_0, y_0)$ et calculer la fonction de densité de la quantité de non-codant, en utilisant l'algorithme RL quand c'est possible, ou l'algorithme LL sinon. Cette fonction de densité, évaluée à chaque limite verticale de subdivision, nous permet de calculer la contribution de la subdivision de départ à toutes les subdivisions d'arrivée qui correspondent à $n_f$.

- $n_f < (b + n_{min}(y_0) + 100 - 1)/a$ : on ne peut pas utiliser l'algorithme RL du tout.
- $(b + n_{min}(y_0) + 100 - 1)/a \leq n_f < \min\{n_f^{whole}, n_{min}(y_0)\}$ : on utilise RL pour les lignes de départ $n_{min}(y_0) \leq n_0 \leq an_f - b$ et LL pour les autres.
- $n_f^{whole} \leq n_f < n_{min}(y_0)$ : on utilise RL pour toutes les lignes de départ contenues dans $(x_0, y_0)$.
- $n_{min}(y_0) \leq n_f \leq n_{max}(y_0) - 100$ : si $n_f \geq (100 + b)/(a - 1)$, on utilise LL pour $n_0 = n_f$ et RL pour les lignes au-dessus, sinon on utilise LL pour toutes les lignes de départ.
- $n_f \geq n_{max} - 100 + 1$ : on utilise LL pour toutes les lignes de départ contenues dans $(x_0, y_0)$.

Dans le programme, il faut répéter cette boucle pour toutes les subdivisions $(x_0, y_0)$ de départ possibles

## 3.4 Agrégation des duplications

### 3.4.1 Agrégation ligne vers ligne, cas où le génome initial ne contient pas de gène (LL_empty)

S'il n'y a pas de gène initialement, le génome $(L_0, 0)$ contiendra après duplication un nombre de bases non codantes distribuées uniformément entre $L_0 + 1$ et $2L_0$. Il s'agit ici d'agréger ces distributions pour $L_0$ compris entre $L_{min}$ et $L_{max}$, correspondant à la ligne du bas de la subdivsion $(x_0, 1)$. Soit $p_1 = 1/\Delta L(x_0).\sum_{L_{min} \leq L \leq L_{max}} 1/L$. Si $x_0 = 1$, on a un

cas particulier car $2L_{min} = 0 < L_{max}$. Ce cas sera traité à la fin. Si $x_0 > 1$, $L_{max} \leq 2L_{min}$, la transition moyenne allant vers $L$ vaut

- $\forall L \in [L_{min}+1, L_{max}]$,

$$f(L) = \frac{1}{\Delta L(x_0)} \sum_{L_{min} \leq L \leq L-1} 1/k = p_1 - \frac{1}{\Delta L(x_0)} \sum_{L \leq k \leq L_{max}} 1/k.$$

- $\forall L \in [L_{max}+1, 2L_{min}]$, $f(L) = p_1$.
- $\forall L \in [2L_{min}+1, 2L_{max}]$, $f(L) = 1/\Delta L(x_0). \sum_{\lceil L/2 \rceil \leq k \leq L_{max}} 1/k$

La probabilité de rester sur la ligne du bas de la subdivision $(x_0, 1)$ sachant que l'état de départ était sur cette ligne vaut

$$p(x_0) = (\Delta L(x_0)-1)p_1 - \frac{1}{\Delta L(x_0)} \sum_{k=1}^{\Delta L-1} \sum_{L_{min}+k}^{L_{max}} \frac{1}{L} = (\Delta L(x_0)-1)p_1 - (1-L_{min}p_1) = L_{max}p_1 - 1$$

Le second terme a été simplifié en réarrangeant les termes de la somme.

En échelle logarithmique, les transitions qui ne restent pas en sur la ligne du bas de la case $(x_0, 1)$ vont sur la ligne du bas de la case $(x_0+1, 1)$, on a donc tout simplement $p(x_0+1) = 1 - p(x_0)$

En échelle linéaire, les supports des distributions uniformes pour chaque point de départ contiennent le segment de $L_{max}+1$ à $2L_{min}-1$, ce qui correspond aux subdivision $x = x_0+1$ jusqu'à $x = 2x_0 - 2$. Dans ces subdivisions, le profil moyen est constant, on a simplement $p(x) = \Delta L p_1$. Tout ce qui reste sont les contributions entre $2L_{min}$ et $2L_{max}$ qui sont réparties entre les subdivisions $2x_0 - 1$ et $2x_0$. Commençons par calculer la contribution pour ces deux subdivisions ensemble

$$p(2x_0-1) + p(2x_0) = p_1 + 2\frac{1}{\Delta L} \sum_{k=1}^{\Delta L-1} \sum_{L_{min}+k}^{L_{max}} \frac{1}{L} = p_1 + 2(1 - L_{min}p_1) = 2 - (2L_{min}-1)p_1$$

Le point de départ le plus petit qui contribue pour la subdivision $2x_0$ est $L_{mid} = L_{min} + \Delta L/2$ (car $L_{min}(2x_0) = (2x_0-1)\Delta L = 2(L_{min} + \Delta L/2)$) et il ne contribue que sur un point d'arrivée. Soit $p_{mid} = 1/\Delta L. \sum_{L_{mid} \leq L \leq L_{max}} 1/L$. Alors

$$p(2x_0) = p_{mid} + 2\frac{1}{\Delta L} \sum_{k=1}^{\Delta L/2-1} \sum_{L_{mid}+k}^{L_{max}} \frac{1}{L} = p_{mid} + 2\left(\frac{\Delta L/2}{\Delta L} - L_{mid}p_{mid}\right) = 1 - (2L_{mid}-1)p_{mid}$$

D'où

$$p(2x_0-1) = 1 + (2L_{mid}-1)p_{mid} - (2L_{min}-1)p_1$$

Il reste le cas $x_0 = 1$ (valable pour les deux échelles). Tous les points de départ sous $L_{mid} = \Delta L/2$ contribuent exclusivement à la subdivision de départ. Ensuite, comme dans le cas précédent, les transitions qui atteignent la subdivision suivante somment à

$$p(2) = 1 - (2L_{mid} - 1)p_{mid}$$

d'où on déduit $p(1) = 1 - p(2)$.

### 3.4.2 Calcul de l'agrégation des transitions en pratique

**Condition pour l'utilisation de l'algorithme RL** Il s'agit d'exactement la même condition que pour les délétions : l'existence du plateau.

**Utilisation optimale de l'algorithme RL pour agréger les duplications** Pour utiliser l'algorithme RL de manière optimale, il faut voir combien de lignes de $(x_0, y_0)$ peuvent y participer. Disons que pour que l'algorithme RL soit préféré à une sommation naïve via les algorithmes LL, il faut qu'il y ait au mois 100 lignes de départ qui participent à l'algorithme RL. Si on connaît les conditions *a priori*, on gagne du temps à l'implémentation.

**Condition pour que l'algorithme RL s'applique avec toutes les lignes de** $(x_0, y_0)$ Si $n_f \geq (b + n_{max}(y_0))/a$, toutes les lignes vérifient $n_f \geq (b + n_0)/a$. Soit $n_f^{whole} = \lceil (b + n_{max}(y_0))/a \rceil$. Si $n_f^{whole} \leq n_{max}(y_0)$, on pose $n_{whole} = n_{max}(y_0) + 1$. L'algorithme RL peut être utilisé pour toutes les lignes de départ de $(x_0, y_0)$ pour $n_f^{whole} \leq n_f \leq 2n_{min}(y_0) - 1$ ($2n_{min}(y_0)$ n'est pas éligible car le profil venant de $n_{min}(y_0)$ est tronqué). Si $n_f^{whole} \geq 2n_{min}(y_0)$, l'algorithme RL ne peut pas être utilisé avec le rectangle de départ en entier.

**Condition pour utiliser RL avec une partie du rectangle de départ** D'après la condition pour le plateau, on voit que si $n_0$ ne remplit pas la condition pour un certain $n_f$ alors toutes les lignes situées au-dessus de $n_0$ ne la remplissent pas non plus. Si on veut qu'au moins 100 lignes remplissent la condition, il faut donc que les lignes $n_{min}(y_0) \leq n_0 \leq n_{min}(y_0) + 100 - 1$ participent à l'algorithme. C'est le cas en théorie dès que $n_f \geq (b + n_{min}(y_0) + 100 - 1)/a$.

Si $n_f < 2n_{min}(y_0)$, RL fonctionnera dès que $n_f \geq (b + n_{min}(y_0) + 100 - 1)/a$ (condition de plateau pour les profils provenant de $n_{min}(y_0) \leq n_0 \leq n_{min}(y_0) + 100 - 1$) et $n_f \geq n_{min}(y_0) + 100$ (on s'assure que ces lignes ont bien le droit de participer à l'agrégation).

Si $n_f \geq 2n_{min}(y_0)$, les choses se compliquent un peu. En effet, à partir $n_f = 2n_{min}(y_0)$, certaines lignes de départ ne participent plus de fait car leur profil est tronqué ou inexistant, ce qui a été ignoré dans la condition de plateau. C'est le cas pour toutes les lignes

de départ inférieures à $\lfloor n_f/2 \rfloor + 1$. Il faut ici vérifier deux conditions. Si $a > 0.5$

$$\lfloor an_f - b \rfloor - (\lfloor n_f/2 \rfloor + 1) + 1 \geq 100 \quad \Leftarrow \quad n_f \geq \frac{100 + b + 0.5}{a - 0.5}$$

et

$$n_f - 1 - (\lfloor n_f/2 \rfloor + 1) + 1 \geq 100 \quad \Leftarrow \quad n_f \geq 2 \times 100 + 3$$

La première condition dit qu'au-dessus de $\lfloor n_f/2 \rfloor + 1$, il y a effectivement 100 lignes susceptibles de participer. La deuxième condition gère le cas particulier où $2n_{min}(y_0)$ est à l'intérieur du rectangle de départ : on vérifie que ces 100 lignes sont en dessous de $n_f$. Enfin, pour être sûr que les 100 lignes sont bien à l'intérieur du rectangle de départ, on se restreint à $n_f \leq 2(n_{max}(y_0) - 100 + 1)$.

Globalement, on peut alors utiliser RL entre $\min\{(b+n_{min}+100-1)/a, (100+b+0.5)/(a-0.5), 2100+3, n_{min}+100\}$ et $2(n_{max} - 100 + 1)$. D'après la condition de plateau, on voit que à chaque fois que $n_f$ augmente, le nombre de lignes participant augmente de $a$. Si $a > 0.5$, ce nombre s'accroît assez vite pour englober des lignes de départ de $(x_0, y_0)$ et même quand $n_f \geq 2n_{min}$, le nombre de lignes qui se mettent à participer compensent celles qui s'arrêtent.

**Bilan : calcul des contributions** La partie difficile avec toutes ces conditions est de répertorier correctement toutes les transitions pour lesquels il faut utiliser les algorithmes ligne à ligne. Pour chaque nombre de gènes final, on va considérer les lignes de la subdivision (méta-état) de départ $(x_0, y_0)$ et calculer la fonction de densité de la quantité de non-codant, en utilisant l'algorithme RL quand c'est possible, ou l'algorithme LL sinon. Cette fonction de densité, évaluée à chaque limite verticale de subdivision, nous permet de calculer la contribution de la subdivision de départ à toutes les subdivisions d'arrivée qui correspondent à $n_f$.

- si $a \leq 0.5$ : on ne peut pas utiliser l'algorithme RL du tout.
- $n_f < n_{min}(y_0) + 100$ : on ne peut pas utiliser l'algorithme RL du tout.
- $n_{min}(y_0) + 100 \leq n_f < 2n_{min}(y_0)$ :
  - si $n_f < (b + n_{min}(y_0) + 100 - 1)/a$ : on ne peut pas utiliser RL du tout.
  - si $(b + n_{min}(y_0) + 100 - 1)/a \leq n_f < n_{whole}$ : RL pour $n_{min}(y_0) \leq n_0 \leq an_f - b$ et LL pour les autres lignes de départ de $(x_0, y_0)$.
  - si $n_f \geq n_f^{whole}$ : RL pour toutes les lignes de départ de $(x_0, y_0)$.
- $2n_{min}(y_0) \leq n_f \leq 2(n_{max}(y_0) - 100 + 1)$ :
  - si $n_f < \max\{(100 + b + 0.5)/(a - 0.5), 2 \times 100 + 3\}$ : on ne peut pas utiliser RL du tout.

  - si $n_f \geq \max\{(100 + b + 0.5)/(a - 0.5), 2 \times 100 + 3\}$ : on utilise RL pour $\lfloor n_f/2 \rfloor + 1 \leq n_0 \leq \min\{n_{max}(y_0), an_f - b, n_f - 1\}$ et LL pour les autres lignes de départ de $(x_0, y_0)$.

- $n_f > 2(n_{max}(y_0) - 100 + 1)$ : on ne peut pas utiliser RL du tout.

Remarque : comme mentionné plus haut, si la condition RL est remplie pour $n_f < 2n_{min}(y_0)$, elle l'est automatiquement pour $n_f \geq 2n_{min}(y_0)$, ce qui permet d'éviter de faire un test à ce moment-là.

Dans le programme, il faut répéter cette boucle pour toutes les subdivisions $(x_0, y_0)$ de départ possibles.