Dynamique des génomes bactériens : Une étude expérimentale in silico avec la plateforme Aevol

Marco Foley

N°d’ordre NNT : 2023ISAL0130

**THESE de DOCTORAT DE L’INSA LYON,
membre de l’Université de Lyon**

**Ecole Doctorale InfoMaths
ED 512**

Informatique et Application

Soutenue publiquement le 19/12/2023 par :

Marco Foley

Dynamique des génomes bactériens : une étude expérimentale in silico avec la plateforme Aevol

---

Devant le jury composé de :

Lafontaine, Ingrid	Professeur des Universités	Sorbonne Université	Examinatrice
Bredeche, Nicolas	Professeur des Universités	Sorbonne Université	Rapporteur
Junier, Ivan	Directeur de recherche	Université Grenoble Alpes	Rapporteur
Beslon, Guillaume	Professeur des Universités	INSA-LYON	Directeur de thèse
Rouzaud-Cornabas, Jonathan	Maître de conférences	INSA-LYON	Co-encadrant

INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Département de la Formation par la Recherche et des Études Doctorales (**FEDORA**)

Bâtiment INSA direction, 1er étage
37, av. J. Capelle
69621 Villeurbanne Cédex
fedora@insa-lyon.fr

Référence : TH1055_FOLEY Marco

L'INSA Lyon a mis en place une procédure de contrôle systématique via un outil de détection de similitudes (logiciel Compilatio). Après le dépôt du manuscrit de thèse, celui-ci est analysé par l'outil. Pour tout taux de similarité supérieur à 10%, le manuscrit est vérifié par l'équipe de FEDORA. Il s'agit notamment d'exclure les auto-citations, à condition qu'elles soient correctement référencées avec citation expresse dans le manuscrit.

Par ce document, il est attesté que ce manuscrit, dans la forme communiquée par la personne doctorante à l'INSA Lyon, satisfait aux exigences de l'Etablissement concernant le taux maximal de similitude admissible.

INSA LYON
Campus LyonTech La Doua
20, avenue Albert Einstein - 69621 Villeurbanne cedex - France
Tél. +33 (0)4 72 43 83 83 - Fax +33 (0)4 72 43 85 00
www.insa-lyon.fr

membre de UNIVERSITÉ DE LYON

Département FEDORA – INSA Lyon - Ecoles Doctorales

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
ED 206 CHIMIE	**CHIMIE DE LYON** https://www.edchimie-lyon.fr Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage secretariat@edchimie-lyon.fr	**M. Stéphane DANIELE** C2P2-CPE LYON-UMR 5265 Bâtiment F308, BP 2077 43 Boulevard du 11 novembre 1918 69616 Villeurbanne directeur@edchimie-lyon.fr
ED 341 E2M2	**ÉVOLUTION, ÉCOSYSTÈME, MICROBIOLOGIE, MODÉLISATION** http://e2m2.universite-lyon.fr Sec. : Bénédicte LANZA Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.e2m2@univ-lyon1.fr	**Mme Sandrine CHARLES** Université Claude Bernard Lyon 1 UFR Biosciences Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69622 Villeurbanne CEDEX e2m2.codir@listes.univ-lyon1.fr
ED 205 EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTÉ** http://ediss.universite-lyon.fr Sec. : Bénédicte LANZA Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.ediss@univ-lyon1.fr	**Mme Sylvie RICARD-BLUM** Laboratoire ICBMS - UMR 5246 CNRS - Université Lyon 1 Bâtiment Raulin - 2ème étage Nord 43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex Tél : +33(0)4 72 44 82 32 sylvie.ricard-blum@univ-lyon1.fr
ED 34 EDML	**MATÉRIAUX DE LYON** http://ed34.universite-lyon.fr Sec. : Yann DE ORDENANA Tél : 04.72.18.62.44 yann.de-ordenana@ec-lyon.fr	**M. Stéphane BENAYOUN** Ecole Centrale de Lyon Laboratoire LTDS 36 avenue Guy de Collongue 69134 Ecully CEDEX Tél : 04.72.18.64.37 stephane.benayoun@ec-lyon.fr
ED 160 EEA	**ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE** https://edeea.universite-lyon.fr Sec. : Philomène TRECOURT Bâtiment Direction INSA Lyon Tél : 04.72.43.71.70 secretariat.edeea@insa-lyon.fr	**M. Philippe DELACHARTRE** INSA LYON Laboratoire CREATIS Bâtiment Blaise Pascal, 7 avenue Jean Capelle 69621 Villeurbanne CEDEX Tél : 04.72.43.88.63 philippe.delachartre@insa-lyon.fr
ED 512 INFOMATHS	**INFORMATIQUE ET MATHÉMATIQUES** http://edinfomaths.universite-lyon.fr Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 infomaths@univ-lyon1.fr	**M. Hamamache KHEDDOUCI** Université Claude Bernard Lyon 1 Bât. Nautibus 43, Boulevard du 11 novembre 1918 69 622 Villeurbanne Cedex France Tél : 04.72.44.83.69 direction.infomaths@listes.univ-lyon1.fr
ED 162 MEGA	**MÉCANIQUE, ÉNERGÉTIQUE, GÉNIE CIVIL, ACOUSTIQUE** http://edmega.universite-lyon.fr Sec. : Philomène TRECOURT Tél : 04.72.43.71.70 Bâtiment Direction INSA Lyon mega@insa-lyon.fr	**M. Jocelyn BONJOUR** INSA Lyon Laboratoire CETHIL Bâtiment Sadi-Carnot 9, rue de la Physique 69621 Villeurbanne CEDEX jocelyn.bonjour@insa-lyon.fr
ED 483 ScSo	**ScSo**[1] https://edsciencessociales.universite-lyon.fr Sec. : Mélina FAVETON Tél : 04.78.69.77.79 melina.faveton@univ-lyon2.fr	**M. Bruno MILLY** (INSA : J.Y. TOUSSAINT) Univ. Lyon 2 Campus Berges du Rhône 18, quai Claude Bernard 69365 LYON CEDEX 07 Bureau BEL 319 bruno.milly@univ-lyon2.fr

[1] ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Sciences politiques, Sociologie, Anthropologie

Remerciements

Avant toute chose, je tiens à remercier toutes les personnes sans qui ce travail de thèse n'aurait pas été possible. J'adresse mes plus sincères remerciements...

... à Nicolas Bredèche et Ivan Junier pour avoir accepté d'être mes rapporteurs et pour vos remarques constructives sur ce manuscrit,

... à Philippe Lopez pour avoir également accepté d'être rapporteur même si les règles administratives en ont décidé autrement,

... à Ingrid Lafontaine pour avoir accepté de participer à mon jury,

... à tous les membres de l'équipe Beagle pour veiller collectivement à ce que notre recherche soit une aventure humaine avant tout,

... à mes directeurs de thèse Guillaume et Jonathan, pour votre présence, pour m'avoir guidé dans les moments d'égarement, pour votre soutien indéfectible dans les moments difficiles et pour m'avoir initié à la grimpe,

... à Paul, Julie, Laurent, Théotime, Lisa, Lisa, Ju', Nathan, Romain et tous les autres doctorants de l'antenne pour tous ces merveilleux moments passés ensemble, au labo comme en dehors,

... à ma famille pour son soutien moral et matériel sans lequel rien de ceci n'aurait existé,

... à Angelo, Maria, Adrien et Noé pour avoir été mon havre de paix ici à Lyon,

... à Antoine pour ces quatre années vécues ensemble,

... à Pierre et Arthur, mes amis de toujours,

... à tous les autres, que je n'ai pas cités ici mais qui sauront se reconnaître.

Table des matières

Table des figures

Chapitre I

Introduction

Vu très globalement, le sujet de cette thèse est d'utiliser le modèle Aevol pour étudier l'évolution des génomes bactériens par une approche de simulation computationnelle. Aevol est une plateforme de simulation de vie artificielle conçue pour étudier l'évolution de la structure des génomes. Parmi les modèles utilisés en biologie, Aevol occupe une place singulière de part sa conception qui adopte le point de vue de la vie artificielle. De ce fait, les racines épistémologiques d'Aevol sont plus proches des modèles utilisés par les informaticiens que ceux classiquement utilisés par les biologistes. D'un point de vue conceptuel, Aevol hérite donc des modèles de type automate-cellulaire, conçus pour étudier les propriétés émergentes, où l'objet d'étude est isolé et uniquement gouverné par les règles du modèle (Symons, 2008). Ce qui est, en clair, contraste avec les sciences du vivant qui se reposent principalement sur des modèles quantitatifs et prédictifs, comme des systèmes d'équations différentielles ou des modèles statistiques, dans lesquels sont injectés des données issues d'expériences (Chen *et al.*, 2005; Dougherty et Braga-Neto, 2006).

Depuis sa conception, Aevol a pu s'affirmer comme une plateforme offrant une approche computationnelle très pertinente pour étudier l'évolution en biologie. Ceci, en permettant de nombreuses études améliorant la compréhension de différents phénomènes évolutifs, particulièrement en génomique. Comme, par exemple l'effet des taux de mutation sur l'évolution de la taille des génomes (Knibbe *et al.*, 2007), ou encore l'évolution de la complexité en biologie (Liard *et al.*, 2020; Beslon *et al.*, 2021). Malgré cela, les résultats scientifiques apportés grâce à Aevol peinent encore aujourd'hui à diffuser dans la communauté de la biologie évolutive. Selon nous, ce manque de rayonnement est principalement dû à deux limitations du modèle qui rendent difficile le comblement du faussé épistémologique séparant Aevol des modèles plus classiquement utilisés en biologie. Ces deux limites sont les suivantes :

- Premièrement, jusqu'à très récemment, les temps de calculs étaient trop longs dans Aevol. Même si cela n'a pas empêché le modèle de fournir des résultats souvent très intéressant , cela a limité l'étendue des simulations classiquement effectuées avec Aevol, aussi bien en termes de temps d'évolution qu'en nombre de réplicats. Or, les processus d'évolution sont des processus longs nécessitant de nombreuses générations pour atteindre un état d'équilibre et aussi très variables car reposant sur une succession d'évènements aléatoires. Les expériences n'avaient donc jusqu'à présent pas l'envergure nécessaire pour permettre une exploration paramétrique large, mais

surtout pour atteindre des états d'équilibre permettant d'isoler les différentes forces évolutives en présence.

- Deuxièmement, bien que les génomes des organismes simulés dans Aevol soient réalistes à l'échelle structurelle (l'organisation du chromosome en séquences transcrites puis traduites), dans le modèle la séquence de bases composant ces génomes est une séquence binaire, c'est-à-dire un enchainement de 0/1 au lieu des bases ACTG des génomes biologiques. Ce choix de bases binaires est en parti-pris historique qui répondait aux contraintes de simplicité, nécessaires à toute modélisation, et de coûts calculatoires, nécessaires à toute simulation. Ces choix remontent donc à la création du modèle, il y a maintenant plus de 15 ans. Si, du point de vue du modélisateur, ces contraintes correspondent à une idéalisation parfaitement légitime des génomes « réels » — puisque le modèle est destiné à étudier l'évolution de la structure des génomes et non leur séquence, il a eu une conséquence négative forte en créant des difficultés de communications avec les biologistes de l'évolution, et avec ceux la biologie moléculaire, limitant là encore la porté des résultats obtenus grâce au modèle, mais aussi la diversité des questions que ce modèle pouvait aborder.

Dans cette thèse, nous visons à adresser chacun des deux problèmes soulevés plus haut. Pour cela cette thèse est divisée en trois chapitres.

Le chapitre II présente le modèle Aevol, préambule indispensable à la compréhension des chapitres suivants. Nous présenterons dans un premier temps comment de l'approche mécanistique de la conception du modèle découle une genotype-to-phenotype map réaliste. C'est ce principe qui permet aux organismes dans Aevol d'avoir des contraintes évolutives comparables à celle de l'évolution en biologie (du moins à l'échelle de la structure du génome). Ensuite nous présenterons le modèle de population, le processus de sélection et les opérateurs génétiques de mutation. Ces derniers correspondent à l'ensemble des opérateurs de variation permettant l'exploration du fitness landscape. Ils sont eux aussi comparables, en diversité, à ceux opérant sur le génome d'organismes biologiques « réels , incluant en particulier une relativement large gamme d'opérateurs de variation structurelle, les réarrangements chromosomiques. Nous avons déjà brièvement évoqué la « genotype-to-phenotype map . C'est ce deuxième niveau de réalisme, celui des opérateurs de variation, qui permet d'étudier avec Aevol les différentes forces évolutives opérant à l'échelle de la structure des génomes. Nous présenterons enfin la démarche classique d'usage d'Aevol, c'est-à-dire les types de protocoles expérimentaux classiquement utilisés ainsi que les traitement de données post-évolution permettant d'identifier les forces en présence. Enfin, nous ferons un historique – non-exhaustif – des principales études menées sur Aevol permettant d'illustrer par l'exemple la pertinence du modèle pour l'étude de l'évolution en biologie.

Dans le chapitre III nous profiterons des améliorations récentes des performances de calculs et de la parallélisation du modèle pour répondre à la première des limitations exprimée plus haut. Les études conduites jusqu'ici avec Aevol – et présentées dans le chapitre II pour les principales d'entre elles, concourent à suggérer l'existence de mécanismes de régulation pour la partie non-codante des génomes, et ce, contre toute attente puisque dans Aevol, les conditions d'évolution de ces séquences sont telles (neutralité parfaite, pas de biais mutationnels) qu'elles devraient être gouvernées exclusivement par la dérive. Pour confirmer l'existence de cette régulation et en identifier l'origine, nous avons effectué des

simulations au long court – 14 millions de générations – et avec un très large échantillon statistique – 250 réplicats – afin de permettre une observation fine de la dynamique du non-codant. Cette étude constitue à ce jour la plus large expérience conduite avec Aevol.

Après avoir introduit le contexte biologique et les différentes hypothèses classiquement formulées pour expliquer l'existence et la régulation potentielle de la quantité d'ADN non-codant dans les génomes, nous montrerons que la quantité de non-codant dans Aevol n'est jamais nulle et évolue dans un espace borné. Ce résultat n'est pas attendu étant donné que les séquences non-codantes sont totalement non fonctionnelles dans le modèle. Ensuite grâce à une combinaison de méthodes incluant des expériences modifiant la taille des populations, des expériences d'accumulation de mutations neutres, de l'ingénierie de génomes artificiels, des quantifications de la distribution des effets de fitness par des méthodes de Monte-Carlo et un modèle mathématique de la robustesse aux mutations, nous montrerons que l'espace borné dans lequel évolue les séquences non-codantes est le résultat d'une régulation par deux forces contraires. La première de ces forces est une sélection pour des génomes réduits du fait d'une robustesse réplicative plus favorable. La seconde est un biais dans les probabilités d'occurrence neutre entre grandes duplications et grandes délétions, menant à l'accumulation de non-codant par dérive génétique. Enfin, nous recontextualiserons nos résultats avec ceux de la littérature dans le domaine. Bien qu'encore en chantier, ce chapitre reprend et développe un article en préparation [1], raison pour laquelle il respecte, autant que faire ce peu pour un chapitre de thèse, la structure classique d'un article en biologie évolutive et moléculaire. Enfin, dans le chapitre IV nous présenterons une nouvelle version du modèle Aevol, "Aevol_4b", utilisant des séquences génétiques à quatre bases plutôt qu'un génome binaire. Ce chapitre adresse donc directement la deuxième limite exprimée plus haut. Ces travaux ayant été réalisés dans le contexte d'un projet ANR visant à produire des données de benchmarks pour les méthodes en évolution moléculaire, nous commencerons par présenter des travaux fondateurs ayant utilisé Aevol en tant que tel. Puis nous contextualiserons la validation des méthodes en évolution moléculaire. Ensuite, après avoir présenté les difficultés que représente un modèle à quatre bases, nous décrirons les différentes la démarche et les étapes de développement d'Aevol_4b. En particulier l'intégration du code génétique standard de la biologie. Enfin nous présenterons deux jeux de données de complexité croissante que nous avons produits avec le nouveau modèle et montrerons comment, en utilisant des séquences génétiques ayant divergé, dans Aevol, à partir d'un ancêtre commun, il est possible de reconstruire très fidèlement l'arbre d'espèces sur lequel se sont produites les divergences en utilisant des outils de biologie moléculaires classiques. Ce résultat, qui, à notre connaissance, constitue la première expérience de phylogénie sur organismes artificiels, permet une validation partielle, mais enthousiasmante et très encourageante, de Aevol_4b.

1. Marco Foley, Victor Lezaud, Paul Banse, Jonathan Rouzaud-Cornabas, Guillaume Beslon (2023) The Danaïds genome — Duplications/deletions neutrality bias maintains non-coding sequences in genomes despite constant leak, *In prep.*

Chapitre II

Aevol, un modèle pour étudier l'évolution par approches expérimentales *in silico*

Note : Ce chapitre reprend plusieurs éléments issus d'un article en préparation : Paul Banse, Juliette Luiselli, David P. Parsons, Théotime Grohens, Marco Foley, Leonardo Trujillo, Jonathan Rouzaud-Cornabas et Guillaume Beslon (2023) *Forward-in-time simulation of chromosomal rearrangements : The invisible backbone that sustains long-term adaptation*. L'article est actuellement en révision pour publication dans le journal *Molecular Ecology* (dans le cadre du numéro spécial "A genomic update on the evolutionary impact of chromosomal rearrangements").

1 Introduction

Aevol (https://www.aevol.fr) est une plate-forme d'évolution expérimentale *in silico* développée depuis 2005 par l'équipe INRIA/LIRIS Beagle à Lyon. La plate-forme simule l'évolution d'une population d'organismes haploïdes par un processus de variation et de sélection (Knibbe *et al.*, 2007; Frénoy *et al.*, 2013; Batut *et al.*, 2013; Parsons *et al.*, 2010). Initialement, la plate-forme était destinée à étudier l'évolution de la structure des génomes bactériens (en particulier les structures opéroniques). De ce fait, la conception du modèle a été axée sur le réalisme de la structure du génome, mais aussi sur le réalisme des processus de mutation susceptibles de modifier cette structure. Nous verrons (section 5) que l'étude du génome a effectivement été au cœur de nombreux travaux utilisant Aevol. Cependant, on note aussi que, devant le regain d'intérêt récent pour l'étude des variants structurels, Aevol peut aussi être utilisé pour déchiffrer l'effet des réarrangements chromosomiques sur l'évolution du génome, y compris leurs interactions avec d'autres types d'événements mutationnels Banse *et al.* (2023).

En bref, Aevol est composé de trois éléments (Figure II.1) :

— Une « genotype-to-phenotype-to-fitness map [1] » qui décode la séquence génomique d'un individu en un phénotype et calcule la valeur phénotypique correspondante. Dans Aevol, ce « mapping » est un algorithme respectant autant que possible les caractéristiques structurelles du codage de l'information génétique.

— Une population d'organismes, chacun possédant un génome, et donc son propre phénotype et sa propre fitness. À chaque génération, les organismes sont en concurrence pour constituer la génération suivante. Classiquement la compétition est locale (dans un voisinage 3×3) mais le modèle peut être paramétré pour simuler une compétition globale panmictique.

— Un processus de réplication du génome au cours duquel les génomes peuvent subir plusieurs types d'événements mutationnels, notamment des réarrangements chromosomiques et des mutations locales. Sept types de mutations sont modélisés par défaut, incluant trois mutations locales : substitutions, petites insertions, petites délétions (classiquement regroupées sous le terme d'InDels), deux réarrangements équilibrés (qui conservent la taille du génome), les inversions et les translocations, et deux réarrangements déséquilibrés, les duplications et grandes délétions.

1. Ce terme étant quasiment intraduisible en français, nous choisissons de conserver la terminologie anglaise.

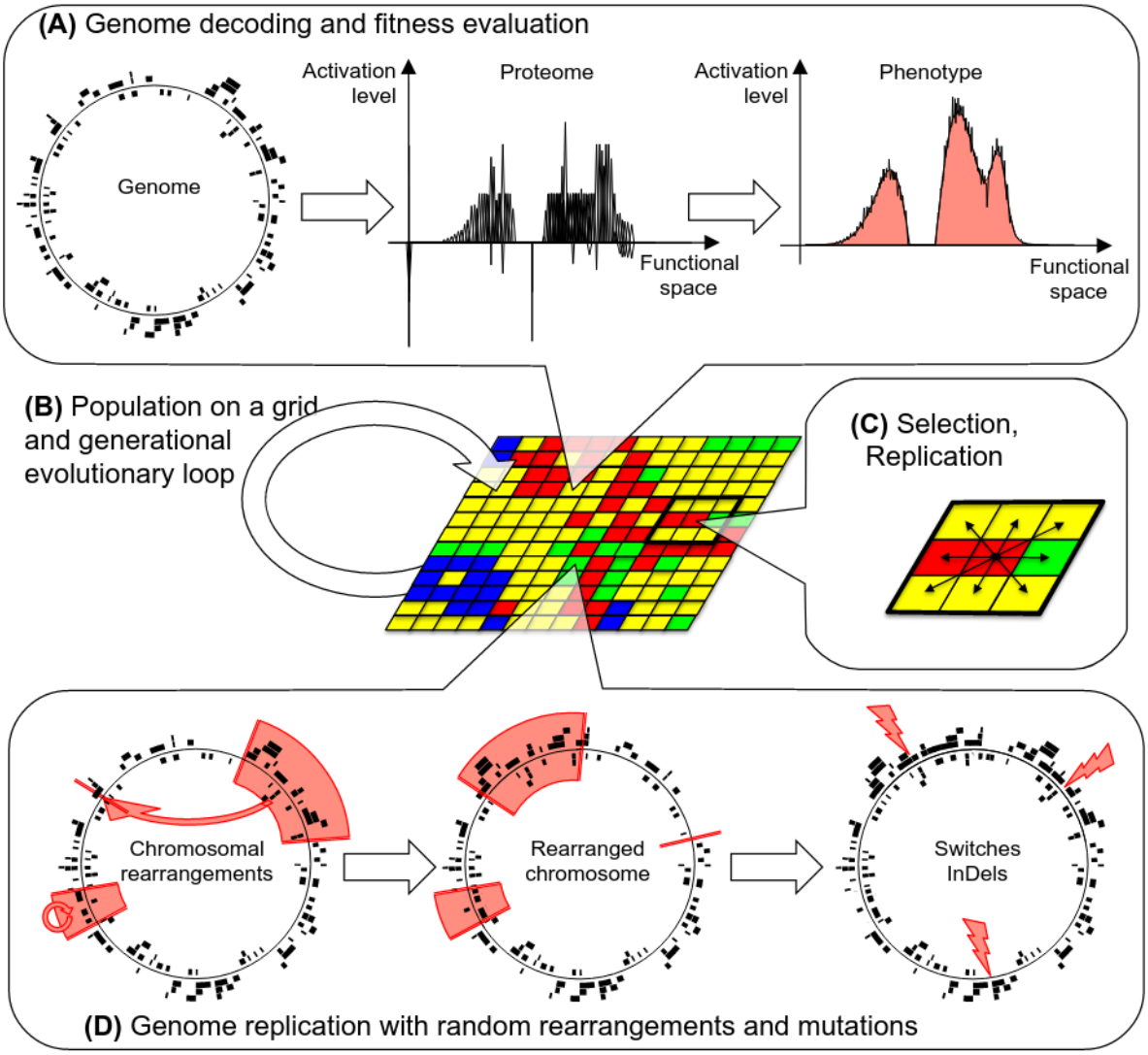


FIGURE II.1 – **Le modèle Aevol.** A) Vue d'ensemble du mapping genotype-to-phenotype. L'organisme présenté ici est un organisme réel ayant évolué dans Aevol pendant dix millions de générations avec une cible phénotypique typique. Il contient donc de nombreux Cadres de Lecture Ouverts (ORF) sur les deux brins (à gauche). Ces gènes code pour de nombreuses protéines (correspondant dans le modèle à des fonction triangulaires – au centre) et il est bien adapté à son environnement (au sens où sa fonction phénotypique — courbe noire — est très proche de la fonction de la cible — courbe unie rouge clair). (B) La population est représentée par une grille, généralement carrée, de taille paramétrable. Elle est entièrement renouvelée à chaque génération. (C) Exemple d'un processus de sélection locale se produisant avec un voisinage de 3×3. (C) Les opérateurs de mutation comprennent les réarrangements chromosomiques (duplications, délétions, translocations et inversions — ici, une translocation et une inversion sont représentées) et les mutations locales (substitutions et InDels).

2 Description du modèle

2.1 La Genotype-to-Phenotype-to-Fitness map

Représentation du génome. Chaque organisme artificiel, à l'instar des procaryotes, est asexué, haploïde et possède un seul chromosome circulaire. Le génome est codé sous la forme d'une chaîne binaire à double brin contenant un nombre variable d'ORF (Open-Reading Frame, ou gènes), séparés par des séquences non-codantes (Figure II.2). Les gènes sont délimités par des séquences signal prédéfinies indiquant les débuts de transcription et de traduction. Le nombre d'ARN et le nombre de protéines que possède un organisme dépend donc de ses séquences signal et peut évoluer à la suite d'événements mutationnels.

Le mode de décodage basé sur des séquences signal permet à de nombreuses structures d'apparaître au cours de l'évolution et de se maintenir (ou pas) en fonction des paramètres. On peut ainsi observer des ARN et des ORF chevauchants ou des structures polycistroniques (opérons). En utilisant le modèle, Parsons *et al.* (2010) ont ainsi montré que, paradoxalement, dans Aevol la longueur des opérons évolue de façon inversement proportionnelle à la longueur du génome.

Algorithme de décodage du génome. La transcription commence au niveau des promoteurs, qui sont définis dans le modèle comme des séquences suffisamment proches d'une séquence consensus choisie arbitrairement (01010110011110010010110 dans toutes les simulations, avec au maximum $d_{max} = 4$ mésappariements). Le niveau d'expression e d'un ARNm est déterminé par la similarité entre son promoteur et la séquence consensus : $e = 1 - \frac{d}{d_{max}+1}$ avec d le nombre de mésappariements ($d \leq d_{max}$). Ce calcul modélise l'interaction de l'ARN polymérase avec le promoteur, sans régulation supplémentaire [1].

Lorsqu'un promoteur est trouvé, la transcription démarre et se poursuit jusqu'à ce qu'un terminateur soit atteint. Les terminateurs sont définis comme des séquences capables de former une structure en tige-boucle, comme le font les terminateurs bactériens ρ indépendants. La taille de la tige est ici fixée à 4 bases et celle de la boucle à 3 bases.

Le signal d'initiation de la traduction est 011011$****$000 (avec $*$ une base quelconque), ce qui correspond à une séquence de type Shine-Dalgarno (ou RBS pour Ribosome-Binding Site) suivie d'un codon Start 000 quatre bases plus loin. Lorsque ce signal est trouvé sur un ARNm, le cadre de lecture (Open-Reading-Frame – ORF) en aval est lu jusqu'à ce que le signal de terminaison (le codon Stop 001) soit trouvé sur le même cadre de lecture. Chaque codon situé entre les signaux d'initiation et de terminaison est traduit en un « acide aminé » abstrait à l'aide d'un code génétique artificiel, donnant ainsi naissance à la séquence de la protéine (Figure II.2). Les séquences transcrites (ARNm) peuvent donc contenir un nombre arbitraire d'ORF. Certains ARNm peuvent ne contenir aucun ORF (ARNm non codant) tandis que d'autres pouvent contenir plusieurs ORF (ARNm polycistronique ou opéron). Il est important de noter que les fractions relatives d'ARNm non codants, monocistroniques et polycistroniques ne sont pas prédéfinies, mais résultent

1. Une version de Aevol intégrant un modèle de régulation a été développée dans le cadre de la thèse de Yolanda Sachez-Dehesa en 2009. Cependant les temps de calcul cumulés de l'évolution et de la dynamique des réseaux sont tel que ce modèle a été peu utilisé jusqu'à présent (Beslon *et al.*, 2010a,b; Vadée-Le-Brun *et al.*, 2015, 2016).

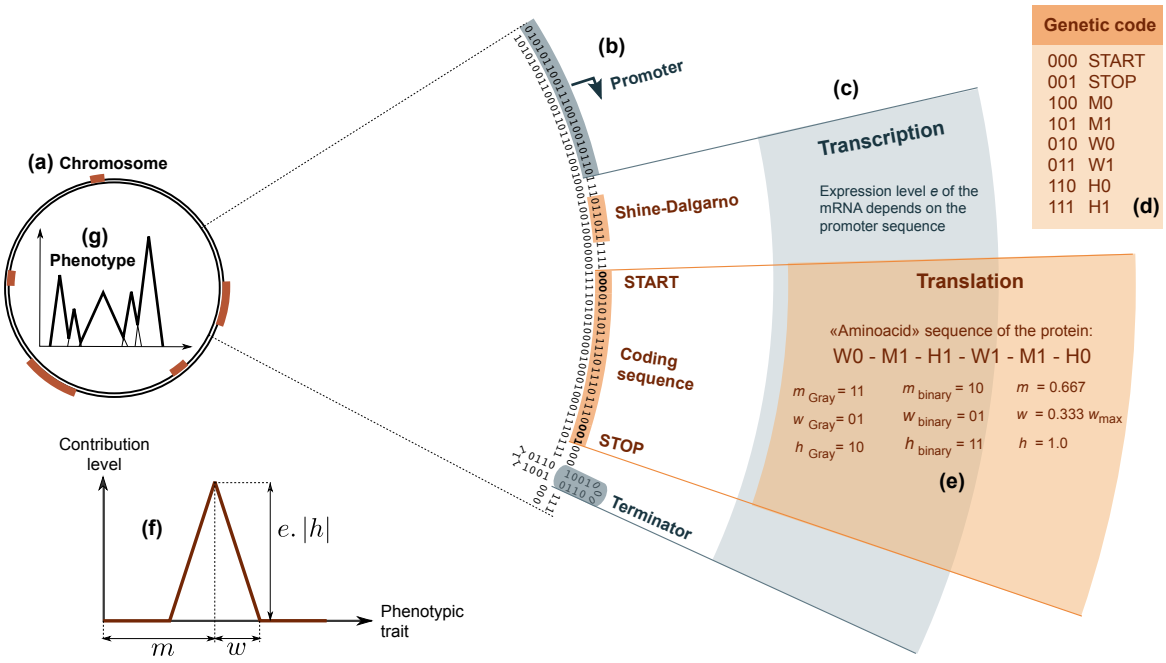


FIGURE II.2 – Dans le modèle, chaque organisme possède un chromosome binaire double brin circulaire **(a)** le long duquel les gènes sont délimités par des séquences signal prédéfinies **(b)**. Parmi ces séquences, les promoteurs et les terminateurs marquent les limites des ARN **(c)**, à l'intérieur desquels des séquences codantes peuvent à leur tour être identifiées entre un signal Shine-Dalgarno (ou RBS – Ribosome Binding Site) suivi, quelques bases plus loin, d'un codon START. La lecture d'un gène est alors initiée jusqu'au prochain codon STOP situé sur le même cadre de lecture. Chaque séquence codante est ensuite traduite en séquence primaire de protéine à l'aide d'un code génétique prédéfini **(d)**. Cette séquence primaire est décodée en trois paramètres réels appelés m, w et h **(e)**. Les protéines, le phénotype et l'environnement sont représentés par des fonctions mathématiques qui associent un niveau dans $[-1, 1]$ à chaque trait phénotypique abstrait (lui-même dans $[0, 1]$). Pour des raisons de simplicité, la contribution d'une protéine est une fonction linéaire par morceaux de forme triangulaire : les paramètres m, w et h correspondent respectivement à la position, à la demi-largeur et à la hauteur du triangle **(f)**. Toutes les protéines (*i.e.* les fonctions triangles) codées dans le chromosome sont ensuite additionnées pour calculer le phénotype **(g)** qui, une fois comparé à une « cible environnementale » prédéfinie, peut être utilisé pour calculer la fitness de l'individu. Figure tirée de (Parsons, 2011).

de la dynamique évolutive et sont susceptibles d'être influencées par les conditions évolutives comme l'ont montré Parsons *et al.* (2010). Le tableau II.1 présente les différentes séquences signal utilisées dans le modèle.

Fonction protéique et calcul du phénotype. Nous venons de voir que Aevol représente relativement fidèlement la structure des génomes procaryotes. Cependant, pour des raisons d'efficacité computationnelle, le calcul des fonctions moléculaires, lui, s'écarte de tout réalisme et, dans Aevol, tous les mécanismes fonctionnels sont décrits dans un espace mathématique abstrait.

Séquence signal	Séquence consensus	Fonction
Promoteur	0101011001110010010110	Initiation et régulation de la transcription
Terminateur	$abcd***\overline{d}\overline{c}\overline{b}\overline{a}$	Fin de la transcription
RBS + Codon START	$011011****000$	Initiation de la traduction
Codon STOP	001	Fin de la traduction

Table II.1 – Séquences consensus et fonction des séquences signal dans Aevol. Notation : $*$ = nucléotide quelconque, $\overline{x}$ = nucléotide complémentaire du nucléotide x.

Nous définissons un espace unidimensionnel continu abstrait $\Omega = [0, 1]$ de traits phénotypiques. Dans cet espace, chaque protéine est modélisée comme une fonction mathématique qui associe un niveau de contribution compris entre -1 et 1 à un sous-ensemble de ces traits. L'étendue maximale des caractères phénotypiques auxquels une même protéine peut contribuer est limité par $2 \times W_{max}$, où W_{max} définit le degré maximal de pléiotropie. W_{max} est un paramètre du modèle qui, lorsqu'il augmente, réduit indirectement le nombre total de protéines nécessaires pour couvrir l'ensemble de l'espace phénotypique. De même que le paramètre K du classique NK-fitness landscape proposé par Kauffman et Levin (1987), l'augmentation de W_{max}, en accroissant le niveau de pléiotropie, accroît indirectement l'épistasie et, par conséquent, la rugosité du paysage adaptatif.

Par souci de simplicité et d'efficacité, Aevol utilise des fonctions linéaires par morceaux de forme triangulaire symétrique pour modéliser la contribution des protéines aux traits phénotypiques (Figure II.2). De cette manière, seuls trois paramètres sont nécessaires pour caractériser la contribution d'une protéine donnée : la position $m \in \Omega$ du triangle sur l'axe, sa demi-largeur $w \in W_{max}$ et sa hauteur $h \in [-1, 1]$. Cela signifie que cette protéine contribue aux caractères phénotypiques dans $[m-w, m+w]$, avec une contribution maximale h pour les caractères situés en m et une décroissance linéaire de la contribution (jusqu'à une contribution nulle en $m - w$ et $m + w$). Ainsi, différents types de protéines peuvent coexister, de très efficaces (h élevé) à peu efficaces (h faible) voire inhibitrices (h négatif) et de très spécialisées (w faible) à très polyvalentes (w élevé).

Pour calculer les trois paramètres de la contribution fonctionnelle d'une protéine, la séquence primaire de celle-ci est interprétée en termes de trois sous-séquences binaires entrelacées qui seront à leur tour décodées comme les valeurs des paramètres m, w et h. Pour cela, un code génétique associe chaque codon (à l'exception du Start et du Stop) à un acide aminé parmi les six possibles (Figure II.2). Ces six acides aminés sont regroupés par paires ($M0/M1$, $W0/W1$ et $H0/H1$) ce qui permet, à partir de la séquence primaire de la protéine, d'extraire trois sous-séquences binaires correspondant à la suite des acides aminés de type M, W et H respectivement. Par exemple, le codon 010 (resp. 011) est traduit en un seul acide aminé $W0$ (resp. $W1$), ce qui signifie qu'il concatène un bit 0 (resp. 1) au code binaire du paramètre w. Les mutations dans les séquences codantes, y compris bien sûr les mutations locales, mais aussi les réarrangements chromosomiques, vont donc pouvoir modifier les codons et donc la contribution de la protéine au phénotype. Une fois les trois sous-séquences binaires extraites, elles sont converties en trois entiers puis normalisées (en fonction de leur longueur – qui détermine l'entier maximal possible

pour la sous-séquence) pour obtenir les valeurs de m, w et h. On notera que ce codage permet d'obtenir des gènes de longueur quelconque, l'allongement d'un gène augmentant la précision de la normalisation, donc la précision de la fonction protéique associée.

Une fois toutes les fonctions triangles calculées, les contributions de toutes les protéines codées dans le génotype d'un organisme sont combinées pour obtenir le niveau final de chaque trait phénotypique de Ω. Pour ce faire, toutes les contributions des protéines sont d'abord ré-échellonnées par le taux de transcription e de l'ARNm correspondant (voir ci-dessus), puis les fonctions mathématiques de toutes les protéines sont additionnées, avec des limites en 0 et 1. La fonction linéaire par morceaux qui en résulte $f_P : \Omega \to [0, 1]$ correspond au phénotype de l'organisme, c'est-à-dire à l'ensemble des traits activés par son patrimoine génétique.

Calcul de la fitness. Dans le modèle, la fitness dépend uniquement de la différence entre les niveaux des traits phénotypiques et les niveaux des traits définis par une « cible phénotypique » représentant indirectement l'environnement. La cible phénotypique est une fonction mathématique définie par l'utilisateur $f_E : \Omega \to [0, 1]$. Elle indique le niveau optimal de chaque trait phénotypique dans Ω. Dans les expériences Aevol habituelles, f_E est la somme de plusieurs courbes Gaussiennes dont l'écart-type, la hauteur maximale et les moyennes sont définies par un fichier de paramètres. Elle peut être stable au cours de l'évolution ou changer de manière stochastique. L'utilisation de Gaussiennes permet d'assurer que le phénotype (qui correspond, lui, à une fonction linéaire par morceaux) ne pourra pas être parfaitement adapté à la cible, ce qui garantit que la limite de l'évolution sera liée à la dérive et non à la sélection.

La différence entre f_P et f_E est définie mathématiquement par une intégrale $\Delta := \int_\Omega |f_E(x) - f_P(x)|dx$, $\forall x \in \Omega$ et est appelée « erreur métabolique ». Elle est utilisée pour mesurer l'adaptation d'un organisme en pénalisant à la fois la sous-réalisation et la sur-réalisation de ses traits phénotypiques. Compte tenu de l'erreur métabolique d'un individu, sa fitness f est donnée par $f := \exp(-k\Delta)$ avec k un paramètre fixe régulant la force de la sélection (plus k est élevé, plus l'effet des variations de l'erreur métabolique sur les valeurs de fitness est important).

Pour illustrer ce calcul, nous pouvons examiner les organismes de la Figure II.3 : à la génération 0 (à gauche), il n'y a qu'un seul gène, donc une seule protéine dans le protéome. De ce fait l'erreur métabolique est relativement grande (de l'ordre de 0.15, soit l'aire de la différence entre la somme des Gaussiennes – en gris – et le triangle noir). Dans notre expérience nous utilisons un paramètre $k = 1000$, soit une fitness de $f = \exp(-1000 \times 0,1519) \approx 1,1 \times 10^{-66}$. À titre de comparaison, l'organisme évolué (à droite sur la Figure II.3) possède 58 gènes, de sorte que son phénotype est la somme des 58 triangles représentés dans le protéome. On voit (à droite en bas) que la courbe phénotypique de cet organisme ne s'écarte que très peu de la cible représentant l'environnement. La somme des petits écarts à la cible correspond à une aire totale de $0,00296$ environ, soit une fitness de $f = \exp(-1000 \times 0,00296) \approx 0,0517$. Cet organisme a évolué pendant un million de générations au cours desquelles il a acquis puis optimisé ses gènes, produisant un gain de fitness substantiel (dans Aevol, la fitness maximale correspond à une erreur métabolique nulle, donc à une fitness de $\exp(0) = 1$).

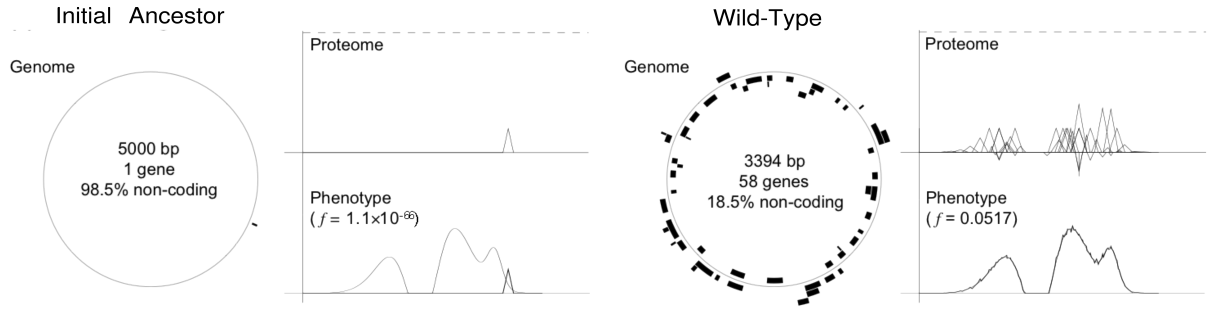


Figure II.3 – Représentation graphique des génomes et protéomes : À gauche, un ancêtre à la génération zéro. À droite, un organisme ayant évolué pendant un million de générations en conditions constantes.

2.2 Modèle de population et processus de sélection

La population est modélisée sur une grille torique, généralement carrée, avec un individu par cellule de la grille. À chaque génération, la fitness de chaque individu est calculée et les individus entrent en concurrence pour peupler chaque cellule de la grille à la génération suivante. Cette compétition peut être entièrement locale (les 9 individus du voisinage d'une cellule donnée sont en concurrence pour la peupler à la génération suivante, Figure II.1.C) ou englober une sous-population plus large. Si le champ de sélection englobe l'ensemble de la population, tous les individus sont en concurrence pour toutes les cellules de la grille. Il est important de noter que plus la sélection est locale, plus le modèle de population diverge du modèle panmictique de Wright-Fisher, car la sélection locale augmente la taille effective de la population N_e pour une taille de population donnée (Waples, 2010).

Compte tenu de la portée de la sélection, les individus du voisinage $\mathcal{N}$ d'une cellule donnée sont en concurrence par le biais d'un schéma de sélection classique, directement proportionnel à la fitness : la probabilité p_j, pour un individu j ayant une fitness f_j, de peupler une cellule donnée à la génération suivante est donnée par $p_j = f_j / \sum_{i \in \mathcal{N}} f_i$.

2.3 Opérateurs génétiques

Par rapport à la plupart des modèles d'évolution, une des spécificités de Aevol est la relativement grande diversité d'opérateurs génétiques modélisés. En effet, comme nous l'avons montré (section 2.1) l'utilisation d'un algorithme de décodage du génome (par opposition à la plupart des simulateurs qui définissent a priori l'ensemble des gènes d'un organismes et qui simulent l'évolution de ces gènes) permet de calculer le phénotype d'un organisme, quelle que soit sa séquence (même s'il est possible qu'aucun caractère ne soit activé, par exemple si il n'y a pas d'ORF sur la séquence) et, donc, de simuler n'importe quel opérateur de modification de cette séquence. Ainsi, au cours de leur réplication, les génomes de Aevol peuvent subir différentes modifications de leur séquence génétique (Figure II.1).

Dans l'utilisation classique du simulateur, sept types de mutations sont modélisés

(illustrés sur la Figure II.4). Trois mutations sont locales (substitutions et petites insertions ou délétions), et quatre sont des réarrangements chromosomiques, soit équilibrés (sans changement de la taille du génome) : les translocations et les inversions, soit déséquilibrés : les duplications et les délétions. D'autres opérateurs peuvent facilement être ajoutés au modèle (transfert horizontal, duplications en tandem, ...). Les sept opérateurs présentés ici sont ceux utilisés dans toutes les expériences présentées dans ce manuscrit.

Les mutations locales se produisent à une position quelconque, choisie par tirage uniforme sur le génome. Les substitutions modifient un seul nucléotide. Les InDels insèrent (ou suppriment) une petite séquence de longueur aléatoire — et de composition aléatoire pour les insertions. La longueur de la séquence est tirée uniformément entre 1 et une valeur maximale (6 par défaut). Notamment, les InDels se produisant au sein d'un ORF peuvent décaler le cadre de lecture ou simplement ajouter/supprimer des codons, ce qui se traduit par des résultats très différents sur le plan de l'évolution.

Les points de cassure des réarrangements chromosomiques sont tirés uniformément sur le chromosome, le nombre de points de cassure dépendant du type de réarrangement (Figure II.4). Ainsi, les réarrangements chromosomiques peuvent être de n'importe quelle taille entre 1 et la taille totale du génome, ce qui permet d'étudier l'effet de variants structurels de toutes tailles, y compris les petits variants structurels qui sont effectivement observés *in vivo* (Musumeci *et al.*, 2000; Audrézet *et al.*, 2004; Blakely *et al.*, 2006).

Les taux μ_t auxquels les différents types t de mutations génétiques se produisent sont définis comme une probabilité par base et par réplication. Cela signifie que le nombre d'événements spontanés dépend linéairement de la longueur du génome. Cependant, la probabilité de fixation d'un type de mutation donné dépend de son effet phénotypique (par exemple, une mutation affectant exclusivement une région non transcrite est très susceptible d'être neutre car la probabilité qu'elle crée spontanément un ORF est extrêmement faible). Il est important de noter que, dans Aevol, la distribution des effets de fitness (DFE, en anglais *distribution of fitness effect*) de tout type de mutation n'est pas prédéfinie, mais dépend de son effet sur la séquence et de la structure du génome. Par exemple, la fraction des séquences codantes ou la distribution spatiale des gènes le long du chromosome sont susceptibles de modifier la probabilité qu'une mutation donnée modifie le phénotype et donc la fitness. Si cet effet est globalement bien compris pour les mutations locales (qui, en première approximation, seront neutres si leur locus correspond à une séquence non-codante), il est bien moins connu pour les réarrangements chromosomiques. Le fait de disposer d'une DFE émergente au lieu d'une DFE prédéfinie permet donc d'étudier les effets complexes, directs et indirects, des réarrangements chromosomiques sur la dynamique de l'évolution. Nous verrons au chapitre suivant que ces effets peuvent être pour le moins inattendus.

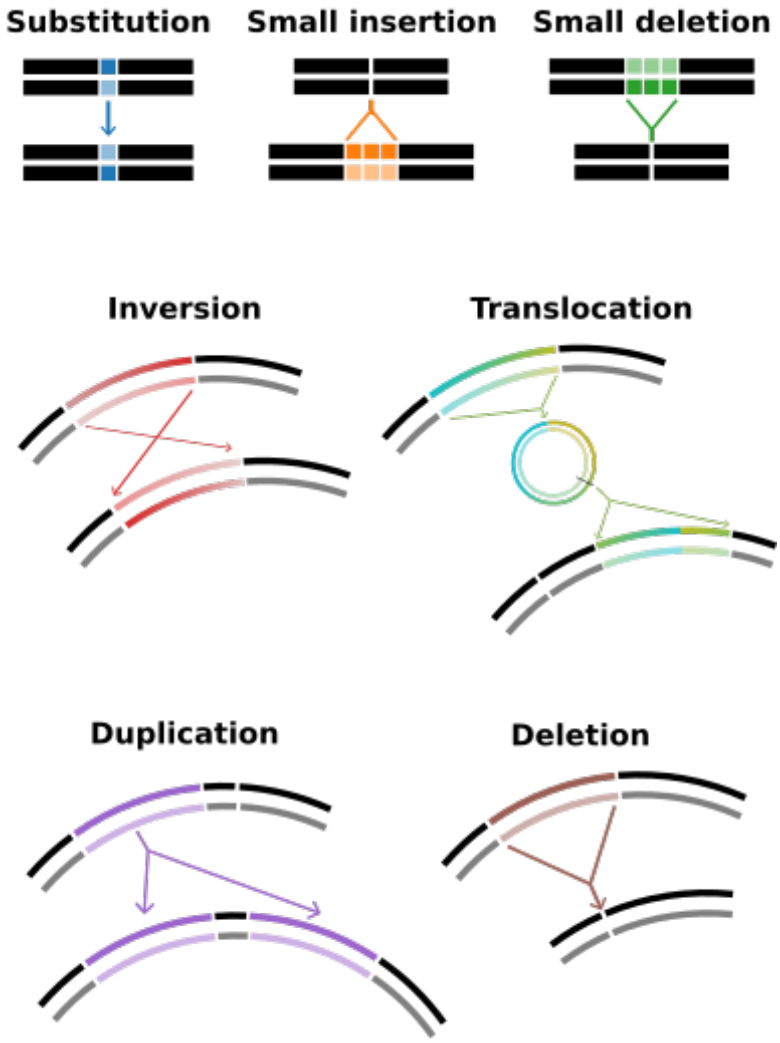


Figure II.4 – Représentation stylisée des opérateurs de mutation dans Aevol. (en haut) Mutations locales : substitution (une paire de bases est remplacée par une autre), petite insertion et petite délétion (une à six paires de bases sont insérées ou supprimées). (au milieu) Réarrangements chromosomiques dits « équilibrés » (*balanced* – ne modifiant pas la taille du chromosome) : inversion (deux points sont tirés au hasard sur le chromosome et le segment intermédiaire entre les deux est inversé entre les deux brins) et translocation (deux points sont choisis au hasard, le segment intermédiaire est excisé, circularisé (pour former un plasmide), recoupé et inséré en un point aléatoire du le génome). (en bas) Réarrangements chromosomiques dits « déséquilibrés » (*unbalanced* – modifiant la taille du chromosome) : duplication (un segment situé entre deux points tirés au hasard est copié et inséré en un point aléatoire) et délétion (un segment situé entre deux points tirés au hasard est excisé).

3 Protocoles expérimentaux typiques

Aevol est basé sur l'exécution et l'analyse de simulations dites « forward-in-time ». C'est-à-dire qu'il va simuler l'évolution future d'une séquence initiale, par opposition aux modèles « backward-in-time » qui cherchent à reconstituer les événements ancestraux ayant conduit à une séquence génétique particulière. Plus précisément, toute expérience avec Aevol est divisée en quatre étapes principales. La première étape consiste à préparer une simulation à l'aide de la commande `aevol_create`. Celle-ci lit le fichier de paramètres (voir Figure II.7 et Tableau II.2) et utilise une fonction de boostrap pour créer une population d'organismes à la génération zéro selon les valeurs spécifiées (voir section suivante). `aevol_run` simule ensuite l'évolution de cette population initiale, génération après génération, jusqu'à une génération spécifiée à l'appel. `aevol_run` offre aussi la possibilité de repartir d'une population sauvegardée, ce qui permet, par exemple, de dupliquer une population (`aevol_propagate`), de modifier des paramètres (`aevol_modify`), et de tester expérimentalement l'effet de la modification.

`aevol_run` produit plusieurs fichiers de données : des statistiques sommaires concernant le meilleur individu à chaque génération (fitness, taille du génome, nombre de gènes...) et, régulièrement, des fichiers de sauvegarde (pour reprendre une simulation). Il peut aussi sauvegarder des fichiers d'arbres contenant les « replication reports ». Ceux-ci enregistrent tous les événements de réplication et de mutation se produisant au cours de la simulation. Ainsi, en analysant les arbres, on peut reconstruire avec précision la série d'événements qui a abouti à la population finale, reconstruire la lignée de descendance et extraire les caractéristiques de toutes les mutations fixées. À cette fin, `aevol_post_lineage` part de la population finale et lit les fichiers de l'arbre à rebours dans le temps pour reconstruire la lignée de descendance et extraire les replication reports correspondants. En effet, à chaque génération, les arbres contiennent les données de chacune des N réplications (N étant la taille de la population), mais pour des organismes aploïdes et assexués, seule une de ces réplications va effectivement conduire à la génération finale (jusqu'au point de coalescence ; ensuite plusieurs lignées peuvent coexister). Enfin, la quatrième étape, `aevol_post_ancestor_stats`, utilise les replication reports de la lignée pour calculer les statistiques de la lignée ancestrale et la liste des événements mutationnels qui se sont fixés le long de cette lignée.

Il est important de signaler que, bien que les utilisateurs puissent être tentés d'arrêter les expériences après l'étape `aevol_run`, les statistiques des meilleurs individus au fil des générations, si ils sont représentatifs de la tendance globale de la simulation, ne doivent pas être confondues avec les statistiques de la lignée ancestrale, car les événements mutationnels portés par le meilleur individu peuvent ne pas être fixés à long terme.

3.1 Utilisation de base : Partir d'un individu naïf

Aevol est un modèle expérimental, ce qui signifie que l'utilisation typique consiste à répliquer des simulations d'évolution avec des paramètres différents et d'analyser l'effet de divers paramètres évolutifs (typiquement les taux de mutation, les biais mutationnels, la taille de la population...) sur les génomes (ou sur la fitness) en comparant des simulations dans les différents scénarios (voir le Tableau II.2 pour une liste des principaux paramètres

testables). Une fois le plan expérimental choisi (valeurs des paramètres, nombre de répétitions, durée des expériences), deux types d'utilisation sont possibles. L'utilisation la plus simple de Aevol consiste à tester l'effet de ces paramètres directement, à partir d'individus « naïfs » (voir section suivante pour l'utilisation avancée).

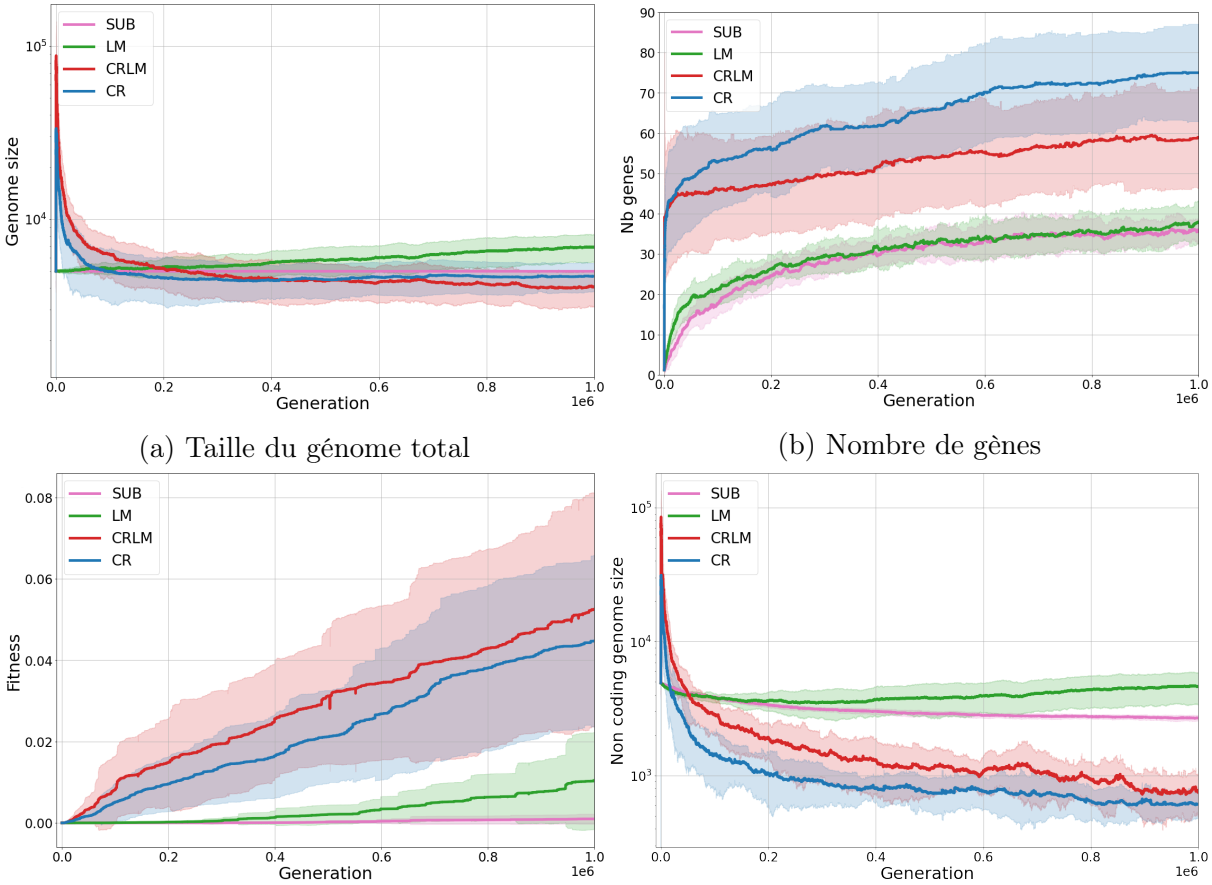
Dans ce cas, on utilise `aevol_create` pour lancer un mécanisme de « bootstrap » permettant de créer un premier organisme. Pour cela `aevol_create` génère des séquences aléatoires d'une longueur prédéfinie (typiquement 5 000 paires de bases) jusqu'à ce qu'il trouve un génome dont la fitness est meilleure que celle d'un organisme qui n'aurait aucun gène. Ce génome a typiquement un ou (rarement) deux gènes fonctionnels mais sa fitness est très faible (voir Figure II.3) car il est situé très loin de l'optimum (organisme « naif »). La population est alors entièrement initialisée avec ce premier organisme (population clonale). Cette approche permet d'étudier les trajectoires évolutives en partant loin de l'optimum. Toutefois, dans ce cas, la dynamique évolutive est fortement dominée par le recrutement de gènes, avec une variation massive de la taille du génome, en particulier au début de l'évolution comme le montre la Figure II.5. Cette utilisation de Aevol met donc l'accent sur une dynamique évolutive très spécifique ce qui peut masquer des effets plus subtils. Si l'on souhaite étudier de tels effets, cette utilisation de base n'est donc pas appropriée et l'on doit se tourner vers un plan expérimental plus avancé basé sur la création de « Wild-Types ».

La Figure II.3 présente un exemple de plan expérimental typique pour ce type d'utilisation. Partant d'un organisme naif, le plan expérimental consiste ici à tester l'effet des différents types de mutations. Quatre conditions expérimentales sont testées : substitutions seules (SUB), mutations locales seules (LM) réarrangements chromosomiques et mutations locales (CRLM) et réarrangements chromosomiques seuls (CR), avec 10 répétitions par condition. Sans entrer dans une analyse détaillée de ces simulations, les résultats montrent que, en partant loin de l'optimum, les réarrangements chromosomiques sont les principaux contributeurs de l'évolution. Ce résultat illustre l'importance des mécanismes de duplications de gènes pour l'acquisition de fonctions. On constate cependant qu'en présence des deux types de mutations, les génomes comportent moins de gènes mais aboutissent à une meilleure fitness. Ceci montre que la duplication de gènes est plus efficace en présence de mutations locales facilitant l'adaptation (la divergence) des gènes dupliqués (Banse *et al.*, 2023).

3.2 Utilisation avancée : Wild-Typing

Comme expliqué dans la section précédente et illustré par la Figure II.5, l'initialisation avec des organismes naïfs permet de distinguer des effets forts entre les simulations. Cependant ces effets peuvent en masquer d'autres, plus subtils. Pour étudier des effets fins, il est nécessaire d'initialiser les simulations en évitant la phase de construction initiale du génome qui est susceptible d'introduire de nombreux biais ainsi qu'une très forte variabilité entre les réplicats. Pour cela, Aevol permet de générer des « Wild-Types », c'est-à-dire des individus ayant évolué suffisamment longtemps (de quelques centaines de milliers de générations à plusieurs millions de générations en fonction des paramètres [1]) dans des conditions constantes. Ces individus possèdent alors un ensemble de gènes rela-

1. Voir `https://www.aevol.fr/doc/user-doc/` pour plus de détails.



(a) Taille du génome total

(b) Nombre de gènes

(c) Fitness

(d) Taille du génome non-codant

FIGURE II.5 – Évolution à partir d'organismes naïfs. Évolution de la taille du génome total (a), du nombre de gènes (b), de la fitness (c) et de la quantité d'ADN non-codant (d) sur 1 000 000 de générations à partir d'un individu naïf. Quatre conditions sont testées. SUB : Substitutions uniquement. LM : Mutations locales uniquement (substitutions et InDels). CR : Réarrangements chromosomiques uniquement. CRLM : Réarrangements chromosomiques et mutations locales (toutes les mutations). Lorsque les réarrangements chromosomiques sont présents, ils conduisent à une explosion initiale de la taille du génome. Ce phénomène est dû au recrutement massif de gènes par un mécanisme de duplication-divergence (voir panel b). La quantité d'ADN non-codant augmente simultanément par effet d'auto-stop (d). Figure issue de (Banse *et al.*, 2023).

tivement stables et sont bien adaptés à leur environnement. Pour générer un wild-type, on simule l'évolution d'une population pendant un temps très long (typiquement 10 100 000 générations dans les expériences présentées au chapitre III) puis on extrait un ou plusieurs individus dans la lignée coalescente de la population finale (à la génération 10 000 000 dans l'exemple précédent – cette procédure permet de s'assurer que le wild-type correspond bien à un individu fixé, ce qui ne sera pas nécessairement le cas du meilleur individu de la dernière génération). Cet individu est alors transmis à `aevol_create` qui va l'utiliser pour créer une population clonale et lui associer des paramètres évolutifs, possiblement différents de ceux utilisés pour l'évolution du wild-type. Il est alors possible de lancer des expériences testant l'influence de différents paramètres sur un même individu initial.

Le Wild-Typing permet d'étudier la réponse d'un organisme bien adapté à différents types de perturbations, et donc d'analyser les trajectoires évolutives de scénarios plus réalistes sur le plan biologique (Batut *et al.*, 2013; Banse *et al.*, 2023).

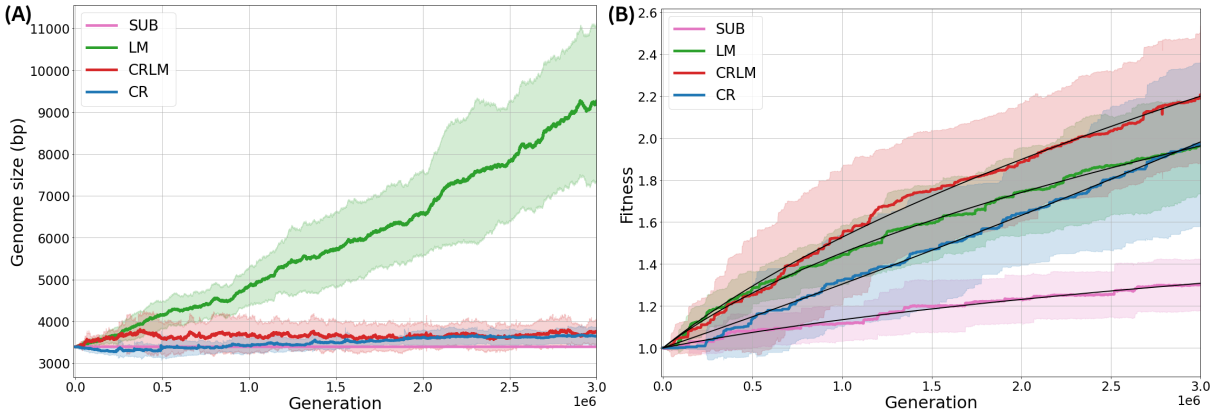
La Figure II.6 présente un tel cas d'utilisation (quoique peu réaliste biologiquement en l'occurrence). Partant d'un organisme pré-évolué pendant 1 000 000 générations, les quatre mêmes conditions évolutives que précédemment ont été testées pendant 3 millions de générations. Les résultats suggèrent que la présence de réarrangements régule la taille des génomes alors qu'en présence de mutations locales seules, la taille des génomes croît continuement pendant les 3 millions de générations de l'expérience, ce résultat soutenant l'hypothèse de « border-induced selection » formulée par (Loewenthal *et al.*, 2022). Sur le plan de la fitness, les résultats montrent que les substitutions seules sont très limitées mais que l'adjonction de mutations plus complexes (InDels ou réarrangements chromosomiques) permet des gains de fitness continus. Par rapport aux simulations présentées précédemment, on constate ici que les InDels font quasiment jeu égal avec les réarrangements. En effet, comme le répertoire génique des individus est déjà constuit, les InDels suffisent à optimiser les séquences existantes. Cependant, on constate aussi qu'en temps long, la présence de réarrangements permet de soutenir la dynamique évolutive en limitant le « diminishing return » (courbe noire). Ce résultat n'était pas visible sur la Figure II.5 car masqué par les fortes différences dans la dynamique d'acquisition des gènes (Banse *et al.*, 2023).

3.3 Analyse post-évolution

Une fois les simulations terminées [1], les caractéristiques générales des génomes des ancêtres de la population finale sont disponibles (taille du génome, nombre de gènes, proportion de codage, etc.), ainsi que la liste de toutes les mutations fixées avec leurs types, leurs loci et leurs effets sur la fitness. Cependant, ces données ne sont généralement pas suffisantes pour estimer le rôle relatif des différentes forces évolutives (sélection directe et indirecte, dérive, et les différents événements mutationnels - événements locaux, réarrangements chromosomiques équilibrés et déséquilibrés) dans les observations.

Aevol fournit plusieurs outils pour aider l'utilisateur à analyser les individus le long de la lignée de descendance en estimant en particulier leur robustesse, leur évolvabilité et la distribution des effets de fitness (DFE) pour tous les types de mutations. Pour cela,

1. donc à l'issue de l'exécution de `aevol_post_ancestor_stats`



(a) Taille du génome (b) Fitness

FIGURE II.6 – Évolution à partir d'organismes wild-type. Les simulations ont été initialisées à partir d'un organisme avant évolué pendant 1 000 000 générations dans les conditions CRLM. Évolution de la taille du génome (a) et de la fitness (b), pendant 3 000 000 générations à partir d'un individu wild-type. Quatre conditions sont testées. SUB : Substitutions uniquement. LM : Mutations locales uniquement (substitutions et InDels). CR : Réarrangements chromosomiques uniquement. CRLM : Réarrangements chromosomiques et mutations locales (toutes les mutations). Lorsque les réarrangements chromosomiques sont présents la taille du génome est régulée et la fitness à long terme est meilleure. Figure issue de (Banse *et al.*, 2023).

le modèle utilise une approche de type Monte-Carlo : pour un individu donné, il génère un très grand nombre de descendants indépendants et, en analysant la fitness de ces descendants, calcule la robustesse réplicative et l'évolvabilité des ancêtres. La robustesse réplicative est ici définie comme la fraction de descendants ayant conservé la fitness de l'individu initial. L'évolvabilité est considérée comme le « potentiel évolutif » (Blount *et al.*, 2012) d'un individu et correspond ici à l'espérance de gain de fitness parmi tous les individus générés (c'est-à-dire la somme des gains de fitness divisée par le nombre d'individus testés).

De même, Aevol peut générer et analyser un grand nombre de mutants afin d'estimer la distribution d'effet de fitness (DFE) et la robustesse mutationnelle pour n'importe quel individu et pour n'importe quel type de mutation. On distingue ici la robustesse mutationnelle de la robustesse réplicative (telle que définie ci-dessus). La robustesse mutationnelle correspond à la fraction d'individus conservant leur fitness initiale après avoir subi une et une seule mutation.

4 Fichier de paramètres typique

Nous présentons ici un fichier de paramètres typique (ici un fichier ayant servi à générer une population de type CRLM dans la Figure II.5). La description des principaux paramètres et de leurs conséquences est présentée Table II.2.

```
###################################
#         AEVOL PARAMETERS        #
###################################
##### 0. Initial setup ############
STRAIN_NAME             Wild-Type-Evolution
SEED                    4575654216
INIT_METHOD             ONE_GOOD_GENE CLONE
CHROMOSOME_INITIAL_LENGTH   5000
### 1. Genotype-to-Fitness map ####
# Target function (H, mu, sigma)
ENV_ADD_GAUSSIAN        1.2   0.52  0.12
ENV_ADD_GAUSSIAN       -1.4   0.5   0.07
ENV_ADD_GAUSSIAN        0.3   0.8   0.03
# W_Max
MAX_TRIANGLE_WIDTH      0.01
### 2. Population and selection ###
INIT_POP_SIZE           1024
WORLD_SIZE              32 32
SELECTION_SCOPE         local 3 3
SELECTION_SCHEME        fitness_proportionate   1000
##### 3. Mutation rates ############
# Local events
POINT_MUTATION_RATE     1e-6
SMALL_INSERTION_RATE    1e-6
SMALL_DELETION_RATE     1e-6
# Balanced chromosomal rearrangements
INVERSION_RATE          1e-6
TRANSLOCATION_RATE      1e-6
# Unbalanced chromosomal rearrangements
DUPLICATION_RATE        1e-6
DELETION_RATE           1e-6
#### 4. Recording #################
BACKUP_STEP             100000
RECORD_TREE             true
TREE_STEP               1000
```

Figure II.7 – Fichier de paramètres typique dans Aevol.

Section	Parameter	Usual range	Description
Genotype to Phenotype to Fitness map	Maximal pleiotropy (W_{max}) `MAX_TRIANGLE_WIDTH`	0.01 – 1 (default : 0.033333)	Largest range of phenotypic values a single protein can impact. Regulates the mean pleiotropy degree and impacts the maximal phenotypic contribution of a single gene (Knibbe *et al.*, 2007)
	Target function `ENV_ADD_GAUSSIAN`	Sum of 1 to 3 Gaussian functions	The target function is a linear combination of n Gaussian function G_i, each with a weight $\mathcal{H}_i$, a mean μ_i and a standard deviation σ_i : $Target = \sum_{i<n} \frac{\mathcal{H}_i}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2\right)$
	Length of a randomly generated genome `CHROMOSOME_INITIAL_LENGTH`	5000	Initial size of the chromosome when starting from a naive individual (see section 3.1)
Population / Selection	population size (N) `INIT_POP_SIZE`	256 – 4096	Census population size. Correlated with the effective population size N_e hence influencing the efficiency of the selection
	Grid size `WORLD_SIZE`	16x16 – 64x64	Shape of the grid. The grid shape influences the speed at which an individual can invade the population (Misevic *et al.*, 2015)
	Selection neighborhood `SELECTION_SCOPE`	Local 3x3 – Global	Type of selection (local or global), and, in the local case, size of the neighborhood used for competition. Local selection slows down the spreading of favorable mutants and increases the effective population size (Zhang, 2003)
	Intensity of the selection (k) `SELECTION_SCHEME`	fitness proportionate 250 – 2500	The selection strength influences the genome size of individuals by increasing/decreasing the indirect selection for robustness (Batut *et al.*, 2013). Note that $k = 0$ suppresses the selection
Replication process	`POINT_MUTATION_RATE`	10^{-4} – 10^{-7}	Per base mutation rates for each kind of mutation. Changes in mutation rates have been shown to impact both the genome length and the genome structure (Knibbe *et al.*, 2007; Rutten *et al.*, 2019)
	`SMALL_INSERTION_RATE`		
	`SMALL_DELETION_RATE`		
	`DUPLICATION_RATE`		
	`DELETION_RATE`		
	`INVERSION_RATE`		
	`TRANSLOCATION_RATE`		
	`MAX_INDEL_SIZE`	6	Maximal size of small insertion or deletion

TABLE II.2 – Paramètres principaux du modèle Aevol.

5 Pourquoi Aevol ?

Dans cette dernière section de description du modèle, nous allons présenter un historique des principaux résultats obtenus avec Aevol [1]. Le but ici n'est pas d'être exhaustif, mais de présenter plusieurs exemples d'études montrant la pertinence d'Aevol pour répondre à des questions concernant l'évolution de la structure et la complexité des génomes et des organismes. Nous verrons en particulier comment ces résultats ont ouvert la voie aux deux études présentées dans ce manuscrit.

5.1 Régulation de la taille des génomes en fonction du taux de mutation

Le premier résultat important obtenu avec le modèle concerne la régulation spontanée de la taille des génomes en fonction des taux de mutations. Dans leur étude, Knibbe *et al.* (2007) ont réalisé des simulations dans Aevol qui montrent une corrélation négative entre la quantité de séquences non-codantes (et donc de la taille des génomes) et les taux de mutations. Ce résultat est expliqué par un mécanisme de sélection pour la robustesse : Étant donné que les taux de mutations sont exprimés par paires de bases, le nombre de mutations par réplication dépend de la longueur totale du génome, incluant les sections non-codantes. Par conséquent, la quantité de séquences non-codantes peut être soumise à une sélection visant à optimiser le niveau de variabilité des descendants d'un individu. Knibbe *et al.* (2007) montrent qu'en augmentant les taux de mutations, des génomes plus courts sont sélectionnés, car ils augmentent la capacité des individus à produire plus de descendants neutres (en d'autres termes, les génomes plus courts sont plus robustes). Cependant, si ce mécanisme explique bien le lien entre robustesse et taille des génomes, il n'explique pas pourquoi une quantité minimale de non-codant est toujours conservée quels que soient les taux de mutations. Les auteurs supposent donc que les génomes excessivement courts sont contre-sélectionnés car ils ne sont pas propices à la génération de nouvelles mutations favorables. En d'autres termes, une faible évolvabilité serait contre-sélectionnée. La quantité de non-codant conservé résulterait donc d'un compromis entre robustesse et évolvabilité. Bien que, en raison des faibles vitesses de calcul de l'époque, les simulations soient trop courtes pour être totalement affirmatifs sur les résultats, l'étude de Knibbe *et al.* (2007) reste fondatrice dans le développement des recherches réalisées avec le logiciel Aevol, non seulement parce qu'elle est parmi les premières études menées avec ce simulateur, mais aussi et surtout surtout parce qu'elle a apporté une justification essentielle à la poursuite de l'utilisation d'Aevol pour répondre à des questions en biologie. En effet, en montrant que le non-codant bien que non fonctionnel pouvait être régulé par l'évolution, ces travaux ont mis en lumière la capacité du simulateur à générer des comportements d'intérêt.

1. Nous ne présenterons pas ici les résultats obtenus avec Aevol dans le cadre de la génération de jeux de tests synthétiques pour la validation de méthodes de bio-informatique. Ces résultats seront en effet présentés en détails dans le chapitre IV.

5.2 Régulation des structures polycistroniques

Dans leur étude, Parsons *et al.* (2010) examinent l'impact des taux de réarrangements chromosomiques sur la structure des génomes dans Aevol. Ils constatent que des taux de réarrangements plus élevés sont associés à des génomes de taille réduite et plus riches en séquences codantes. Cette compaction des génomes avec l'augmentation des taux de réarrangements se manifeste par plusieurs observations. Tout d'abord, la taille des génomes diminue, avec une réduction à la fois des régions codantes et non codantes (conformément aux résultats fondateurs de Knibbe *et al.*), mais avec une concentration accrue des séquences codantes. En outre, la structure des séquences codantes est aussi affectée, avec une augmentation de la proportion d'opérons et du nombre moyen de gènes par opéron. En particulier, Parsons *et al.* montrent que, de manière contre-intuitive, plus les génomes sont compacts, plus les ARN codants sont longs.

L'étude montre qu'il existe une corrélation entre le nombre de descendants neutres au sein des populations et les conditions expérimentales (les taux de réarrangements). En moyenne, plus les taux de réarrangements sont élevés, plus la fraction de descendants neutres est faible. Cependant, il existe une limite inférieure imposée par la sélection et la fraction de descendants neutres reste toujours au-dessus d'une valeur seuil permettant aux meilleurs individus d'avoir au moins un descendant neutre. Cette limite correspond à un seuil d'erreur, essentiel pour maintenir la stabilité de l'information génétique au fil des générations. En résumé, l'augmentation des taux de réarrangements entraîne une sélection en faveur de génomes de taille réduite en raison de leur nombre de mutations par réplication plus faible, ce qui leur permet d'avoir un plus grand nombre de descendants neutres et donc une meilleure robustesse. Cette réduction de taille affecte à la fois les régions non codantes et codantes, et provoque avec une densification de ces dernières une augmentation du nombre et de la longueur des opérons. On notera que, même si leur explication est différente et inclue des mécanismes de régulation, par la suite Nuñez *et al.* (2013) ont empiriquement confirmé cette relation négative entre la taille des génomes et le nombre et la taille des opérons dans les génomes procaryotes.

5.3 Effet de l'intensité de la sélection

Batut *et al.* (2013) étudient l'effet de la force de sélection sur les génomes. Tout comme (Knibbe *et al.*, 2007), ce travail est fondateur car Batut *et al.* y introduisent le principe de l'étude expérimentale *in silico* à partir de wild-types (voir ci-dessus). En l'occurrence, partant de population wild-types, les auteurs ont testé l'effet du paramètre k contrôlant la force de la sélection. En termes de génétique des populations, cela revient à modifier, dans une population, la distribution des coefficients de sélection s. Empiriquement, diminuer k revient à diminuer l'impact des gènes ou des mutations sur la fitness et réciproquement. Les résultats de cette étude montrent qu'en diminuant k (de $k = 750$ à $k = 250$), la taille des génomes est réduite d'environ 33%. Cette diminution se fait majoritairement par une diminution du compartiment non codant (55%), accompagnée d'une petite perte de gènes (10%). Les gènes perdus sont ceux codant pour des protéines ayant une plus faible surface dans l'espace phénotypique. Donc *a priori* les gènes apportant le moins de contribution à la fitness. Ces gènes ont probablement dégénéré, car leur valeur sélective est passée sous la barrière de dérive génétique.

5.4 Dynamique des génomes chez les hypermutants

Rutten *et al.* (2019) étudient l'évolution des populations hypermutantes. Ces populations ont des taux de mutations ponctuelles très élevés en raison de la perte d'une partie de leur machinerie de réparation de l'ADN. Elles sont fréquemment observées dans la nature et dans les expériences d'évolution (Couce *et al.*, 2017). Dans leur expérience, Rutten *et al.* simulent dans Aevol des populations avec des taux de mutation ponctuelle multipliés par 100 et étudient l'effet de cette augmentation. Ils montrent que l'augmentation des taux de mutation a des répercussions sur les génomes et la fitness des individus. Sur le plan génomique, une diminution de la taille du codant et une augmentation du non-codant sont observés. Les lignées sélectionnées perdent initialement en fitness, ce qui est attendu en raison de l'augmentation de la pression mutationnelle. Cependant, contre toute attente, ils regagnent ensuite leur fitness initiale et la dépassent même dans certaines populations. La dynamique évolutive de ces populations est particulièrement intéressante. L'augmentation des taux de mutation ponctuelle favorise la sélection de génomes avec une cible mutationnelle plus petite, c'est-à-dire moins de codant. La diminution du codant couplée à l'augmentation du non-codant modifie la distribution des effets de fitness de sorte que la proportion de mutations faiblement délétères diminue et celle des mutations fortement délétères augmente. Cet effet affecte la distribution de la fitness des populations de façon inattendue : bien que la lignée finisse par regagner sa fitness initiale, la fitness moyenne de la population en revanche ne fait que diminuer au cours des expériences présentées. Ce phénomène, qualifié d'anti-robustesse par les auteurs, augmente le nombre de descendants des meilleurs individus et leur assure une descendance viable malgré les forts taux de mutation.

5.5 The complexity ratchet

Comme nous l'avons vu précédemment, les cibles phénotypiques classiquement utilisées dans Aevol sont composées de sommes de Gaussiennes. Ce type de cibles contraint les organismes à se complexifier en accumulant des gènes puisqu'une somme de Gaussiennes ne peut être parfaitement reproduite par une somme finie de triangles. Liard *et al.* (2020) et Beslon *et al.* (2021) ont proposé une expérience « impossible » (O'Neill, 2003) originale. En changeant le type de cible phénotypique pour lui donner la forme d'un unique triangle, ils offrent aux individus la possibilité d'être parfaitement adaptés à leur environnement avec un seul gène. En tant que simulateur, Aevol offre ici l'avantage de pouvoir étudier l'évolution de la complexité dans des conditions impossibles à réaliser dans le monde biologique. Les résultats de ces expériences montrent que bien que la simplicité soit ici la solution la plus efficace (celle produisant la meilleure fitness, et la plus rapide évolutivement), dans la très grande majorité des simulations les populations s'engagent dans une évolution complexe, accumulant de nombreux gènes (68% à 86% des simulations selon les taux de mutation). Les auteurs montrent que si les gènes initiaux ne sont pas proches de la solution simple, une épistasie négative apparaît du fait des duplications de gènes qui rendent la solution simple inaccessible. L'évolution des organismes s'engage alors dans un « cliquet de complexité » (complexity ratchet) augmentant graduellement leur fitness mais au prix d'une complexification croissante et sans jamais atteindre la fitness des organismes simples.

5.6 Evolution des gènes dupliqués

La duplication de gènes est un mécanisme important pour l'évolution des organismes. En effet, la plupart des gènes semblent être organisés en familles de gènes proches dont l'origine évolutive est un gène initial ayant subi des duplications, à partir desquelles les copies ont ensuite divergé. Le devenir des copies après duplication peut être cependant très variable. Par exemple, après une duplication, une copie peut dégénérer, diverger fonctionnellement, les deux copies peuvent changer de fonction ou encore les deux copies peuvent se spécialiser pour réaliser chacune différentes sous-fonctions du gène d'origines. Comme toujours en évolution, savoir avec certitude l'histoire évolutive est impossible, Kalhor *et al.* (2023) ont donc classifié le devenir des gènes après leur duplication dans Aevol. Sans entrer dans les détails de leurs résultats, les proportions observées pour les différentes trajectoires post-duplication semblent cohérentes avec les observations en biologie, ce qui suggère que la méthodologie de classification est pertinente et, par la même occasion, que Aevol peut se révéler utile pour tester les outils de la bio-informatique.

5.7 Conclusion

En conclusion de cet historique, deux points saillants apparaissent.

Premièrement, grâce à sa conception basée sur un algorithme de décodage du génome, Aevol permet l'étude de différentes propriétés de l'évolution des génomes qui ne sont pas explicitement incluses dans le modèle. Ainsi, l'existence d'opérons n'est pas codée dans le modèle. Les opérons peuvent apparaître simplement parce que l'algorithme de décodage l'autorise implicitement. Ces propriétés implicites du modèle découlent de la genotype-to-phenotype map et permettent aux organismes simulés d'évoluer sur un fitness landscape comparable à celui d'organismes biologiques, avec des contraintes d'explorations similaires. Cela permet l'émergence, dans le modèle, de propriétés ou de dynamiques qui sont aussi observées dans les organismes biologiques. C'est cette caractéristique forte qui justifie l'utilisation du modèle comme générateur de jeux de tests pour la biologie évolutive et moléculaire. En effet, Aevol permet de simuler des dynamiques émergentes (au sens où elles ne sont pas spécifiées *a priori* mais où elles sont issues des interactions entre mutation et sélection au sein de la population simulée). L'idée que nous exposerons et exploiterons au chapitre IV est d'utiliser cette propriété pour permettre la production de jeux de tests plus stringents que ceux habituellement générés par des simulations plus ou moins ad hoc en évolution moléculaire.

En second lieu, nous avons vu que plusieurs des résultats obtenus jusqu'ici avec Aevol concourent à suggérer que les éléments non-codants du génome sont régulés, sans qu'il ait été possible jusqu'ici d'identifier précisément par quels mécanismes cette régulation opère. Nous avons supposé que la difficulté à identifier ces mécanismes provenait de deux facteurs. D'une part, le faible nombre de répétitions effectuées dans les travaux présentés (généralement 10 par paramètre au maximum), d'autre part, les limites de durées des simulations qui ne permettaient pas aux génomes d'atteindre des états d'équilibre structurels. Nous avons donc utilisé Aevol pour mener une campagne de simulations à très large échelle, à la fois en nombre de générations (14 millions) et en nombre de répétitions (250 par condition expérimentale). Ces expériences permettent une étude sur des génomes ayant atteint une taille d'équilibre (après environ 8 millions de générations) et

donc d'observer les variations de taille du non-codant sans biais lié aux conditions initiales. Le nombre de répétitions permet quant à lui une observation fine de la dynamique du non-codant. Aucune expérience de cette échelle n'avait jusqu'à présent été réalisée avec Aevol, essentiellement en raison de l'important coût de calcul que cela représente. Cependant, des améliorations récentes en termes de performance de calcul et de parallélisation du modèle ont rendu la réalisation de ces expériences possible en un temps raisonnable. En outre, une partie de ce temps a été (déraisonnablement) libéré par un certain virus ... mais ceci est une autre histoire.

Chapitre III

Évolution de la quantité de non-codant dans les génomes bactériens

1 Introduction

Comparativement aux génomes eucaryotes, les génomes bactériens sont de petites tailles. Ils sont par ailleurs compacts au sens où leur densité en gènes dépasse généralement les 70%. De ce fait, les variations de taille de génome entre espèces bactériennes sont principalement expliquées par la quantité et la taille de leurs gènes (Koonin et Wolf, 2008). Cependant, quelle que soit la taille d'un génome bactérien, des séquences non-codantes d'ADN sont toujours présentes, et ce, dans une proportion majoritairement restreinte entre 8 et 15% (Figure III.1)(Rogozin, 2002; Kuo *et al.*, 2009; Giovannoni *et al.*, 2014).

La raison de cette proportion relativement constante de non-codant et même son maintien quasi-systématique reste un phénomène inexpliqué. Ce constat peut même sembler contre-intuitif dans la mesure où l'évolution de la taille des génomes bactériens est généralement qualifiée d'« évolution réductrice » en raison d'un biais apparent à la délétion (Kuo et Ochman, 2009; Wolf et Koonin, 2013). De fait, il est bien établi que les bactéries sont sujettes à un biais mutationnel en faveur des délétions (Mira *et al.*, 2001; Petrov, 2001; Gregory, 2004; Kuo et Ochman, 2009; Sela *et al.*, 2016; Long *et al.*, 2018; Senra *et al.*, 2018), ce qui laisse supposer que, par la simple action de la dérive génétique, des lignées bactériennes sans séquences non-codantes devraient être observées — sous l'hypothèse que ces séquences soient, au plus, neutres d'un point de vue fonctionnel. Cette réduction devrait même être renforcée par la sélection contre les effets faiblement délétères des séquences non-codantes, tels que le coût (ou la vitesse) de réplication ou le fardeau mutationnel (Lynch et Conery, 2003; Lynch, 2006).

Cependant, même si certaines bactéries présentent des fractions de non-codant significativement plus faibles que la moyenne (Giovannoni, 2005; Giovannoni *et al.*, 2014), aucune bactérie totalement dépourvue d'ADN non-codant n'a été observée, ce qui soulève une question : dans un contexte où les principales forces agissant sur les séquences non-codantes semblent tendres à leur élimination, pourquoi les séquences non-codantes sont-elles maintenues chez les bactéries ?

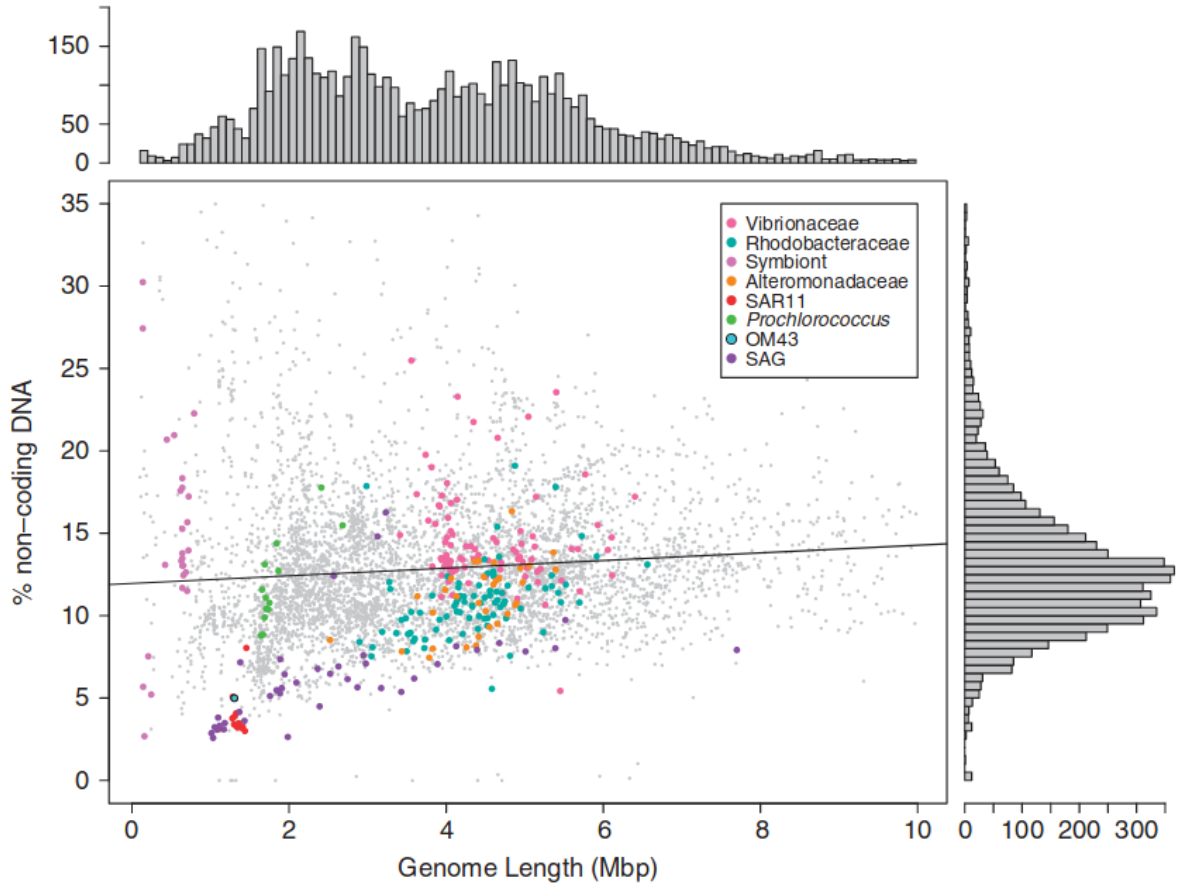


Figure III.1 – Figure provenant de Giovannoni *et al.* (2014). Pourcentage d'ADN non-codant en fonction de la taille des génomes pour 5 689 génomes bactériens publiés dans la base de données IMG 400 et dans Swan *et al.* (2013). Sont surlignés en couleur différents taxa : methylotrophes obligatoires (OM43), symbiontes obligatoires, cyannobactéries marines (*Vibrionaceae*, *Rhodobacteraceae*, *Alteromonadaceae*, SAR11 (*Pelagibacterales*), *Prochlorococcus* et SAG ("single amplified genomes" issues de Giovannoni *et al.*)). Les histogrammes représentent la distribution des points sur les deux axes et montrent que, malgré une distribution de tailles de génomes très large, la distribution des fraction non-codante est au contraire centrée autour de 12%.

En évolution biologique, la tendance « naturelle » est d'invoquer une hypothèse adaptative : « Ces séquences non-codantes n'auraient-elles pas un rôle fonctionnel conférant un avantage face à la sélection naturelle ? ». De nombreuses espèces, bactériennes ou non, présentent en effet des corrélations entre la taille de leur génome (partiellement modulée par le non-codant) et différents traits comme la taille de leurs cellules (Cavalier-Smith, 1982), leur vitesse de réplication (Gregory et Hebert, 1999) ou leur activité métabolique (Vinogradov, 1997). Bien que des cas spécifiques où l'hypothèse adaptative semble convaincante existent Cavalier-Smith (2005); Keeling et Slamovits (2005), dans l'ensemble cette hypothèse ne suffit pas à expliquer l'ensemble des variations observées de tailles de génomes (Lynch, 2006). L'hypothèse adaptative ne semble ici qu'un retour des travers panglossiens de la biologie évolutive, souvent balayables par les mêmes arguments de contraintes allométriques évoqués par Gould et Lewontin (1979) 27 ans avant les objections de Lynch sur le sujet.

Les hypothèses restantes sont donc non adaptatives. Trois d'entre elles se dégagent de la littérature.

La première, l'hypothèse des séquences « égoïstes », suppose que la taille des séquences non-codantes est principalement régulée par l'activité de séquences capables de se répliquer activement au sein des génomes et qui ont de ce fait été qualifiées d'« ADN égoïste » par Orgel et Crick (1980). Ces séquences sont nommées éléments transposables (*Transposable Elements* – TE) chez les eucaryotes et séquences d'insertions (*Insertion Sequences* – IS) chez les bactéries. Il est généralement admis que les éléments transposables sont une composante majeure de l'évolution de la taille des génomes eucaryotes (Kidwell, 2002; Sessegolo *et al.*, 2016; Wells et Feschotte, 2020), chez qui leur présence est presque ubiquitaire (Wells et Feschotte, 2020). Cependant, dans une étude récente chez le genre *Caenorhabditis*, les éléments transposables n'expliquent pas les variations de taille de génomes (Adams *et al.*, 2023). Par ailleurs, cette hypothèse peine à expliquer la présence de séquences non-égoïstes, en particulier chez les bactéries (Lynch, 2006). Enfin, les séquences d'insertions sont absentes chez les bactéries symbiotiques. Or, celles-ci présentent un taux d'ADN non-codant souvent élevé malgré leurs génomes extrêmement réduits (Moran et Plague, 2004). Ces observations remettent en cause le caractère généralisable de l'hypothèse « égoïste ».

La deuxième hypothèse est l'hypothèse de l'équilibre mutationnel (Petrov, 2002). Elle repose sur l'idée que ce sont les taux relatifs entre les différents types de mutations (augmentant ou réduisant le nombre de bases) qui définissent l'équilibre de la taille des génomes. Plus particulièrement, (Petrov, 2002) suggère que les génomes croissent par de rapides explosions dues à de longues insertions ou à l'activité d'éléments transposables, puisque les séquences ainsi insérées sont éliminées progressivement par le biais de petites délétions. Cette hypothèse n'a reçu, depuis, qu'une attention limitée et restreinte à des modèles eucaryotes (Canapa *et al.*, 2015; Mueller et Jockusch, 2018; Loewenthal *et al.*, 2022). Ceci est en grande partie dû à la difficulté d'étudier les séquences non-codantes à partir de données de séquençage « short reads ». Difficulté qui n'est surmontable que depuis récemment grâce à la popularisation des techniques « long reads ». Celles-ci ont d'ailleurs commencé à porter leurs fruits sur l'étude des réarrangements chromosomiques dans les séquences non-codantes (Noureen *et al.*, 2019; Weigand *et al.*, 2019; Seferbekova *et al.*, 2021; Schenk *et al.*, 2022; D'Iorio et Dewar, 2023). En outre, cette hypothèse, qui pourtant repose largement sur des phénomènes de dérive génétique, n'explique pas pourquoi, pratiquement aucun génome bactérien n'a pas de séquences non-codantes.

La troisième hypothèse est l'hypothèse de hasard mutationnel (*Mutational Hazard Hypothesis* – MHH) qui repose sur l'idée que les séquences non-codantes sont un fardeau mutationnel, c'est-à-dire que ces séquences sont des sites potentiels pour des mutations faiblement délétères (par exemple via la dérégulation de gènes voisins ou la création de nouveaux codons d'initiation dont la transcription aurait des effets néfastes). Ces séquences peuvent par conséquent ségréger dans les petites populations où l'efficacité de la sélection est réduite (Lynch et Conery, 2003; Lynch, 2006, 2007) mais pas dans les grandes populations où le seuil de dérive est réduit, provoquant leur élimination. Cette hypothèse repose sur un principe de génétique des populations, la taille effective de population [1](Ne)(Wright, 1931). La taille effective de population Ne est une mesure théorique de l'efficacité de la sélection dans une population idéalisée correspondant aux critères d'applicabilités des modèles de génétique des populations. Une population réelle de taille donnée va se comporter (du point de vue de la sélection) effectivement comme une population théorique d'une autre taille (en général plus faible, même s'il existe des situations où la taille effective est plus grande que la taille « réelle »). Plus Ne est faible, moins la sélection est capable de purger les mutations faiblement délétères, donc les séquences non-codantes selon l'hypothèse MHH. Celle-ci a donc l'avantage de reposer sur un principe général de l'évolution (issu de la théorie de la génétique des populations) et est donc applicable à l'ensemble des organismes, elle a d'ailleurs reçu de nombreux supports depuis son énonciation (Koonin, 2009; Ågren et Wright, 2011; Boussau *et al.*, 2011; Wolf et Koonin, 2013; Lefébure *et al.*, 2017; Lower *et al.*, 2017). Elle présente par contre le défaut de n'offrir aucune explication quant à l'origine de l'accumulation du non-codant lorsque qu'aucune pression à l'augmentation (par exemple les éléments transposables) n'est présente, comme c'est le cas chez certaines bactéries et eucaryotes (Moran et Plague, 2004; Wells et Feschotte, 2020).

Les hypothèses présentées ci-dessus, qu'elles soient adaptatives ou non, ne sont pas mutuellement exclusives, ce qui montre bien la complexité de l'étude de l'évolution de la taille des génomes et des séquences non-codantes. La difficulté à trancher cette question vient en grande partie de la difficulté qu'il y a à concevoir des expériences concluantes. Il serait en effet nécessaire d'étudier l'évolution de la taille du génome d'un organisme en contrôlant un maximum de facteurs impliqués dans les différentes hypothèses formulées dans la littérature (taille efficace, biais mutationnels, facteurs sélectifs...).

Plusieurs expériences d'évolution *in vitro* existent dans lesquelles certains de ces facteurs peuvent être contrôlés, mais elles sont en général dédiées à l'étude des phénomènes d'adaptation à un nouvel environnement et non à l'étude de la structure des génomes (Lenski *et al.*, 1991; Barrick *et al.*, 2009; Lenski, 2017; Couce *et al.*, 2022). En outre, même dans ces expériences *in vitro*, il est impossible de maîtriser – voire de mesurer – l'ensemble des biais mutationnels ou sélectifs. Enfin, les expériences d'évolution *in vitro* sont probablement trop courtes pour observer des variations significatives de la taille des génomes. La « Long Term Evolution Experiment » de Richard Lenski (Lenski, 2017, 2023), dans laquelle 12 réplicats d'*E. coli* sont suivis depuis 1988 (ce qui représente plus de 75

1. Le terme normalement consacré en français est « taille efficace de population » qui est à mon avis une mauvaise traduction historique depuis l'anglais « effective population size », où « effective » peut effectivement être traduit littéralement par « efficace ». Cette traduction a pour défaut que le terme « efficace » ne communique pas, à mon avis, la bonne intuition sur le concept. Je prends donc la liberté d'utiliser ici le terme « effective » qui me semble plus approprié en plus d'être plus transparent.

000 générations à ce jour), est probablement l'une des seules à être assez longues pour permettre l'observation de variations significatives de taille de génome (Tenaillon *et al.*, 2016). Cependant, cette expérience est menée dans des conditions très spécifiques qui permettent probablement l'accumulation de délétions de gènes inutiles (voire délétères) dans le milieu de culture utilisé (milieu « glucose-minimal »). C'est le cas, par exemple, de l'opéron ribose qui est très rapidement perdu dans la majorité des réplicats (Cooper *et al.*, 2001).

Les variations de taille de génome se produisant sur des temps longs qu'il est difficile d'approcher expérimentalement, une solution alternative consiste à comparer des espèces proches et à relier les variations de tailles de génomes aux changements de traits d'histoire de vie (Schaack *et al.*, 2010; Hess *et al.*, 2014; Sessegolo *et al.*, 2016; Lefébure *et al.*, 2017; Tristan *et al.*, 2019). Cependant, outre que, là encore, les biais sont difficilement quantifiables, cette approche peut se révéler ardue en raison des difficultés, déjà évoquées plus haut, concernant les techniques de séquençage. Même si, aujourd'hui, avec les techniques de séquençage les plus récentes, combinées à des développements analytiques et des expériences d'évolution, il est potentiellement possible de démêler les hypothèses évoquées les une des autres (Blommaert, 2020), l'approche reste cependant trop récente pour avoir été mise en application à large échelle.

Face aux difficultés liées à l'étude des variations de taille de génome dans des organismes réels, il peut sembler pertinent de se tourner vers la modélisation pour évaluer les effets relatifs des différentes forces en présence. Ainsi, les modèles mathématiques pourraient sembler être une bonne solution pour étudier l'évolution des génomes, ayant fait leurs preuves dans d'autres domaines de l'évolution comme la biologie des populations ou en épidémiologie où ils sont très largement utilisés (Brauer *et al.*, 2001). Cependant, ces modèles reposent sur des hypothèses souvent très simplificatrices et il est difficile d'intégrer de trop nombreux facteurs. Pour cette raison les approches numériques sont souvent préférées en génomique.

Une autre approche est l'utilisation de l'évolution expérimentale *in silico*, dont le principe est le même que celui de l'évolution expérimentale *in vitro*, mais cette fois en utilisant des organismes virtuels simulés par ordinateur (Hindré *et al.*, 2012). Comme nous l'avons déjà montré dans le chapitre II, cette approche a déjà permis de montrer qu'il existe une dépendance entre la taille des génomes, les taux de mutations et la taille de population, grâce à des expériences sur la plate-forme de simulation Aevol (Knibbe *et al.*, 2007; Carde *et al.*, 2019). Les conditions de ces expériences, en particulier l'absence de biais mutationnels, l'absence de séquences d'insertions et la neutralité fonctionnelle parfaite des séquences non-codantes, permettent de soulever un point clé. À savoir, que des variations non aléatoires de la quantité de séquences non-codantes sont observables dans des conditions où l'hypothèse d'équilibre mutationnel, l'hypothèse des séquences égoïstes et les hypothèses adaptatives sont parfaitement inopérantes. Même si Carde *et al.* montrent que la taille des génomes varie négativement avec la taille de population comme prédit par Lynch et Conery (2003), la MHH, ne semble pas non plus applicable dans Aevol étant donné que des séquences « faiblement » délétères peuvent difficilement s'accumuler dans le non-codant « pour » [1] être ensuite purgées par la sélection. En effet, si l'on considère les exemples de mutations faiblement délétères cité par Lynch (2006). (changements de

1. Finalisme à unique fonction de communication.

régulation depuis des mutations du non-codant ou décalages de cadre de lecture) celles-ci ne s'appliquent pas – ou mal – aux organismes d'Aevol.

L'hypothèse proposée par Knibbe *et al.* (2007) pour expliquer la réduction du non-codant est assez proche de la MHH. Elle repose sur l'observation que les individus dont les génomes comportent plus de séquences non-codantes ont une plus faible fraction de descendants neutres (autrement dit plus de descendants avec des mutations délétères). Ce résultat est interprété au travers du concept de robustesse. La robustesse est en effet la propriété d'un système à maintenir sa fonction malgré des perturbations internes et/ou externes. Elle est considérée comme essentielle aux organismes vivants (Kitano, 2004) en raison de l'omniprésence de telles perturbations. Ici, les génomes avec plus de non-codant seraient donc moins robustes face aux perturbations que sont les mutations accumulées lors de leur réplication. Il est à noter que nous distinguerons ici la robustesse « réplicative », c'est-à-dire la robustesse face à l'ensemble des mutations accumulées lors de la réplication, de la robustesse « mutationnelle » qui caractérise la robustesse face à un et un seul événement mutationnel. Cette distinction est d'importance puisque c'est la robustesse mutationnelle qui est la plus étudiée à l'échelle génomique (de Visser *et al.*, 2003; Wagner, 2005; Fares, 2015). La différence sous-entendue par la perspective de la robustesse réplicative, comparativement à la MHH, est que, à fitness égale, les séquences non-codantes augmentent la probabilité de mutations à fort effet délétère, réduisant ainsi la viabilité des descendants des individus les moins robustes – en l'occurrence ceux dont les génomes comportent le plus de séquences non-codantes. Comme le soulignent Knibbe *et al.*, étant donné que les taux de mutations dans Aevol sont exprimés par paires de bases, l'excèdent de bases non-codantes augmente le nombre de réarrangements chromosomiques aux effets potentiellement plus délétères que des mutations locales.

Comme évoqué plus haut, la MHH ne contient pas de mécanisme explicite poussant à l'augmentation de la taille des génomes lorsque la taille de population diminue. Elle se repose donc sur d'autres hypothèses, non adaptatives, pour l'expliquer (typiquement des séquences égoïstes ou des biais mutationnels). Or, comme nous l'avons évoqué précédemment, ces hypothèses sont inapplicables dans le cas des expériences sur Aevol. Knibbe *et al.* (2007) mettent en évidence la diminution de la taille du non-codant lors de l'augmentation des taux de mutations et émettent l'hypothèse que la quantité de non-codant est régulée par un compromis d'allocation entre robustesse (réplicative) et évolvabilité : l'augmentation du non-codant augmenterait le nombre de mutations par réplication donc la variabilité et la vitesse d'adaptation. Le point d'équilibre du compromis de robustesse/évolvabilité serait donc modifié par le changement des taux de mutation. Cependant, les expériences menées par Carde *et al.* (2019) remettent en cause cette hypothèse. En effet, Carde *et al.* effectuent leurs expériences à partir de populations qui ont évolué pendant un temps long en environnement constant, populations appelées Wild-Type (WT). Cette démarche permet aux populations testées d'atteindre une taille de génome stable. Si cet état d'équilibre était régulé par le compromis robustesse/évolvabilité, la diminution de la taille de population ne devrait pas entraîner une augmentation de la taille de non-codant, comme c'est le cas dans (Carde *et al.*, 2019). La diminution devrait seulement augmenter la variation de cette taille autour de l'équilibre déjà atteint, du fait de la diminution de l'efficacité de la sélection, donc de la capacité à se maintenir sur l'optimum.

En résumé, l'étude de l'évolution des séquences non-codantes est une entreprise com-

plexe dans laquelle démêler l'action des différents facteurs est presque impossible dû à la difficulté de mener des expériences *in vitro* ou *in vivo* concluantes. Des expériences *in silico* sur la plate-forme de simulation Aevol ont pu montrer que, bien que de nombreuses hypothèses existent dans la littérature pour expliquer la régulation du non-codant, aucune d'entre elles ne permet d'expliquer la dynamique observée de ces séquences dans les conditions d'un modèle nul, à savoir, en l'absence de biais mutationnels, en l'absence de séquences égoïstes et en l'absence totale d'effet sélectif des séquences non-codantes (qui sont ici parfaitement neutres). Dans le contexte de ce modèle nul, l'hypothèse de la sélection pour la robustesse réplicative semble bien expliquer la limitation du nombre de bases non-codantes. Néanmoins, le mécanisme à l'origine de l'accumulation des séquences non-codantes reste inconnu. Si on considère les expériences menées jusqu'ici sur Aevol, les raisons pour lesquelles elles ne permettent pas de conclure sont diverses, allant de difficultés liées à la durée des expériences, par exemple dans (Knibbe *et al.*, 2007), à des plans expérimentaux orientés vers la réduction des génomes et ne permettant donc pas d'identifier les mécanismes d'accumulation de non-codant (Carde *et al.*, 2019). En outre, pour des raisons computationnelles, le nombre de réplications est souvent resté trop limité. Nous proposons donc ici une étude expérimentale permettant d'examiner plus en profondeur la dynamique des séquences non-codantes dans Aevol. Pour ce faire, nous avons conduit des expériences en temps long (14 millions de générations au total) et sur un très grand nombre de réplicats (250). Nous verrons que l'ampleur de ces expériences – inédite sur ce modèle – nous permet d'identifier précisément les mécanismes de régulation de l'ADN non-codant dans le modèle.

Les résultats présentés dans ce chapitre sont l'objet d'un article en cours de rédaction et destiné à être soumis à une revue de biologie [1]. C'est pourquoi ils sont présentés ici en suivant le plan classique d'un article de revue : Matériel et Méthode (section 2), Résultats (section 3) et Discussion (section 4). Pour vérifier les hypothèses formulées au cours de ce travail, nous avons comparées ces dernières à un modèle mathématique développé spécifiquement. Ce modèle ayant été développé par Paul Banse, il n'est expliqué que brièvement dans le présent chapitre (section 2.6), et est présenté entièrement en Annexes VI de ce manuscrit.

1. Marco Foley, Victor Lezaud, Paul Banse, Jonathan Rouzaud-Cornabas, Guillaume Beslon (2023) The Danaïds genome — Duplications/deletions neutrality bias maintains non-coding sequences in genomes despite constant leak, *In prep.*

2 Matériel et Méthodes

2.1 Wild-Types

Classiquement, dans Aevol, les expériences sont initialisées avec des individus clonaux générés par bootstrap. Pour ce faire, des génomes de 5 000 paires de bases sont testés jusqu'à trouver un génome comportant au moins un gène favorable et c'est cet individu qui va être utilisé pour ensemencer la population (voir chapitre II). Bien que simple d'utilisation, cette approche n'est pas du tout adaptée à l'étude de la variation de taille du non-codant. En effet, partant d'un organisme quasiment naïf, l'évolution va prioritairement agir sur les éléments codants et le non-codant va donc principalement varier par auto-stop. C'est d'ailleurs ce qui est observé dans les premières phases de l'évolution : l'accumulation de séquences codantes (par duplication de gènes) entraîne une croissance très rapide du non-codant, car celui-ci est dupliqué en même temps que les gènes. C'est pourquoi, dans l'étude présentée ici, nous initialiserons les expériences avec des individus « Wild-Types » ayant évolué suffisamment longtemps pour que leur génome soit structurellement stable.

Pour construire ces Wild-Types, nous avons laissé évoluer dix populations de 1 024 individus pendant 10 millions de générations en environnement constant [1] (le même pour les 10) avec des taux de chaque type de mutation de 10^{-6} mutations par paire de bases par génération (pour chacun des types de mutations simulés dans Aevol). En pratique, les simulations ont été prolongées de 100 000 générations supplémentaires afin d'isoler les individus coalescents de la génération 10 millions (autrement dit l'ancêtre commun à la 10 millionième génération des individus de la génération 10 100 000). Cet isolement des coalescents permet d'assurer l'ensemencement de nouvelles populations à partir d'individus ayant été effectivement sélectionnés. Nous avons ensuite sélectionné les 5 wild-types les plus proches de la médianne (en termes de taille de génome) afin d'éviter d'initialiser nos simulations avec des « outliers ». Sur les 10 simulations initiales, ces 5 médians étaient les Wild-Types WT0, WT1, WT3, WT4 et WT5. Ce sont ces 5 coalescents que l'on nommera les Wild-Types (WTs) par la suite. Le Tableau III.1 présente les caractéristiques génomiques des 10 Wild-Types simulés.

2.2 Expériences contrôle

À partir de chacun des 5 WTs, nous avons ensemencé 50 nouvelles populations (250 populations au total). L'ensemencement consiste à créer une population clonale à partir d'un WT, ceci permettant d'assurer que tous les réplicats d'une expérience sont initialisés dans les mêmes conditions [2]. Ces 250 réplicats ont ensuite évolué pendant 4 millions de générations supplémentaires dans les mêmes conditions (environnement, taille de population, taux de mutation) que celles des WTs et nous avons mesuré les variations de taille de génome, de taille codante et de taille non-codante au cours de l'expérience.

Afin de caractériser la dynamique d'évolution du non-codant dans les 250 réplicats,

1. Des expériences préliminaires nous ont montré que, dans les conditions évolutives utilisées, la taille des génomes se stabilise après 6 à 8 millions de générations.

2. Initialiser des réplicats à partir de copies de populations entières pourrait, à cause d'effets stochastiques, mener à des résultats inconsistants dû à des effets fondateurs.

Wild-Type	Taille génome (pb)	% non-codant	Nombre d'ARN	Nombre de Gènes
WT0	**13 599**	**47,4**	**66**	**148**
WT1	**13 660**	**50,6**	**55**	**141**
WT2	7 792	21.5	74	99
WT3	**14 171**	**49,7**	**60**	**137**
WT4	**14 507**	**48,7**	**82**	**168**
WT5	**14 290**	**41,8**	**80**	**206**
WT6	16 811	40,3	106	132
WT7	19 126	35,0	133	196
WT8	16 077	33,5	88	167
WT9	16 968	33,2	169	175

TABLE III.1 – Caractéristiques génomiques des dix wild-types. Les cinq wild-types les plus proches de la médianne de taille de génome (14398.5), sont ceux utilisés pour les expériences (en gras).

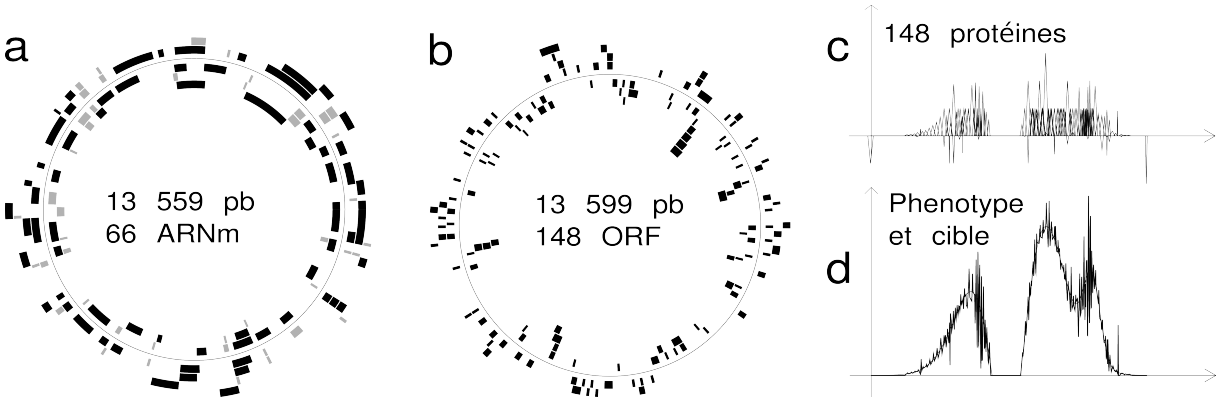


FIGURE III.2 – Exemple de Wild-Type (WT0). (a) Génome (cercle en trait fin) et ARNs (segments noirs : ARN codants, segments gris : ARN non-codants). (b) Génome et Open-Reading Frames (ORFs, segments noirs). (c) Protéines activatrices (activité supérieure à zéro) et inhibitrices (activité inférieure à zéro) encodées par les 148 ORFs (voir chapitre II). (d) Phénotype et fonction cible.

nous avons calculé la *Mean-Square Displacement* (MSD), à savoir la variation, au cours des générations, de la moyenne des écarts à l'origine au carré. Cet indicateur, classique en physique, permet de caractériser les processus de diffusion : il est en effet supposé varier de façon linéaire lors d'une diffusion non-contrainte (hypothèse nulle ici). Une MSD infra-linéaire ou supra-linéaire caractérise respectivement une diffusion sous- ou sur-diffusive. Enfin, si, au bout d'un certain temps, la MSD devient constante, cela permet de caractériser une diffusion bornée.

2.3 Variation de taille de population

Pour tester l'effet de la taille de population, nous avons mené des expériences d'évolution analogue à l'expérience contrôle mais pour deux conditions de test différentes, une taille de population divisée par 4 (256 individus) et une taille de population multipliée

par 4 (4 096 individus). L'expérience d'évolution précédente avec une population de 1 024 individus servira ici de condition contrôle. Pour chaque condition test, comme pour le contrôle, nous avons ensemencé 50 nouvelles populations à partir de chacun des 5 WTs et les populations ont évolué pendant 4 millions de générations avec les mêmes taux de mutation et dans le même environnement. Le choix des tailles de populations testées est un compromis entre trois contraintes : la maximisation de la modification de l'effet de la sélection, la limitation du temps de calcul en grande population et la forme de l'environnement spatial dans Aevol qui impose que les tailles de population soient des carrés parfaits ($256 = 16^2$, $1024 = 32^2$ et $4096 = 64^2$).

2.4 Expérience d'accumulation de mutations neutres

Parmi l'arsenal d'outils utilisés en évolution expérimentale, l'expérience d'accumulation de mutations en est un particulièrement efficace, qui permet de partiellement s'affranchir du biais du survivant, si prépondérant en évolution. Dès 1964, Mukai (1964) propose pour étudier les mutations, qu'il qualifie de « matériaux bruts » de l'évolution, une expérience permettant de ne sélectionner qu'un seul reproducteur et ainsi d'obtenir des lignées ayant accumulé des mutations non filtrées par la sélection naturelle. Cette méthode reste aujourd'hui le moyen le plus fiable d'étudier les taux de mutation spontanée et l'évolution neutre (Halligan et Keightley, 2009).

Ici, pour tester l'effet de la robustesse sur l'évolution de la taille des génomes, nous avons conçu une expérience proche de l'expérience d'accumulation de mutations, mais permettant de supprimer la sélection indirecte pour la robustesse sans pour autant accumuler de mutations délétères. Nous avons appelé cette expérience l'« accumulation de mutations neutres ». Le principe de l'expérience est le même, à savoir générer une lignée accumulant des mutations, mais avec une contrainte supplémentaire permise par la simulation qui est de ne conserver que les mutations neutres et de rejeter tous les descendants portant des mutations délétères ou favorables. D'un point de vue algorithmique, le principe est simple, il suffit de générer un descendant à partir d'un individu : si ce descendant est neutre, le processus est répété à partir du descendant, si le descendant n'est pas neutre un nouveau descendant est généré à la place. Le processus est répété jusqu'à avoir le nombre de générations souhaitées. Cette expérience permet des conditions d'évolution très particulière, que sont, une sélection purificatrice absolue [1] et une dérive génétique totale sur le non-codant ce qui supprime les effets de sélection pour la robustesse.

2.5 Estimation de l'impact de la quantité de non-codant

Afin de tester l'effet évolutif du non-codant, nous avons utilisé les cinq WTs pour créer des individus de phénotypes identiques, mais dont les tailles de non-codant sont parfaitement contrôlées. Pour cela, nous avons développé un nouvel outil dans Aevol permettant

1. On notera cependant que, même si elles sont *a priori* rares, des mutations modifiant les séquences codantes sans impacter le phénotype sont possibles dans Aevol comme dans des organismes réels. Par exemple, dans Aevol des mutations « synonymes » à l'échelle des gènes sont possibles même en l'absence de redondance du code génétique. Il est aussi possible de créer des chevauchements de gènes sur des cadres de lectures différents, ce qui peut modifier la taille du codant de façon neutre.

de faire varier la taille du non-codant d'un individu sans modifier son phénotype ni, par conséquent, sa fitness.

Le post-traitement de modification neutre de la taille des génomes fonctionne de manière très similaire au processus d'accumulation de mutations neutres, mais en ajoutant une condition supplémentaire. À partir d'un individu et d'une taille cible, le post-traitement génère un descendant (par une réplication classique, donc incluant possiblement des mutations). Si ce descendant est neutre et que la taille de son génome le rapproche de la taille cible (par rapport à son ancêtre), alors le descendant est conservé et le processus est répété jusqu'à atteindre la taille cible. Sinon, le descendant est rejeté et le processus est répété jusqu'à trouver un descendant admissible.

Ce processus permet d'atteindre la taille cible. Cependant, une fois celle-ci atteinte, 50 000 générations supplémentaires sont simulées afin que la distribution de la taille des séquences intergéniques soient comparables à celles observées au cours d'expériences d'évolution standard, qui sont largement influencés par l'effet des inversions (Biller *et al.*, 2015). Ce processus a l'avantage de générer des génomes à structures variables ayant été soumis aux mêmes pressions mutationnelles que des génomes issus d'évolution classique et donc d'avoir des structures comparables. Le principal désavantage est que la taille du codant peut varier marginalement sous l'effet de mutations synonymes. Néanmoins, cette méthode s'est révélée plus efficace à la réduction des génomes que des méthodes itératives retirant les bases non-codantes une par une.

À partir de ces individus « génétiquement modifiés », nous avons estimé l'influence du non-codant sur la robustesse réplicative (toutes mutations confondues) et sur la robustesse mutationnelle pour les différents types de mutation.

Dans l'introduction, nous avons défini la robustesse réplicative comme la capacité d'un génome à maintenir son phénotype associé, après un évènement de réplication. En pratique, la robustesse réplicative détermine la probabilité qu'a un individu de produire un descendant avec le même phénotype que le sien, autrement dit la probabilité de produire un descendant neutre. Les deux façons de produire un descendant neutre sont, soit d'avoir un descendant qui n'a pas subi de mutations, soit d'avoir un descendant ne portant que des mutations neutres. La robustesse mutationnelle est donc une composante de la robustesse réplicative, car c'est elle qui détermine la probabilité de mutation neutre.

Dans Aevol, nous pouvons estimer numériquement la robustesse d'un individu (ou d'un génome) en générant un très grand nombre de descendants, parmi lesquels nous pouvons comptabiliser le nombre d'occurrences neutres. Cette estimation pouvant rapidement devenir coûteuse en temps de calcul si appliquée sur de nombreux génomes, nous avons restreint son calcul à un nombre limité de génomes, mais couvrant une plus large taille de non-codant que dans les expériences précédentes. Nous avons effectué ces estimations sur les 5 WT déjà présentés, mais en modifiant leur taille de non-codant par post-traitement (voir ci-dessus). Nous avons modifié les génomes des 5 WTs en génomes comportant 21 tailles de non-codant différentes, allant de 0 pb (paires de bases) à 20 000 pb toutes les 1 000 pb, avec 50 réplicats par taille et par WT [1]. Pour chacun des 5 250 génomes (5 WTs x 21 tailles x 50 réplicats) nous avons généré un million de descendants pour estimer la

1. Comme énoncé ci-dessus, les génomes résultant du post-traitement peuvent avoir des structures légèrement différentes, notamment une distribution de la taille du non-codant variable, d'où la nécessité de faire des réplicats afin de diminuer le bruit sur l'estimation de la robustesse.

robustesse réplicative.

Enfin, parmi ce million de descendants, nous avons isolé ceux ne comportant qu'une seule mutation et nous avons calculé, pour chaque type de mutation, la probabilité de neutralité (robustesse mutationnelle) et, pour les mutations modifiant la taille du génome, la taille moyenne des mutations neutres. Ces deux paramètres (neutralité mutationnelle et taille moyenne des mutations neutres) nous permettent d'évaluer la contribution de chaque type de mutation aux variations neutres de la taille des génomes.

2.6 Modèle mathématique

Afin de valider nos analyses et de vérifier que les deux forces identifiées dans les simulations expliquent bien les variations de non-codant, nous avons construit un modèle mathématique de ces deux forces et nous avons comparé les résultats de ce modèle avec les données de robustesse et de biais estimées pour les génomes de taille variable. Le modèle mathématique a été défini par Paul Banse et un article est en cours de rédaction (en collaboration avec Juliette Luiselli et Olivier Mazet [1]. Ce modèle n'ayant donc pas été développé dans le cadre de cette thèse – seulement utilisé – il ne sera pas décrit ici, mais simplement présenté dans l'annexe VI. En quelques mots, il s'agit d'un modèle probabiliste permettant de calculer, pour tous les types de mutations, la probabilité de neutralité d'une mutation en fonction de la structure du génome (longueur totale, nombre de séquences sécables, taille moyenne des séquences sécables et fraction non-codante). Nous utilisons ici le terme de « séquence sécable » car la définition d'une séquence codante (donc « insécable ») dans le modèle mathématique est structurelle et diffère trop de la définition d'un gène (ou d'un ARN) pour pouvoir être comparable. Une séquence insécable est en effet une séquence qui ne peut pas être coupée par une délétion, une inversion ou une translocation, ni par l'insertion d'un segment dupliqué ou déplacé. Ainsi, deux gènes chevauchants (sur le même brin ou sur les brins complémentaires) forment un seul segment insécable. Par construction, le génome de Aevol étant circulaire, le nombre de séquences sécables et le nombre de séquences insécables sont identiques (ce qui n'est pas le cas pour le nombre de gènes et le nombre de séquences non-codantes en raison des chevauchements). Notez que, dans ce qui suit et dans la description du modèle mathématique, on utilisera parfois le terme de « gène » pour désigner les séquences insécables et de « séquence non-codante » pour désigner les séquences sécables. Cet abus de langage ne nuit en effet pas à la compréhension et permet de considérablement alléger la lecture.

Pour pouvoir utiliser le modèle mathématique, il est nécessaire d'estimer, pour chaque génome, les valeurs des trois variables d'entrée du modèle : la taille du codant (γ), la taille du non-codant (λ) et le nombre de séquences insécables (g). Les deux premières valeurs sont mesurées par Aevol et sont donc déjà connues. Le nombre de gènes est aussi connu mais comme nous venons de le préciser, il peut différer du nombre de séquences non-codantes. En outre, le nombre de ces segments non-codants (ou insécables) est susceptible de varier selon le degré de compactions des génomes (*i.e.* selon le pourcentage de codant). De fait, il paraît naturel que, lorsque la fraction codante est très fortement réduite (en utilisant la procédure décrite ci-dessus), le nombre de gènes chevauchant augmente et

1. Paul Banse et Juliette Luiselli sont tous deux doctorants dans l'équipe Beagle. Olivier Mazet est maître de conférences à l'INSA de Toulouse

donc que le nombre de séquences sécables diminue.

Afin de pouvoir comparer les estimateurs mathématiques et numériques, nous avons développé un post-traitement permettant d'estimer expérimentalement le nombre de segments sécables pour les 5 250 génomes testés (5 WTs x 21 tailles x 50 réplicats). En effet, la complexité structurale et de décodage des génomes dans Aevol sont telles que ce calcul, qui semble n'être qu'un simple décompte, n'est en fait pas trivial. Bien que ce fait semble négligé en biologie évolutive, on notera en particulier que la notion de neutralité d'une base est tout sauf triviale. En effet, une base peut être neutre par substitution, mais pas par insertion (par exemple dans le cas de la séquence de taille fixe, mais de composition quelconque séparant le RBS du codon START).

Le processus retenu pour ce post-traitement est le suivant : entre chaque base d'un génome sont insérés indépendamment deux terminateurs [1] symétriques. Si dans un des deux cas le phénotype n'est pas modifié, alors les deux bases entourant la position d'insertion sont considérées comme insécables. Un ensemble de bases contiguës appartenant à une séquence insécable est alors comptabilisé comme un segment insécable. On notera que, selon cette définition, un segment est considéré comme sécable s'il est formé d'au moins deux positions sécables contigües (la longueur du segment sécable étant égale à $N_s - 1$, où N_s est le nombre de positions sécables qu'il contient).

Une fois le nombre de séquences sécables estimé pour tous les génomes, nous avons utilisé une régression linéaire afin d'obtenir le nombre moyen de séquences insécables en fonction de la taille du non-codant. Les valeurs issues de cette régression linéaire sont ensuite injectées dans le modèle mathématique.

1. Pour rappel, les terminateurs dans Aevol sont des séquences palindromiques de quatre bases sans queue PolyA, ils sont donc terminateurs sur les deux brins.

3 Résultats

3.1 Évolution des tailles de génome dans les expériences contrôle

Dans cette première série d'expériences, nous avons utilisé cinq wild-types pré-évolués (voir section 2.1) que nous avons clonés 50 fois chacun et laissé évoluer de nouveau pendant 4 millions de générations dans des conditions strictement constantes et identiques à celles des wild-types. La Figure III.3 présente l'évolution des caractéristiques génomiques de ces 250 répétitions au cours des 4 millions de générations.

Le premier constat de ces expériences est que, même si nous observons d'importantes fluctuations, la taille des génomes est globalement constante pour chaque wild-type. Ce résultat confirme que les génomes de nos wild-types étaient bien stables en début d'expérience. Ensuite, nous constatons que la taille du codant est presque constante au cours des 4 millions de générations d'évolution, ce qui est attendu étant donné que les WTs sont déjà très adaptés à leur environnement. Le codant est donc principalement contraint par la sélection purificatrice, même si, comme on peut le voir sur la Figure III.4, les populations gagnent toujours en fitness. Ces gains semblent cependant trop faibles pour provoquer des variations notables du codant. Les fluctuations observées de taille de génomes sont donc majoritairement expliquées par la variation de la quantité de non-codant au cours de l'expérience. On notera que, pour certains wild-types, une variation marquée de la moyenne de taille des génomes est visible au début de l'expérience (à la baisse pour le WT1, à la hausse pour les WT3 et 4). Ces variations s'expliquent par l'initialisation clonale des expériences. En effet, il est possible que certains wild-types aient été relativement loin de l'état moyen lorsqu'ils ont été extraits à la génération 10 millions (voir section 2.1). Dans ce cas, il est tout à fait normal d'observer un retour à la moyenne dans les 50 simulations ensemencées avec ces wild-types.

Deuxièmement, on constate que, pour les 250 réplicats et au cours des 4 millions de générations, une quantité minimale de non-codant est toujours maintenue dans les génomes, et ce, sans même que cette quantité s'approche de zéro. De plus, si l'on observe la dynamique générale du non-codant, les variations semblent même comprises dans un intervalle restreint, ce qui suggère que, dans Aevol, le non-codant est régulé alors que *a priori* aucune force connue n'est susceptible de conduire à ce résultat, comme nous l'avons déjà expliqué dans le chapitre II et en introduction de ce chapitre.

Pour mieux caractériser cette observation, nous avons calculé la moyenne des carrés des déviations (MSD pour *Mean Square Displacement* en anglais) à la taille initiale (Figure III.5) sur l'ensemble des 250 réplicats. La MSD est en effet une mesure classique permettant de caractériser les processus de diffusion en physique statistique. Or, ici, en l'absence de force sélective ou de biais mutationnel, la quantité d'ADN non-codant est censée être uniquement soumise à la dérive, ce qui revient à un processus de diffusion en 1D dans l'espace des tailles de génomes. Un tel processus devrait conduire à une MSD à croissance linéaire. De fait, au cours de 350 000 premières générations, la MSD croît linéairement en fonction du temps. Cependant, à partir de la génération 350 000 environ, on observe un ralentissement de la MSD qui atteint un plateau aux alentours de la génération 1 million avant même de redescendre à partir de la génération 2 millions (Figure III.5). Ce comportement est caractéristique d'une marche aléatoire bornée dans les

deux sens (vers les petites tailles de non-codant et vers les grandes tailles de non-codant) : tant que les bornes ne sont pas atteintes, la diffusion suit un comportement normal. À partir du moment où quelques-unes des 50 trajectoires atteignent une borne (supérieure ou inférieure), ces trajectoires ne suivent plus une diffusion normale et la MSD ralentit. Lorsque toutes les trajectoires subissent le confinement, la MSD atteint son plateau et la redescente s'explique simplement par un effet mémoire, lié au rebond sur les bornes, qui a tendance à créer des corrélations à très long terme sur le comportement des 50 trajectoires.

En conclusion, les génomes partant d'une même taille initiale commencent par varier de façon neutre, étant donné le codant constant et la marche aléatoire du non-codant. Puis les tailles de génomes atteignent des bornes, signe de contraintes évolutives. L'origine de la borne inférieure étant encore non identifiée, commençons par nous focaliser sur la borne supérieure. Comme énoncé par Knibbe *et al.* (2007), l'hypothèse expliquant l'origine de la borne supérieure est que la sélection pour la robustesse réplicative favorise les génomes plus petits. Cette hypothèse étant sélective (quoique indirecte), la méthode simple et directe pour la tester est de modifier l'intensité de la sélection. En effet, une sélection plus forte devrait entraîner une augmentation de la pression à la robustesse et donc une réduction de la quantité de non-codant. Inversement, une sélection plus faible devrait au moins augmenter l'intervalle de fluctuation du non-codant. Pour modifier l'intensité de la sélection, nous avons donc simulé des expériences en modifiant la taille de population.

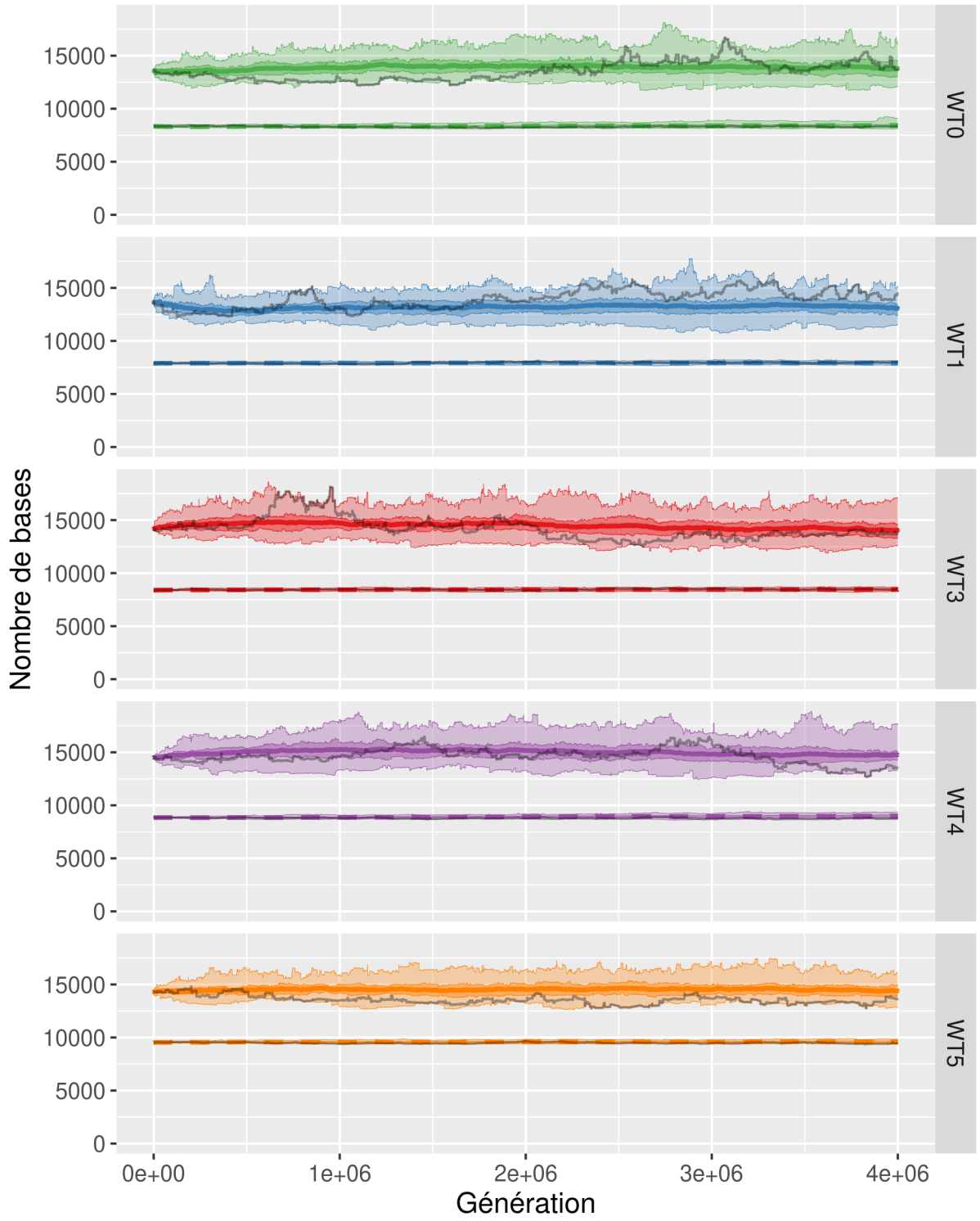


Figure III.3 – Pour 50 réplicats des 5 WTs, évolution de la moyenne de la taille des génomes (traits pleins en gras) et de la taille du codant (traits pointillés en gras) pendant 4 millions de générations avec leurs intervalles inter-quartiles (ruban sombre) et bornes minimales et maximales (ruban clair) respectifs. En traits gris, les valeurs d'un réplicat pris au hasard. Au vu de la stabilité du codant, la différence entre la taille du codant et la taille totale des génomes permet de visualiser la variation de la taille totale du non-codant en moyenne (traits pleins en gras) ou sur une trajectoire type (traits gris).

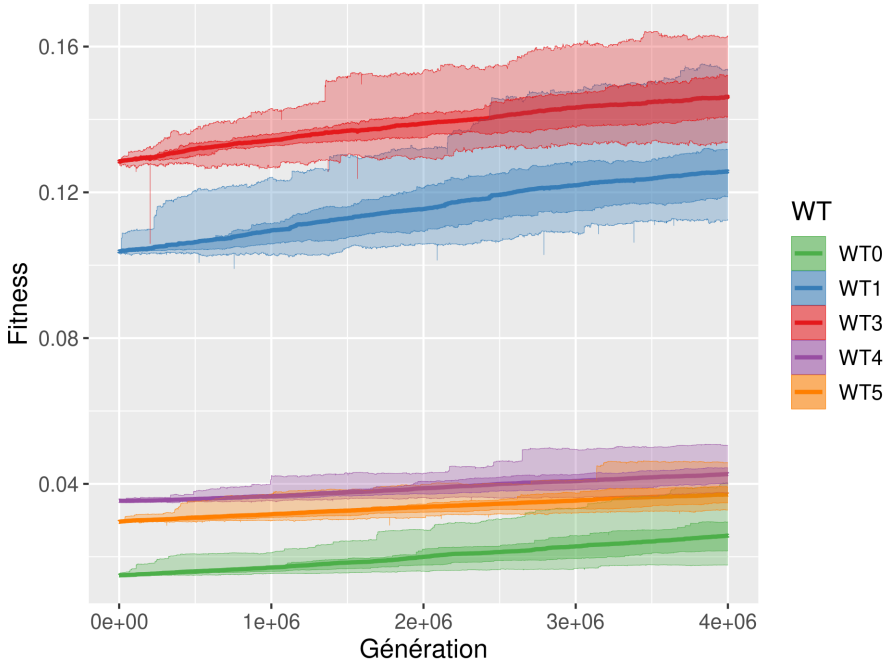


FIGURE III.4 – Pour 50 réplicats de 5 WTs, évolution de la moyenne des fitness (traits pleins en gras) pendant 4 millions de générations avec l'écart inter-quartiles (ruban sombre) et les bornes minimales et maximales (ruban clair).

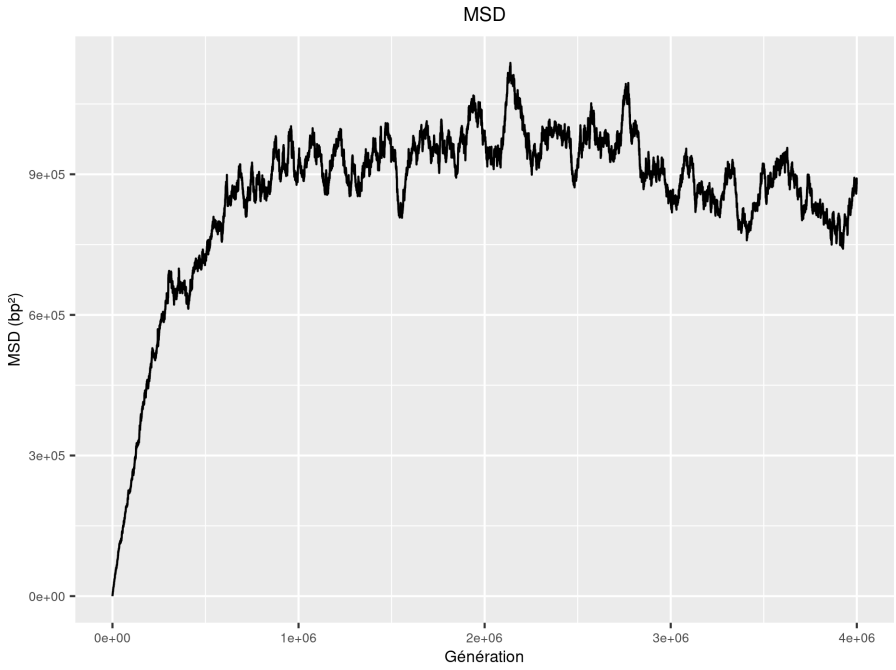


FIGURE III.5 – Moyenne des carrés des déviations (MSD) à la taille initiale des génomes, pour 250 réplicats (50 par WT) sur 4 millions de générations. La MSD attendu pour une marche aléatoire est linéaire en fonction du temps. Une MSD avec un plateau est caractéristique d'un mouvement borné.

3.2 Expérience sur les tailles de population

Les résultats des simulations sous trois tailles de population différentes (256, 1 024 et 4 096 individus) montrent clairement que la taille des génomes est inversement corrélée à la taille de la population : plus la population est grande, plus la taille des génomes est petite et inversement (Figure III.6). En outre, cet effet est parfaitement répétable pour chaque WT, dans les mêmes proportions, avec un intervalle de confiance très inférieur à l'effet de la taille de population.

Comme le montre la Figure III.7, cette variation de la taille des génomes est expliquée par la variation du non-codant. L'effet, bien que beaucoup plus faible, est même inversé sur le codant puisque la quantité de codant apparaît, elle, positivement corrélée à la taille de population. Ce résultat est attendu : une population plus grande permet une exploration plus rapide de l'espace génomique et l'augmentation de la sélection permet de retenir des séquences bénéfiques avec de plus faibles effets sur la fitness. Inversement, la diminution de la taille de population fait basculer les séquences à plus faible effet sur la fitness sous la barrière de dérive génétique et celles-ci sont alors éliminées par un effet de type cliquet de Muller. Cet effet sur le codant permet de confirmer une augmentation (respectivement diminution) de l'effet de la sélection lors de l'augmentation (respectivement diminution) de la taille de population. L'augmentation de l'intensité de la sélection entraîne donc bien une diminution de la quantité de séquences non-codantes tandis que la baisse de l'intensité de la sélection entraîne leur augmentation.

Les résultats de cette expérience confirment bien que la sélection favorise des génomes avec moins de séquences non-codantes. Comme ces séquences sont parfaitement neutres dans Aevol, cet effet ne peut être dû qu'à un mécanisme de sélection indirecte, possiblement pour la robustesse comme proposé par Knibbe *et al.* (2007). Afin de pouvoir valider cette hypothèse, il reste cependant à vérifier deux points clés. Premièrement que la taille des génomes et, surtout, la quantité de non-codant, influencent bien la robustesse réplicative, et ce, dans le bon sens (donc vérifier que les grands génomes sont *moins* robustes que les petits, même si la différence de taille n'est liée qu'à des séquences non-codantes). Deuxièmement, que la différence de robustesse est bien susceptible d'entraîner des différences de sélection indirecte suffisantes pour avoir un effet sur l'évolution des tailles de génomes, donc sur le non-codant.

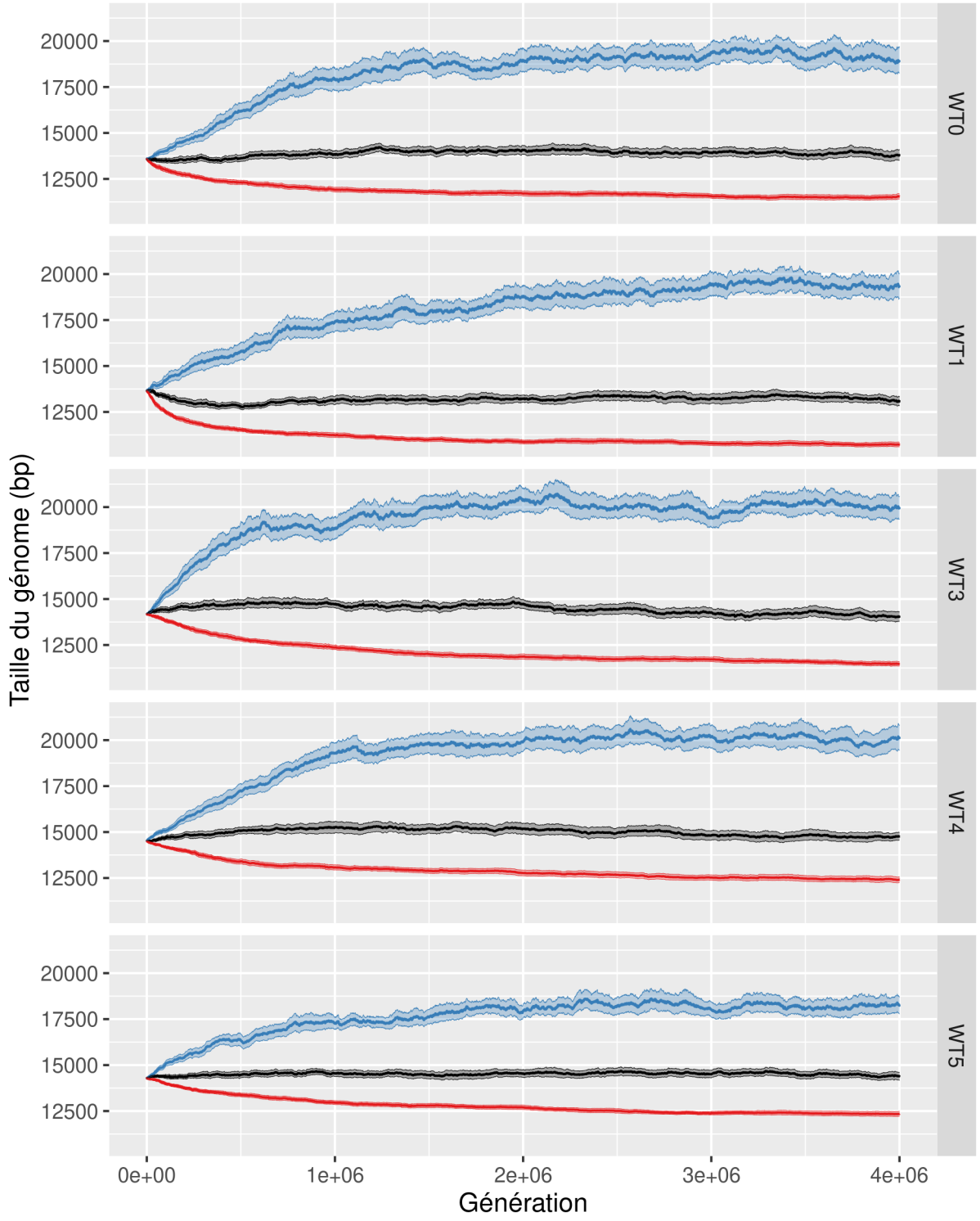


FIGURE III.6 – Évolution de la taille moyenne des génomes pour la taille de population contrôle (en noir, $n = 1024$), la taille de populations divisée par 4 (en bleu, $n = 256$) et multipliée par 4 (en rouge, $n = 4096$). Les bandeaux autour des moyennes correspondent aux intervalles de confiances à 95% calculés à partir de 10 000 bootstraps. Taille d'échantillon par moyenne : 50.

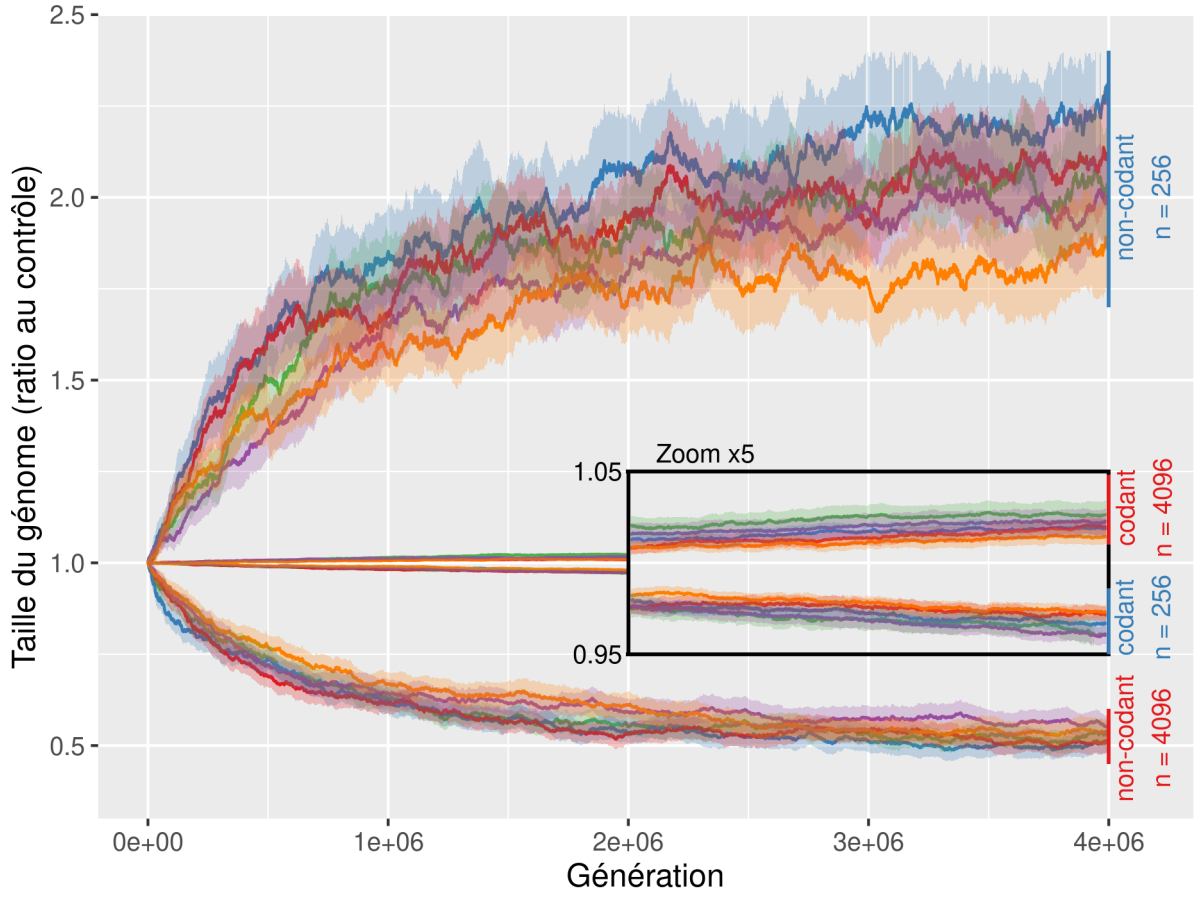


Figure III.7 – Pour chaque WT, variation relative (par rapport aux expériences contrôle, $n = 1024$), de la taille moyenne du codant (courbes au centre et dans la partie zoomée) et de la taille moyenne du non-codant (courbes supérieures et inférieures), pour des tailles de populations divisées par 4 ($n = 256$, barres bleues à droite) et multipliées par 4 ($n = 4096$, barres rouges à droite). Les bandeaux autour des moyennes correspondent aux intervalles de confiances à 95% calculés à partir de 10 000 bootstraps. Taille d'échantillon par moyenne : 50. Les couleurs correspondent aux différents wild-types (vert : WT0 ; bleu : W1 ; rouge : WT3 ; violet : WT4 ; orange : WT5).

3.3 Relation entre taille des génomes et robustesse

Afin de mieux comprendre comment la quantité de non-codant influe sur la robustesse réplicative, nous avons modifié les génomes des 5 wild-types pour générer 5 250 nouveaux génomes caractérisés par des tailles de non-codant définies (de 0 pb à 20 000 pb avec 50 réplicats par wild-type et par taille de non-codant – voir section 2.5). Ces 5 250 génomes vont nous permettre de quantifier la relation entre robustesse réplicative et quantité de non-codant, mais aussi de décomposer la robustesse réplicative en ses différentes composantes mutationnelles pour comprendre l'origine de la relation (voir Matériel et Méthodes, section 2.5 pour la méthode d'estimation de la robustesse réplicative et de la robustesse mutationnelle).

La Figure III.8 montre une claire corrélation négative entre la quantité de non-codant et la robustesse réplicative (c'est-à-dire la fraction de descendants neutres attendus pour

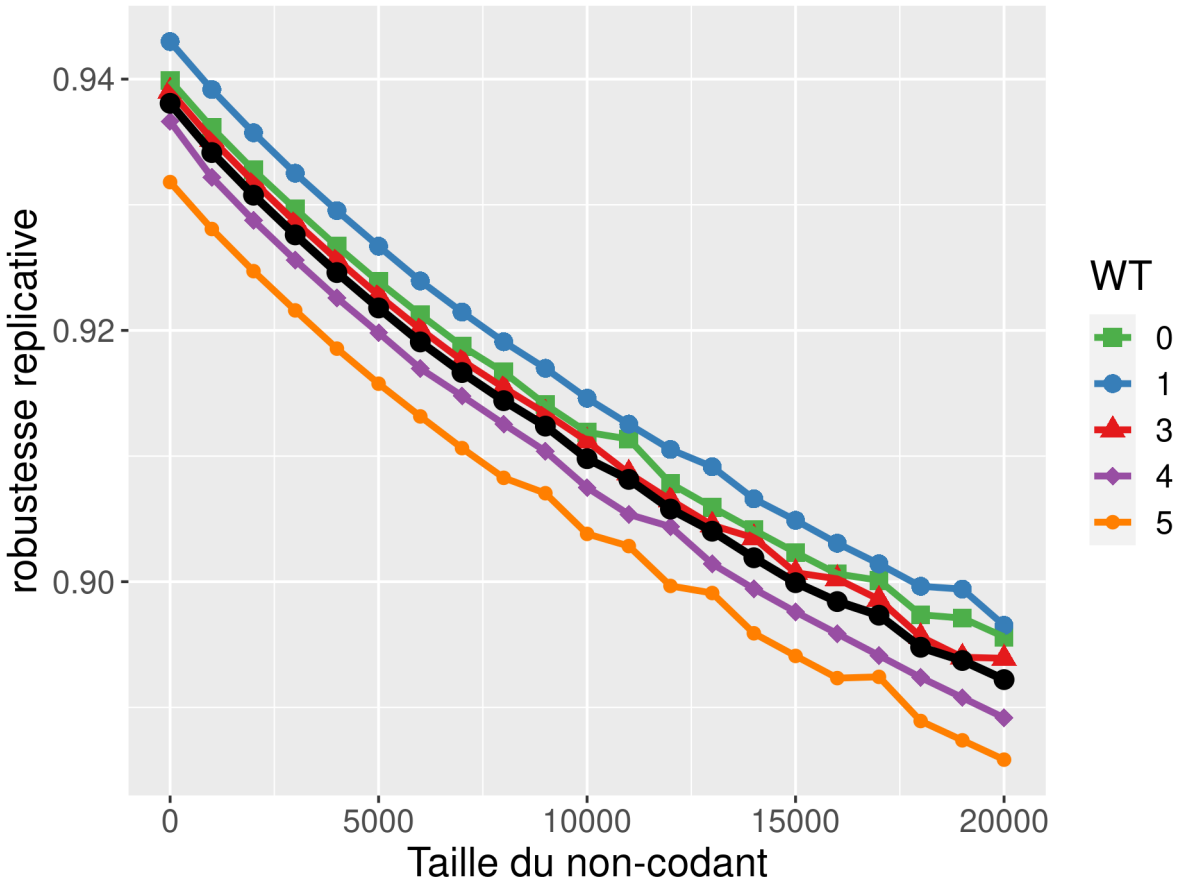


FIGURE III.8 – Robustesse réplicative en fonction de la quantité de non-codant pour en fonction de la quantité de non-codant pour les génomes modifiés à partir des 5 WTs. Chaque point en couleur est une moyenne sur 50 estimations indépendantes faites à partir d'un million de descendants. La moyenne globale est en noire.

un génome donné). Ce résultat contraste singulièrement avec les variations de robustesse mutationnelle pour les mêmes génomes (Figure III.9). En effet, comme l'intuition le suggère naturellement, l'ajout de bases non-codantes augmente la neutralité lorsque le génome subit une et une seule mutation (puisque celle-ci a plus de chance de toucher le non-codant, protégeant ainsi indirectement le codant). Ce paradoxe apparent s'explique pourtant assez simplement. En effet, si l'ajout de bases augmente la robustesse mutationnelle, il augmente aussi le nombre moyen de mutations par réplication, donc la probabilité de mutations délétères par réplication. Or, la robustesse réplicative dépend précisément de ces deux facteurs (robustesse mutationnelle et nombre moyen de mutations par réplication). De plus, l'augmentation du nombre de mutations est linéaire avec la taille du génome tandis que l'augmentation de la robustesse mutationnelle est sous-linéaire (voir Figure III.9). La conséquence de ces deux tendances est que la relation entre robustesse réplicative et non-codant est décroissante et non linéaire.

Si l'on regarde plus en détails la robustesse mutationnelle par type de mutations (Figure III.10), nous pouvons constater que la robustesse aux mutations locales (switch et InDels) est bien supérieure à celle des réarrangements. Cet effet est attendu (les réarran-

gements sont en moyenne plus délétères que les mutations locales) mais permet de mettre le doigt sur l'élément essentiel expliquant la relation entre robustesse réplicative et non-codant. En effet, comme le montre la Figure III.10, le codant est le principal déterminant du caractère délétère des mutations locales, l'ajout de non-codant augmente donc fortement la robustesse, dans des proportions telles que la probabilité de subir une mutation délétère ne change pas malgré l'augmentation du nombre de mutations (Figure III.11). L'effet du non-codant est cependant très différent sur les réarrangements chromosomiques, soit parce qu'ils ont plusieurs points de cassure (pour les réarrangements équilibrés – inversions et translocations), soit parce qu'ils ajoutent ou suppriment des segments de taille non bornée (contrairement aux InDels) pour les réarrangements non équilibrés [1] (duplication et délétion de segments chromosomiques). De ce fait, même si l'ajout de séquences non-codantes augmente la neutralité face aux réarrangements chromosomiques (Figure III.10), le résultat net en prenant en compte l'augmentation induite du nombre de mutations est une forte augmentation de la probabilité de subir une mutation délétère, comme le montre la Figure III.11.

Les observations précédentes confirment que la variation de non-codant influe sur la robustesse réplicative [2]. Reste à vérifier que les variations de robustesse observées sont suffisantes pour permettre la sélection indirecte de la taille des séquences non-codantes dans les génomes.

Si l'on compare l'amplitude des variations de la quantité d'ADN non-codant dans les expériences contrôle (Figure III.3) et l'impact des variations de non-codant sur la robustesse réplicative (Figure III.8), on constate que le non-codant varie de l'ordre de 5 000 paires de bases, ce qui représente, une variation de robustesse de l'ordre de 1 à 2%. Cela soulève la question de savoir si cette variation de robustesse est visible par la sélection ou si elle est inférieure au seuil de dérive. Sous l'approximation que les mutations sont majoritairement fortement délétères, la fitness à long terme d'un individu peut se calculer comme le produit de sa fitness (phénotypique) et de sa robustesse réplicative, ce qui signifie qu'une mutation modifiant de $x\%$ la fitness ou de $x\%$ la robustesse sont équivalentes d'un point vue sélectif (mais à une génération d'écart). Les modèles de génétique des populations prédisent que si le produit de la taille effective de population Ne et du coefficient de sélection s d'une mutation est supérieur à 1 (Si $Ne \times s > 1$) alors la mutation est visible par la sélection tandis que dans le cas contraire, l'effet de la mutation est masqué par la dérive ($Ne \times s = 1$ correspondant au seuil de dérive). La taille de population dans nos expériences est de 1 024 individus, mais, du fait de la spatialisation de la population dans Aevol (voir chapitre II), cela représente un Ne d'environ 1 550 (Zhang *et al.*, 2014). Donc si l'on suppose une perte de robustesse de 1% et que l'on calcule $s \times Ne$ avec $Ne = 1550$, on obtient $1550 \times 0,01 = 15,5$ ce qui est bien supérieur au seuil de dérive ($Ne \times s = 1$) et donc soumis à la sélection. Cependant, il n'est pas raisonnable de supposer que des

1. Nous utilisons ici les termes de réarrangements équilibrés et non-équilibrés pour traduire les termes anglais "*balanced*" (qui conservent la taille du génome) et "*unbalanced*" (qui modifient la taille du génome). Voir, par exemple (Mérot *et al.*, 2020).

2. Elles expliquent aussi pourquoi cet effet est essentiellement passé inaperçu jusqu'ici (en dehors des résultats de Knibbe *et al.* (2007) bien sûr). En effet, cet effet est lié à la présence de réarrangements chromosomiques et n'apparaîtrait pas en présence de mutations locales seules (substitutions et InDels). Or, comme l'énoncent Mérot *et al.* (2020), les effets évolutifs des réarrangements chromosomiques sont encore très mal connus.

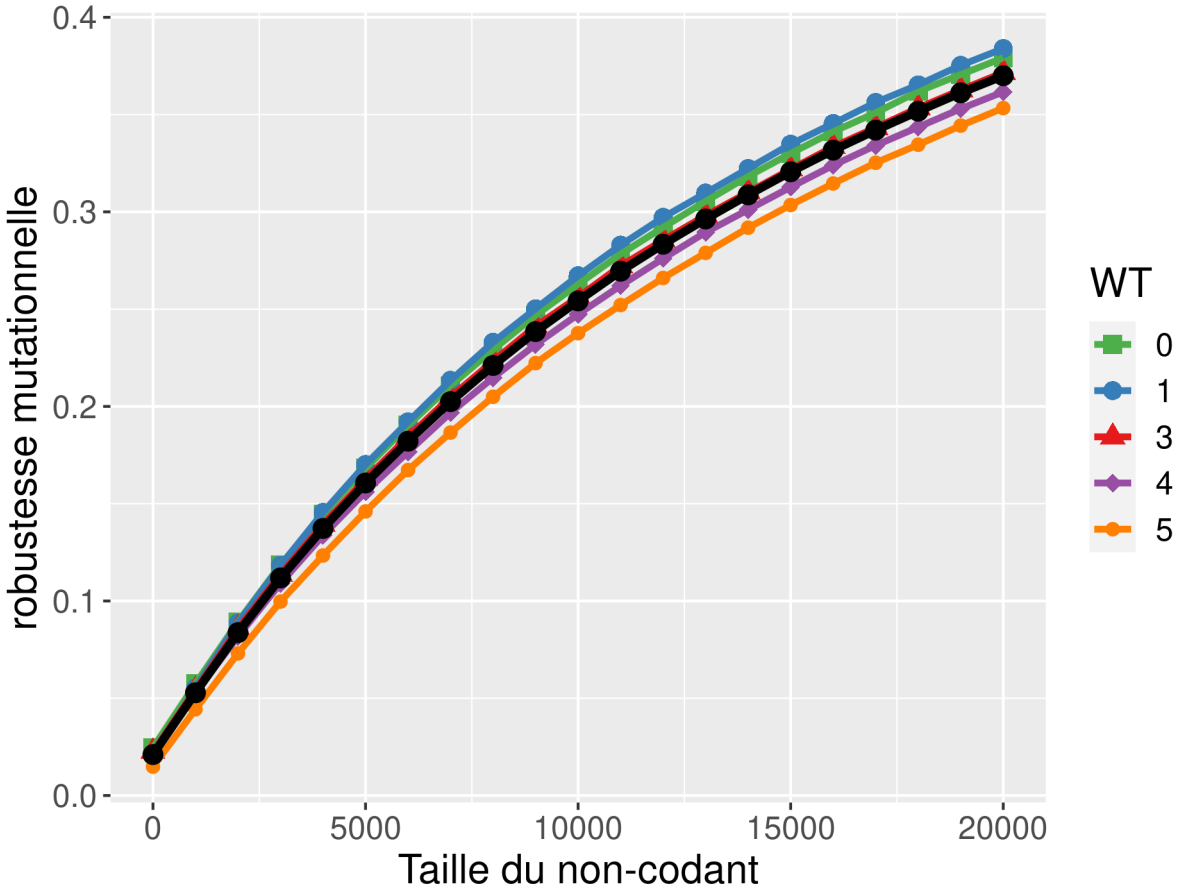


FIGURE III.9 – Robustesse mutationnelle en fonction de la quantité de non-codant pour les génomes modifiés à partir des 5 WTs. Chaque point en couleur est une moyenne sur 50 estimations indépendantes faites à partir d'un million de descendants chacune. La moyenne globale en noire.

variations de plus ou moins 5 000 pb existent au sein d'une population à une génération donnée. Il est donc plus judicieux d'estimer la plus petite variation de robustesse (donc la plus petite variation de quantité d'ADN non-codant) sélectionnable. Dans les conditions de l'expérience contrôle, le seuil de sélection d'une variation de robustesse est de $6,45e^{-4}$ $(1/Ne)$, ce qui, si l'on approxime la relation entre robustesse et non-codant présentée dans la Figure III.8 par une régression linéaire ($y = -2,229e^{-6}x + 0,934$), correspond à une variation d'environ 289 pb ($2,229e^{-6}/1550 = 289,38$). Une variation de 289 pb est aisément réalisable en une seule mutation par réarrangement chromosomique et correspond donc à une taille de divergence qui semble raisonnablement atteignable (environ 2% de la taille initiale des génomes des WTs) par accumulation de mutations entre des lignées divergentes au sein d'une population. On notera que, si notre argumentaire se concentre ici sur Ne, il permet aussi d'expliquer les corrélations entre taille du non-codant et les taux de mutation telles qu'observées par Knibbe *et al.* (2007). En effet, augmenter le taux de mutations (μ) revient à augmenter le coût en robustesse [1] par paire de bases, donc à

1. Plus haut, le taux de mutation est masqué dans la relation entre robustesse réplicative et taille du non-codant.

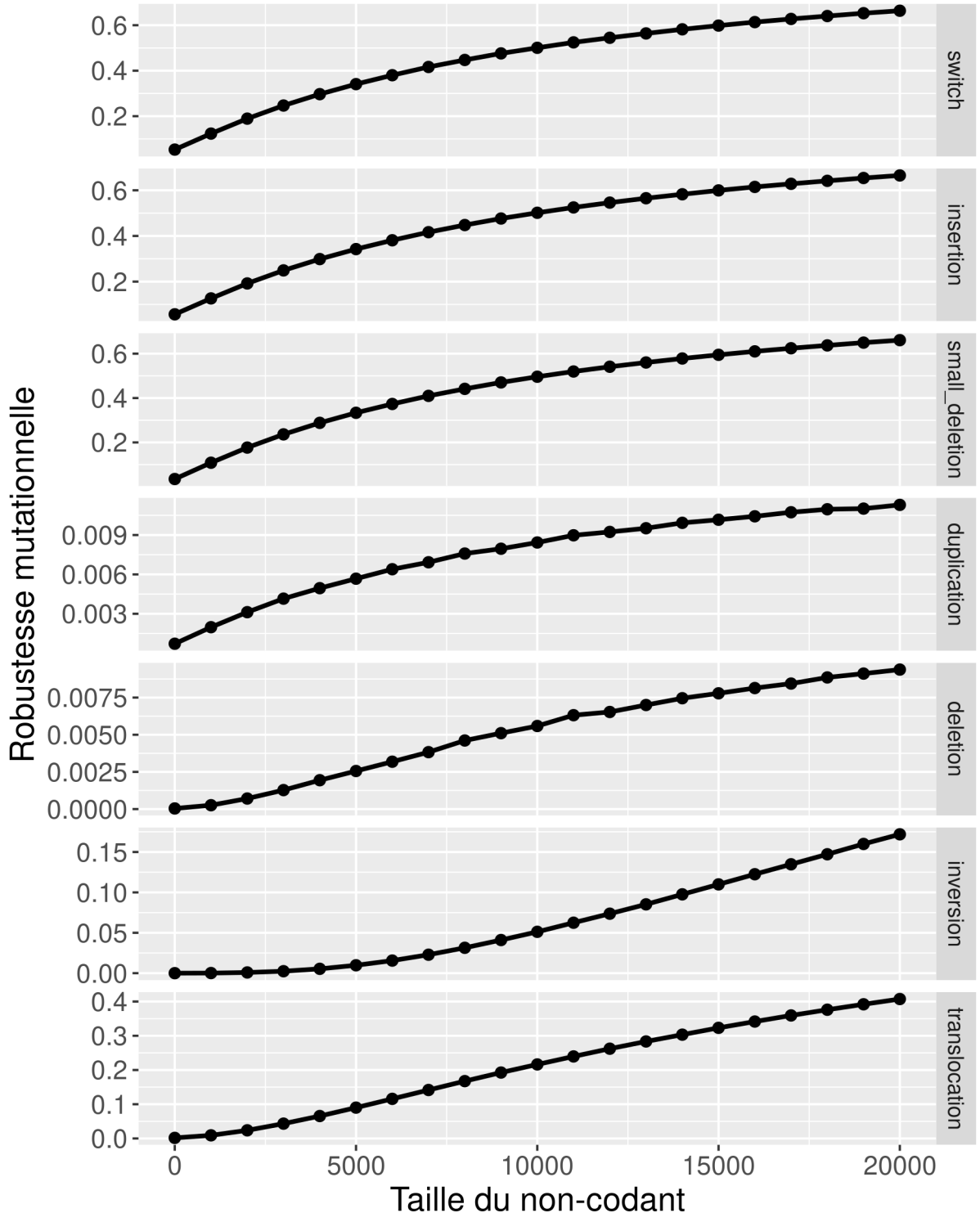


Figure III.10 – Moyenne globale de la robustesse mutationnelle en fonction de la quantité de non-codant pour les sept types de mutations dans Aevol. Chaque point est une moyenne sur 250 estimations (50 par WT et par taille de non-codant) indépendantes faites à partir d'un million de descendants chacune.

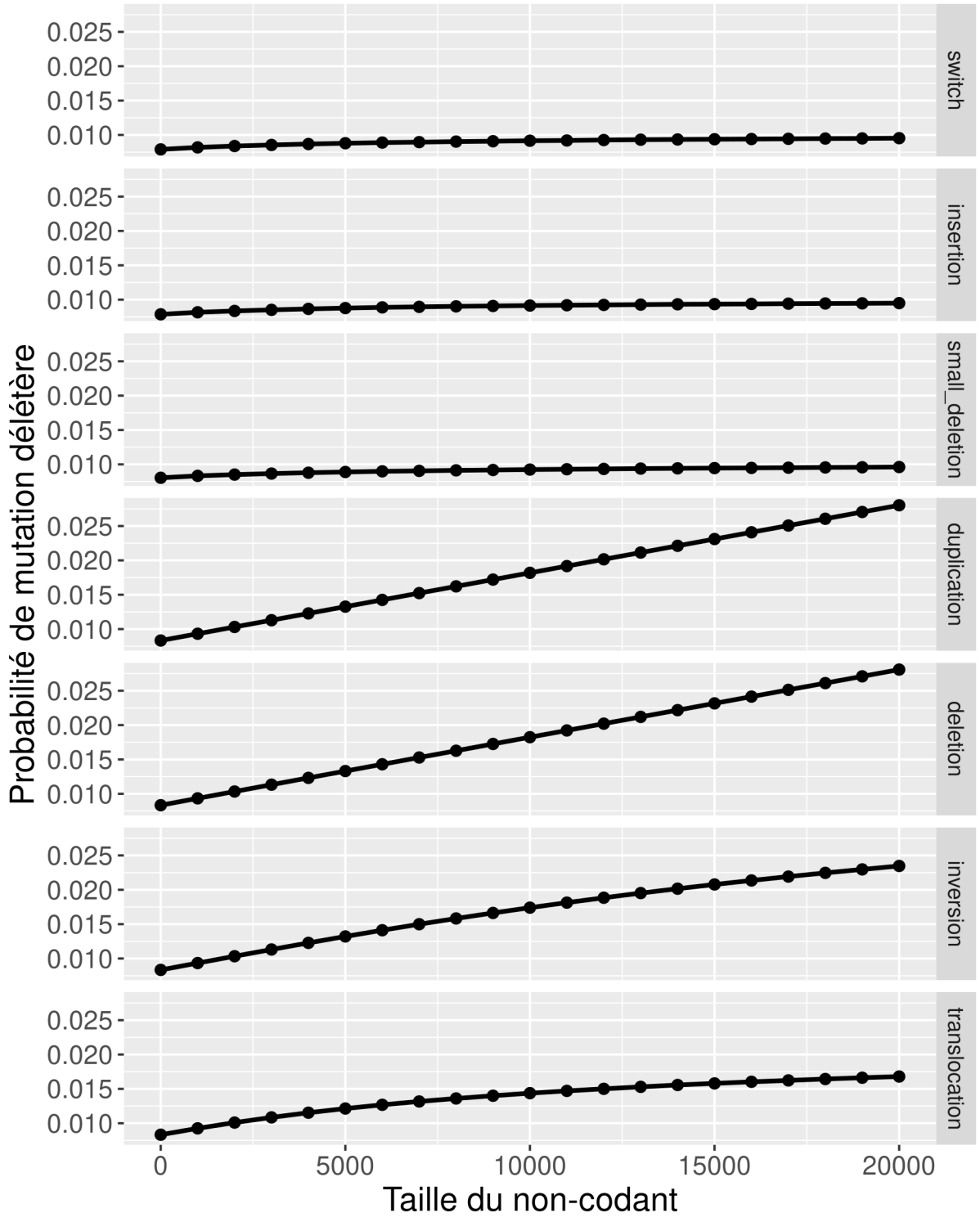


FIGURE III.11 – Moyenne globale de la probabilité de subir une mutation délétère, pour chacun des sept types de mutation de Aevol, en fonction de la quantité de non-codant. Chaque point est une moyenne sur 250 estimations (50 par WT et par taille de non-codant) indépendantes faites à partir d'un million de descendants chacune.

diminuer le nombre de bases nécessaires pour dépasser le seuil de sélection ce qui réduit mécaniquement la taille des génomes.

En résumé, la robustesse réplicative corrèle négativement avec la quantité de non-codant et le minimum de variation de robustesse, potentiellement visible par la sélection, correspond à une variation de quantité de non-codant qui semble facilement atteignable au sein d'une population. Ces deux résultats soutiennent donc l'hypothèse de sélection pour la robustesse réplicative comme force de régulation de la taille du non-codant dans nos expériences. Cette force sélective permet donc d'expliquer en partie les variations de taille de génomes observées sur la Figure III.6. Cependant, la sélection pour la robustesse réplicative n'agit que dans le sens de la réduction de la quantité d'ADN non-codant. Elle ne peut donc pas expliquer pourquoi une quantité minimale de non-codant est toujours maintenue dans nos expériences (qu'il s'agisse des expériences contrôle, Figure III.3, ou des expériences de variation de tailles de population, Figure III.6). En outre, elle ne permet pas d'expliquer pourquoi la diminution de la taille de population provoque une augmentation de la taille des génomes. Pour expliquer ces observations, nous devons faire l'hypothèse qu'une deuxième force existe dans nos simulations, force qui pousserait à l'augmentation de la quantité d'ADN non-codant. En outre, les résultats présentés Figure III.6 suggèrent que cette force n'est pas sélective.

Une observation plus attentive de la Figure III.10 (robustesse mutationnelle par type de mutation) permet d'émettre une hypothèse sur l'origine de cette force. En effet, on constate que la robustesse mutationnelle des duplications est différente de celle des délétions. La Figure III.12 présente la différence entre ces deux robustesses pour tous les Wild-Types et pour toutes les tailles de non-codant testées. Elle montre que, quelle que soit la quantité d'ADN non-codant, les génomes sont toujours plus robustes aux duplications qu'aux délétions (même si la relation n'est pas du tout linéaire). Nous pouvons donc supposer qu'un plus grand nombre de duplications que de délétions peuvent arriver à fixation par dérive génétique, ce qui expliquerait la croissance des génomes lors de la réduction de la taille de population. Dans la section suivante, nous présenterons une expérience permettant de vérifier cette hypothèse.

3.4 Expérience d'accumulation de mutations neutres

Les expériences d'accumulation de mutations neutres (décrites dans la section 2.4) permettent d'identifier l'existence de biais mutationnels indirects, c'est-à-dire de biais liés au taux de fixation de mutations neutres, même en l'absence de biais mutationnel spontané. Ici, pour chaque Wild-Type, nous avons effectué 50 répétitions d'accumulation de mutations neutres sur 500 000 générations. Les résultats de l'expérience sont présentés sur la Figure III.13. Le constat est sans équivoque : les tailles moyennes de génomes ont crû au minimum d'environ 33%, ce qui est particulièrement important pour 500 000 générations (rappelons que les expériences précédentes ont été conduites sur 4 millions de générations). Si on étudie plus en détails quelles mutations ont provoqué cette augmentation (Tableau III.2), on constate que les mutations neutres ajoutant des séquences sont plus fréquentes et plus longues en moyenne que celles qui suppriment des séquences, et ce pour les deux types de mutations concernés (InDels et réarrangements chromosomiques non-équilibrés). Cependant, bien que les InDels neutres soient environ 50 fois plus fréquents

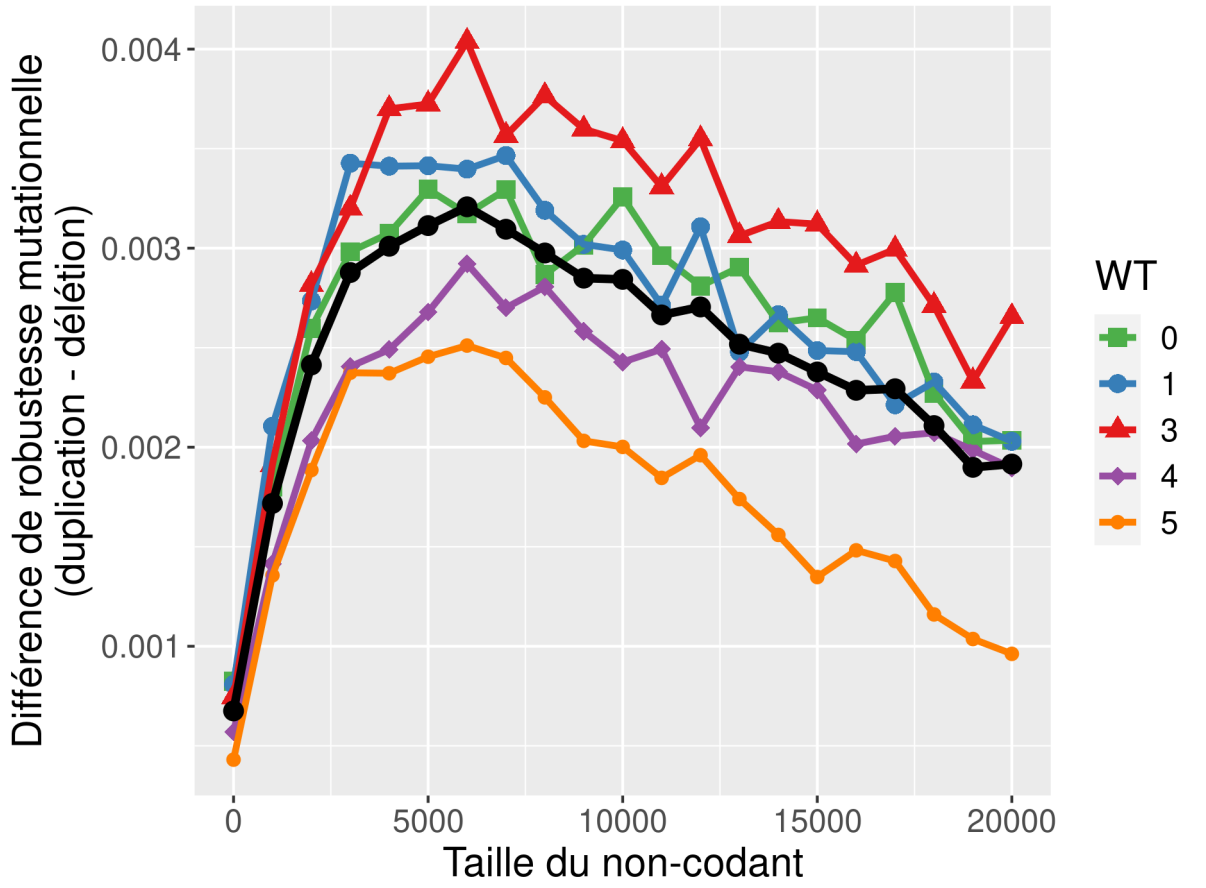


FIGURE III.12 – Différence de robustesse mutationnelle entre les duplications et les délétions en fonction de la quantité de non-codant pour les génomes modifiés à partir des 5 WTs. Chaque point en couleur est une moyenne sur 50 estimations indépendantes faites à partir d'un million de descendants chacune. La moyenne globale en noire.

que les réarrangements neutres, ce sont ces derniers qui expliquent la grande majorité de la croissance des génomes (95,6%).

Les expériences d'accumulation de mutations neutres confirment que le biais à la neutralité identifié sur la Figure III.12 engendre une force augmentant la taille des génomes. En outre, les résultats présentés Table III.2 montrent que cette augmentation n'est pas seulement due à la différence de robustesse entre duplications et délétions, mais aussi à une différence de taille moyenne de ces mutations lorsqu'elles sont neutres. En revanche, même si un biais entre les InDels est aussi présent, ses effets semblent négligeables en comparaison du biais entre les réarrangements. Cependant, les spécificités de l'expérience d'accumulation de mutations neutres font que l'effet des réarrangements augmente au cours de l'expérience. En effet, la taille moyenne des réarrangements est proportionnelle à la taille des génomes (qui, ici, augmente au cours de l'expérience). Pour vérifier si l'effet des InDels est bel et bien négligeable, nous avons estimé le gain moyen par génération dû au biais entre les InDels et au biais entre les réarrangements. Ces estimations ont été réalisées sur les génomes modifiés déjà utilisés dans la section précédente (voir aussi section 2.5). Les résultats présentés sur la Figure III.14 confirment que le biais des InDels

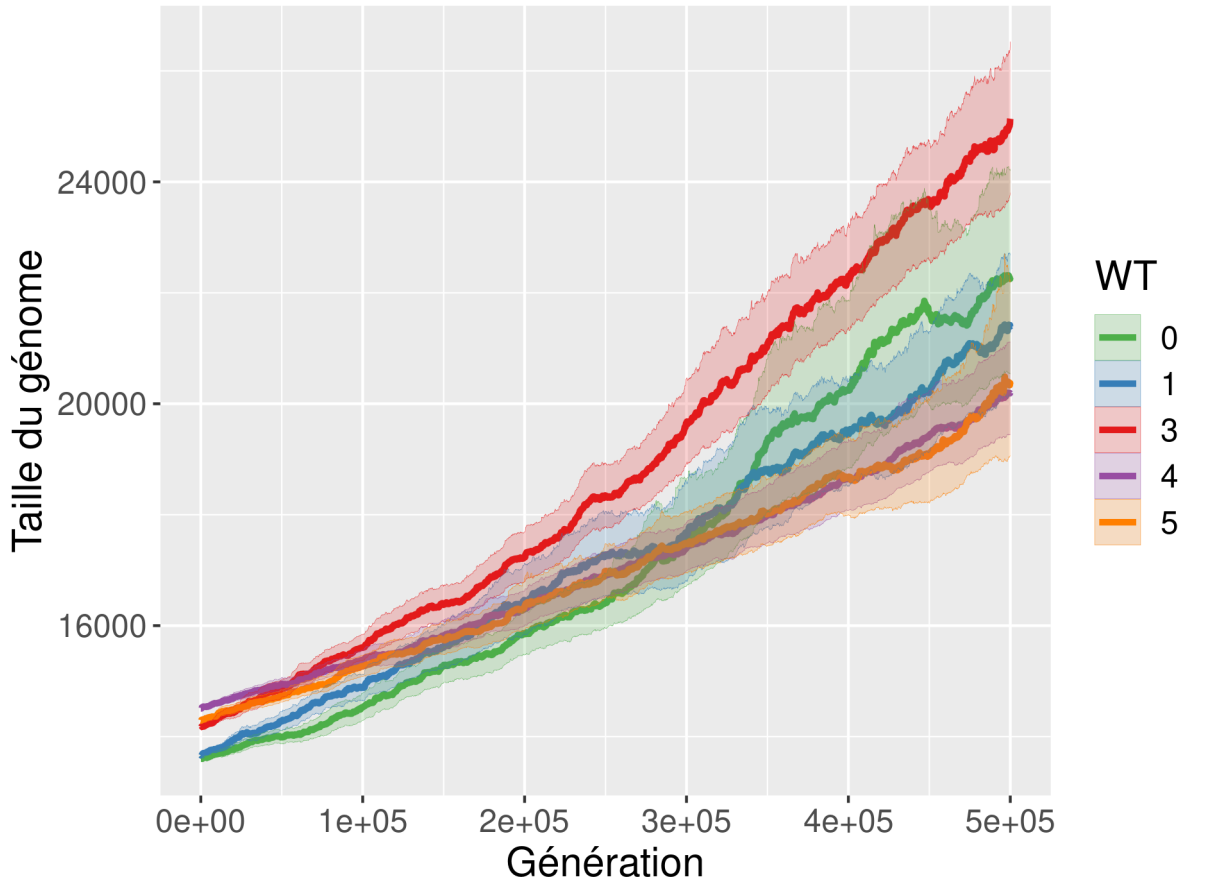


Figure III.13 – Pour chacun des 5 WTs, évolution de la moyenne des tailles de génomes au cours de 500 000 générations d'accumulation de mutations neutres. Les bandeaux autour des moyennes représentent leurs intervalles de confiances à 95% calculés à partir de 10 000 bootstraps. 50 répétitions par expérience.

est négligeable au regard du biais des réarrangements (mis à part dans des conditions où les génomes sont très compacts et comportent très peu de bases non-codantes). Cette estimation permet aussi de quantifier le biais à l'augmentation, dans les conditions contrôle (Figure III.3) où les WTs possèdent environ 6 000 pb non-codantes, le gain en bases est en effet de l'ordre de 1 à 1,5 bases toutes les 100 générations selon le WT. La force de ce biais peut donc sembler faible par rapport à la sélection pour la robustesse, mais, en temps longs, un tel apport de bases non-codantes représente en fait une force non négligeable (dans les expériences contrôle, l'ajout représente en effet un cumul de +40 000 pb sur l'ensemble de l'expérience, soit environ trois fois la taille des génomes des WTs).

3.5 Origine mécanique du biais

Maintenant qu'un biais en faveur des duplications a été identifié comme force poussant à l'augmentation du non-codant, nous allons nous intéresser à l'origine de ce biais. Les deux sections précédentes montrent deux aspects concernant ce biais : les duplications neutres sont plus fréquentes que les délétions neutres et sont aussi plus grandes en

Type de mutation	Nombre de mutations fixées	Taille moyenne des mutations fixées	Contribution totale
Délétions	3836 ± 130	$3,476 \pm 0,0032$	13323 ± 456
Insertions	3915 ± 133	$3,489 \pm 0,0038$	13663 ± 461
Diff. InDels	$\mathbf{83,9 \pm 11.2}$	$\mathbf{0,011 \pm 0.0046}$	$\mathbf{339 \pm 44,5}$
Grandes délétions	$45,6 \pm 2,6$	$247,6 \pm 12,45$	12931 ± 1495
Duplications	$69,0 \pm 3,6$	$264,6 \pm 10,97$	20381 ± 2000
Diff. Réarrangements	$\mathbf{23,5 \pm 1,5}$	$\mathbf{16,94 \pm 6,05}$	$\mathbf{7450 \pm 673}$

TABLE III.2 – Comparaison de l'impact des mutations modifiant la taille des génomes pendant les expériences d'accumulation de mutations neutres. Toutes les valeurs sont des moyennes (n = 250, 50 répétitions par WT, 5 WTs) avec des intervalles de confiance à 95% calculés à partir de 10 000 bootstraps. Les différences sont faites pair à pair au sein de chaque réplicat avec les délétions soustraites aux duplications, une différence positive indique donc un biais en faveur de l'ajout de séquences.

moyenne. Ces deux points suggèrent que des contraintes structurelles différentes s'appliquant aux duplications et délétions. De fait, les conditions de neutralité de ces deux types de mutations sont différentes : une délétion est neutre si elle ne supprime aucune base codante tandis qu'une duplication est neutre si elle ne modifie pas une séquence codante existante ou si elle n'en crée pas de nouvelle. Ces conditions de neutralité créent une asymétrie entre les mutations qui permet aux duplications de copier des parties de gènes dans des zones non-codantes tout en restant neutres (si l'insertion du segment copié a lieu dans une séquence non-codante bien sûr), tant que l'ensemble des séquences d'initiations (promoteur, RBS et codon START) ne sont pas copiées. Ceci permet donc aux duplications d'être plus grandes que les délétions, car elles peuvent chevaucher un gène et une séquence intergénique (les codons STOP et les terminateurs pouvant être dupliqués de façon neutre). Cette observation permet d'expliquer une composante du biais, à savoir, la différence de taille de mutation. Cependant, étant donné que les duplications nécessitent un point d'insertion où insérer le segment dupliqué, comparativement aux délétions qui ne font que supprimer un segment, il n'est pas évident de savoir si cette différence structurelle de mutation suffit à elle seule à expliquer le biais de robustesse mutationnelle en faveur des duplications.

Afin d'explorer l'effet de la structure des mutations sur les probabilités de neutralité, nous pouvons estimer mathématiquement la probabilité qu'une duplication (ou une délétion) soit neutre en fonction des principes mécaniques des réarrangements. Pour cela, nous considérerons un modèle de génome simplifié dans lequel les gènes sont tous orientés dans le même sens, sont tous de taille égale, et sont tous séparés par des séquences intergéniques de taille égale. En outre, nous négligerons la différence entre ARNs et gènes pour ne considérer que les séquences dites « insécables » (voir section 2.6 pour une explication détaillée des notions de séquences sécables et insécables).

Sous les simplifications énoncées ci-dessus, nous pouvons calculer les probabilités de neutralité ν pour les duplications et les délétions, en fonction de la taille totale du génome (L), de la taille du non-codant (λ), de la taille du codant (γ) et du nombre de gènes (g). L'ensemble des développements mathématiques sont exposés en annexe VI.

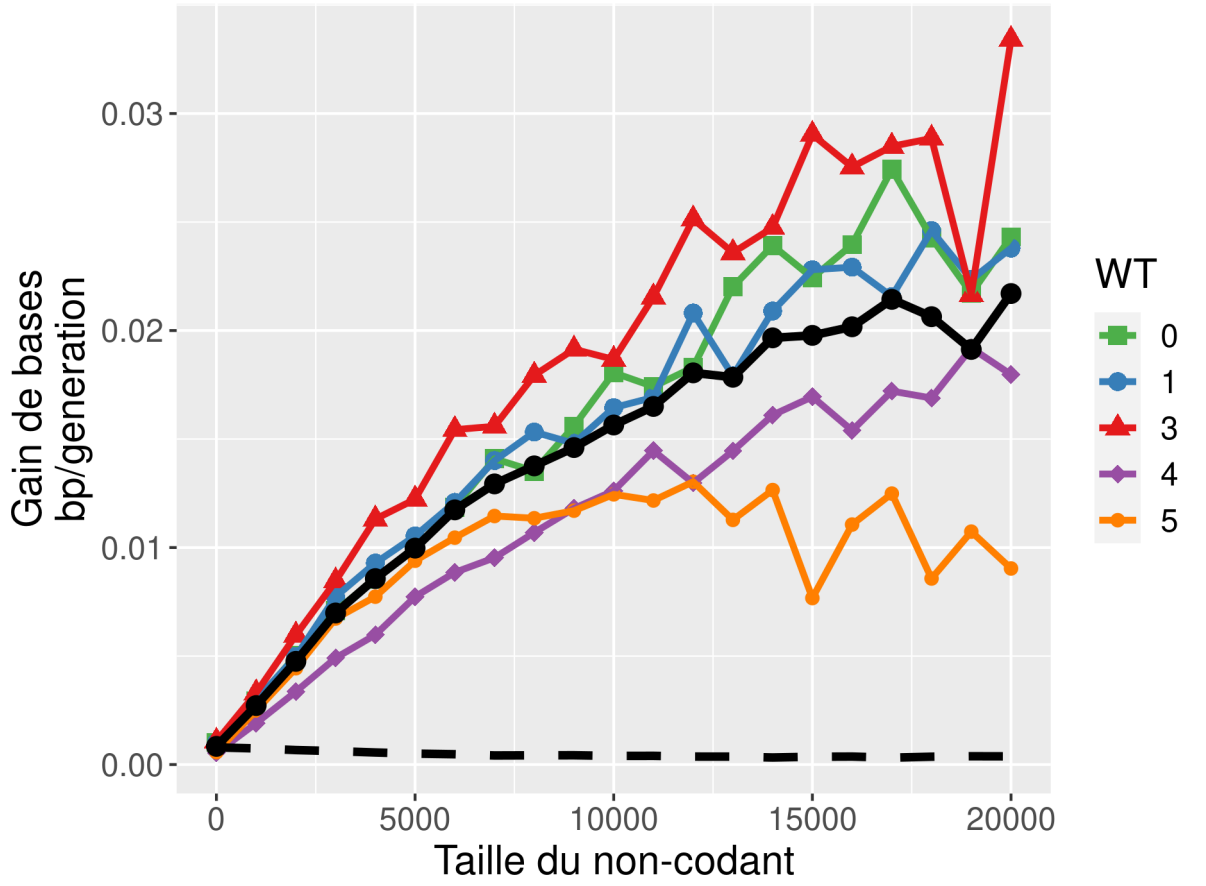


Figure III.14 – Gain moyen de paires de bases par génération en fonction de la quantité de non-codant pour les génomes modifiés à partir des 5 WTs. Ce gain est dû à la différence de robustesse et de taille de mutations entre duplications et grandes délétions. En trait plein noir, la moyenne globale. Pour comparaison, le gain moyen dû aux InDels est présenté en pointillés noirs.

$$\nu_{del} = \frac{(\lambda + g)\lambda}{2gL^2} \tag{III.1}$$

$$\nu_{dupl} = \frac{(\lambda + g)(L - \lambda + \lambda - g)}{2gL^2} \tag{III.2}$$

Soit :

$$\nu_{dupl} = \nu_{del} + \frac{(\lambda + g)(\gamma - g)}{2gL^2} \tag{III.3}$$

D'après l'équation III.3, étant donné que tous les paramètres sont strictement positifs, la probabilité de duplication neutre est toujours supérieure à la probabilité de délétion neutre, sous réserve que $g < \gamma$ (autrement dit, si le nombre de gènes est inférieur à la taille du codant, ce qui est à l'évidence toujours vraie, aussi bien dans Aevol et chez tout organisme biologique). Ceci semble bien indiquer que les contraintes structurelles

des mutations suffisent à expliquer la différence de robustesse entre ces deux mutations et donc le biais.

Les calculs de neutralité correspondants aux équations III.1, III.2 et III.3 reposent sur un modèle comportant de nombreuses simplifications. Afin de vérifier leur bien-fondé, nous les avons donc confrontées aux données d'Aevol. Pour cela, nous avons étendu le modèle mathématique afin de comparer les estimations mathématiques et numériques de quatre mesures différentes (déjà présentées dans les sections précédentes) : la différence de robustesse entre duplications et délétions (Figure III.12), la robustesse réplicative (Figure III.8), la robustesse mutationnelle (Figure III.8) et la différence d'impact sur le changement de taille du codant (Figure III.14).

Pour pouvoir estimer ces mesures dans le modèle mathématique, nous devons estimer le nombre de segments insécables (g) pour les différentes tailles de non-codant. Nous avons pour cela dû développer un estimateur (voir section 2.5). La Figure III.15 présente l'estimation de g pour les 5 250 génomes testés. Afin d'éliminer le bruit très important sur cette mesure, les valeurs de g retenues sont fonction de la quantité de non-codant, la relation est définie par une régression linéaire (Figure III.15).

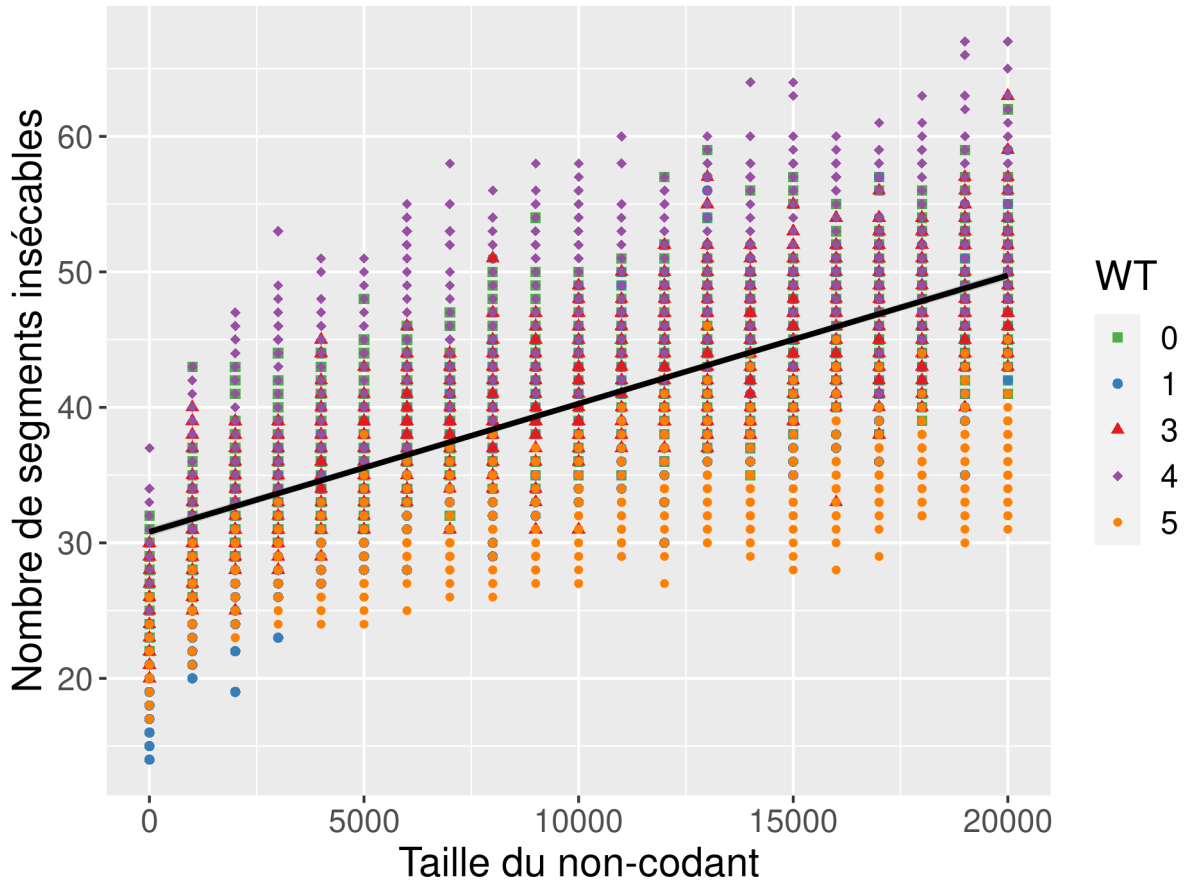


FIGURE III.15 – Nombre de segments insécables estimé par post-traitement pour tous les génomes à taille de non-codant modifiée (n = 5 250). En noir, régression linéaire sur l'ensemble des points ($y = 0,000946x + 30,814771$, R = 0,66, R^2 = 0,43).

En combinant les valeurs de g estimées avec les valeurs déjà connues de γ (codant)

et λ (non-codant), nous pouvons calculer les estimateurs mathématiques et les comparer aux résultats numériques. Sur la Figure III.16, les comparaisons des estimateurs sont visualisées avec des valeurs de λ allant de 0 à 20 000, de g égale à $0,000946\lambda + 30,814771$ et de γ valant 8 342 bases correspondant à la moyenne de la taille du codant des 5 WT. Pour les quatre mesures testées ici, on constate que les estimateurs mathématiques sont très proches des moyennes des estimateurs numériques. Ce résultat est très encourageant quant à la pertinence du modèle mathématique. En effet, en réduisant la définition du génome à une simple succession de séquences sécables et insécables, nous pouvons prédire les tendances générales des quatre paramètres distincts, tous déterminés par la structure des génomes. Ceci permet d'affirmer que les conditions de neutralité des duplications et des délétions suffit à expliquer à la fois la sélection pour la robustesse — et donc la borne supérieure de la taille du non-codant — et les biais de robustesse et de taille entre ces mutations — qui agissent de concert pour augmenter la taille des génomes par un apport lent, mais régulier de bases non-codantes.

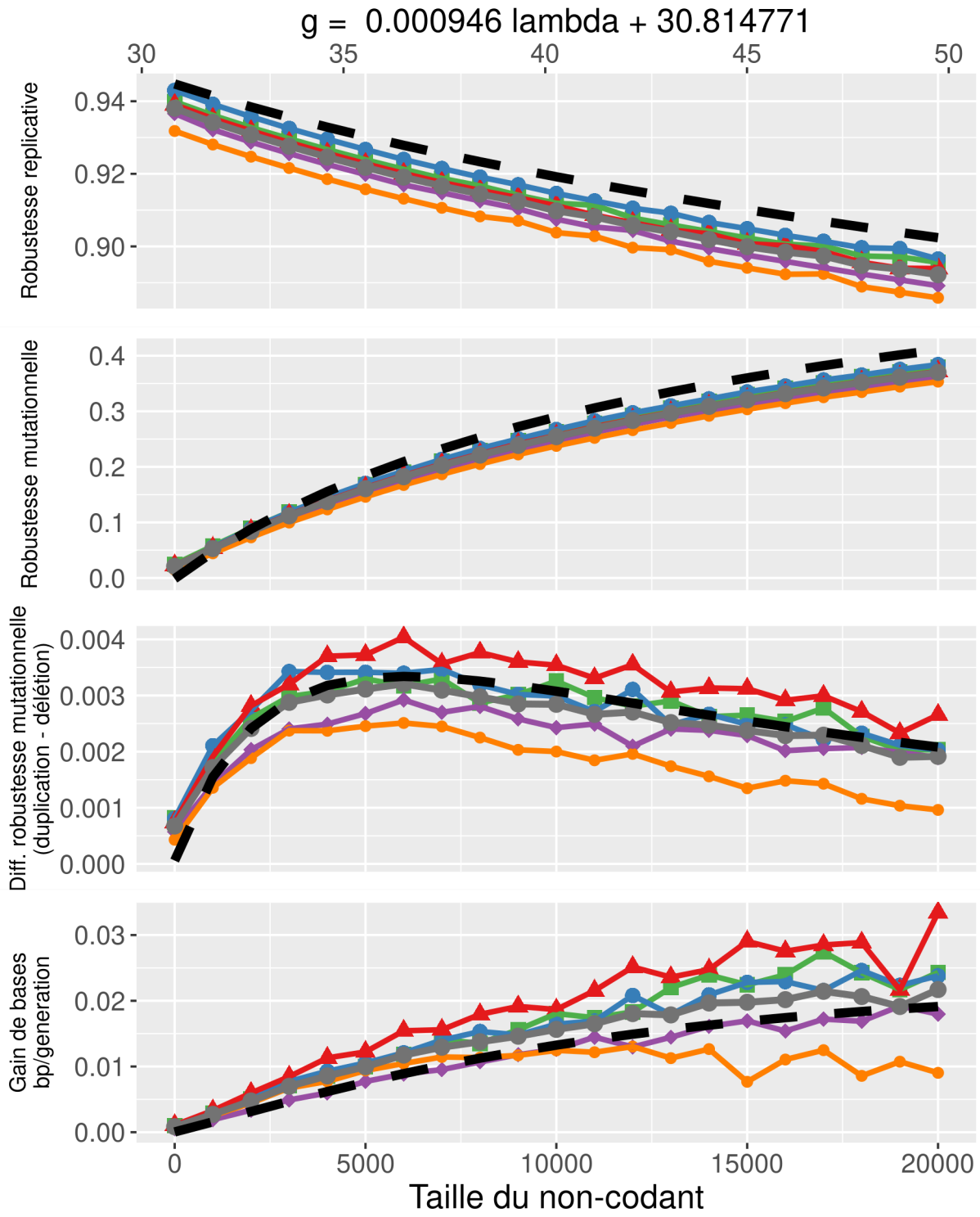


FIGURE III.16 – Comparaison entre les estimations numériques et mathématiques pour quatre mesures : de haut en bas, la robustesse réplicative, la robustesse mutationnelle, le biais duplications/délétions et le gain moyen de bases par génération. En couleur, les estimations numériques pour chaque WT (WT0 : carré vert ; WT1 : rond bleu ; WT3 : triangle rouge ; WT4 : losange violet ; WT5 : rond orange). La moyenne des cinq WTs est représentée en rond gris. En pointillé noir, les estimations mathématiques avec comme valeurs de paramètres : λ de 0 à 20 000 ; $g = 0,000946\lambda + 30,814771$; $\gamma = 8342$ (correspondant à la moyenne des tailles codantes des cinq WTs).

4 Discussion

Nos résultats montrent que, dans les conditions d'un modèle nul où les séquences non-codantes sont totalement non fonctionnelles et en l'absence de biais mutationnels, le non-codant est systématiquement conservé dans les génomes et qu'il varie de façon non-aléatoire (où, pour être plus précis, qu'il varie selon une loi de diffusion bornée). Nous avons montré que cette dynamique provient de deux forces aux effets opposés :

— La première de ces forces est une sélection de second ordre pour des génomes réduits en taille. Ces génomes sont en effet plus robustes face aux mutations intervenant lors des évènements de réplication. En conséquence, les petits génomes ont plus de descendants viables (robustesse réplicative).

— La seconde force est un biais de robustesse mutationnelle en faveurs des duplications vis-à-vis des délétions. En effet, les probabilités d'occurrence neutre et de taille moyenne, lorsque neutre, de ces deux types de réarrangements sont déséquilibrées en faveur des duplications. Il en résulte une accumulation de séquences non-codantes sous l'action de la dérive génétique.

De l'interaction entre ces deux forces résulte un équilibre de la quantité de non-codant. Cet équilibre peut être modifié par des changements de taux de mutation (Knibbe *et al.*, 2007) ou des changements de taille effective de populations (Figures III.7 et III.6). Les modifications de taille effective de population provoquent d'ailleurs des changements de tailles de génomes qualitativement similaires à ceux prédits par les modèles de génétique des populations (Lynch et Walsh, 2007). Cependant, les mécanismes énoncés ci-dessus ne nécessitent aucune des conditions classiquement nécessaires à l'application de ces modèles (effet faiblement délétère du non-codant) et portent l'attention sur les réarrangements chromosomiques plutôt que les mutations ponctuelles ou les éléments transposables.

Maintenant que nous avons caractérisé la dynamique du non-codant dans Aevol — donc dans un modèle computationnel, il est indispensable d'évaluer la pertinence des deux forces exposées ici dans le contexte de génomes biologiques, en particulier bactériens.

4.1 La dynamique des pseudogènes, un marqueur pertinent ?

Selon nous, un des points essentiels de comparaison est celui de la dynamique temporelle de la structure des génomes causée par les deux mécanismes décrits ici. Le biais aux duplications neutres permet, par la copie partielle de gènes, un flux de séquences du codant vers le non-codant ; non-codant qui est ensuite progressivement purgé par la sélection pour la robustesse réplicative. Ces copies partielles et inactives de gènes dans les génomes d'Aevol sont analogues aux pseudogènes décrits dans les génomes biologiques. Les pseudogènes sont des gènes incomplets et inactifs avec un homologue intact et sont présents dans la quasi-totalité des génomes bactériens (Andersson et Andersson, 2001; Liu *et al.*, 2004; Lerat, 2004, 2005; Karro *et al.*, 2007).

En étudiant les vitesses d'extinction des pseudogènes, Kuo et Ochman (2010) décrivent une dynamique comparable à celle présentée ici chez des génomes de *Salmonella*. Kuo et Ochman montrent que, dans les génomes de *Salmonella*, les pseudogènes sont éliminés plus rapidement qu'attendu selon un modèle neutre, donc qu'une sélection purificatrice

s'applique sur ces séquences. Les auteurs supposent que la sélection purifie contre des effets faiblement délétères causés par les pseudogènes, en rejetant l'hypothèse de hasard mutationel proposé par Lynch (2007) (mais sans comparaison avec les taux d'élimination des séquences intergéniques non pseudogénéisées pour pouvoir l'affirmer). Concernant l'inactivation des gènes en pseudogènes, Kuo et Ochman rapportent que 91% des pseudogènes sont inactivés par une seule mutation et que la majorité d'entre elles sont des délétions totales ou partielles de l'ORF (41% des mutations). Notre modèle de duplication neutre, car n'incluant qu'une partie du gène ayant perdu son caractère codant, serait une bonne explication à cette observation. En effet, étant donné que les duplications de gènes complets sont majoritairement délétères (Katju et Bergthorsson, 2013), les individus portant ces duplications ont donc de fortes chances d'être contre-sélectionnés avant que ces copies soient inactivées. Nos résultats suggèrent donc que ces « délétions partielles d'ORF » pourraient être en réalité des non-duplications totales d'ORF, ce qui expliquerait la prépondérance de cette forme d'inactivation. Dans l'ensemble, les dynamiques des génomes observées chez les organismes biologiques et ceux d'Aevol sont donc au moins en partie comparables et notre modèle semble bien avoir un pouvoir explicatif sur les observations biologiques.

4.2 Sélection pour la robustesse réplicative

L'étude de la dynamique des pseudogènes ne permet pas de valider l'hypothèse de sélection pour la robustesse réplicative. En effet, certains pseudogènes sont fonctionnels (Goodhead et Darby, 2015; Cheetham *et al.*, 2020), ce qui rend leur statut de « séquence neutre » problématique, car la sélection contre d'éventuels effets délétères des pseudogènes serait donc un facteur confondant vis-à-vis de la sélection pour la robustesse. La détection de la sélection pour la robustesse nécessite donc d'étudier les dynamiques des séquences intergéniques. Cependant, ces séquences se situent dans un angle mort de la bio-informatique en raison de à la difficulté de reconstitution à partir de données de séquençages « short reads », données qui représentent l'écrasante majorité des données de séquençage.

Bien que la robustesse soit une propriété ubiquitaire des systèmes biologiques (Kitano, 2004; Wagner, 2007) et que la sélection pour la robustesse soit bien soutenue théoriquement (Wilke *et al.*, 2001; Wilke et Adami, 2003; Kitano, 2007; Whitacre, 2012; Rao et Leibler, 2022), les supports empiriques pour ce type de sélection sont limités. Ceci est d'autant plus prononcé à l'échelle de la séquence que les niveaux d'organisation supérieurs (repliement des protéines, réseaux de régulations, réseaux métaboliques, etc.) lui sont souvent préférés pour étudier la robustesse (Wagner, 2007). Or, la robustesse à ces plus hauts niveaux d'organisation pourrait tamponner les perturbations à l'échelle de la séquence. Les études sur la sélection pour la robustesse sont donc le plus souvent conduites sur des organismes ne possédant pas ces niveaux d'organisations supérieurs, qu'ils soient numériques (Elena *et al.*, 2007; Knibbe *et al.*, 2007) ou bien viraux (Sanjuán *et al.*, 2007; Lauring *et al.*, 2013).

La robustesse est une propriété relativement bien comprise théoriquement (Kitano, 2004; Wagner, 2007) mais pour laquelle la question de savoir si elle est le produit de la sélection naturelle ou simplement une propriété émergente de ces systèmes reste non résolue (Wilke *et al.*, 2001; de Visser *et al.*, 2003; Wilke et Adami, 2003; Kitano, 2007;

Whitacre, 2012; Rao et Leibler, 2022). La robustesse à l'échelle de la séquence, autrement dit la robustesse génétique, est dans la littérature presque exclusivement considérée comme une robustesse mutationnelle (de Visser *et al.*, 2003; Whitacre, 2012; Fares, 2015). À notre connaissance, seul Rao et Leibler (2022) utilise un concept de robustesse équivalent au nôtre appelé « *robustness of reproduction to variations* » bien que dans leur modèle, la robustesse soit une propriété émergente du système. Le terme de robustesse réplicative comme nous le présentons dans cette étude n'est en réalité pas un concept totalement nouveau, car sous l'hypothèse que la majorité des mutations sont délétères, il est l'inverse du taux de mutation par génomes par réplications comme ont pu l'utiliser Drake (1991) ou Lynch *et al.* (2016) ; Lynch *et al.* utilisent aussi le terme « fidélité de réplication » plus souvent utilisé en biochimie, terme qui suscite plus intuitivement la possible action de la sélection naturelle.

Concernant la sélection pour la robustesse mutationnelle, Wilke (2001) montrent, chez des organismes numériques, que des individus plus robustes mutationnellement peuvent envahir une population avec une plus grande vitesse de réplication (donc une meilleure fitness) mais une plus faible robustesse. En référence à la célèbre phrase de Spencer (1864) « *survival of the fittest* » en parlant des travaux de Darwin, Wilke (2001) appellent ce processus « *survival of the flattest* », « *flattest* » faisant ici référence à l'analogie du fitness landscape, car les individus plus robustes se situent sur de larges plateaux neutres où les mutations déplacent les génotypes sans modifier leur fitness. Des études sur des virus ont pu montrer expérimentalement que ce principe de « *survival of the flattest* » semble pouvoir s'appliquer en biologie (Codoñer *et al.*, 2006; Sanjuán *et al.*, 2007). Cependant, Lauring *et al.* (2013) remarquent qu'étant donné que la vitesse de réplication est utilisée comme proxy de la fitness, ces résultats peuvent être artificiels, car ce paramètre peut être influencé à la fois par la fitness « réelle » et par la robustesse.

4.3 Biais aux duplications neutres

Nous avons présenté rapidement dans l'introduction le modèle d'équilibre mutationnel de Petrov (2002) qui repose sur le fait que les grandes duplications sont plus fréquentes que les grandes délétions. Ce fait est cependant énoncé par Petrov sans source comme si ce biais était bien connu. En effet, il semble que ce biais soit une connaissance (partiellement) oubliée lors la transition des techniques de microscopie optique, de cytométrie et de chromatographie vers les techniques de séquençages car celles-ci permettaient d'observer les grands réarrangements chromosomiques plus efficacement. Une revue de Starlinger (1977) portant sur les réarrangements chez les procaryotes confirme bien l'existence du biais à la duplication affirmée par Petrov. De même, dans une revue plus récente, Katju et Bergthorsson (2013) rapportent qu'au cours d'expériences d'accumulations de mutations chez *S. cerevisiae* et *C. elegans* les duplications sont plus fréquentes que les délétions chez ces deux espèces ainsi que chez *D. melanogaster*. En outre, les duplications sont en moyenne plus longues. Des confirmations récentes de l'existence de ce biais chez les bactéries manquent cependant, car les réarrangements chromosomiques sont peu étudiés chez ces organismes (Periwal et Scaria, 2015). Toutefois, les mécanismes à l'origine des réarrangements étant identiques chez les bactéries et les eucaryotes (Hastings *et al.*, 2009), cette information reste pertinente.

Jusqu'à récemment, peu d'études ont cherché à expliquer la dynamique de la taille des

génomes sous l'angle du modèle d'équilibre mutationnel. En étudiant, les probabilités de neutralité des InDels avec une approche similaire à la nôtre (bien que formulés différemment), Loewenthal *et al.* (2022) montrent que les insertions sont plus fréquentes que les délétions. À partir de ces résultats, ils proposent un modèle qui prédit la distribution de tailles des introns chez l'humain. Dans leur modèle, Loewenthal *et al.* ont décidé d'ignorer les réarrangements, car considérés comme trop rares. Or, nos simulations montrent que bien que les réarrangements neutres soient effectivement très rares comparativement aux InDels neutres (50 fois moins fréquents, voir Table III.2), ils ont bien plus d'impact sur la dynamique de la taille des génomes. Ceci semble même être suggéré par leur modèle : la taille maximale des InDels y est controlée par un paramètre et, lorsque celui-ci augmente, le biais à la duplication augmente mécaniquement, car les délétions sont d'autant plus contre-sélectionnées qu'elles sont grandes. De plus, toutes leurs prédictions échouent à prédire la distribution de taille des introns au-dessus de 10^4 pb, ce qui suggère la nécessité d'inclure d'autres types de mutations, par exemple les réarrangements.

4.4 Taux spontanés vs taux fixés des réarrangements chromosomiques

La décision de Loewenthal *et al.* (2022) d'ignorer les réarrangements, car considérés comme trop rares, est classique en évolution moléculaire. Nous voudrions ici suggérer que ce choix est une erreur dictée par une mauvaise interprétation de la différence entre les taux de mutations spontanés et fixés. En effet, la faible robustesse mutationnelle des génomes aux réarrangements (Figure III.10) rend leurs observations rares même si leur taux spontané est aussi élevé que celui des mutations locales. C'est d'ailleurs le cas dans nos simulations où tous les taux de mutations sont égaux afin de ne pas donner plus de poids *a priori* à certaines mutations. Cependant, chez les bactéries, il semble que les taux spontanés des réarrangements soient en réalité plus élevés que les taux spontanés des mutations ponctuelles. En effet, des analyses par génomique comparative (Koonin et Wolf, 2008; Puigbò *et al.*, 2014), d'évolution expérimentale (Raeside *et al.*, 2014) et de séquençage « long-reads » de bactéries à faible Ne (Seferbekova *et al.*, 2021; Schenk *et al.*, 2022) appuient toutes en ce sens. Ceci laisse suggérer que le biais à la duplication et la pression de sélection pour la robustesse réplicative pourraient être plus élevé chez les bactéries que dans nos simulations.

D'autres analyses récentes par séquençage « long-reads » montrent aussi que, parmi les réarrangements, les inversions sont les plus fréquemment observées (Weigand *et al.*, 2019; D'Iorio et Dewar, 2023) mais restent rarement fixées (Noureen *et al.*, 2019). Cette observation est cohérente avec notre modèle, étant donné que la robustesse mutationnelle aux inversions est plus élevée que celles des autres types de réarrangements [1], elles sont plus fréquemment fixées. Elles sont cependant comparativement moins souvent fixées que les mutations locales pour lesquelles les génomes sont très robustes, pour peu qu'ils contiennent une proportion de non-codant suffisante.

1. Contrairement aux bactéries, dans Aevol les taux de transcriptions des gènes ne dépendent pas de leur distance à l'origine de réplications OriC. Les inversions dans notre modèle sont donc plus souvent neutres qu'en biologie.

4.5 Sélection pour la robustesse et taux de mutations

Selon la MHH, plus la taille effective de population est élevée, plus la taille de génomes devrait être faible. Or, si l'on observe la distribution de la taille des génomes bactériens, les plus petits génomes ne sont pas seulement représentés par les bactéries marines avec les plus grosses tailles de populations (par exemple *pelagibacter* ou *prochlorococcus M.*), mais aussi par les endosymbiontes (Giovannoni, 2005; Giovannoni *et al.*, 2014). Cette observation reste encore aujourd'hui un argument majeur en défaveur de la MHH. En effet, étant donné leur mode de vie, les endosymbiontes subissent des « bottlenecks » de population lors de leur transmission verticale et ont donc une taille effective de population extrêmement faible, ce qui pû amener à la conclusion que leur faible taille de génomes est une conséquence de la dérive (Kuo *et al.*, 2009). Il semble cependant que cette faible taille de génome soit dictée par un fort taux de mutation comme l'ont montré relativement récemment Bourguignon *et al.* (2020). Dans cette étude, les auteurs montrent que chez la cyanobactérie marine *Prochlorococcus* et chez des espèces endosymbiotiques, la réduction de taille de génome est causée par un fort taux de mutation, lui-même causé par une absence de machinerie de réparation de l'ADN.

Étonnamment, jusqu'alors, la relation entre taux de mutation et tailles de génomes chez les endosymbiontes est passé inaperçue. Pourtant, la relation négative entre taille des génomes et taux de mutation est connue depuis longtemps chez les microorganismes (Drake, 1991). De plus, les bactéries (Lynch, 2010) et les endosymbiontes étaient connue pour, manquer d'une partie de leur machinerie de réparation de l'ADN (McCutcheon et Moran, 2012), et pour, être régulés par leur hôte via des mécanismes de stress oxydatif (Nyholm et Graf, 2012) augmentant, eux aussi, les taux de mutation (Torres-Barceló *et al.*, 2013). Bourguignon *et al.* (2020) présentent une série d'hypothèses afin d'expliquer l'origine de la relation entre taux de mutation et taille de génome. Nous souhaitons ajouter à celles-ci l'hypothèse de la sélection pour la robustesse réplicative que nous présentons ici. Cette hypothèse a l'avantage de permettre de rassembler sous un même cadre théorique l'effet des taux de mutations et de la taille efficace de population. En effet, comme déjà évoqué plus haut, la force de sélection pour la robustesse réplicative dépend de deux facteurs, la force de la sélection (*i.e.* taille effective de population) et le coût en robustesse par paire de bases (*i.e.* taux de mutation spontané).

Chapitre IV

Aevol comme benchmark pour l'évolution moléculaire

1 Introduction

Dans les deux chapitres précédents, nous avons tout d'abord présenté Aevol, le simulateur d'évolution moléculaire que nous utilisons tout au long de cette thèse, puis nous avons montré comment l'utilisation de ce modèle nous a permis d'identifier un nouveau mécanisme de régulation de l'ADN non-codant dans les génomes bactériens. Ces deux premiers chapitres correspondent donc à une utilisation « classique » d'un tel modèle, à savoir la découverte de nouvelles connaissances en biologie évolutive ou en génomique, au même titre que, par exemple, l'évolution expérimentale (Hindré *et al.*, 2012).

Un simulateur tel que Aevol peut toutefois être utilisé dans un autre but que la découverte de connaissance. En effet, ses caractéristiques particulières (structure des génomes et diversité des opérateurs mutationnels en particulier) en font une base particulièrement intéressante pour générer, par la simulation, des jeux de test destinés à valider les algorithmes de bio-informatique (en particulier dans le domaine de la phylogénie). Ainsi, dès 2016, Biller *et al.* ont utilisé ce modèle pour tester l'efficacité des algorithmes d'estimation de distance d'inversion en phylogénie. Les résultats obtenus ont crûment montré que les preuves théoriques et les tests menés pour valider les principaux estimateurs proposés dans la littérature surestimaient, parfois très largement, les performances (Biller *et al.*, 2016b). Une analyse plus poussée a permis de montrer que les estimateurs (et les simulateurs utilisés pour les tester) reposaient tous sur un présupposé fallacieux ; une « interprétation naturelle » (Feyerabend, 1975) partagée par l'ensemble de la communauté. Suite à ce résultat, Biller *et al.* ont corrigé ce présupposé ce qui leur a permis de proposer un estimateur de distance d'inversion plus performant (Biller *et al.*, 2016a).

Malheureusement, la « belle histoire » du simulateur qui permet de découvrir une erreur dans les processus de raisonnement reste d'une portée limitée. En effet, comme nous l'avons présenté au Chapitre II, Aevol utilise une représentation de génome binaire à la place de l'alphabet à quatre bases/lettres de l'ADN. Si cette différence n'interdit pas d'étudier l'évolution de la structure des génomes (comme nous l'avons montré au chapitre III), elle interdit en revanche d'utiliser cette plate-forme pour générer des jeux de tests intéressants. En ce sens, le test des distances d'inversion présenté par Biller *et al.* (2016b)

est une des seules situations où Aevol peut se révéler pertinent — ce qui relève d'une certaine logique puisqu'il s'agit en l'occurrence de tester la vitesse d'évolution de la structure du génome. En outre, comme nous l'avons exposé au chapitre II, Aevol repose sur l'idée que la structure macroscopique du fitness landscape « vrai » est en grande partie déterminée par l'interaction entre la structure de l'information codée sur le génome et les opérateurs de mutations agissant sur cette structure. Cependant, le génome binaire de Aevol s'accompagne d'un code génétique restreint avec 8 « codons » et 6 « acides aminés » non redondants, ce qui constitue une entorse au mode de codage de l'information génétique. En résumé, le codage de l'information dans Aevol est macroscopiquement correct (organisation des ARN et des gènes sur le génome) mais microscopiquement trop simplifié pour pouvoir étudier l'évolution de paramètres aussi classiques en évolution moléculaire que le rapport dN/dS (rapport du nombre de mutations synonymes, dN sur le nombre de mutations non-synonymes, dS). Ce rapport permet d'estimer l'intensité relative des sélections positives et purificatrices.

Pour les raisons qui viennent d'être énoncées, la modification du modèle Aevol visant à passer d'un génome binaire à un génome à quatre bases (A, C, T, G), utilisant un code génétique redondant est à l'étude depuis plusieurs années dans l'équipe Inria Beagle. Pourtant, jusqu'à maintenant, toutes les tentatives pour passer à un génome quatre bases s'étaient toutes révélées infructueuses, en particulier en raison de la complexité computationnelle induite par le passage de 6 à 20 acides aminés qui entraînent une refonte profonde du modèle. Dans ce chapitre, nous présenterons un nouveau prototype que nous avons développé, **Aevol_4b**, qui, pour la première fois, permet de faire évoluer des séquences quaternaires. Pour ce faire, nous avons employé une méthode différente des précédentes tentatives : l'idée ici a été de limiter autant que possible les modifications apportées à Aevol de façon à maîtriser la complexité de Aevol_4b et de tester les modifications successives selon une approche inspirée des méthodes agiles en développement logiciel (Beck *et al.*, 2001).

Dans ce chapitre, nous commencerons par montrer pourquoi/comment les outils de simulation peuvent être utiles pour tester les méthodes et outils en évolution moléculaire [1]. Nous présenterons ensuite plus en détails la preuve de concept évoquée ci-dessus ainsi que d'anciens prototypes d'une version quatre bases du simulateur. Cela nous permettra de planter le décor afin de présenter notre propre version quatre bases, Aevol_4b, ainsi qu'un premier jeu de test. Enfin, afin de montrer que Aevol_4b est effectivement fonctionnel et peut être utilisé pour générer des jeux de tests pertinents, nous présenterons une première campagne de génération de benchmarks au cours de laquelle, à partir d'un ancêtre commun, 40 séquences ont été simulées le long d'un arbre d'espèces prédéfini. Nous monterons comment, à partir des 40 séquences « feuilles », il a été possible de reconstruire l'arbre initial, illustrant l'intérêt et le potentiel de cette nouvelle version du simulateur.

1. Ce travail a été réalisé dans le cadre du projet ANR *Evoluthon* visant précisément à utiliser Aevol pour développer des benchmarks pour la phylogénie moléculaire. C'est pourquoi, dans ce chapitre, nous adopterons le point de vue du benchmarking plutôt que celui de l'évolution expérimentale *in silico* que nous avions adopté au chapitre précédent. En effet, une grande partie de la justification du travail effectué ici est directement issue du projet ANR.

2 La validation des méthodes en évolution moléculaire

Les méthodes d'évolution moléculaire, de génomique et de phylogénétique sont largement appliquées dans l'ensemble des sciences biologiques. Par exemple, elles sont utilisées pour découvrir l'importance fonctionnelle des gènes pour les espèces d'intérêt (Liu *et al.*, 2015), prédire les souches virales saisonnières contre lesquelles un vaccin doit être développé (Łuksza et Lässig, 2014), comprendre les migrations humaines sur terre (Slatkin et Racimo, 2016), améliorer la gestion des agro-systèmes (Thrall *et al.*, 2011), annoter des variants médicalement pertinents (Cooper et Shendure, 2011). Elles sont même utilisées dans les enquêtes judiciaires (Scaduto *et al.*, 2010).

Parce que ces méthodes infèrent des processus historiques, elles se heurtent à un problème de validation : il n'est pas possible de voyager dans le temps et vérifier les hypothèses et les prédictions, qui concernent des événements pouvant remonter jusqu'à 4 milliards d'années. Une validation expérimentale possible serait de faire évoluer des organismes en laboratoire (Randall *et al.*, 2016). Cependant, les expériences ne sont que de court termes, coûteuses et n'ont jamais permis d'être discriminantes entre différentes méthodes. Une validation croisée peut être réalisée en comparant les registres fossiles (Szöllősi *et al.*, 2012; Romiguier *et al.*, 2013) à de l'ADN ancien (Duchemin *et al.*, 2015), mais les échantillons sont rares, en particulier dans le monde microbien et l'ADN ancien n'est pas conservé au-delà d'un million d'années. Les protéines ancestrales reconstruites peuvent être synthétisées pour vérifier si elles sont toujours fonctionnelles (Groussin *et al.*, 2016) mais même des méthodes simplistes comportant des défauts bien connus semblent tout de même reconstituer des protéines ancestrales fonctionnelles. Les considérations théoriques concernant les modèles et les méthodes peuvent également aider à choisir parmi les approches concurrentes (par exemple la cohérence statistique ou la complexité informatique) et il existe des moyens d'évaluer la robustesse des résultats (par exemple, en ré-échantillonnant les données), mais cela ne permet pas de valider les hypothèses et les choix de modélisation sous-jacents (Felsenstein, 2004).

Dans la littérature scientifique, l'approche de validation la plus populaire reste celle des simulations informatiques. L'évolution de génomes peut être relativement facile à simuler *in silico* pour un nombre de générations beaucoup plus élevé que dans le cas de l'évolution expérimentale, à un coût beaucoup plus faible. Les résultats des simulations peuvent alors être utilisés comme instances pour tester des méthodes d'inférence.

La réalisation de simulations à des fins de validation nécessite une réflexion épistémologique et organisationnelle. En effet, très souvent une méthode individuelle est testée avec une simulation ad hoc, c'est-à-dire une simulation créée spécifiquement pour tester la méthode. Dans cette situation, certains éléments et hypothèses de la méthode sont inévitablement intégrés dans le simulateur, qui est alors susceptible de ne générer que des données faciles pour cette méthode et n'a aucune chance d'atteindre la complexité des données biologiques réelles. Même lorsque les simulations sont basées sur un logiciel généraliste qui n'a pas été conçu pour une étude spécifique (Edgar *et al.*, 2011; Dalquen *et al.*, 2012; Sjöstrand *et al.*, 2013; Arenas et Posada, 2014; Mallo *et al.*, 2016), certains principes sous-jacents importants demeurent partagés entre les méthodes de simulation et d'inférence, simplement parce qu'ils sont largement acceptés (et souvent implicites) dans la communauté de bio-informatique. Ces principes sont les « interprétations naturelles »

de la communauté (Feyerabend, 1975) : par exemple, les gènes sont considérés comme des unités évolutives et les réarrangements intragéniques sont négligés – alors que, comme nous l'avons montré dans le chapitre précédent, ils peuvent avoir des conséquences importantes ; les simulations sont effectuées au niveau interspécifique et ignorent les processus au niveau de la population où se produisent les mutations, la dérive et la sélection ; les loci évoluent indépendamment les uns des autres et indépendamment des contraintes des échelles supérieures (par exemple, structurelles). Les espèces éteintes ou non échantillonnées sont ignorées et ne sont pas simulées. Dans une telle situation, les méthodes ne sont testées que dans un monde conçu pour elles, ce qui ne permet pas d'évaluer leur efficacité dans le monde réel.

Un effort de coopération est donc nécessaire pour organiser et normaliser les jeux de tests. C'est cette nécessité qui a, par exemple, mené à l'ajout d'une section dédiée au benchmarking dans *PLoS Computational Biology* et à la publication d'un numéro spécial sur le développement de benchmarks dans *Genome Biology* (Robinson et Vitek, 2019).

Les simulations informatiques sont utilisées dans de nombreux contextes en génomique. Elles servent, par exemple, à évaluer l'adéquation du modèle par des simulations prédictives *a posteriori* pour inférer les paramètres de modèles en utilisant la méthode ABC (Approximate Bayesian Computation) ou pour calibrer des approches d'apprentissage automatique sur des données synthétiques (Chan *et al.*, 2018). En ce qui concerne la validation des modèles et méthodes d'inférence en évolution, on peut identifier deux types de simulateurs selon qu'ils aient été développés spécifiquement pour tester une méthode (simulations ad hoc) ou qu'ils présentent un caractère plus générique (simulations généralistes). Les deux prochaines sections présentent ces deux approches. Nous montrerons ensuite comment les approches issues de la « vie artificielle » peuvent présenter un intérêt particulier pour la génération de benchmarks.

2.1 Simulations ad hoc

Il est très fréquent qu'un simulateur soit développé *a posteriori* afin d'évaluer une méthode précise. C'est d'autant plus fréquent que les éditeurs et reviewers exigent souvent, pour publier une méthode, qu'elle soit d'abord testée sur des simulations, sans exigence précise quant au type de simulations, si ce n'est qu'elles doivent être "réalistes", ce qui n'est pas très prescriptif. C'est ce qu'on appelle un simulateur « ad hoc » au sens où il est développé pour un usage précis. Malgré le sens négatif que peut prendre l'expression « ad hoc » en science (une hypothèse ad hoc étant une hypothèse arbitraire), les simulations ad-hoc ne sont pas inutiles. Elles permettent de traiter des problèmes d'identifiabilité ou d'aider à estimer la plage de paramètres dans laquelle, sous un certain modèle, la méthode est efficace. Elles invalident parfois la méthode d'une équipe concurrente (sous les hypothèses de la nouvelle méthode toutefois – on retrouve ici le sens négatif). Cependant, du fait même de leur caractère ad hoc, elles ne peuvent en aucun cas être considérées comme une validation, ni même comme un test sérieux des limites d'un modèle. Des centaines de références pourraient être citées ici, où une méthode est prétendument, mais trompeusement « validée » par une simulation ad hoc.

Par exemple, les méthodes de détection de l'introgression génétique, c'est-à-dire l'échange de gènes entre différentes espèces via des hybrides (Rosenzweig *et al.*, 2016), sont systématiquement testées avec des simulations incluant des introgressions entre les ancêtres des

espèces échantillonnées. Or, si c'est précisément ce que les méthodes d'inférence peuvent détecter, il est gravement trompeur de l'inclure dans les simulations, car cela revient à faire l'hypothèse absurde que les introgressions passées n'ont eu lieu qu'entre des espèces dont les descendants sont tous échantillonnés des millions d'années plus tard (Szöllősi *et al.*, 2012). Ce type d'hypothèse est représentatif d'une simulation ad hoc. Elle importe les principes de la méthode d'inférence, même si cette importation est incompatible avec la théorie de l'évolution. Une simulation sans cette importation pourrait bien invalider les résultats, comme rapporté pour la phylogénie des anophèles (Davín *et al.*, 2019).

Un autre exemple est la reconstruction d'arbres phylogénétiques lorsqu'on soupçonne que des événements de spéciation se sont produits à des périodes proches les unes des autres. Il a été proposé que certaines méthodes de reconstruction utilisant les arbres de gènes comme inputs seraient supérieures aux approches concurrentes, et ce, essentiellement pour deux raisons : la méthode est statistiquement cohérente, c'est-à-dire qu'étant donné un nombre infini d'arbres de gènes vrais, elle renvoie un arbre des espèces correct, et la méthode fonctionne parfaitement sur les simulations (Liu *et al.*, 2010). Ces résultats ont toutefois suscité de nombreuses discussions, centrées sur l'impact des erreurs dans les arbres de gènes (Song *et al.*, 2012; Gatesy et Springer, 2013; Mirarab *et al.*, 2016), point qui avait été complètement négligé dans les simulations initiales.

Enfin, pour un dernier exemple de simulation ad hoc, on peut citer (Nelson-Sathi *et al.*, 2015). Dans cet article, les auteurs ont utilisé une simulation pour tester la méthode de calcul de similarité de deux ensembles d'arbres qu'ils avaient eux-mêmes développés. Leur simulation consistait à intervertir au hasard les branches de chaque arbre d'un ensemble et de vérifier que leur méthode permettait effectivement de voir que l'ensemble d'origine et l'ensemble avec les branches permutées étaient différents. Groussin *et al.* (2016) ont souligné que cette simulation ad hoc a été spécialement conçue pour faire la publicité d'une méthode, au risque d'utiliser des conditions particulières et irréalistes.

Le point commun de tous ces exemples – et il serait possible de multiplier les exemples à l'infini – est que les méthodes ont été testées sur un simulateur conçu et développé par l'équipe qui a développé la méthode elle-même. Dans ces simulations, les organismes virtuels obéissent donc à des lois issues non pas de l'observation de la nature, mais de la méthode d'inférence elle-même. En résumé, sans accuser les auteurs de malversation, ces simulations sont souvent conçues pour défendre une méthode ou un résultat plus que pour les tester en profondeur.

2.2 Simulations généralistes

Tous les programmes de simulation ne sont pas directement liés à une méthode ou à un outil spécifique. Certains sont en effet conçus pour un usage plus large et traitent d'un problème général sur lequel plusieurs méthodes peuvent être testées. Ainsi, Seq-Gen (Strope *et al.*, 2009) produit des séquences simulées le long d'un arbre phylogénétique ; BottleSim (Kuo et Janzen, 2003) simule le processus de goulots d'étranglement dans les populations ; Simphy (Mallo *et al.*, 2016) produit des arbres de gènes par un processus de duplication, de perte, de transfert et de triage incomplet des lignées.

Le National Cancer Institute du NIH dispose d'un site web (non exhaustif), le GSR

(pour *Genetic Simulation Resources* [1]), regroupant, lors de notre dernière consultation, 227 programmes de simulation informatique publiés pour les études génétiques telles que phylogénie, génétique des populations, repliement de l'ARN ou des protéines, simulation du séquençage de prochaine génération (Peng *et al.*, 2013). Dans le même ordre d'idées, Carvajal-Rodriguez (2008) a comparé les propriétés de 25 simulateurs et les prérequis nécessaires à la constitution d'une méthode de simulation généraliste sont discutés dans Carvajal-Rodriguez (2010), où est proposé un « document d'exigences de simulation ». Dans celui-ci, il est conseillé d'inclure les spécifications du logiciel de simulation (comme d'ailleurs dans tout projet de développement logiciel). Cependant, il ne figure toujours pas parmi les exigences, qu'il est préférable d'éviter de construire des simulateurs spécifiquement conçus pour les méthodes d'inférence. En d'autres termes, même si ces simulateurs généralistes sont conçus par des équipes différentes de celles qui les utilisent (ce qui n'est d'ailleurs pas toujours le cas — voir ci-dessous le cas du logiciel Evolver), ces équipes sont issues de la même communauté scientifique. En conséquence, les hypothèses simplificatrices généralement incluses dans les méthodes d'inférence sont transposées implicitement dans les simulations (d'autant plus implicitement qu'une grande partie de ces hypothèses sont elles-mêmes implicites). L'un des exemples les plus frappants d'une hypothèse simplificatrice universelle, commune à l'inférence et à la simulation, est que le gène est pris comme unité évolutive. Bien qu'il soit évident pour tout évolutionniste, qu'au cours de l'évolution les gènes sont combinés, fusionnés, fissionnés, coupés, étendus, diversement transcrits... ces événements sont négligés dans les simulateurs tout comme ils sont absents des modèles utilisés pour l'inférence. Une définition empirique du gène utilisée pour les regrouper en familles est ainsi transposée en une définition *a priori* des « familles de gènes », ce qui est une façon d'inclure la méthode d'inférence (où les gènes seront effectivement regroupés en familles) dans la simulation des génomes ancestraux.

Un cas particulièrement intéressant est celui du logiciel Evolver (Edgar *et al.*, 2011). Evolver est censé faire évoluer des génomes entiers à l'échelle des nucléotides, afin de produire des histoires qui ressemblent le plus possible à l'histoire supposée des génomes de mammifères. Il n'est pas dédié à tester une méthode en particulier. Il peut servir de référence pour diverses méthodes, telles que la recherche de gènes, le regroupement de gènes, l'alignement multiple, les réarrangements du génome, la phylogénie, etc. Cependant, étant développé par des auteurs de méthodes d'inférence célèbres, Evolver inclut — intentionnellement ou non — les mêmes hypothèses simplificatrices que ces méthodes : à l'intérieur des gènes, aucune duplication et aucun réarrangement ne sont autorisés, ce qui reflète le fait que dans les programmes d'alignements multiples, ce type d'événement n'est pas pris en compte. En adoptant une position humoristique, on est en droit de se demander comment les réarrangements font pour éviter les séquences codantes ! En outre, dans Evolver, les effets de population, y compris la sélection, sont moyennés et non simulés, et les sites évoluent sans contraintes structurelles. Là encore, ces phénomènes sont absents du simulateur car ils ne sont pas directement utilisés par les méthodes à tester. Malgré le caractère généraliste d'un tel simulateur, on constate que les méthodes sont toujours testées sur un monde conçu pour elles, sans beaucoup d'espace pour l'inattendu et sans chercher à simuler directement l'évolution « réelle » (si tant est que ce terme puisse dire

1. `//https://surveillance.cancer.gov/genetic-simulation-resources/`, consulté le 9 octobre 2023

quelque chose !).

2.3 Utiliser la vie artificielle

Face aux simulateurs présentés ci-dessus, qui sont tous directement issus de la communauté « bio-informatique », nous proposons de développer des benchmarks par une approche totalement nouvelle et originale. Reconnaissant que les biais de conception des simulations sont en partie inconscients et largement partagés par la communauté, nous proposons que les simulations générant les benchmarks obéissent à certains principes qui peuvent sembler évidents lorsqu'ils sont formulés, mais qui n'ont cependant quasiment jamais été mis en œuvre. Le principe central est que les simulations ne doivent pas être conçues par les mêmes équipes que les méthodes d'inférence, ni même par la même communauté. Elles devraient au contraire être issues d'une communauté scientifique séparée, aussi isolée que possible de la bio-informatique.

Cette situation, qui peut sembler impossible *a priori*, est en fait rendue possible par l'existence d'une communauté scientifique dont l'intérêt est plus tourné vers la simulation du vivant que vers la simulation de séquences. Il s'agit de la communauté de la « vie artificielle » et, au sein de cette communauté, de l'évolution expérimentale *in silico* (Hindré *et al.*, 2012; Batut *et al.*, 2013), également appelée génétique numérique (Adami, 2006). Lehman *et al.* (2020) affirment qu'une caractéristique notable de l'évolution expérimentale *in silico* est que les organismes artificiels produisent des comportements inattendus et surprenants. En d'autres termes, les systèmes sont suffisamment complexes pour être imprévisibles. L'influence croisée de nombreux objets ou processus (gènes, ARN, protéines, mutations, sélection, phénotype, environnement, etc) rend ces systèmes beaucoup plus riches que les simulateurs développés pour répondre à une question précise et, *a fortiori*, pour tester une méthode de bio-informatique précise. Cette caractéristique est particulièrement intéressante dans le cadre de la génération de benchmarks. En effet, les simulations destinées à tester des méthodes d'inférence sont intéressantes principalement si elles mettent en évidence des comportements inattendus par la méthode plutôt que de se concentrer sur ses comportements attendus.

Il existe de nombreuses plateformes de génétique numérique, elles ont toute une propriété commune qu'il est possible de formuler, de façon provocatrice ainsi : *de la même façon que la vie n'a pas évolué pour être étudiée par les évolutionnistes, ces plateformes n'ont pas été conçues pour produire des benchmarks.* En tant que telles, elles sont peut-être, et paradoxalement, plus appropriées pour produire des benchmarks d'intérêt. De plus, provenant de disciplines complètement différentes (principalement la biophysique ou l'informatique bio inspirée), leurs développeurs ne sont que très peu connectés à la bio-informatique (souvent pas du tout). La conséquence négative est que plupart des plateformes de génétique numérique ne sont pas facilement utilisables pour la production de benchmarks en évolution, car elles font évoluer des objets trop éloignés des séquences biologiques (Hindré *et al.*, 2012). Par exemple, la plateforme bien connue Avida (Wilke, 2001; Lenski *et al.*, 2003; Adami, 2006) fait évoluer du code pseudo-assembleur. D'autres font évoluer, par exemple, des circuits électroniques numériques (Kashtan et Alon, 2005), des graphes ou des réseaux (Crombach et Hogeweg, 2008). De ce fait, certains de ces modèles se sont avérés utiles pour déchiffrer des règles macro-évolutives, mais ils ne peuvent pas pour autant être directement utilisés pour générer des données de benchmarks.

Dans ce contexte, Aevol (Knibbe *et al.*, 2007) occupe une position intéressante. Comme nous l'avons vu dans les deux chapitres précédents, dans Aevol, la structure du paysage adaptatif est fortement déterminée par la structure du codage de l'information. Par conséquent, Aevol imite précisément la structure des génomes biologiques et l'organisation multicouches de la « genotype-to-phenotype map » telle que décrite, par exemple, par le « dogme central de la biologie moléculaire » énoncé par Francis Crick en 1958 (voir Crick 1970). La conséquence immédiate de ce choix est que, dans Aevol, les entités moléculaires sont directement inspirées de la biologie moléculaire, dans le sens où les génomes sont des séquences de nucléotides, que ces séquences sont transcrites en ARNm puis en protéines, et que les organismes sont en compétition Darwinienne en ce qui concerne l'adéquation d'un phénotype non trivial avec un environnement. Il est intéressant de noter que dans Aevol, très peu des simplifications habituellement importées depuis les méthodes d'inférence vers les simulateurs « classiques » sont présentes. Par exemple, dans Aevol les gènes sont des entités évolutives à part entière qui peuvent se combiner, se chevaucher, subir des réarrangements, des duplications partielles, etc., tout comme dans « la réalité ». Et, tout comme dans « la réalité », l'observation de l'évolution en action dans Aevol, montre que de tels réarrangements intragéniques sont largement contre-sélectionnés et très rares dans les lignées (alors qu'ils sont relativement fréquents en moyenne). Cependant, il arrive régulièrement qu'ils finissent par se fixer, avec des conséquences pour la phylogénie qu'il est difficile, voire impossible, de prévoir.

Contrairement à la plupart des simulateurs utilisés pour valider les méthodes bioinformatiques, Aevol ne simule donc pas l'évolution d'une séquence, mais celle de populations d'organismes virtuels codés par des séquences. Bien que cette différence puisse sembler purement sémantique, elle conduit à des principes de simulation très différents. D'une part, Aevol ne simule pas une lignée unique ; il simule une population et la lignée fixée est identifiée *a posteriori* ; d'autre part, il ne simule pas uniquement les mutations fixées (ce qui nécessiterait un modèle de substitution). Il simule les mécanismes biophysiques des variations aléatoires apparaissant lors de la copie de l'ADN et le processus de sélection/dérive qui agissent au niveau de la population et mène (ou non) à fixer ces variations. Une conséquence directe est que le modèle de substitution (ou plus largement de mutation) n'est pas donné. Il s'agit d'un observable qui émerge des interactions complexes entre le modèle biophysique des mutations et le processus de sélection qui conduit à leur propagation dans la population.

En conclusion, bien que Aevol n'ait pas été conçu pour la production de benchmarks, il possède des propriétés intéressantes pour être utilisé en tant que tel. En particulier, il sauvegarde un enregistrement fossile parfait de tous les événements à chaque étape de l'évolution. De plus, il peut être utilisé pour simuler plusieurs types de génomes, puisqu'il a été montré qu'en paramétrant différemment les taux de mutation, la taille et la structure des génomes peuvent varier, allant de petits génomes très compacts de type viral, ou génomes beaucoup moins denses de type bactérien (Knibbe *et al.*, 2007).

2.4 Preuve de principe

Dans une étude expérimentale, Biller *et al.* (2016b) ont montré qu'il était possible d'utiliser Aevol pour produire des benchmarks destinés à tester les méthodes de génomique comparative et que ces données permettaient de détecter efficacement certaines erreurs

conceptuelles largement partagées par les méthodes d'inférence habituelles. En outre, il a été montré qu'en corrigeant ces erreurs conceptuelles, les méthodes d'inférence pouvaient être améliorées.

Le test conduit par Biller *et al.* (2016b) portait sur l'estimation des distances d'inversion dans les chromosomes. Ce problème consiste à comparer l'ordre des gènes sur les chromosomes de deux espèces différentes (synténie) afin d'estimer le nombre d'inversions chromosomiques qui se sont produites au cours de l'évolution depuis le dernier ancêtre commun à ces deux espèces. Il s'agit d'un problème très étudié en génomique comparative, avec des solutions combinatoires et statistiques pour l'estimation de cette distance génomique (Eriksson, 2004; Fertin, 2009). Tous ces modèles représentent l'ordre des gènes sur les chromosomes sous forme de permutations et utilisent ces permutations pour estimer le nombre d'inversions. De ce fait, la taille des séquences intergéniques n'est pas utilisée par les estimateurs statistiques. Des simulations ad hoc ou généralistes ont été produites pour valider ces estimateurs, dans lesquelles les séquences intergéniques sont systématiquement négligées (les séquences intergéniques ne sont même pas mentionnés, car leur présence n'est tout simplement pas envisagée dans les modèles d'inversion). Comme nous l'avons énoncé ci-dessus, si un paramètre n'est pas utilisé par les méthodes d'inférence, les concepteurs de simulations n'envisagent même pas de l'intégrer dans les simulateurs, ce qui renforce l'idée qu'il n'était pas nécessaire d'en tenir compte !

Pour le problème de l'inversion des chromosomes, Biller *et al.* ont testé une douzaine d'estimateurs statistiques sur des données issues de simulations Aevol. Contrairement aux simulateurs ad hoc classiquement utilisés pour tester ces méthodes, Aevol inclut des séquences intergéniques, tout simplement parce qu'elles participent à la biophysique des mutations. En ce sens, il est agnostique aux méthodes d'inférence. Il génère donc de nombreuses dynamiques qui ne sont ni inférées, ni exploitées par les estimateurs statistiques. Ainsi, la présence de séquences intergéniques n'est pas considérée (puisque les inversions ne sont détectées que sur la base de données de synténie – l'ordre relatif des gènes sur le génome). Pour autant, même si elles ne sont pas exploitées, ces données peuvent interférer avec les processus évolutifs, par exemple en introduisant des biais, et donc modifier les inférences. Dans ce cas, les résultats peuvent radicalement diverger par rapport aux tests effectués avec des simulateurs ad hoc.

Dans le cas de l'inférence de distance d'inversion, aucune des méthodes testées sur les données produites par Aevol n'est parvenue à atteindre la moitié des performances annoncées par les simulations ad-hoc (Biller *et al.*, 2016b). Ce constat a permis aux auteurs d'identifier la cause de la différence et de remonter à une interférence entre les séquences intergéniques et les probabilités de réarrangements chromosomiques. Étonnamment, *a posteriori*, ce résultat peut même paraître simpliste : il énonce simplement qu'un segment intergénique plus long a plus de chance de subir une inversion. Cependant, ce simple énoncé interdit de considérer les inversions comme équiprobables, hypothèse classique dans les estimateurs.

Partant de ce constat, Biller *et al.* (2016a) ont conçu de nouveaux estimateurs de distance d'inversion qui se sont révélés plus performants d'un ordre de grandeur sur les données d'Aevol ainsi que sur les simulations ad hoc. Ce résultat était totalement inattendu : l'influence de la taille intergénique n'était pas soupçonnée et n'a pu être découverte que grâce à la procédure de test en aveugle. En outre, il est important de souligner que les séquences intergéniques n'ont pas été incluses dans Aevol dans le but de tester leur

effet sur les distances d'inversion, mais simplement parce qu'elles étaient nécessaires pour modéliser les cassures double-brin de l'ADN dont la réparation conduit à des inversions.

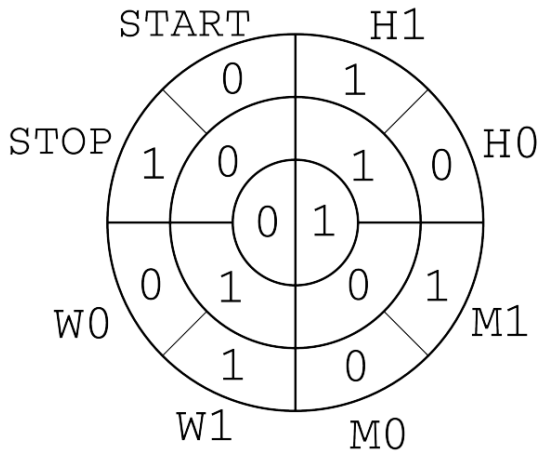
En dehors de cette preuve de principe, à notre connaissance, une telle procédure consistant à impliquer différentes communautés scientifiques pour tester, entre elles, une méthode en double aveugle n'a jamais été tentée. C'est ce constat qui a conduit les équipes Beagle (INRIA-LIRIS) et le Cocon (CNRS-LBBE) à proposer le projet ANR Evoluthon qui est à l'origine des travaux présentés dans ce chapitre.

3 Aevol 4 bases

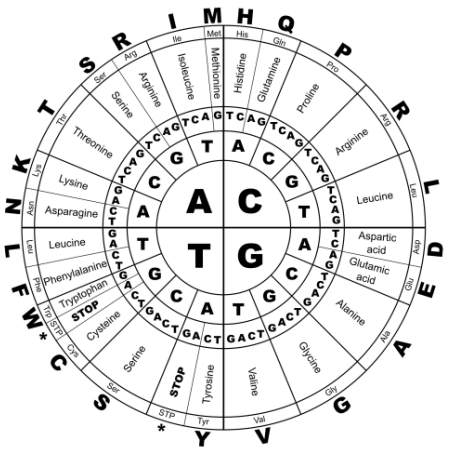
3.1 Les limites d'Aevol

Le choix de l'estimation des distances d'inversions chromosomiques comme preuve de principe de l'utilisation d'Aevol pour la génération de benchmark n'est pas un hasard. C'est en fait un choix presque par défaut imposé par les limites du modèle. Aevol ayant été conçu pour l'étude de l'évolution de la structure des génomes, il est bien adapté à un problème d'inférence portant sur l'évolution de cette structure comme c'est le cas avec les inversions chromosomiques. Cependant, la majorité des problèmes de l'évolution moléculaire sont centrés autour de la question de l'inférence et la reconstitution de séquences ancestrales d'ADN. Ceci souligne la limite principale d'Aevol quant à la simulation de benchmarks pour l'évolution moléculaire : les séquences dans Aevol sont binaires. Les séquences produites par les simulations d'Aevol sont donc inutilisables pour la majorité des problèmes de la phylogénie moléculaire.

Le choix des séquences binaires remonte à la conception du modèle. Il a été fait, car il offrait plusieurs avantages : une plus grande facilité de conception et d'implémentation, une meilleure interprétabilité de ses résultats et une plus grande vitesse de calcul. Aevol a depuis fait l'objet de nombreuses optimisations calculatoires, ce qui a levé les limites de vitesse de calcul pour la transition en séquences à quatre bases. En revanche, le surcoût de complexité amené par cette transition reste bel et bien présent, à la fois pour la conception du modèle et pour l'interprétation des résultats. En effet, étant donné qu'Aevol se conforme au « dogme central de la biologie moléculaire », et en particulier au principe de traduction de codons (constitués de séquences de 3 bases) en acides aminés, la simple transition d'un code binaire (0, 1) à un code quaternaire (A, C, T, G) augmente mécaniquement le nombre de codons de 8 (2^3) à 64 (4^3). En outre, dans la continuité de la logique de conception d'Aevol, cette modification amène naturellement à intégrer le code génétique standard de la biologie, c'est-à-dire que les 64 différents codons codent pour 20 acides aminés (dont la méthionine (Met) qui code aussi pour le START de traduction) et trois codons STOP (Figure IV.1). Ceci pose un problème pour l'attribution des acides aminés aux paramètres de l'espace fonctionnel (voir chapitre II). En effet, dans la version canonique (binaire) d'Aevol, parmi les huit codons, deux codons sont assignés aux signaux START et STOP et les six autres sont assignés par paires à trois paramètres M, W et H (respectivement les codons M0, M1, W0, W1 et H0, H1). Ce principe nous permet de disposer d'un code binaire simple pour calculer la valeur des paramètres fonctionnels des protéines. Ces paramètres sont ceux des fonctions triangles représentant les protéines dans l'espace fonctionnel, à savoir : M la position des triangles, W la largeur des triangles et H la hauteur des triangles. Même si le calcul est pour le moins abstrait, ce principe permet une interprétation simple des séquences protéiques. Avec la transition vers un code protéique de 6 à 20 lettres (les acides aminés), l'interprétation de la séquence d'acides aminés des protéines devient mécaniquement plus complexe et l'attribution des acides aminés à des « fonctions phénotypiques » devient difficile à poser.



Code génétique Aevol : $2^3 = 8$ codons



Code génétique réel : $4^3 = 64$ codons

Figure IV.1 – À gauche, le code génétique canonique de Aevol, reposant sur un génome binaire et comportant donc 8 codons (soit 6 acides aminés, un codon START et un codon STOP). À droite, le code génétique standard reposant sur un génome quaternaire et comportant 64 codons partiellement redondants (soit 20 acides aminés, dont la méthionine (Met) qui code aussi pour le START, et trois codons STOP).

3.2 Approche fonctionnelle : un prototype

La volonté au sein de l'équipe Beagle de modifier Aevol avec des séquences d'ADN à quatre bases est en réalité bien antérieure au projet Evoluthon pour la production de benchmarks. Un code à quatre bases était en effet une évolution naturelle du modèle afin de permettre une meilleure représentation du codage de l'information génétique. De ce fait, depuis la création d'Aevol, plusieurs prototypes ont été développés pour essayer d'étendre le modèle à des génomes quaternaires. Afin de résoudre le problème de correspondance du nombre d'acides aminés au nombre de paramètres fonctionnels, tous ces prototypes ont utilisé une approche visant à augmenter la complexité de l'espace fonctionnel. En effet, en supposant que le code binaire soit conservé pour le codage de l'information dans le décodage des protéines, 20 acides aminés permettent de coder 10 paramètres et il semblait donc naturel de remplacer les triangles protéiques de Aevol par des structures mathématiques codées sur $20/2 = 10$ paramètres.

Pour illustrer cette approche, nous allons présenter succinctement le prototype réalisé par Nicolas Comte, un élève-ingénieur de l'INSA de Lyon ayant réalisé un stage de 5 mois au sein de l'équipe Beagle durant le premier semestre 2016.

Dans ce prototype, pour augmenter le nombre de paramètres de l'espace fonctionnel, le choix de modélisation a été d'ajouter une dimension spatiale, donc de passer d'un

phénotype en 2D à un phénotype en 3D et d'ajouter des paramètres d'épistasie [1] aux gènes en s'inspirant du modèle mathématique proposé par Hansen et Wagner (2001). Le degré d'épistasie est ici contrôlé par deux paramètres. Ces deux facteurs permettent d'augmenter le nombre de paramètres de 3 à 6 : X, Y et Z pour la position et la hauteur des cônes représentant les protéines, SZ pour la largeur des cônes, E pour le nombre de gènes influencés par l'épistasie et SE pour le degré d'influence épistatique. En outre, là où, dans la version canonique d'Aevol chaque paramètre des protéines est codé en binaire (deux codons ou acides aminés assignés à chaque paramètre), dans ce prototype, la base numérique de codage des différents paramètres est variable. Ainsi, X, Y et Z sont codés en quaternaire (4 acides aminés assignés à chacun), SZ et E en tertiaire et SE en binaire. La Figure IV.2 présente la répartition des acides aminés dans ce prototype. Cette répartition des acides aminés est une façon assez élégante de conserver une propriété intéressante du code génétique puisque regroupant les acides aminés aux propriétés physico-chimiques proches (polarité, hydrophobicité, poids moléculaires, etc.) dans le codage d'un même paramètre.

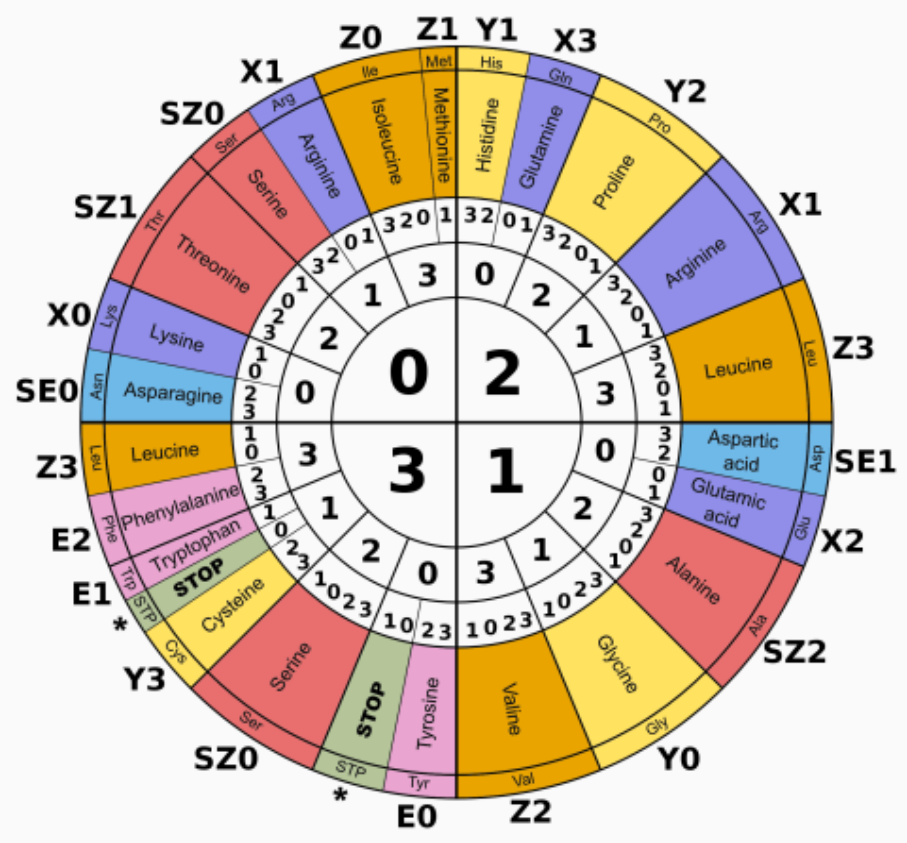
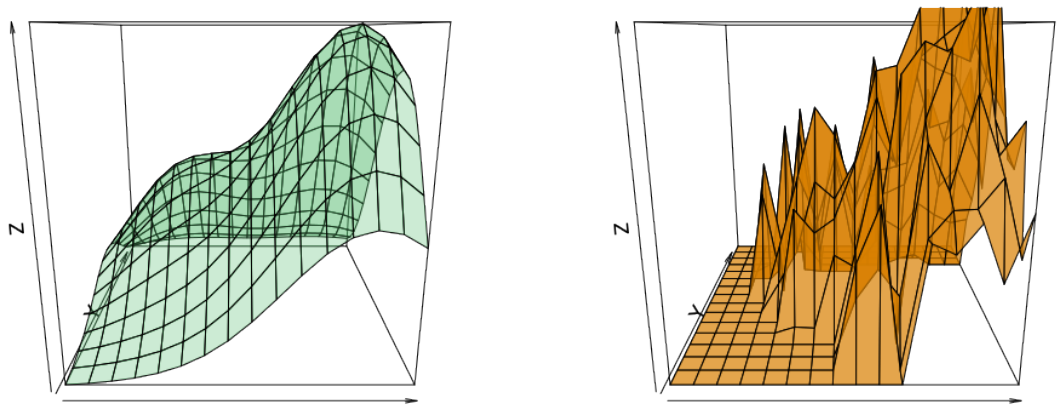


FIGURE IV.2 – Code génétique adopté pour le prototype de Aevol 4 bases développé par Nicolas Comte.

Sans entrer dans les détails des résultats du prototype, la complexification de l'espace fonctionnel soulève un certain nombre de difficultés. Premièrement, pour l'utilisateur,

1. L'épistasie caractérise l'interaction entre deux ou plusieurs gènes, autrement dit, à quel point un gène peut modifier – ou être modifié par – l'activité d'autres gènes.

l'observation des phénotypes est plus difficile en trois dimensions qu'en deux dimensions. Deuxièmement, la cible phénotypique devient très complexe, ce qui rend l'évolution des organismes lente et difficile (Figure IV.3). Troisièmement, le coût calculatoire de ces phénotypes est très important, coût qui, combiné à l'évolution lente, rend certains protocoles expérimentaux trop longs pour être réalisés [1]. Un dernier problème du prototype est que les génomes évoluaient vers des structures dites « en escargot » (voir Figure IV.4). Ces structures sont appelées ainsi à cause la forme qu'ont les génomes transcrits dans la représentation graphique d'Aevol. Cette structure résulte d'une accumulation de promoteurs en amont d'un gène, formant des ARN chevauchants, tous terminés par un unique terminateur. Ceci a pour effet d'augmenter mécaniquement le niveau de transcription de ce gène (puisqu'il est transcrit plusieurs fois). Cette structure est un problème, car c'est un artefact du modèle qui ne serait *a priori* pas possible dans un génome biologique (du fait des conflits biomécaniques entre les ARN polymérases transcrivant simultanément ces brins d'ARNm). Même si, comme nous le verrons plus tard, il s'est avéré que cet artefact n'était pas directement dû à la complexification de l'espace fonctionnel, ce prototype s'est avéré trop complexe et ajoutant trop d'éléments non maîtrisés à la fois pour comprendre, à l'époque, l'origine de cet artefact. Pour toutes ces raisons, le prototype développé par Nicolas Comte n'a pas été conservé. Il a cependant jeté les bases de la conception et du développement de Aevol_4b, comme nous allons le voir maintenant.



(a) Objectif phénotypique (b) Phénotype à la génération 30 000

Figure IV.3 – Représentation 3D de la cible phénotypique et du meilleur individu d'une simulation à la génération 30 000 pour prototype d'Aevol 4 bases développé par Nicolas Comte.

1. Ce point est cependant à relativiser, la version d'Aevol à l'époque de conception du prototype ne profitait pas des nombreuses optimisations présentes aujourd'hui.

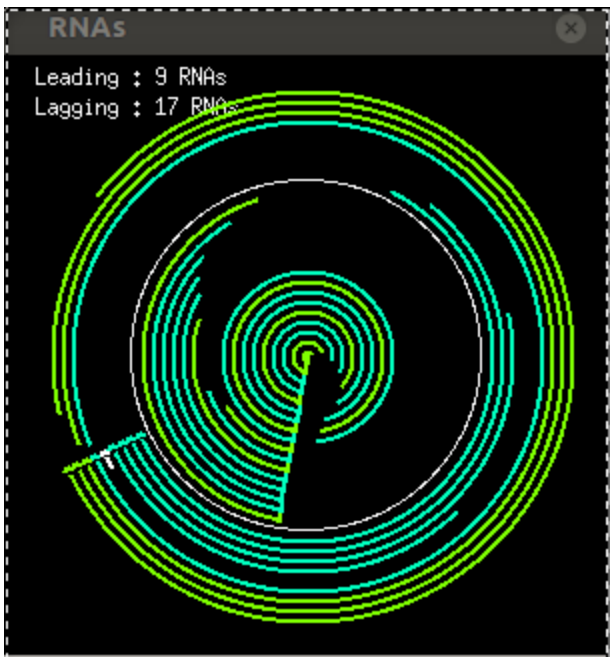


FIGURE IV.4 – Visualisation graphique des ARN dans Aevol. Organisme issu d'une simulation faite avec le prototype de Nicolas Comte. Sur chacun des deux brins, tous les ARN se terminent sur le même terminateur, présentant donc une structure dite « en escargot ».

3.3 Aevol_4b

3.3.1 Démarche de développement

Le prototype développé par Nicolas Comte nous a permis d'identifier les difficultés et de proposer une méthodologie pour le développement d'Aevol_4b, une version fonctionnelle d'Aevol utilisant un génome quaternaire. Il nous a en effet montré que les difficultés causées par un gain de complexité trop abrupt ne sont pas à sous-estimer. En particulier, la perte de repères liée à une modification trop importante et trop rapide du modèle représente une difficulté non négligeable. En conséquence, nous avons choisi de développer Aevol_4b selon deux modalités. D'une part, contrairement aux choix opérés jusqu'à présent, nous avons décidé de ne pas complexifier le phénotype, mais d'utiliser le plus grand nombre d'acides aminés pour coder les trois paramètres (m, w et h) de Aevol en utilisant des bases 6 ou 7. Ce choix est conservateur, au sens où il vise à développer une version 4 bases de Aevol en ne modifiant que les étapes de décodage de l'information génétique, mais sans modifier la structure fonctionnelle du modèle. D'autre part, nous avons choisi d'effectuer une transition depuis la version classique d'Aevol vers Aevol_4b par une succession d'étapes incrémentales entrecoupées de phases de tests. Ces incréments, librement inspirés de la méthode agile en ingénierie logicielle (Beck *et al.*, 2001), ont pour but de conserver la compréhension du modèle tout au long des étapes de développement, mais aussi de dissocier la complexité due aux modifications du modèle de la complexité aux modifications de l'implémentation du modèle. Autrement dit, dissocier la complexité logicielle de la complexité induite sur la dynamique évolutive. Les étapes qui ont été choisies sont les suivantes :

1. Ajout de deux nouvelles bases redondantes en termes d'information. Dans cette première version du modèle 4 bases, les nucléotides sont donc parfaitement redondants deux à deux.
2. Modifier les séquences de signalisation (promoteur, terminateur, RBS) pour utiliser des séquences à quatre bases.
3. Supprimer la redondance des bases 2 et 3 en intégrant un code génétique totalement redondant. Ainsi, les 64 codons possibles d'un code à quatre bases sont repartis en huit groupes de huit codons codant tous la même information que les huit codons du modèle initial ($START$, $STOP$, $M0$, $M1$, $W0$, $W1$, $H0$ et $H1$).
4. Traduction des codons en 20 acides aminés et intégration du code génétique universel.

3.3.2 Résultats

Comme on pouvait s'y attendre puisqu'elles n'intégraient que de la redondance non fonctionnelle, les étapes 1 et 3 n'ont pas eu d'impact sur le comportement du modèle. Leur utilité a été de préparer les étapes 2 et 4 d'un point de vue logiciel en garantissant, justement, que l'introduction de cette redondance n'était pas biaisée. Ainsi, l'étape 1 a nécessité de modifier tous les opérateurs de mutations tandis que l'étape 3 a nécessité l'implémentation d'une table de traduction. Dans les deux cas, introduire ces modifications sans en propager les conséquences au niveau fonctionnel nous a permis de vérifier que la dérive génétique jouait bien son rôle (par exemple que les bases redondantes étaient bien équiprobables) et donc de vérifier l'absence de biais dans les processus de mutation et/ou de traduction. Lors des tests des étapes 2 et 4, nous avons ainsi pu nous assurer que les différences causées par les modifications fonctionnelles étaient bien la conséquence des changements du modèle.

L'étape 2 consistait à modifier les séquences consensus classiquement utilisées dans Aevol pour les promoteurs et pour les « ribosome binding sites » (RBS) en séquences à quatre bases. Cette modification pourrait permettre, par la même occasion, d'utiliser des séquences consensus connues en biologie moléculaire, même si nous avons pour le moment choisi de limiter la taille des séquences consensus afin de permettre la découverte « rapide » de nouveaux gènes lors de la phase d'initialisation du modèle (voir chapitre II). La séquence consensus des promoteurs est donc passée de :

0101011001110010010110

à

GGCCACCATAGTAGGGTTCATG

tandis que la séquence consensus des Ribosome Binding Sites (parfois appelés séquences de Shine-Dalgarno) est, elle, passée de :

011011****

à

ATTATT****

où * est une base quelconque.

La séquence terminatrice est toujours une structure de type *hairpin* de forme $abcd**** \bar{d}\bar{c}\bar{b}\bar{a}$ avec $\bar{x}$ la base complémentaire de x, analogue aux séquences terminatrices ρ-indépendantes chez les bactéries.

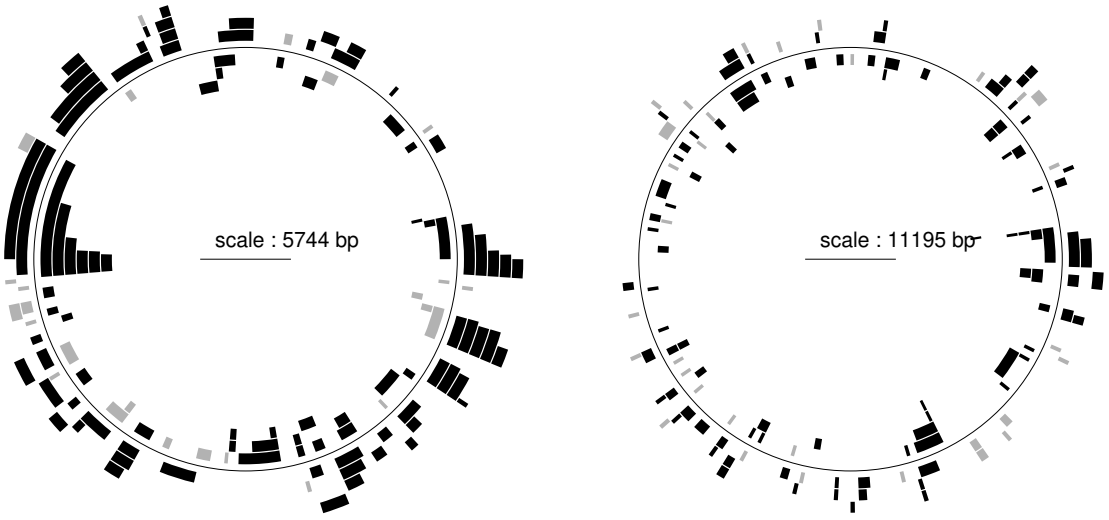
Cette étape, qui pourrait sembler anodine au premier abord a en pratique une conséquence importante. En effet, elle diminue la probabilité de formation spontanée de séquences consensus dans les génomes. La conséquence directe en a été la réapparition des structures en « escargot » présentées dans les résultats du prototype.

En première analyse, ces structures semblent spécifiquement dues à la diminution de la probabilité de formation de terminateurs en particulier, en effet la probabilité de formation d'un motif de type hairpin avec des branches de L bases est de $1/L^B$ ou B est le nombre de bases utilisées pour coder l'ADN. Ce calcul très simple montre que le passage de deux à quatre bases conduit à réduire la probabilité d'occurrence des terminateurs de $1/4^2 = 1/16$ (soit en terminateur en moyenne toutes les 16 bases) à $1/4^4 = 1/256$ (un terminateur en moyenne toutes les 256 bases). Ce changement de probabilité se traduit dans les génomes par une raréfaction du nombre de terminateurs, ce qui augmente la probabilité que, lorsqu'un nouveau promoteur se forme en amont d'un promoteur existant, il n'y à aucun terminateur intermédiaire. Dans ce cas, mécaniquement, le taux de transcription des gènes transcrits par ce promoteur va augmenter. Ce mécanisme étant plus probable, il est plus facilement recruté par la sélection, particulièrement dans les premiers stades de l'évolution où les organismes d'Aevol comportent souvent des gènes peu exprimés. Ces « escargots » persistent alors en raison de la prévalence des effets fondateurs en évolution. Ce type de structure est *a priori* possible dans le modèle binaire, mais n'est que rarement observé.

L'augmentation du taux de transcription par juxtaposition de promoteurs successifs est possible dans Aevol, car le modèle n'intègre pas de nombreux éléments de la biochimie cellulaire : seules les séquences de signalisation sont prises en compte pour modéliser la transcription. Dans des organismes biologiques, les processus d'encombrement cellulaire, d'interférence, de transcription ou encore de compétition entre promoteurs pour le recrutement des ARN polymérases diminueraient fortement, voire annuleraiten, l'effet d'accumulation de promoteurs. Modéliser ces processus biochimiques n'étant pas une solution envisageable, nous avons dû faire un choix de modélisation pour limiter la formation de ce type de structures. La solution retenue a été de borner le taux de transcription total en ne retenant qu'un seul des promoteurs. Afin de limiter la complexité du modèle, nous avons donc choisi, lorsque plusieurs promoteurs transcrivent la même séquence et sont terminés par un même terminateur, de ne retenir que l'ARN formé à partir du promoteur le plus en amont (option "farthest" dans Aevol). Ceci empêche tout effet de promoteurs multiples et reflète un peu mieux les contraintes de la réalité biologique, même si la solution reste partiellement arbitraire [1]. La Figure IV.5 présente deux génomes ayant évolué dans Aevol_4b, l'un avec l'option farthest et l'autre sans. Les structures globales des génomes sont radicalement différentes. Sans cette option, nous pouvons observer de nombreux « escargots » alors que ceux-ci ne se forment pas – ou moins – lorsque cette

1. Une solution qui pourrait sembler moins arbitraire aurait été de retenir le promoteur le plus actif. Cependant, outre que ce choix peut être ambigu si plusieurs promoteurs ont la même activité, il entraîne des effets indésirables puisque dans ce cas, on pourrait voir apparaître des opérons dont les taux de transcription ne sont pas constants le long du transcrit.

option est activée. On notera qu'avec l'option farthest, nous pouvons quand même observer quelques successions de promoteurs. De fait, cette structure n'est pas interdite. Cependant, étant non-fonctionnelle, elle a tendance à disparaître par dérive, ce qui, *in fine*, limite la persistance de ce type de structures.



(a) Sans l'option farthest (b) Avec l'option farthest

Figure IV.5 – Visualisation des ARN à la génération 100 000 avec et sans l'option farthest. Cette option permet de limiter l'accumulation des transcrits en ne retenant, lorsque plusieurs promoteurs pourraient conduire à la transcription d'un même gène, que le taux de transcription du promoteur le plus en amont. En noir, les ARN codant. En gris, les ARN non-codant.

L'étape 4 est l'étape la plus cruciale dans la conception du modèle à quatre bases. Copier le code génétique universel est en soi assez facile, la difficulté consiste à assigner ensuite un effet à chacun des acides aminés dans le monde fonctionnel mathématique de Aevol, sans pour autant complexifier à outrance ce monde fonctionnel.

De fait, pendant longtemps, nous avons pensé que cette étape nécessiterait de fortement complexifier le modèle mathématique des protéines et du phénotype pour augmenter le nombre de paramètres utilisés dans le calcul du phénotype (à l'instar de ce qui avait été proposé dans le prototype de Nicolas Comte). Pour cela, dans un premier temps, deux solutions ont été proposées. Soit complexifier l'espace fonctionnel, par exemple en augmentant le nombre de dimensions du phénotype, soit conserver un phénotype à deux dimensions (fonction-activité) mais utiliser des fonctions-protéines plus complexes que les triangles de Aevol pour qu'elles puissent dépendre de paramètres plus nombreux. La première solution est celle adoptée par le prototype de Nicolas Comte et nous avons vu qu'elle s'est révélée infructueuse. La seconde semblait initialement plus prometteuse et il a été, par exemple, proposé d'utiliser des fonctions linéaires par morceaux ou d'ajouter des moments aux fonctions protéiques telles que l'asymétrie (skewness) ou la kurtosis (en plus des deux moments actuellement utilisés : la moyenne et la variance). Ces deux so-

lutions sont des complexifications du modèle fonctionnel qui diminuent notre capacité à comprendre la trajectoire fonctionnelle de l'évolution (avec comme conséquence indirecte la difficulté à détecter des bugs ou problèmes de paramétrage lors de l'observation des phénotypes des individus – en plus d'être aussi plus coûteux en calcul). Pour éviter cette complexification, nous avons décidé de conserver exactement le modèle fonctionnel de Aevol (protéines-triangle, espace fonctionnel à deux dimensions – voir chapitre II) mais de modifier le mécanisme de « repliement » utilisé pour calculer les paramètres des protéines à partir de leur séquence primaire. Pour ce faire, nous avons étendu le principe de codage des paramètres par une base non-binaire. Dans le prototype, les différents paramètres des fonctions protéiques n'ont pas le même nombre d'acides aminés assignés et ne sont pas calculés avec la même base numérique. Par exemple, le paramètre X était calculé en quaternaire, les acides aminés assignés à x codaient donc pour quatre digits $x_0 = 0$, $x_1 = 1$, $x_2 = 2$ et $x_3 = 3$. Ainsi assignés, les acides aminés codent un nombre en base 4 qui est ensuite normalisé pour donner une valeur entre 0 et 1 déterminant la valeur du paramètre x. Nous avons repris cette idée et « simplement » utilisé des bases numériques plus larges pour coder les paramètres M, W et H des fonctions triangles représentant les protéines.

La Figure IV.6 présente le code génétique standard (nous verrons ci-dessous que celui-ci peut être modifié pour tester des mécanismes d'épistasie) retenu : les paramètres M et H sont codés en base sept (septénaire, sept acides aminés assignés) et le paramètre W en base 6 (sénaire, six acides aminés assignés). Au total, le codage de ces trois paramètres utilise bien les $7 + 7 + 6 = 20$ acides aminés du code génétique, sans modifier le modèle fonctionnel de Aevol. En outre, ce code génétique rassemble grossièrement les acides aminés de mêmes familles en les assignant au calcul des mêmes paramètres. Il s'agit cependant là probablement plus d'un choix esthétique arbitraire que d'une fonctionnalité réelle ; la volonté première de la modélisation retenue est de proposer un modèle fonctionnel le plus simple possible et, surtout, de permettre une interprétation simplifiée des dynamiques évolutives en s'écartant le moins possible du modèle canonique d'Aevol qui, d'une part, a fait ses preuves, et d'autre part, est bien maîtrisé par l'équipe de développement.

L'assignation arbitraire des acides aminés aux paramètres des fonctions protéiques triangle présente l'avantage d'être simple et de « faire le job », c'est-à-dire, d'utiliser le code génétique universel dans Aevol tout en conservant la capacité d'adaptation des organismes dans les simulations. Ce choix a en particulier permis de débloquer le développement d'une version 4 bases de Aevol pour un coût *in fine* limité puisque seule l'interprétation des séquences a été modifiée. De fait, comme nous le verrons dans la section suivante, cela a permis d'obtenir un *modèle* fonctionnel (quoique encore en version de test) en conservant la grande majorité du code du *programme* Aevol. En outre, ce choix ouvre des possibilités expérimentales inattendues. En effet, la combinatoire des assignations possibles est immense (nous pourrions tout à fait assigner 16 acides aminés au paramètre M et deux à W et à H) et, en l'état, cette caractéristique du modèle est probablement largement sous-exploitée. En particulier, rien n'interdit d'assigner un même acide aminé à deux, voire trois, paramètres, augmentant d'autant l'épistasie et, donc, la rugosité du fitness landscape implicite engendré par Aevol.

Afin d'explorer cette possibilité, nous avons testé les effets d'une « épistasie maximale » dans laquelle tous les paramètres sont encodés en base 20 (M, W et H sont donc dépendants de tous les acides aminés) mais en assignant à chaque acide aminé une valeur différente entre 0 et 19 (sans donner ici le détail du codage utilisé, on peut citer, par

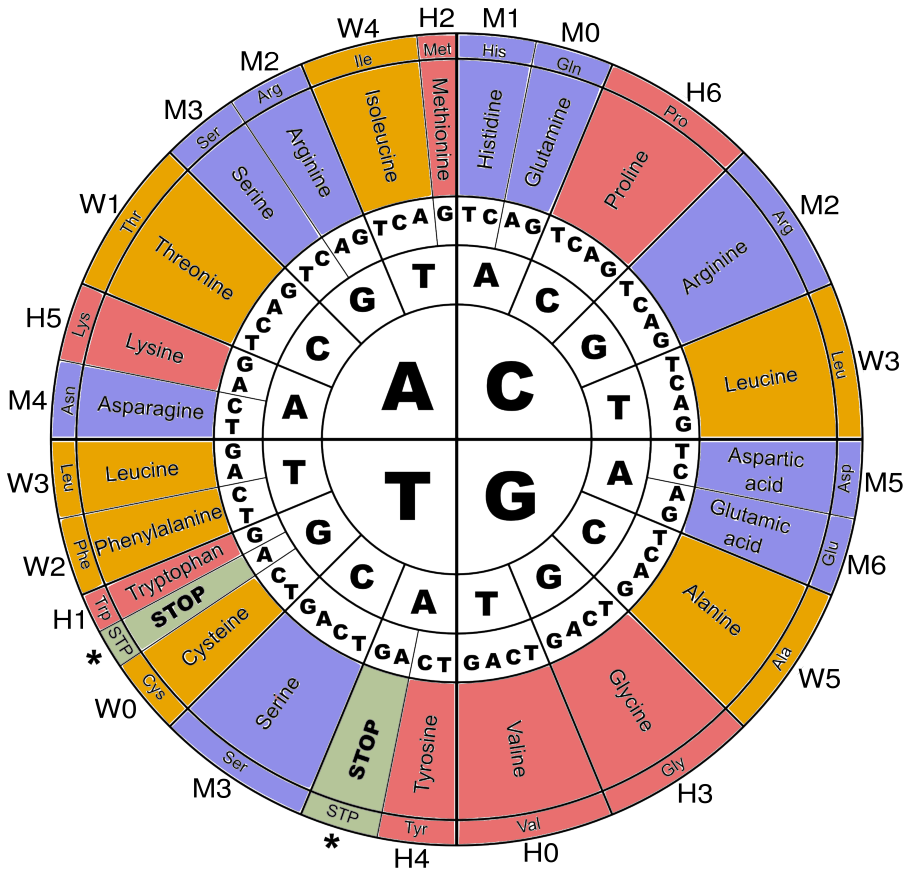


Figure IV.6 – Code génétique de Aevol_4b. 7 acides aminés sont attribués aux paramètres M et H, 6 acides aminés sont attribués aux paramètres W et les trois codons TAA, TAG et TGA correspondent classiquement au $STOP$ de traduction.

exemple la Lysine, qui, dans ce test, codait pour les digits M_{11}, W_1 et H_5 tandis que l'Asparagine codait pour les digits M_4, W_7 et H_2. On voit qu'en mutant, par exemple, un codon AAA (codant pour une Lysine) en un codon AAT (codant pour une Asparagine), les trois paramètres du triangle vont varier, mais l'un (W) à la hausse et les deux autres (M et H) à la baisse [1], entraînant de très forts effets épistatiques sur la composition des gènes. Cette approche a très vite été abandonnée en raison de la lenteur de l'évolution, lenteur probablement due aux fortes contraintes que ce codage impose (la rugosité du fitness landscape rendant rapidement les mutations favorables rarissimes). C'est pourquoi nous avons retenu par défaut une assignation des acides aminés qui limite les contraintes épistatiques. En outre, étant donné le caractère très abstrait de l'espace fonctionnel dans Aevol, l'assignation des acides aminés a probablement peu d'importance pour la production de benchmarks. Il n'en reste pas moins que d'autres assignations mériteraient d'être explorées, notamment de donner un degré d'interaction variable aux différents acides aminés. On pourrait ainsi assigner les acides aminés fréquents dans le code génétique (par

1. L'intensité de la hausse ou de la baisse dépendant de la position de l'acide aminé dans la séquence primaire de la protéine, il n'est pas possible de la déduire dans cet exemple.

exemple la Serine avec ses six codons) à un seul des trois paramètres et les acides aminés rares (par exemple le tryptophane qui n'est codé que par un unique codon) aux trois paramètres [1]. Une telle assignation permettrait de refléter le fait que les acides aminés n'ont pas tous le même impact sur la structure des protéines et d'influencer la fréquence d'apparition des différents acides aminés dans les séquences protéiques. En somme, si les besoins de modélisation viennent à changer ou à s'affiner pour la production de benchmarks, bon nombre de possibilités resteraient à explorer dans cette composante du modèle.

1. Il existe une corrélation entre la fréquence des acides aminés dans le code génétique et différentes propriétés comme leur poids moléculaire. Plus les acides aminés sont rares, plus on suppose qu'ils ont un impact fort sur la structure, et donc sur la fonction, des protéines.

4 Production de Benchmarks avec Aevol_4b

4.1 Protocole de test

Le projet de production de benchmarks avec Aevol repose sur la collaboration entre l'équipe Beagle (Inria de Lyon, LIRIS) produisant les données avec Aevol_4b et l'équipe Le Cocon (LBBE, Université Lyon 1 et CNRS) testant des méthodes de phylogénie moléculaire sur ces données. L'organisation de cette collaboration a été pensée pour tester les benchmarks produits par l'équipe Beagle selon un protocole inspiré du « double aveugle » : les données sont transmises à l'équipe Le Cocon sans information sur les paramètres de simulations (taux de mutation en particulier) et sans information sur la forme de l'arbre d'espèces que les simulations ont suivi. Ainsi, de la même façon que la conception du modèle à quatre bases d'Aevol (et Aevol lui-même) et la production de données sont faites sans connaissance des méthodes de phylogénie moléculaire, l'inférence des histoires évolutives a été faite sans connaissance des principaux choix de modélisation, qu'il s'agisse du modèle lui-même ou des choix effectués pour la génération de données. Nous allons d'ailleurs voir que cette méconnaissance mutuelle (choisie) a parfois pu conduire à quelques incompréhensions.

4.2 Résultats préliminaires

Dans la section 2.4, nous avons montré comment Aevol peut être un bon candidat pour la production de benchmarks en évolution moléculaire. Fort de la nouvelle version de Aevol capable de produire des séquences formées des quatre bases classiques de l'ADN, nous pouvons étendre cette preuve de concept à d'autres méthodes de phylogénie moléculaire. La Figure IV.7 représente un cladogramme régulier, que nous avons utilisé comme arbre d'espèces pour générer des données avec Aevol. Le protocole suivant a été suivi : Nous avons simulé une population naïve pendant 1 100 000 générations, c'est-à-dire jusqu'au premier nœud du cladogramme. Une fois simulée jusqu'à ce nœud, la population a été dupliquée, c'est-à-dire que la population a été copiée à l'identique. Puis ces deux populations ont évolué en parallèle, mais en utilisant des graines pseudo-aléatoires différentes. En d'autres termes, les conditions évolutives sont restées exactement identiques dans toutes les branches de l'arbre, mais les composantes stochastiques de l'évolution [1] ont suivi des trajectoires différentes (mais tirées selon les mêmes distributions). Les simulations ont ainsi été poursuivies jusqu'au prochain nœud où le processus de duplication de population est répété ou bien, jusqu'aux feuilles de l'arbre si plus aucun nœud n'est rencontré sur la branche. Au total, six populations finales ont été générées selon ce processus, toutes ayant évolué pendant 1 350 000 générations (avec 250 000 générations de temps de divergence maximale depuis le premier nœud). Plusieurs processus de branchement plus complexes pourraient être envisagés mais en l'état du modèle, dupliquer les populations apparaissent comme le processus se rapprochant le plus d'une spéciation liée à un isolement reproducteur (spéciation allopatrique par vicariance).

1. Dans Aevol les composantes stochastiques de l'évolution sont les mutations (occurrence et locus/-loci) et la dérive liée à l'échantillonnage aléatoire des reproducteurs.

L'évolution de la fitness le long des six lignées est présentée sur la Figure IV.8 (la fitness de chaque lignée ou de chaque sous-branche étant représentée avec le même code couleur que celui de la Figure IV.7). Nous avons ensuite fourni les séquences des génomes de chacune des feuilles et de chacun des nœuds (identifiés par des numéros sur la Figure IV.7) aux phylogénéticiens de l'équipe Le Cocon du LBBE qui ont utilisé ces onze séquences pour reconstituer l'arbre des espèces suivi par les simulations.

Il est important de souligner que, dans ce processus, aucune information n'a été fournie sur les séquences en question et que celles-ci n'étaient pas annotées. De même, aucune information n'a été fournie sur la structure de l'arbre. En particulier, le fait que certaines séquences correspondaient à des séquences ancestrales (et non à des feuilles de l'arbre) n'a pas été signalé. Enfin, pour ne pas donner d'information, les séquences, ont été renommées avec un identifiant en lettre dont l'ordre a été randomisé. La correspondance lettre - nombre était la suivante : 1-D, 2-C, 3-E, 4-G, 5-J, 6-A, 7-K, 8-H, 9-B, 10-F, 11-I.

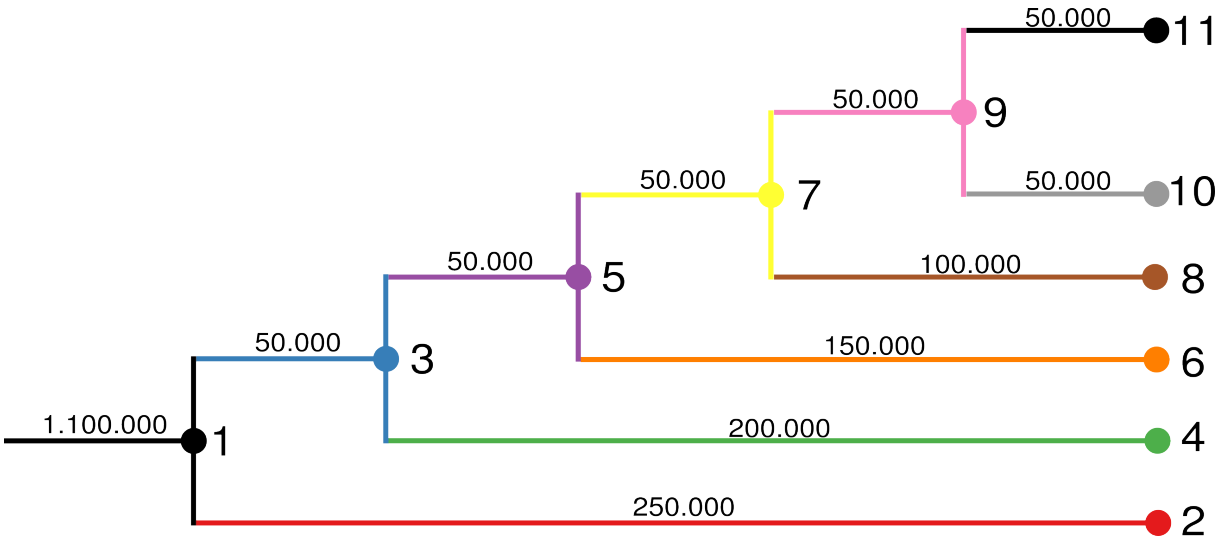


FIGURE IV.7 – Cladogramme suivi pour produire des benchmarks de test. La longueur de chaque branche est indiquée en nombre de générations. À partir d'une population initiale ayant évolué 1 100 000 générations, des duplications de populations ont été effectuées à chacun des nœuds (nombres impairs). En pratique, cela signifie qu'à chaque nœud, la population est copiée puisque chaque copie évolue séparément (avec des graines aléatoires différentes).

Les premiers essais de reconstruction à partir des onze séquences n'ont pas été concluants, montrant que sans un minimum de connaissance sur la nature des séquences fournies, la reconstruction est, sinon impossible, du moins très difficile. Ainsi, aux yeux de phylogénéticiens, les séquences de Aevol, avec leur longueur de 30 Kb environ dans les conditions du test, ne ressemblent pas à des génomes bactériens complets, mais plus à des opérons [1]. En outre, le fait que ces séquences aient pu être massivement réarrangées le long des

1. Richard Lenski, connu pour son expérience unique d'évolution expérimentale – la LTEE – mais aussi pour ses collaborations avec des informaticiens dans le cadre de la génétique digitale, compare les séquences issues de simulateurs tels que Avida ou Aevol à des virus (personal communication).

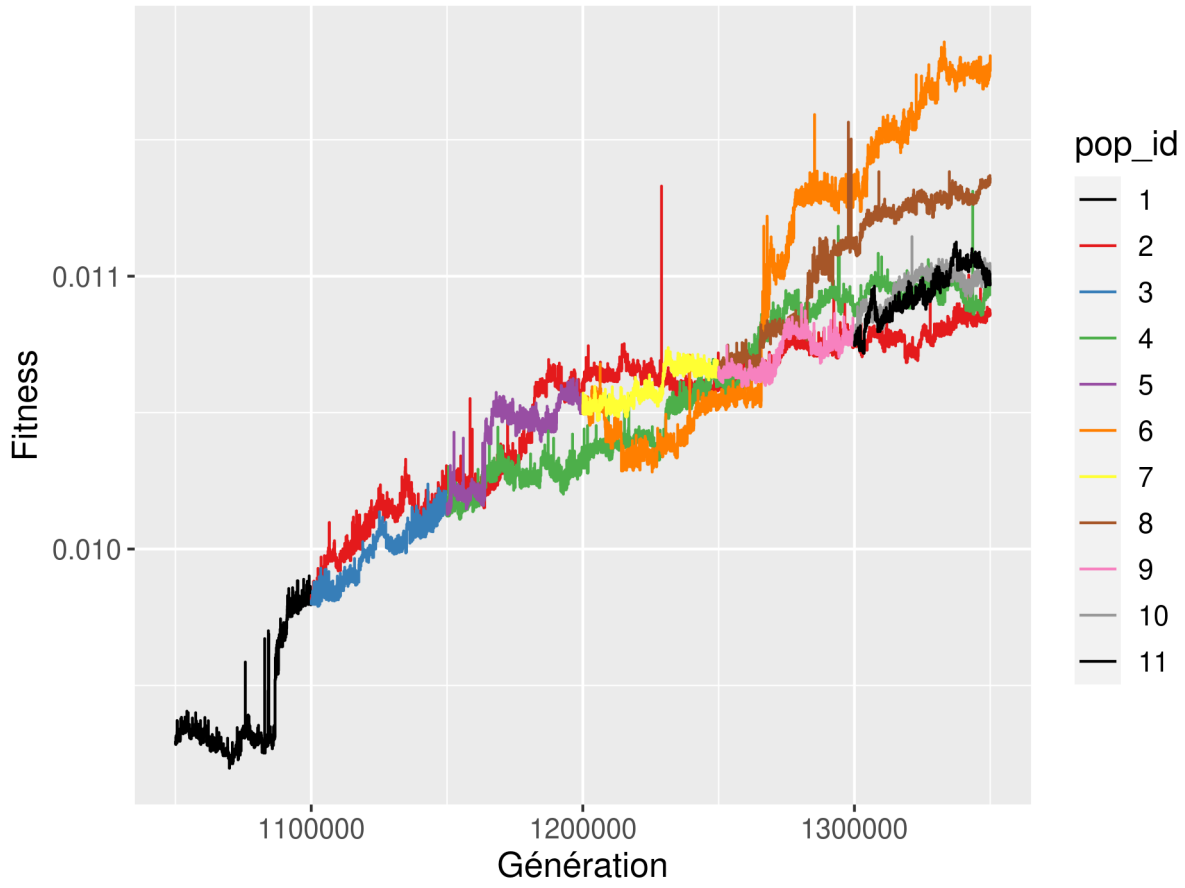


Figure IV.8 – Évolution de la fitness du meilleur individu de chaque branche pour les populations présentées en Figure IV.7. Les codes couleur des deux figures sont identiques.

branches prend à revers le réflexe premier des phylogénéticiens qui est d'aligner les séquences en autorisant exclusivement des substitutions et des « Gaps », d'autant plus que, les organismes d'Aevol fixant souvent plus d'inversions que des organismes biologiques, la synténie est peu conservée. Sans ces connaissances *a priori*, peu des phylogénéticiens ayant testé le jeu de données ont eu l'idée d'utiliser des outils d'alignement incluant les réarrangements chromosomiques.

Une fois les phylogénéticiens informés de la présence de réarrangements, l'alignement puis la reconstruction d'un arbre furent grandement facilités. La Figure IV.9 présente ainsi l'arbre reconstruit avec l'algorithme IQ-TREE (Nguyen *et al.*, 2015) après un alignement multiple par progressiveMauve, un outil d'alignement de séquences prenant en compte les réarrangements (Darling *et al.*, 2010). Malgré la différence de topologie apparente (IQ-TREE considérant que tous les génomes sont des feuilles – d'où la présence de très petites branches et la structure très déséquilibrée de l'arbre inféré), l'arbre inféré ne présente qu'une erreur de branche aux niveaux des points 9-B, 10-F et 11-I, c'est-à-dire là où les temps de divergence entre les feuilles sont les plus faibles. En outre, en comparant les longueurs de branches de l'arbre inféré avec les temps de divergence entre les séquences de l'arbre vrai (Figure IV.10), on constate une excellente corrélation entre les longueurs

inférées et vraies, ce qui semble montrer que l'inférence d'arbres phylogénétiques semble applicable aux données d'Aevol_4b en utilisant des outils d'inférence classique et sans qu'aucune des deux composantes du processus (algorithme de simulation et outils d'inférence) n'aient été développés en concertation. À notre connaissance, cette expérience de reconstruction de données issues d'un algorithme de génétique numérique par des outils de phylogénie « sur étagère » est inédite.

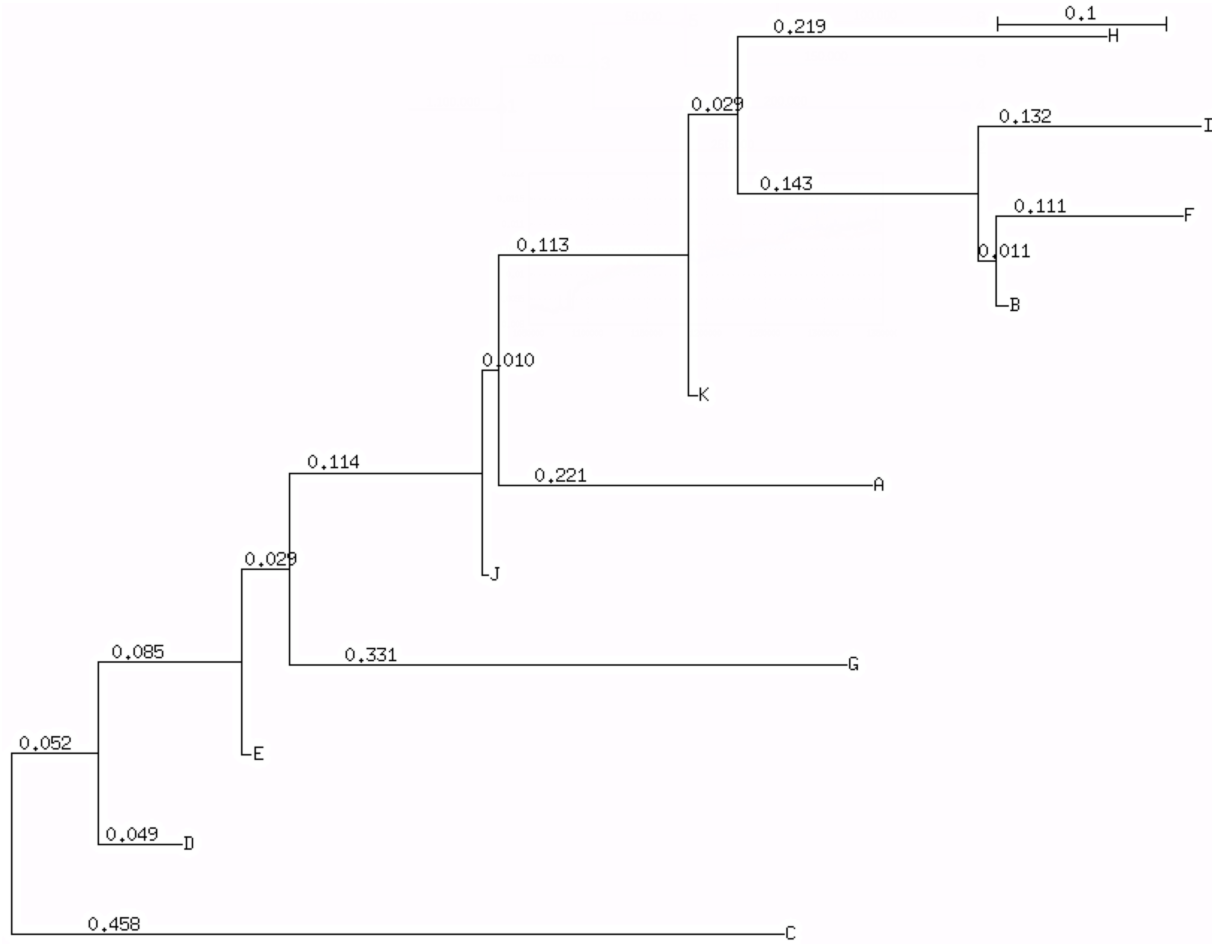


FIGURE IV.9 – Arbre phylogénétique reconstruit par l'équipe Le Cocon à partir des séquences présentées en Figure IV.7. Les correspondances entre identifiants alphabétiques et identifiants numériques sur la Figure IV.7 est : D-1, C-2, E-3, G-4, J-5, A-6, K-7, H-8, B-9, F-10, I-11. Les nœuds ancestraux sont bien tous représentés par des branches courtes (D, E, J, K, B). L'arbre compte une seule erreur de branche (Feuille B).

4.3 Évolution le long d'un arbre aléatoire

Après avoir obtenu des résultats concluants lors de nos premiers tests sur un arbre phylogénétique simpliste, nous avons pris l'initiative de générer de nouvelles données en suivant un cladogramme plus complexe et réaliste, illustré dans la Figure IV.11. Ce cladogramme a été simulé en utilisant un processus de Gillespie (Gillespie, 1976), au cours

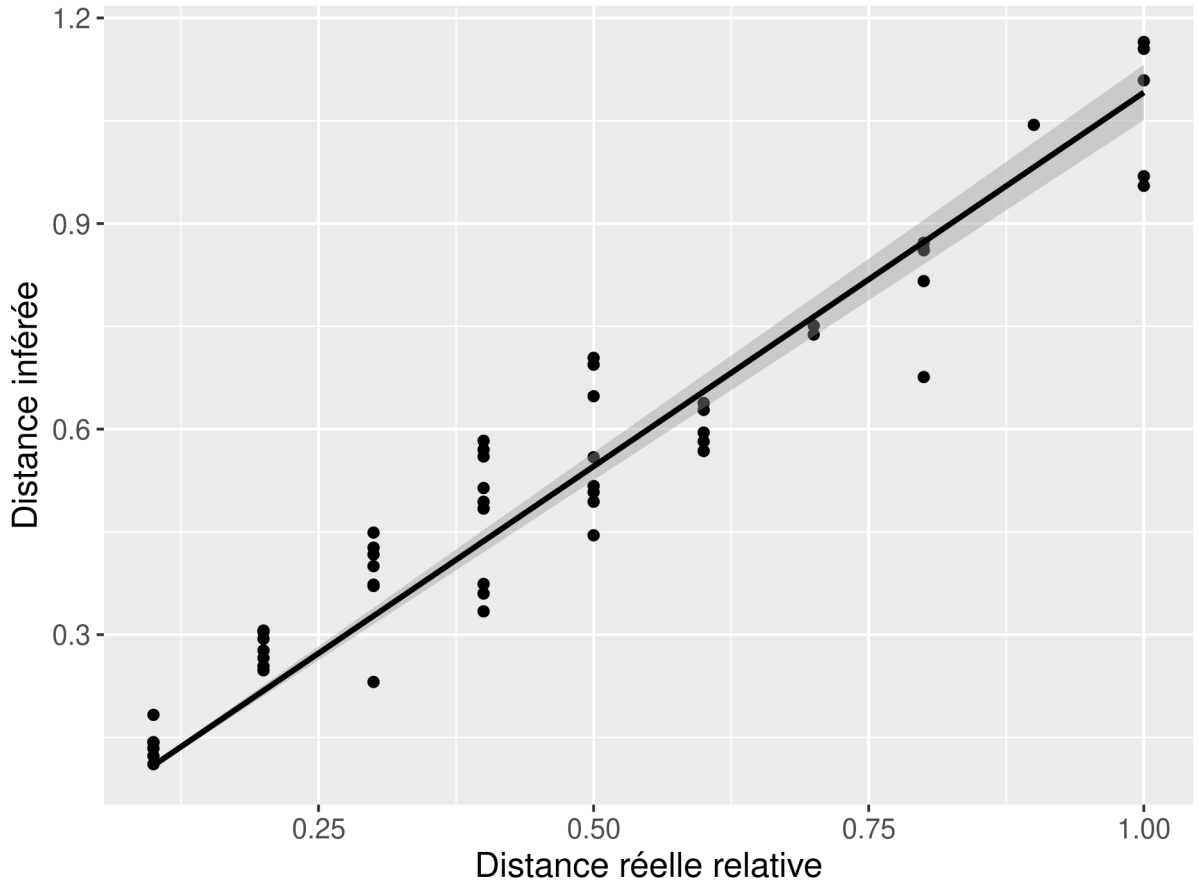


Figure IV.10 – Correspondance entre distance réelle des branches (Figure IV.7) et distance inférée (Figure IV.9). Droite en noire : régression linéaire de pente $y = 1,091$ avec un coefficient de corrélation $r^2 = 0,98$. Zone grisée : écart-type.

duquel les probabilités de branchement sont déterminées selon une distribution de Poisson à paramètres constants. Les paramètres du processus de branchement ont été ajustés afin d'obtenir précisément 40 branches terminales sur le cladogramme.

À partir de ce cladogramme nouvellement généré, nous avons suivi le même protocole que lors de notre premier test pour simuler l'évolution des populations. La Figure IV.12 présente l'évolution de la fitness des meilleurs individus au cours du temps pour chaque population, tandis que la Figure IV.13 présente la diversité des principales caractéristiques génomiques de 40 séquences feuilles. Il convient de noter que, cette fois-ci, seuls les génomes des feuilles ont été extraits. Les 40 séquences ont ensuite été mises à disposition de la communauté et un concours informel a été lancé pour la reconstruction de l'arbre [1]. La Figure IV.14 présente la reconstruction de l'arbre produite par l'équipe Le Cocon en suivant le même protocole que pour l'exemple précédent (alignement par progressiveMauve, reconstruction de l'arbre par IQ-TREE). L'arbre ne présente que deux erreurs sur la topologie de branches très courtes (séparées de 1 000 générations lors de la simulation). Comme dans l'exemple précédent, les temps de divergence inférés sont très

1. `https://project.inria.fr/evoluthon/fr/evoluthon-contest-2/`.

proches des temps simulés. Sur ce point cependant, il est intéressant de signaler que la procédure de reconstruction en double aveugle a ici permis, étonnamment, d'identifier une erreur de l'arbre vrai. En effet, lors des tests, nous considérions naïvement que l'arbre vrai était l'arbre produit par l'algorithme de branchement. Or, comme nous l'avons énoncé ci-dessus, cet arbre détermine des événements de spéciation allopatrique par vicariance. De ce fait, les temps de divergence des espèces ne correspondent pas, en général, au temps de divergence des individus (qui doivent être rallongés des temps de coalescences entre les individus des deux populations). De ce fait, l'arbre vrai ne peut être connu qu'*a posteriori* (en reconstruisant les lignées exactes des feuilles). En outre, certaines branches particulièrement courtes de l'arbre des espèces peuvent se retrouver inversées sur l'arbre vrai.

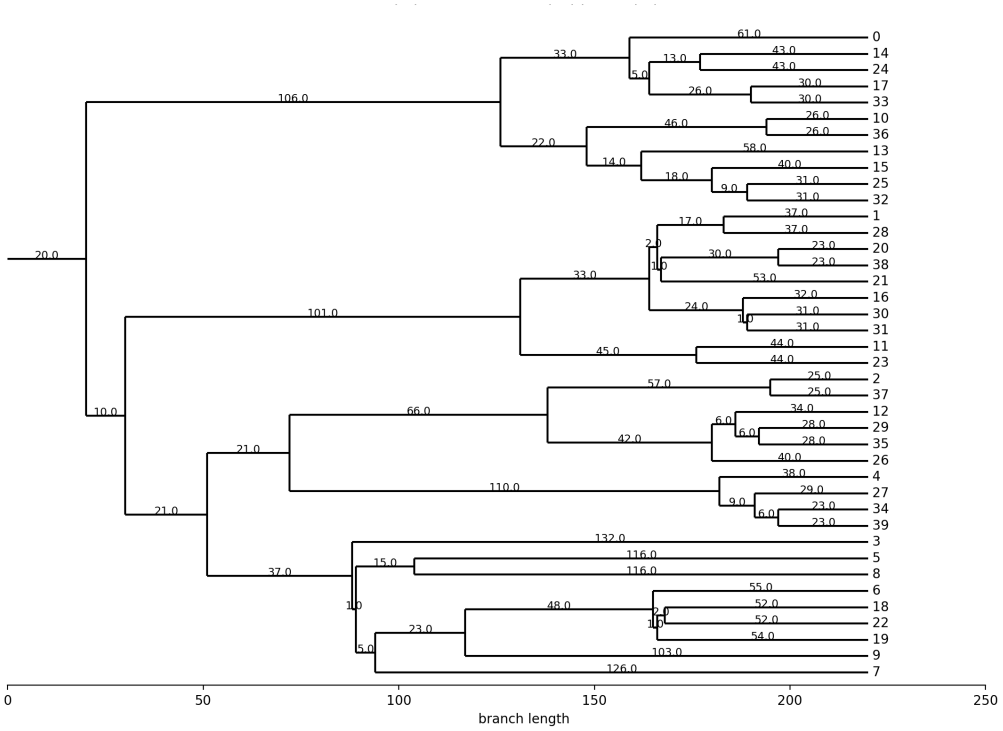


FIGURE IV.11 – Cladogramme construit par un processus de Gillespie pour le deuxième jeu de test. Les probabilités de branchement sont tirées suivant une loi de Poisson. Les paramètres du modèle ont été choisis pour générer 40 branches terminales.

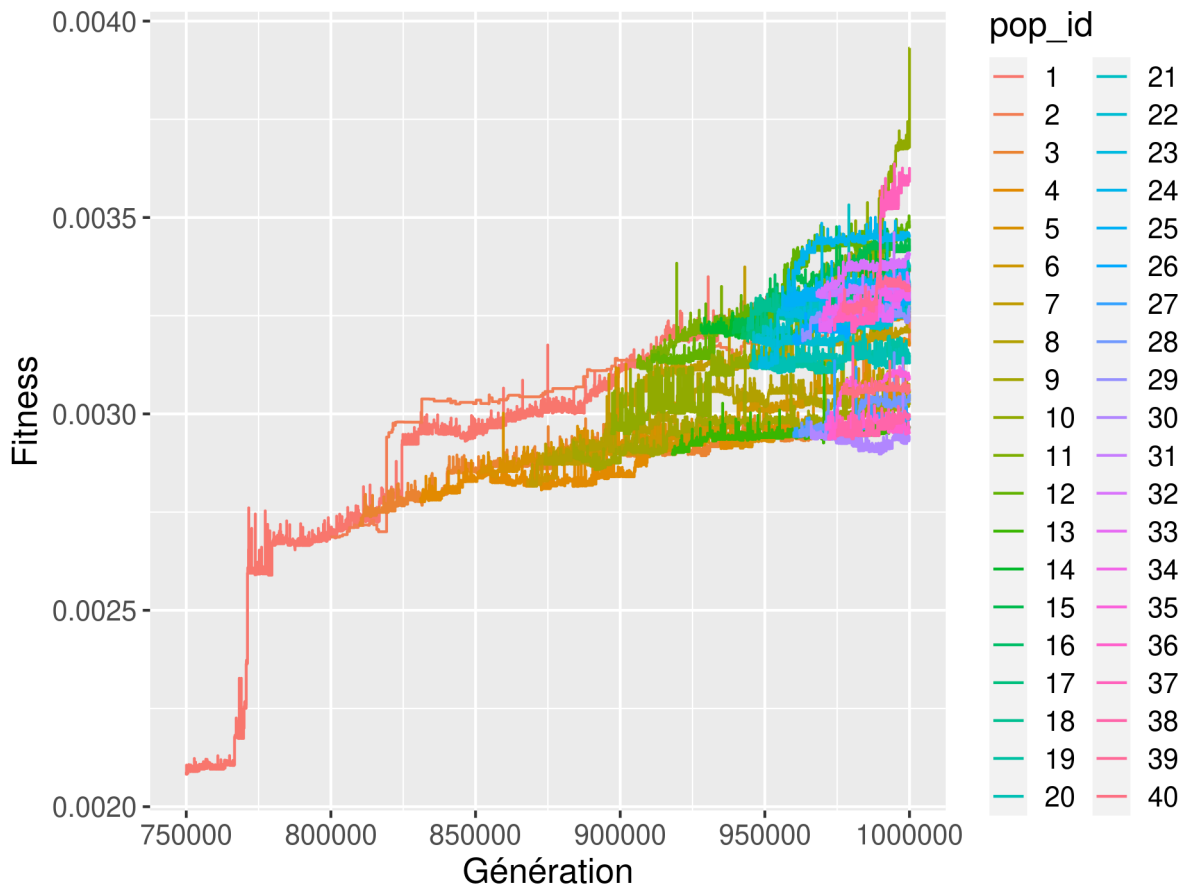


Figure IV.12 – Évolution de la fitness des meilleurs individus des 40 populations présentées en Figure IV.11.

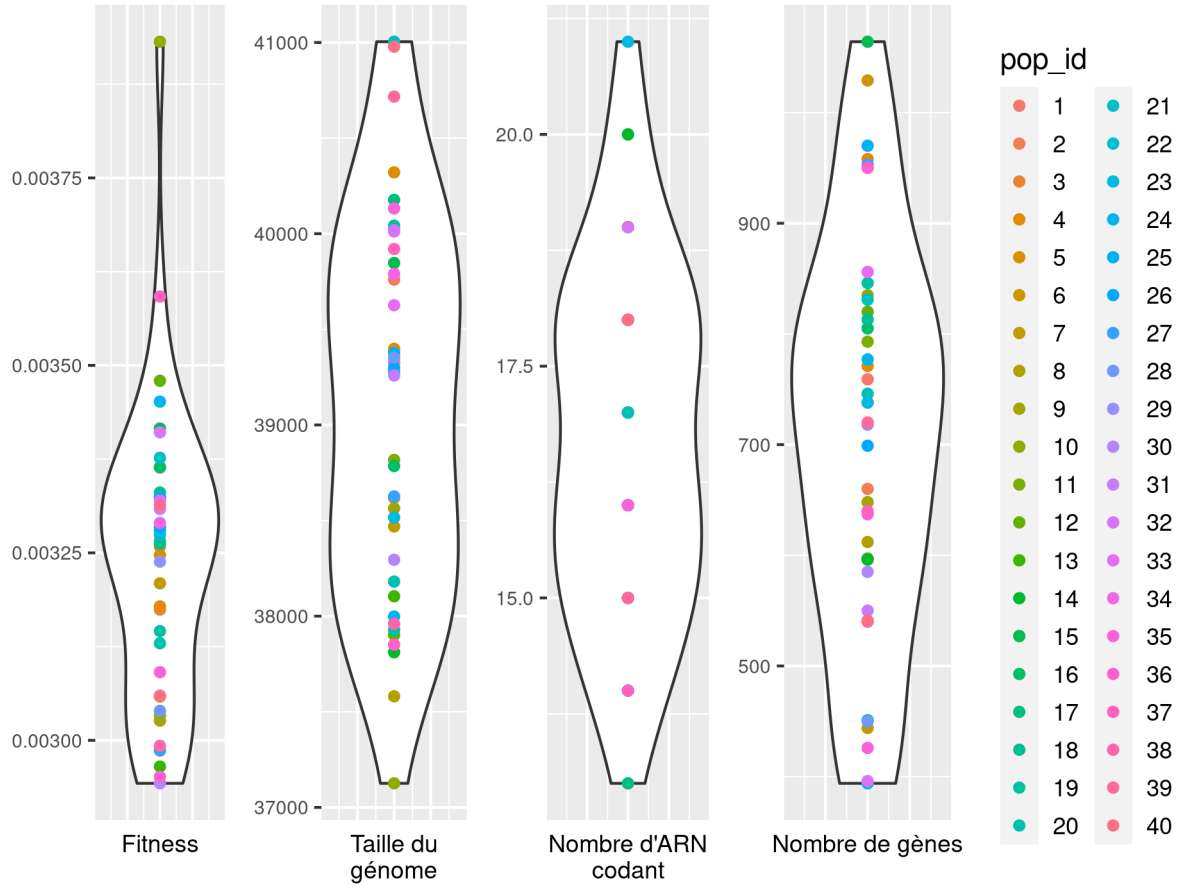
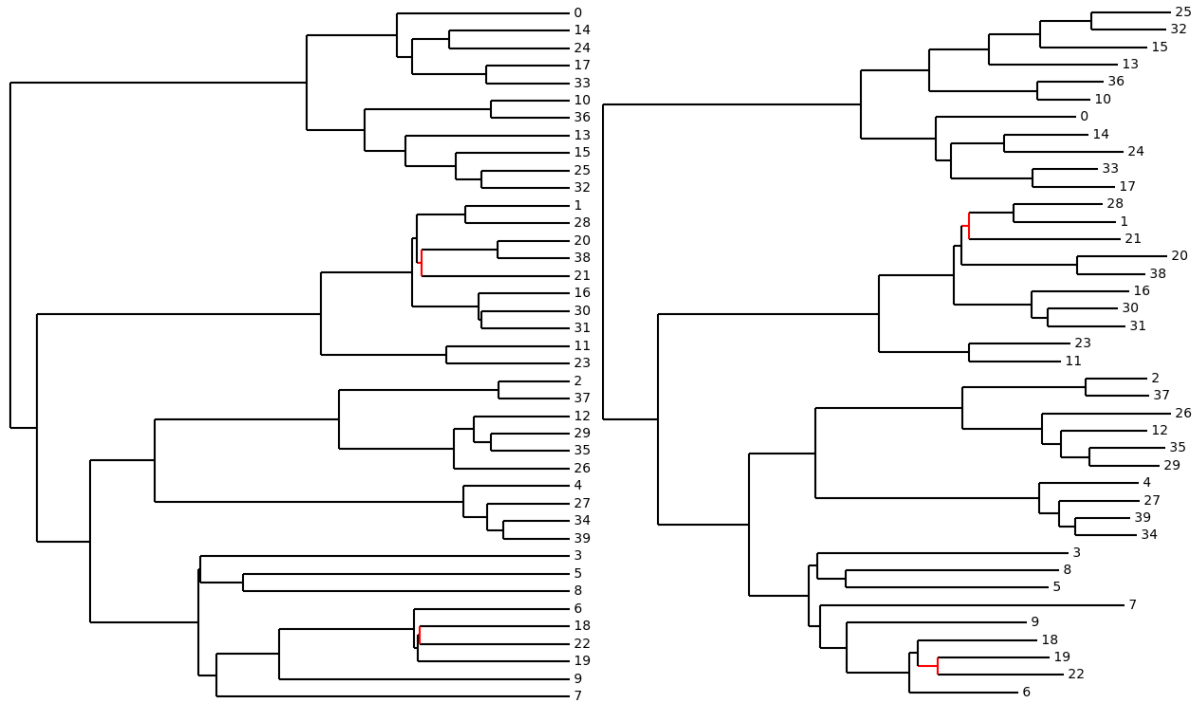


FIGURE IV.13 – Distribution de la fitness, de la taille du génome, lu nombre d'ARN codants et du nombre de gènes à la génération 1 000 000 pour les meilleurs individus des 40 populations terminales.



(a) Arbre vrai (b) Arbre inféré

Figure IV.14 – Comparaison entre l'arbre vrai (a) et l'arbre inféré (b) par l'équipe Le Cocon pour les séquences simulées pour le deuxième jeu de test. Les branches fausses sont indiquées en rouge.

5 Perspectives

Les deux tests présentés dans les sections précédentes sont très encourageants puisqu'ils montrent qu'il est possible de reconstruire l'histoire évolutive d'espèces simulées à partir des séquences terminales. En résumé, « ça marche ! ». Pourtant, dans le même temps, ces résultats se révèlent frustrants, précisément parce que « ça marche... trop bien ». C'est certes une excellente nouvelle, mais au final, plus pour le simulateur (qui s'en trouve renforcé) que pour les algorithmes d'inférence (qui n'en tireront rien). Impossible de se cacher qu'on aurait aimé remettre en cause un petit quelque chose, à l'instar de Biller *et al.* (2016b) et de la remise en cause, majeure celle-ci, des principes méthodologiques de l'inférence de distances d'inversions.

Même si les performances des méthodes d'inférence appliquées aux données d'Aevol_4b peuvent être un peu frustrantes, elles ouvrent de très nombreuses perspectives. En effet, jusqu'à maintenant, les simulations sont restées relativement simplistes et nous n'avons fait « que » simuler des événements de spéciation de façon assez grossière. Ainsi, lors des branchements, aucun des paramètres de simulation (mise à part la graine aléatoire) n'a été modifié. De même, les paramètres de simulation (y compris les probabilités de branchement) sont restés constants le long de toutes les branches. Cela rend, d'une part, l'inférence relativement simple, d'autre part, cela limite les paramètres à inférer à la seule topologie de l'arbre. Or, Aevol permet de modifier différents paramètres d'évolution d'intérêt pour les méthodes de phylogénie. Lors des branchements (ou même le long des branches et indépendamment des branchements) nous pourrions modifier par exemple la taille de population, les taux de mutation, l'environnement (cible phénotypique) ou n'importe quelle combinaison des trois. Ces trois paramètres sont aisément modifiables avec les fonctionnalités actuelles du simulateur (même si la modification de la taille de la population a demandé de mettre en place un processus spécifique afin de garantir un échantillonnage représentatif des individus dans la nouvelle présentation). Or, nous avons vu, dans les chapitres II et III, que ces paramètres sont susceptibles d'influer fortement sur la dynamique évolutive et même sur les caractéristiques structurelles de génomes. La génération de nouveaux jeux de test dans lesquels ces paramètres seraient modifiés régulièrement est donc possible et permettrait, d'une part, de tester les méthodes d'inférences sur des arbres plus compliqués, et d'autre part, de tester les méthodes d'inférence de traits d'histoire de vie ancestraux (par exemple la présence de bottlenecks reproductifs dans la phylogénie de certains clades).

Au cours du projet Evoluthon, l'équipe Le Cocon a montré un très fort intérêt pour les transferts horizontaux. De fait, ces vingt dernières années, les transferts horizontaux se sont révélés comme une composante majeure de l'évolution des organismes (Gogarten et Townsend, 2005; Keeling et Palmer, 2008; Syvanen, 2012; Arnold *et al.*, 2022). Dans Aevol, il est possible de modéliser des transferts horizontaux entre les individus d'une même population, mais la taille limitée des populations et l'exclusion de niche en limitent l'intérêt (Parsons *et al.*, 2012). De fait, les transferts qui intéressent l'équipe Le Cocon sont les transferts interspécifiques. Or, ceci est nettement plus complexe étant donné qu'Aevol n'inclut pas d'échelle écologique. La manière la plus simple de procéder serait probablement de simuler deux populations indépendantes (sans compétition interspécifique) sur

des cibles phénotypiques différentes. Il serait ensuite possible d'implémenter des mécanismes de transfert horizontal entre ces deux populations. Le transfert serait alors vu comme un nouveau type de mutation, similaire aux duplications, mais interspécifique. Ce type d'approche peut se révéler utile pour tester les méthodes d'inférence, surtout si les arbres d'espèces sont échantillonnés de façon incomplète. Cette question constitue une perspective particulièrement prometteuse au vu de l'intérêt de la communauté pour le transfert horizontal.

Enfin, même si nous n'avons évoqué ce point que très succinctement dans les pages précédentes, la transition vers des séquences à quatre nucléotides et, surtout, l'utilisation d'un code génétique redondant, permettent d'utiliser dans Aevol des mesures classiquement utilisées en évolution moléculaire. Une des plus notables est le rapport entre le taux de substitutions non-synonymes (dN) et le taux de substitutions synonymes (dS), couramment appelé dN/dS. En effet, une substitution est dite synonyme, si lorsqu'elle a lieu au sein d'un codon, elle ne modifie pas l'acide aminé codé. Cette propriété, directement liée à la redondance du code, est totalement absente dans la version canonique d'Aevol mais a été introduite mécaniquement par le passage à quatre nucléotides. Or, plus un gène a un effet important sur la fitness, plus le dN est faible au sein de la séquence du gène, car les mutations non-synonymes sont purgées par la sélection. Le dS, lui, ne dépend (en première approximation) que du taux de mutation car une mutation synonyme n'impacte pas (ou peu) la fitness. À taux de mutation égal, le dS devrait donc être constant le long du génome. En conséquence, dN/dS égal à 0 indique une sélection purificatrice parfaite (pas de substitution non-synonyme), un dN/dS de 1 indique de la dérive totale (autant de substitutions synonymes que non synonymes) et un dN/dS supérieur à 1 est un signe de sélection positive (plus de substitutions non-synonymes que synonymes). Le dN/dS est donc un outil classique pour déterminer le régime de sélection sous lequel évolue une séquence. En outre, le dN/dS est utilisé pour inférer de très nombreux traits d'histoire de vie, dont la taille effective de la population. Bien que nous ayons dans Aevol une information parfaite sur tous les événements et mutations subit par les génomes au cours de leur évolution. Tracer précisément l'histoire des gènes pour comprendre leur dynamique évolutive reste complexe. Les mesures de dN/dS étant un outil bien éprouvé en évolution moléculaire, il pourrait être utile de l'utiliser, soit pour analyser les dynamiques évolutives dans Aevol, soit pour mieux comprendre la façon dont cette mesure change au cours du temps ou en fonction des paramètres évolutifs (en particulier la taille effective de la population). Ainsi, un exemple de protocole expérimental à mettre en place serait de faire évoluer plusieurs populations à partir d'un Wild-Type et de regarder la moyenne et la variance de dN/dS de chacun des gènes. Des gènes avec un dN/dS moyen proche de 0 et peu de variance identifieraient des gènes sous forte sélection purificatrice. En remontant la lignée des génomes portant ces gènes (dans le wild-type), nous pourrions par exemple confirmer (ou infirmer) l'hypothèse selon laquelle, les gènes sous forte sélection purificatrice sont des gènes très conservés, apparus tôt dans l'histoire évolutive.

Chapitre V

Conclusion

Dans l'introduction nous avons soulevé la difficulté qu'a Aevol, malgré des résultats pertinents pour la biologie de l'évolution, à propager largement ces résultats. Nous avions avancé deux explications pour expliquer ce manque de rayonnement. Premièrement, les simulations menées jusqu'alors étaient à la fois trop limitées dans la durée et dans le nombre de réplicats. Cela réduisait le potentiel d'analyse et de persuasion des résultats, particulièrement dû au fait que les processus en évolution ont lieu en temps long et sont souvent très variables. Deuxièmement, les séquences génétiques d'Aevol étant binaires, les résultats issus de l'exploitation du modèle peinent à convaincre la communauté scientifique, en particulier les biologistes. En outre, cette caractéristique du modèle limite la généralisation de l'usage d'Aevol à une large variété de questions en biologie évolutive.

Nous pensons que les travaux présentés dans les chapitres III et IV ont fourni des réponses à ces deux limitations :

Dans le chapitre III nous avons mené la plus importante campagne de calcul jamais effectués avec Aevol. Toutes expériences confondues, cela représente plus de cinq mille milliard d'individus simulés. L'ampleur de ces simulations nous a permis d'étudier finement l'évolution de la quantité des séquences non-codantes. Nous avons montré que les patterns de taille de non-codant observés dans les génomes bactériens, notamment le maintien quasi systématique de ces séquences, peuvent être expliqués sans avoir recours aux *a priori* classiquement utilisés dans les hypothèses actuelles. A savoir, le biais vers les courtes délétions et les effets faiblement délétères de l'ADN non-codant. Nous avons aussi montré, que la quantité de non-codant est régi par deux forces contraires. La première est une force de sélection pour des génomes plus courts car plus robustes aux évènements de réplication. La deuxième est un biais de robustesse mutationnelle entre grande duplication et grande délétion, menant à l'accumulation de séquences non-codantes par dérive génétique. Nous sommes, à notre connaissance, les premiers à avoir caractérisé ce biais entre les réarrangements chromosomique déséquilibrés (duplication et délétion). Nous pensons qu'il revêt une importance particulière pour expliquer l'origine de l'accumulation de séquences non-codantes chez les bactéries. En effet, étant donné que ces organismes présentent des biais spontanés vers les courtes délétions et que les séquences dites « égoïstes », telles que les séquences d'insertions (IS), y sont souvent absentes, la biologie évolutive peine à y expliquer l'accumulation de séquences non-codantes. En outre, l'interaction de ces deux forces contraires avec des paramètres populationnels tels que la taille effective (Ne) entraîne une modification de l'équilibre de taille des génomes dans le même sens que celui

prédit par les modèles de génétique des populations ou observé dans de nombreux clades.

Un autre apport de ces travaux, est la mise en lumière de l'importance des réarrangements chromosomiques dans la dynamique évolutive de la structure des génomes. Les deux forces présentées plus haut sont toutes deux des conséquences plus ou moins directes des réarrangements. Cependant, en raison de leurs forts effets délétères, ces mutations sont bien plus rarement fixées que les mutations locales (50 fois moins dans nos expériences). Bien que ceci rende les réarrangements moins directement apparents dans les génomes, il semble que la sélection pour ce prémunir des effets délétères de ces mutations ait un très fort impact sur la structure des génomes, dépassant largement l'effet des mutations locales. En résumé, contrairement au dogme actuel de l'évolution moléculaire qui ignore souvent les réarrangements chromosomiques, nous pensons qu'ils ont une place centrale pour expliquer l'évolution de la structure des génomes car bien qu'invisibles (car fortement contre-sélectionnés), ils exercent des contraintes majeures sur l'évolution.

Dans le chapitre IV nous avons présenté Aevol_4b, une nouvelle version du simulateur incluant les quatre bases ACTG dans les séquences ADN ainsi que le code génétique standard. D'autres tentatives de modification d'Aevol en modèle à quatre base avaient été tentées par le passé mais Aevol_4b est la première version réellement utilisable. En utilisant ce nouveau modèle, nous avons produit des jeux de test pour la phylogénie moléculaire. Ces premiers jeux de données sont encore assez simplistes mais Aevol_4b est encore en phase de bêta et ces premiers résultats sont très prometteurs pour l'avenir de cette nouvelle version du simulateur. L'implémentation d'éléments permettant une complexification des histoires évolutives générées est encore nécessaire. En particulier intégrer la possibilité de paramétrer des modifications de la taille de population en cours de simulation et structurer les populations en sous-populations semble être des étapes essentielles. Du point de vue de leur utilisation, les données de sortie nécessiterons aussi un outil d'annotation des génomes permettant une accessibilité accrue pour tester de nouvelles méthodes.

Nous pourrions probablement encore nous étendre longtemps sur les ajouts qui pourraiens être apportés au modèle pour améliorer ces capacités en tant plateforme de test pour l'évolution moléculaire. Cependant Aevol_4b pourrait s'avérer tout aussi utile pour une utilisation en tant que modèle. En effet, cette version du modèle ouvre tout un champs de possibilités pour l'étude de l'évolution, à l'instar de ce qui a déjà été accompli avec Aevol – y compris dans notre chapitre II. Comme déjà mentionné dans le chapitre IV, nous pouvons maintenant utiliser des mesures classiquement utilisées en évolution moléculaire comme le dN/dS pour étudier les dynamiques évolutives dans les génomes d'Aevol. Le dN/dS est aussi utilisé pour estimer Ne ce qui nous permettrait d'affiner nos mesures de ce paramètre et améliorer nos comparaisons avec les modèles de génétique des populations. Le modèle peut aussi servir de modèle nul pour étudier différentes questions, comme l'évolution du taux de GC ou le biais d'usage du code. Comme montré dans le chapitre III, l'étude approfondie de modèles nuls est loin d'être dénuée d'intérêt. Enfin, Aevol_4b ouvre aussi la porte pour de nouvelles modifications du modèle jusqu'ici ne présentant pas grand intérêt, par exemple l'intégration de génomes diploïdes ou de reproduction sexuée.

A l'issue de cette thèse, nous avons donc pu montrer, d'une part, l'intérêt des très longues campagnes de test lorsqu'il s'agit de démêler des forces évolutives très intriquées, et la possibilité d'étendre la modèle vers un usage plus général en biologie évolutive et en

biologie moléculaire. A ce stade de nos travaux, ces deux approches sont restées dissociées car, ayant été menées en parallèle, il n'était pas envisageable de reprendre, avec Aevol_4b, l'intégralité des simulations présentées dans notre chapitre III et conduites avec la version classique. Nous pensons cependant que l'avenir du modèle passe à la fois par la production de benchmarks pour la biologie moléculaire, mais aussi par la quantification précise des forces évolutives agissant sur les génomes et sur les populations. Nous conjecturons que c'est le couplage entre ces deux approches qui permettra d'abord de révéler, puis de comprendre, les dynamiques observées dans le modèle et dans les populations réelles.

Chapitre VI

Annexes

1 Modèle mathématique de l'évolution de la taille des génomes

Dans le chapitre III, nous utilisons un modèle mathématique pour quantifier l'effet des réarrangements chromosomiques sur la structure des génomes. Ce modèle mathématique a été développé par Paul Banse, doctorant dans l'équipe Beagle et au LIRIS et sera présenté dans son mémoire de thèse.

Nous présentons ici ce modèle, dans la version utilisée dans la présente thèse pour estimer la probabilité d'apparition d'une mutation neutre et pour estimer l'effet moyen de chaque mutation sur la taille du génome en supposant que toute mutation délétère est filtrée par la sélection.

1.1 Propriétés utiles

$\forall Q \in \mathbb{R}$ nous avons :

$$\sum_{s=1}^{Q}\sum_{k=0}^{Q-s} 1 = \sum_{s=1}^{Q} Q - s + 1 = (Q+1)Q - \frac{(Q+1)Q}{2} = \boxed{\frac{(Q+1)Q}{2}}$$

et aussi

$$\sum_{s=1}^{Q}\sum_{k=0}^{Q-s} k = \sum_{s=1}^{Q} \frac{(Q-s)(Q-s+1)}{2} = \frac{1}{2}\sum_{p=0}^{Q-1}(p^2+p) = \frac{Q(Q-1)}{2}\Big(\frac{(2Q-1)}{6} + \frac{1}{2}\Big) = \boxed{\frac{Q(Q^2-1)}{6}}$$

1.2 Notations

Si nous considérons un génome donné de taille L avec g gènes différents, notons le nombre de bases non codantes λ et le nombre de bases codantes γ. Pour simplifier, nous supposerons :

— Qu'il n'y a pas de gènes chevauchants. Les gènes sont codants et de taille égale : $(L - N\lambda)/g$.

— Les sections non-codantes du génome sont également réparties entre les g gènes, et sont donc toutes de taille $\lambda/g \geq 1$.

— Le génome est circulaire.

— Si une mutation change quoi que ce soit dans une section codante, cette mutation est considérée comme étant toujours fortement délétère (pas de mutations synonymes ni de quasi-neutralité).

— À l'inverse, si une mutation n'affecte que les sections non-codantes du génome et si elle ne crée pas de nouveau gène, cette mutation est considérée comme neutre.

— Les séquences promotrices ne peuvent pas se produire au hasard, la première base de chaque gène est considérée comme la séquence promotrice, et seule la duplication de cette base peut créer un nouveau gène. Compte-tenu de ce qui précède, une telle duplication est toujours délétère.

En gardant à l'esprit que le génome est circulaire, sans perte de généralités, nous pouvons supposer que la première base (base d'index 1) est le début d'un gène.

1.3 Duplication et délétion

Dans ce cadre, une délétion est définie comme une mutation qui choisit au hasard deux positions dans le génome et supprime toutes les bases entre ces deux positions. Une délétion est donc neutre si et seulement si la séquence supprimée ne contient aucune partie codante. De même, une duplication est définie comme une mutation qui choisit au hasard trois positions dans le génome et copie la séquence entre les deux premières positions juste après la troisième.

Pour expliquer ce qui rend une duplication neutre, notons ces positions p_1, p_2 et p_3 respectivement. La séquence supprimée ou dupliquée sera d'index $(p_1, p_1+1, p_1+2, ..., p_2-1, p_2)$. La séquence dupliquée est insérée juste après la position p_3. Selon notre hypothèse, si p_3 est dans un gène, la duplication est toujours délétère, sauf si p_3 est la toute dernière base d'un gène, car alors la séquence sera dupliquée dans la partie non codante du génome. Si p_3 se trouve dans une partie non codante du génome, la duplication sera neutre uniquement si il n'y a aucun promoteur entre $p1$ et $p2$. Nous appelons Λ l'ensemble des positions dans le génome où p_3 peut conduire à des duplications neutres. Si $g = 1$ par exemple, la taille de Λ est $\lambda + 1$.

1.4 Probabilité de délétion neutre

Une délétion est neutre si et seulement si les bases supprimées se trouvent toutes dans l'une des g sections non codantes.

$$\begin{aligned}
v_{del} &= \sum_{i=0}^{g-1} \mathbb{P}(p_1, p_2 \in \big[1 + \frac{iL}{g} + \frac{L-\lambda}{g}; \frac{(i+1)L}{g}\big], p_1 \leq p_2) \\
&= \sum_{i=0}^{g-1} \sum_{s=1}^{\frac{\lambda}{g}} \sum_{k=0}^{\frac{\lambda}{g}-s} \mathbb{P}(p_1 = \frac{iL}{g} + \frac{L-\lambda}{g} + s) \times \mathbb{P}(p_2 = \frac{iL}{g} + \frac{L-\lambda}{g} + s + k) \\
&= \frac{1}{L^2} \sum_{i=0}^{g-1} \sum_{s=1}^{\frac{\lambda}{g}} \sum_{k=0}^{\frac{\lambda}{g}-s} 1 \\
&= \frac{1}{L^2} \sum_{i=0}^{g-1} \frac{(\frac{\lambda}{g})(\frac{\lambda}{g}+1)}{2} \\
&= \frac{\lambda}{g^2 L^2} \sum_{i=0}^{g-1} \frac{(\lambda+g)}{2} \\
&= \frac{(\lambda+g)\lambda}{2gL^2}
\end{aligned}$$

1.5 Espérance de taille des délétions neutres

$$\begin{aligned}
\delta_{del} &= \sum_{i=0}^{g-1} \sum_{s=1}^{\frac{\lambda}{g}} \sum_{k=0}^{\frac{\lambda}{g}-s} (p_2 - p_1 + 1)\ \mathbb{P}(p_1 = \frac{iL}{g} + \frac{L-\lambda}{g} + s) \times \mathbb{P}(p_2 = \frac{iL}{g} + \frac{L-\lambda}{g} + s + k) \\
&= \frac{1}{L^2} \sum_{i=0}^{g-1} \sum_{s=1}^{\frac{\lambda}{g}} \sum_{k=0}^{\frac{\lambda}{g}-s} k + 1 \\
&= v_{del} + \frac{1}{L^2} \sum_{i=0}^{g-1} \frac{(\frac{\lambda}{g})(\frac{\lambda^2}{g^2}-1)}{6} \\
&= \frac{(\lambda+g)\lambda}{2gL^2} + \frac{(\lambda^2-g^2)\lambda}{6g^2L^2} \\
&= \frac{\lambda(3\lambda g + 3g^2 + \lambda^2 - g^2)}{6g^2L^2} \\
&= \frac{\lambda(3\lambda g + 2g^2 + \lambda^2)}{6g^2L^2}
\end{aligned}$$

1.6 Probabilité de duplication neutre

Une duplication est neutre si et seulement si elle duplique une séquence sans inclure une base promotrice et la copie à une position dans Λ. Notez que la somme sur s commence à la position 2 (puisque la position 1 est, par définition, un promoteur – voir ci-dessus).

$$
\begin{aligned}
v_{dupl} &= \sum_{i=0}^{g-1} \mathbb{P}(p_1, p_2 \in \big[2 + \frac{iL}{g}; \frac{(i+1)L}{g}\big], p_1 \leq p_2) \times \mathbb{P}(p_3 \in \Lambda) \\
&= \sum_{i=0}^{g-1} \sum_{s=2}^{\frac{L}{g}} \sum_{k=0}^{\frac{L}{g}-s} \mathbb{P}(p_1 = \frac{iL}{g} + s) \times \mathbb{P}(p_2 = \frac{iL}{g} + s + k) \times \mathbb{P}(p_3 \in \Lambda) \\
&= \frac{\lambda + g}{L^3} \sum_{i=0}^{g-1} \sum_{s=2}^{\frac{L}{g}} \sum_{k=0}^{\frac{L}{g}-s} 1 \\
&= \frac{\lambda + g}{L^3} \sum_{i=0}^{g-1} \Big((\sum_{s=1}^{\frac{L}{g}} \sum_{k=0}^{\frac{L}{g}-s} 1) - (\frac{L}{g}) \Big) \\
&= \frac{\lambda + g}{L^3} \sum_{i=0}^{g-1} \Big((\frac{(\frac{L}{g}+1)(\frac{L}{g})}{2}) - (\frac{L}{g}) \Big) \\
&= \frac{\lambda + g}{L^3} \Big(\frac{g(\frac{L}{g}-1)(\frac{L}{g})}{2} \Big) \\
&= \frac{(\lambda + g)(L - g)}{2gL^2} \\
&= \frac{(\lambda + g)(L - \lambda + \lambda - g)}{2gL^2} \\
&= v_{del} + \frac{(\lambda + g)(\gamma - g)}{2gL^2}
\end{aligned}
$$

1.7 Espérance de taille des délétions neutres

$$
\begin{aligned}
\delta_{dupl} &= \sum_{i=0}^{g-1}\sum_{s=2}^{\frac{L}{g}}\sum_{k=0}^{\frac{L}{g}-s}(p_2-p_1+1)\mathbb{P}(p_1=\frac{iL}{g}+s)\times\mathbb{P}(p_2=\frac{iL}{g}+s+k)\times\mathbb{P}(p_3\in\Lambda)\\
&= \frac{\lambda+g}{L^3}\sum_{i=0}^{g-1}\sum_{s=2}^{\frac{L}{g}}\sum_{k=0}^{\frac{L}{g}-s}(k+1)\\
&= v_{dupl}+\frac{\lambda+g}{L^3}\sum_{i=0}^{g-1}\sum_{s=2}^{\frac{L}{g}}\sum_{k=0}^{\frac{L}{g}-s}k\\
&= v_{dupl}+\frac{\lambda+g}{L^3}\sum_{i=0}^{g-1}\Big(\sum_{s=1}^{\frac{L}{g}}\sum_{k=0}^{\frac{L}{g}-s}k-\sum_{k=0}^{\frac{L}{g}-1}k\Big)\\
&= v_{dupl}+\frac{\lambda+g}{L^3}\sum_{i=0}^{g-1}\Big(\frac{\frac{L}{g}(\frac{L^2}{g^2}-1)}{6}-\frac{\frac{L}{g}(\frac{L}{g}-1)}{2}\Big)\\
&= v_{dupl}+\frac{\lambda+g}{L^3}\Big(\frac{L(\frac{L^2}{g^2}-1)}{6}-\frac{L(\frac{L}{g}-1)}{2}\Big)\\
&= v_{dupl}+\frac{(\lambda+g)\big(\frac{L^2}{g^2}-1-3(\frac{L}{g}-1)\big)}{6L^2}\\
&= v_{dupl}+\frac{(\lambda+g)\big(L^2-g^2-3Lg+3g^2\big)}{6L^2g^2}\\
&= v_{del}+\frac{(\lambda+g)(\gamma-g)}{2gL^2}+\frac{(\lambda+g)\big(\lambda^2-g^2+2\lambda\gamma+\gamma^2-3Lg+3g^2\big)}{6L^2g^2}\\
&= v_{del}+\frac{(\lambda+g)3(\gamma-g)}{6gL^2}+\frac{(\lambda+g)\big(\lambda^2-g^2\big)}{6L^2g^2}+\frac{(\lambda+g)\big(2\lambda\gamma+\gamma^2-3Lg+3g^2\big)}{6L^2g^2}\\
&= v_{del}+\frac{(\lambda+g)3(\gamma-g)}{6gL^2}+\frac{\lambda\big(\lambda^2-g^2\big)}{6L^2g^2}+\frac{g\big(\lambda^2-g^2\big)}{6L^2g^2}+\frac{(\lambda+g)\big(2\lambda\gamma+\gamma^2-3Lg+3g^2\big)}{6L^2g^2}\\
&= \delta_{del}+\frac{(\lambda+g)3(\gamma-g)}{6gL^2}+\frac{g(\lambda+g)(\lambda-g)}{6L^2g^2}+\frac{(\lambda+g)\big(2\lambda\gamma+\gamma^2-3Lg+3g^2\big)}{6L^2g^2}\\
&= \delta_{del}+\frac{(\lambda+g)\big(3\gamma-3g)+g(\lambda-g)+2\lambda\gamma+\gamma^2-3Lg+3g^2\big)}{6L^2g^2}\\
&= \delta_{del}+\frac{(\lambda+g)\big(3\gamma-3g+\lambda-g^2+2\lambda\gamma+\gamma^2-3Lg+3g^2\big)}{6L^2g^2}\\
&= \delta_{del}+\frac{(\lambda+g)\big(-2\lambda g-g^2+2\lambda\gamma+\gamma^2\big)}{6L^2g^2}\\
&= \delta_{del}+\frac{(\lambda+g)(\gamma-g)(\gamma+g+2\lambda)}{6L^2g^2}
\end{aligned}
$$

1.8 Probabilité d'InDels neutres

Les probabilités pour les InDels neutres sont une version simplifiée de celles pour les duplications et les délétions. Nous devons garder à l'esprit que la taille des InDels est prise au hasard entre 1 et 6 (on utilise ici les paramètres de Aevol). Pour les délétions, nous supposons que $\frac{\lambda}{g} > 5$.

$$
\begin{aligned}
v_{indel}^{-} &= \sum_{i=0}^{g-1} \mathbb{P}(p_1, p_2 \in \left[1 + \frac{iL}{g} + \frac{L-\lambda}{g}; \frac{(i+1)L}{g}\right] , p_2 \in [p_1 + 0; p_1 + 5]) \\
&= \sum_{i=0}^{g-1} \left(\sum_{s=1}^{\frac{\lambda}{g}} \sum_{k=0}^{5} \mathbb{P}(p_1 = \frac{iL}{g} + \frac{L-\lambda}{g} + s) \times \mathbb{P}(p_2 = p_1 + k) \right. \\
&\quad \left. - \sum_{j=0}^{5} \sum_{k=j}^{5} \mathbb{P}(p_1 = \frac{iL}{g} + \frac{L}{g} - j) \times \mathbb{P}(p_2 = p1 + k) \right) \\
&= \frac{1}{6L} \sum_{i=0}^{g-1} \left(\sum_{s=1}^{\frac{\lambda}{g}} \sum_{k=0}^{5} 1 - \sum_{j=0}^{5} \sum_{k=j}^{5} 1 \right) \\
&= \frac{1}{6L} \sum_{i=0}^{g-1} \left(\frac{6\lambda}{g} - 21 \right) \\
&= \frac{6\lambda - 21g}{6L}
\end{aligned}
$$

Pour les petites insertions, le nombre de bases ajoutées n'a pas d'importance puisque nous supposons que toutes les insertions à l'intérieur des gènes sont délétères.

$$
\begin{aligned}
v_{indel}^{+} &= \mathbb{P}(p_1 \in \Lambda) \\
&= \frac{\lambda + g}{L}
\end{aligned}
$$

1.9 Espérance de taille des InDels neutres

$$
\begin{aligned}
\delta^{-}_{indel} &= \sum_{i=0}^{g-1}(p_2 - p_1 + 1)\ \mathbb{P}(p_1, p_2 \in \left[1 + \frac{iL}{g} + \frac{L-\lambda}{g}; \frac{(i+1)L}{g}\right]\ , p_2 \in [p_1 + 0; p1 + 5]) \\
&= \sum_{i=0}^{g-1}\left(\sum_{s=1}^{\frac{\lambda}{g}}\sum_{k=0}^{5}(k+1)\mathbb{P}(p_1 = \frac{iL}{g} + \frac{L-\lambda}{g} + s)\mathbb{P}(p_2 = p1 + k)\right. \\
&\quad \left. - \sum_{j=0}^{5}\sum_{k=j}^{5}\mathbb{P}(p_1 = \frac{iL}{g} + \frac{L}{g} - j)(k+1)\mathbb{P}(p_2 = p1 + k)\right) \\
&= \frac{1}{6L}\sum_{i=0}^{g-1}\left(\sum_{s=1}^{\frac{\lambda}{g}}\sum_{k=0}^{5}(k+1) - \sum_{j=0}^{5}\sum_{k=j}^{5}(k+1)\right) \\
&= \frac{1}{6L}\sum_{i=0}^{g-1}\left(\frac{21\lambda}{g} - 91\right) \\
&= \frac{21\lambda - 91g}{6L}
\end{aligned}
$$

Là encore, pour les petites insertions, le nombre de bases ajoutées n'a pas d'importance puisque nous supposons que toutes les insertions à l'intérieur des gènes sont délétères. Nous réutiliserons Λ défini ci-dessus.

$$
\begin{aligned}
\delta^{+}_{indel} &= \mathbb{P}(p_1 \in \Lambda)\sum_{k=0}^{5}\frac{(k+1)}{6} \\
&= \frac{3.5(\lambda + g)}{L}
\end{aligned}
$$

1.10 Mutations ponctuelles, inversions et translocations

Les mutations ponctuelles, les inversions et les translocations sont neutres si et seulement si elles affectent des parties non codantes du génome. Nous supposons que la réorganisation des gènes n'a pas d'effet sur la fitness. De plus, ces mutations n'influencent pas la taille du génome, leur espérance de taille est donc sans intérêt dans le cadre de ce modèle.

1.11 Probabilité de mutation ponctuelle neutre

$$\begin{aligned}\nu_{pm} &= \sum_{i=0}^{g-1} \mathbb{P}(p_1 \in \big[1 + \frac{iL}{g} + \frac{L-\lambda}{g}; \frac{(i+1)L}{g}\big]) \\ &= g \times \frac{\lambda}{gL} \\ &= \frac{\lambda}{L}\end{aligned}$$

1.12 Probabilité d'inversion neutre

$$\begin{aligned}\nu_{inv} &= \mathbb{P}(p_1 \in \Lambda) \times \mathbb{P}(p_2 \in \Lambda, p_2 \neq p_1) \\ &= \frac{\lambda+g}{L} \times \frac{\lambda+g-1}{L} \\ &= \frac{(\lambda+g)(\lambda+g-1)}{L^2}\end{aligned}$$

1.13 Probabilité de translocation neutre

Les translocations sont plus complexes, pour faciliter le calcul, on notera $\Lambda_{a,b} = \Lambda \cap [a,b]$, étant donné que notre génome est circulaire, ceci est valable que $a < b$ ou $b < a$, auquel cas l'intervalle inclut la base d'indice 0. La mutation par translocation consiste à séparer le génome en deux parties, puis à couper chaque partie, les nouvelles coupes d'une partie étant jointes à la coupe correspondante sur l'autre partie. On notera également $\mathcal{H}_i = \sum_{k=1}^{i} \frac{1}{k}$ le ième nombre harmonique.

$$\begin{aligned}\nu_{trans} &= \sum_{i_1=1}^{g} \sum_{k_1=0}^{\frac{\lambda}{g}} \mathbb{P}(p_1 = \frac{i_1 L}{g} + \frac{L-\lambda}{g} + k_1) \\ &\quad \Big[\sum_{j=1}^{L} \mathbb{P}(p_2 = p_1 + j, p_2 \in \Lambda) \times \mathbb{P}(p_3 \in \Lambda_{p1,p2}) \times \mathbb{P}(p_4 \in \Lambda_{p2,p1})\Big] \\ &= \sum_{i_1=1}^{g} \sum_{k_1=0}^{\frac{\lambda}{g}} \mathbb{P}(p_1 = \frac{i_1 L}{g} + \frac{L-\lambda}{g} + k_1)\Big[A + B + C\Big]\end{aligned}$$

Ou :

$$A = \sum_{j=1}^{\frac{\lambda}{g}-k_1} \mathbb{P}(p_2 = p_1 + j) \times 1 \times \frac{\lambda - j + g}{L - j}$$

$$B = \sum_{i_2=1}^{g-1} \sum_{k_2=0}^{\frac{\lambda}{g}} \mathbb{P}(p_2 = \frac{(i_1+i_2)L}{g} + \frac{L-\lambda}{g} + k_2) \times \frac{k_2 - k_1 + i_2\frac{\lambda}{g}}{k_2 - k_1 + i_2\frac{L}{g}} \times \frac{k_1 - k_2 + (g-i_2)\frac{\lambda}{g}}{k_1 - k_2 + (g-i_2)\frac{L}{g}}$$

$$C = \sum_{j=1}^{k_1} \mathbb{P}(p_2 = p_1 - j) \times \frac{\lambda - j + g}{L - j} \times 1$$

Développons $A + C$:

$$A + C = \frac{1}{L}\big[(\sum_{j=1}^{\frac{\lambda}{g}-k_1} \frac{\lambda - j + g}{L - j})(\sum_{j=1}^{k_1} \frac{\lambda - j + g}{L - j})\big] \tag{VI.1}$$

$$= \frac{1}{L}\big[(\sum_{j=1}^{\frac{\lambda}{g}-k_1} 1 + \frac{\lambda - L + g}{L - j})(\sum_{j=1}^{k_1} 1 + \frac{\lambda - L + g}{L - j})\big] \tag{VI.2}$$

$$= \frac{1}{L}\big[\frac{\lambda}{g} + (\sum_{J_A=L-\frac{\lambda}{g}+k_1}^{L-1} \frac{\lambda - L + g}{J_A})(\sum_{J_B=L-k_1}^{L-1} \frac{\lambda - L + g}{J_B})\big] \tag{VI.3}$$

$$= \frac{1}{L}\big[\frac{\lambda}{g} + (\sum_{J_A=L-\frac{\lambda}{g}+k_1}^{L-1} \frac{\lambda - L + g}{J_A})(\sum_{J_B=L-k_1}^{L-1} \frac{\lambda - L + g}{J_B})\big] \tag{VI.4}$$

$$= \frac{1}{L}\big[\frac{\lambda}{g} + (\lambda - L + g)(2\mathcal{H}_{L-1} - \mathcal{H}_{L-\frac{\lambda}{g}+k_1} - \mathcal{H}_{L-k_1})\big] \tag{VI.5}$$

Étant donné la complexité des équations, nous utiliserons une version simplifiée dans laquelle nous supposerons que les quatre points de coupures sont indépendants. Ainsi, les probabilités translocations neutre sont :

$$\nu_{trans} = \frac{\lambda}{L^4}$$

1.14 Robustesse mutationnelle

La robustesse mutationnelle correspond à la proportion de mutants neutres :

$$\mathcal{R}_{\text{mutational}} = \overline{\nu} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} (\nu_i)$$

Où $\mathcal{M}$ est l'ensemble de toutes les mutations possibles.

1.15 Robustesse réplicative

La robustesse réplicative est la proportion de descendants neutres. Les mutants multiples seront filtrés dans les données expérimentales. Le nombre de mutations dans une réplication suit une loi binomiale $\mathcal{B}(|\mathcal{M}|L, \mu)$. Instinctivement, cela correspond à un génome sept fois plus long et à un seul type de mutation. Par conséquent, la probabilité d'avoir précisément q mutations d'un type donné est :

$$\binom{|\mathcal{M}|L}{q}(1-\mu)^{|\mathcal{M}|L-q}\mu^q$$

La proportion totale de descendants neutres, sans tenir compte des doubles mutants, est donc :

$$\begin{aligned}\mathcal{R}_{\text{replicative}} &= \frac{(1-\mu)^{|\mathcal{M}|L} + \overline{\nu}|\mathcal{M}|\mu L(1-\mu)^{|\mathcal{M}|L-1}}{(1-\mu)^{|\mathcal{M}|L}} \\ &= \frac{(1-\mu) + \overline{\nu}|\mathcal{M}|\mu L}{(1-\mu) + |\mathcal{M}|\mu L}\end{aligned}$$

1.16 Espérance de gain de bases par réplication

En supposant que $\frac{\lambda}{g} > 5$, l'espérance du nombre de bases gagnées par réplication est :

$$\frac{1}{7}\left(\delta^+_{indel} - \delta^-_{indel} + \delta_{dupl} - \delta_{del}\right)$$

Bibliographie

Adami, C. (2006). Digital genetics : Unravelling the genetic basis of evolution. *Nature Reviews Genetics*, 7(2):109–118.

Adams, P. E., Eggers, V. K., Millwood, J. D., Sutton, J. M., Pienaar, J. et Fierst, J. L. (2023). Genome size changes by duplication, divergence and insertion in *Caenorhabditis* worms. *Molecular Biology and Evolution*, page msad039.

Ågren, J. A. et Wright, S. I. (2011). Co-evolution between transposable elements and their hosts : A major factor in genome size evolution ? *Chromosome research*, 19(6):777.

Andersson, J. O. et Andersson, S. G. E. (2001). Pseudogenes, Junk DNA, and the Dynamics of Rickettsia Genomes. *Molecular Biology and Evolution*, 18(5):829–839.

Arenas, M. et Posada, D. (2014). Simulation of Genome-Wide Evolution under Heterogeneous Substitution Models and Complex Multispecies Coalescent Histories. *Molecular Biology and Evolution*, 31(5):1295–1301.

Arnold, B. J., Huang, I.-T. et Hanage, W. P. (2022). Horizontal gene transfer and adaptive evolution in bacteria. *Nature Reviews Microbiology*, 20(4):206–218.

Audrézet, M.-P., Chen, J.-M., Raguénes, O., Chuzhanova, N., Giteau, K., Maréchal, C. L., Quéré, I., Cooper, D. N. et Férec, C. (2004). Genomic rearrangements in the CFTR gene : Extensive allelic heterogeneity and diverse mutational mechanisms. *Human mutation*, 23(4):343–357.

Banse, P., Luiselli, J., Parsons, D. P., Grohens, T., Foley, M., Trujillo, L., Rouzaud-Cornabas, J., Knibbe, C. et Beslon, G. (2023). Forward-in-time simulation of chromosomal rearrangements : The invisible backbone that sustains long-term adaptation. *in prep.*

Barrick, J. E., Yu, D. S., Yoon, S. H., Jeong, H., Oh, T. K., Schneider, D., Lenski, R. E. et Kim, J. F. (2009). Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature*, 461(7268):1243.

Batut, B., Parsons, D. P., Fischer, S., Beslon, G. et Knibbe, C. (2013). In silico experimental evolution : A tool to test evolutionary scenarios. *BMC Bioinformatics*, 14(S15):S11.

Beck, K., Beedle, M., Van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R. *et al.* (2001).

Manifeste pour le développement agile de logiciel. *En ligne : http ://agilemanifesto.org/iso/fr*.

Beslon, G., Liard, V., Parsons, D. P. et Rouzaud-Cornabas, J. (2021). Of Evolution, Systems and Complexity. *In* Crombach, A., éditeur : *Evolutionary Systems Biology*, pages 1–18. Springer International Publishing, Cham.

Beslon, G., Parsons, D., Sanchez-Dehesa, Y., Peña, J.-M. et Knibbe, C. (2010a). Scaling laws in bacterial genomes : A side-effect of selection of mutational robustness ? *Biosystems*, 102(1):32–40.

Beslon, G., Parsons, D. P., Pena, J.-M., Rigotti, C. et Sanchez-Dehesa, Y. (2010b). From digital genetics to knowledge discovery : Perspectives in genetic network understanding. *Intelligent Data Analysis*, 14(2):173–191.

Biller, P., Guéguen, L., Knibbe, C. et Tannier, E. (2016a). Breaking Good : Accounting for Fragility of Genomic Regions in Rearrangement Distance Estimation. *Genome Biology and Evolution*, 8(5):1427–1439.

Biller, P., Guéguen, L. et Tannier, E. (2015). Moments of genome evolution by Double Cut-and-Join. *BMC Bioinformatics*, 16(S14):S7.

Biller, P., Knibbe, C., Beslon, G. et Tannier, E. (2016b). Comparative Genomics on Artificial Life. *In* Beckmann, A., Bienvenu, L. et Jonoska, N., éditeurs : *Pursuit of the Universal*, volume 9709, pages 35–44. Springer International Publishing, Cham.

Blakely, E. L., Rennie, K. J., Jones, L., Elstner, M., Chrzanowska-Lightowlers, Z. M. A., White, C. B., Shield, J. P. H., Pilz, D. T., Turnbull, D. M., Poulton, J. et Taylor, R. W. (2006). Sporadic Intragenic Inversion of the Mitochondrial DNA MTND1 Gene Causing Fatal Infantile Lactic Acidosis. *Pediatric Research*, 59(3):440–444.

Blommaert, J. (2020). Genome size evolution : Towards new model systems for old questions. *Proceedings of the Royal Society B : Biological Sciences*, 287(1933):20201441.

Blount, Z. D., Barrick, J. E., Davidson, C. J. et Lenski, R. E. (2012). Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature*, 489(7417):513–518. Number : 7417 Publisher : Nature Publishing Group.

Bourguignon, T., Kinjo, Y., Villa-Martín, P., Coleman, N. V., Tang, Q., Arab, D. A., Wang, Z., Tokuda, G., Hongoh, Y., Ohkuma, M., Ho, S. Y., Pigolotti, S. et Lo, N. (2020). Increased Mutation Rate Is Linked to Genome Reduction in Prokaryotes. *Current Biology*, 30(19):3848–3855.e4.

Boussau, B., Brown, J. M. et Fujita, M. K. (2011). Nonadaptive evolution of mitochondrial genome size. *Evolution : International Journal of Organic Evolution*, 65(9):2706–2711.

Brauer, F., Castillo-Chavez, C. et Castillo-Chavez, C. (2001). *Mathematical Models in Population Biology and Epidemiology*, volume 40. Springer.

Canapa, A., Barucca, M., Biscotti, M. A., Forconi, M. et Olmo, E. (2015). Transposons, Genome Size, and Evolutionary Insights in Animals. *Cytogenetic and Genome Research*, 147(4):217–239.

Carde, Q., Foley, M., Knibbe, C., Parsons, D. P., Rouzaud-Cornabas, J. et Beslon, G. (2019). How to reduce a genome ? ALife as a tool to teach the scientific method to school pupils. *In The 2019 Conference on Artificial Life*, pages 497–504, Newcastle, United Kingdom. MIT Press.

Carvajal-Rodriguez, A. (2008). Simulation of Genomes : A Review. *Current Genomics*, 9(3):155–159.

Carvajal-Rodriguez, A. (2010). Simulation of Genes and Genomes Forward in Time. *Current Genomics*, 11(1):58–61.

Cavalier-Smith, T. (1982). Skeletal DNA and the Evolution of Genome Size. *Annual Review of Biophysics and Bioengineering*, 11(1):273–302.

Cavalier-Smith, T. (2005). Economy, Speed and Size Matter : Evolutionary Forces Driving Nuclear Genome Miniaturization and Expansion. *Annals of Botany*, 95(1):147–175.

Chan, J., Perrone, V., Spence, J. P., Jenkins, P. A., Mathieson, S. et Song, Y. S. (2018). A likelihood-free inference framework for population genetic data using exchangeable neural networks. *In* Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N. et Garnett, R., éditeurs : *Advances in Neural Information Processing Systems 31 : Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8603–8614.

Cheetham, S. W., Faulkner, G. J. et Dinger, M. E. (2020). Overcoming challenges and dogmas to understand the functions of pseudogenes. *Nature Reviews Genetics*, 21(3):191–201.

Chen, J., Dougherty, E., Demir, S. S., Friedman, C., Li, C. S. et Wong, S. (2005). Grand challenges for multimodal bio-medical systems. *IEEE Circuits and Systems Magazine*, 5(2):46–52.

Codoñer, F. M., Darós, J.-A., Solé, R. V. et Elena, S. F. (2006). The Fittest versus the Flattest : Experimental Confirmation of the Quasispecies Effect with Subviral Pathogens. *PLoS Pathogens*, 2(12):e136.

Cooper, G. M. et Shendure, J. (2011). Needles in stacks of needles : Finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics*, 12(9):628–640.

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. *Journal of Bacteriology*, 183(9):2834–2841.

Couce, A., Caudwell, L. V., Feinauer, C., Hindré, T., Feugeas, J.-P., Weigt, M., Lenski, R. E., Schneider, D. et Tenaillon, O. (2017). Mutator genomes decay, despite sustained fitness gains, in a long-term experiment with bacteria. *Proceedings of the National Academy of Sciences*, 114(43):E9026–E9035.

Couce, A., Magnan, M., Lenski, R. E. et Tenaillon, O. (2022). Predictability shifts from local to global rules during bacterial adaptation. Preprint, Evolutionary Biology.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

Crombach, A. et Hogeweg, P. (2008). Evolution of Evolvability in Gene Regulatory Networks. *PLoS Computational Biology*, 4(7):e1000112.

Dalquen, D. A., Anisimova, M., Gonnet, G. H. et Dessimoz, C. (2012). ALF—A Simulation Framework for Genome Evolution. *Molecular Biology and Evolution*, 29(4): 1115–1123.

Darling, A. E., Mau, B. et Perna, N. T. (2010). progressivemauve : multiple genome alignment with gene gain, loss and rearrangement. *PloS one*, 5(6):e11147.

Davín, A. A., Tricou, T., Tannier, E., de Vienne, D. M. et Szöllősi, G. J. (2019). Zombi : a phylogenetic simulator of trees, genomes and sequences that accounts for dead linages. *Bioinformatics*, 36(4):1286–1288.

de Visser, J. A. G., Hermisson, J., Wagner, G. P., Meyers, L. A., Bagheri-Chaichian, H., Blanchard, J. L., Chao, L., Cheverud, J. M., Elena, S. F., Fontana, W. *et al.* (2003). Perspective : Evolution and detection of genetic robustness. *Evolution ; international journal of organic evolution*, 57(9):1959–1972.

D'Iorio, M. et Dewar, K. (2023). Replication-associated inversions are the dominant form of bacterial chromosome structural variation. *Life Science Alliance*, 6(1): e202201434.

Dougherty, E. R. et Braga-Neto, U. (2006). Epistemology of computational biology : mathematical models and experimental prediction as the basis of their validity. *Journal of Biological Systems*, 14(01):65–90.

Drake, J. W. (1991). A constant rate of spontaneous mutation in DNA-based microbes. *Proceedings of the National Academy of Sciences*, 88(16):7160–7164.

Duchemin, W., Daubin, V. et Tannier, E. (2015). Reconstruction of an ancestral Yersinia pestisgenome and comparison with an ancient sequence. *BMC Genomics*, 16(S10):S9.

Edgar, RC., Asimenos, G., Batzoglou, S. et Sidow, A. (2011). Evolver : A whole-genome sequence evolution simulator.

Elena, S. F., Wilke, C. O., Ofria, C. et Lenski, R. E. (2007). Effects of Population Size and Mutation Rate on the Evolution of Mutational Robustness. *Evolution*, 61(3):666–674.

Eriksson, K. (2004). Statistical and Combinatorial Aspects of Comparative Genomics*. *Scandinavian Journal of Statistics*, 31(2):203–216.

Fares, M. A. (2015). The origins of mutational robustness. *Trends in Genetics*, 31(7):373–381.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates, Sunderland, Mass.

Fertin, G., éditeur (2009). *Combinatorics of Genome Rearrangements*. Computational Molecular Biology. MIT Press, Cambridge, Mass.

Feyerabend, P. (1975). *Against Method : Outline of an Anarchistic Theory of Knowledge*. NLB ; Humanities Press, London : Atlantic Highlands.

Frénoy, A., Taddei, F. et Misevic, D. (2013). Genetic Architecture Promotes the Evolution and Maintenance of Cooperation. *PLoS Computational Biology*, 9(11):e1003339.

Gatesy, J. et Springer, M. S. (2013). Concatenation versus coalescence versus "concatalescence". *Proceedings of the National Academy of Sciences*, 110(13).

Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22(4):403–434.

Giovannoni, S. J. (2005). Genome Streamlining in a Cosmopolitan Oceanic Bacterium. *Science*, 309(5738):1242–1245.

Giovannoni, S. J., Cameron Thrash, J. et Temperton, B. (2014). Implications of streamlining theory for microbial ecology. *The ISME Journal*, 8(8):1553–1565.

Gogarten, J. P. et Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology*, 3(9):679–687.

Goodhead, I. et Darby, A. C. (2015). Taking the pseudo out of pseudogenes. *Current Opinion in Microbiology*, 23:102–109.

Gould, S. J. et Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm : A critique of the adaptationist programme. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 205(1161):581–598.

Gregory, T. (2004). Insertion–deletion biases and the evolution of genome size. *Gene*, 324:15–34.

Gregory, T. R. et Hebert, P. D. (1999). The modulation of DNA content : Proximate causes and ultimate consequences. *Genome research*, 9(4):317–324.

Groussin, M., Boussau, B., Szöllősi, G., Eme, L., Gouy, M., Brochier-Armanet, C. et Daubin, V. (2016). Gene Acquisitions from Bacteria at the Origins of Major Archaeal Clades Are Vastly Overestimated. *Molecular Biology and Evolution*, 33(2): 305–310.

Halligan, D. L. et Keightley, P. D. (2009). Spontaneous Mutation Accumulation Studies in Evolutionary Genetics. *Annual Review of Ecology, Evolution, and Systematics*, 40(1):151–172.

Hansen, T. F. et Wagner, G. P. (2001). Modeling genetic architecture : a multilinear theory of gene interaction. *Theoretical population biology*, 59(1):61–86.

Hastings, P. J., Lupski, J. R., Rosenberg, S. M. et Ira, G. (2009). Mechanisms of change in gene copy number. *Nature Reviews Genetics*, 10(8):551–564.

Hess, J., Skrede, I., Wolfe, B. E., LaButti, K., Ohm, R. A., Grigoriev, I. V. et Pringle, A. (2014). Transposable element dynamics among asymbiotic and ectomycorrhizal Amanita fungi. *Genome biology and evolution*, 6(7):1564–1578.

Hindré, T., Knibbe, C., Beslon, G. et Schneider, D. (2012). New insights into bacterial adaptation through in vivo and in silico experimental evolution. *Nature Reviews Microbiology*, 10(5):352–365.

Kalhor, R., Beslon, G., Lafond, M. et Scornavacca, C. (2023). Classifying the Post-duplication Fate of Paralogous Genes. *In* Jahn, K. et Vinař, T., éditeurs : *Comparative Genomics*, volume 13883, pages 1–18. Springer Nature Switzerland, Cham.

Karro, J. E., Yan, Y., Zheng, D., Zhang, Z., Carriero, N., Cayting, P., Harrrison, P. et Gerstein, M. (2007). Pseudogene.org : A comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Research*, 35(suppl_1):D55–D60.

Kashtan, N. et Alon, U. (2005). Spontaneous evolution of modularity and network motifs. *Proceedings of the National Academy of Sciences*, 102(39):13773–13778.

Katju, V. et Bergthorsson, U. (2013). Copy-number changes in evolution : Rates, fitness effects and adaptive significance. *Frontiers in Genetics*, 4.

Kauffman, S. et Levin, S. (1987). Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology*, 128(1):11–45.

Keeling, P. J. et Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, 9(8):605–618.

Keeling, P. J. et Slamovits, C. H. (2005). Causes and effects of nuclear genome reduction. *Current Opinion in Genetics & Development*, 15(6):601–608.

Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, 115(1):49–63.

Kitano, H. (2004). Biological robustness. *Nature Reviews Genetics*, 5(11):826–837.

Kitano, H. (2007). Towards a theory of biological robustness. *Molecular Systems Biology*, 3(1):137.

Knibbe, C., Coulon, A., Mazet, O., Fayard, J.-M. et Beslon, G. (2007). A Long-Term Evolutionary Pressure on the Amount of Noncoding DNA. *Molecular Biology and Evolution*, 24(10):2344–2353.

Koonin, E. V. (2009). Evolution of genome architecture. *The international journal of biochemistry & cell biology*, 41(2):298–306.

Koonin, E. V. et Wolf, Y. I. (2008). Genomics of bacteria and archaea : The emerging dynamic view of the prokaryotic world. *Nucleic Acids Research*, 36(21):6688–6719.

Kuo, C.-H. et Janzen, F. J. (2003). Bottlesim : A bottleneck simulation program for long-lived species with overlapping generations. *Molecular Ecology Notes*, 3(4):669–673.

Kuo, C.-H., Moran, N. A. et Ochman, H. (2009). The consequences of genetic drift for bacterial genome complexity. *Genome Research*, 19(8):1450–1454.

Kuo, C.-H. et Ochman, H. (2009). Deletional Bias across the Three Domains of Life. *Genome Biology and Evolution*, 1:145–152.

Kuo, C.-H. et Ochman, H. (2010). The Extinction Dynamics of Bacterial Pseudogenes. *PLoS Genetics*, 6(8):e1001050.

Lauring, A. S., Frydman, J. et Andino, R. (2013). The role of mutational robustness in RNA virus evolution. *Nature Reviews Microbiology*, 11(5):327–336.

Lefébure, T., Morvan, C., Malard, F., François, C., Konecny-Dupré, L., Guéguen, L., Weiss-Gayet, M., Seguin-Orlando, A., Ermini, L., Der Sarkissian, C. *et al.* (2017). Less effective selection leads to larger genomes. *Genome research*, 27(6):1016–1028.

Lehman, J., Clune, J., Misevic, D., Adami, C., Altenberg, L., Beaulieu, J., Bentley, P. J., Bernard, S., Beslon, G., Bryson, D. M., Cheney, N., Chrabaszcz, P., Cully, A., Doncieux, S., Dyer, F. C., Ellefsen, K. O., Feldt, R., Fischer, S., Forrest, S., Fŕenoy, A., Gagńe, C., Le Goff, L., Grabowski, L. M., Hodjat, B., Hutter, F., Keller, L., Knibbe, C., Krcah, P., Lenski, R. E., Lipson, H., MacCurdy, R., Maestre, C., Miikkulainen, R., Mitri, S., Moriarty, D. E., Mouret, J.-B., Nguyen, A., Ofria, C., Parizeau, M., Parsons, D., Pennock, R. T., Punch, W. F., Ray, T. S., Schoenauer, M., Schulte, E., Sims, K., Stanley, K. O., Taddei, F., Tarapore, D., Thibault, S., Watson, R., Weimer, W. et Yosinski, J. (2020). The Surprising Creativity of Digital Evolution : A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities. *Artificial Life*, 26(2):274–306.

Lenski, R. E. (2017). Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. *The ISME journal*, 11(10):2181.

Lenski, R. E. (2023). Revisiting the Design of the Long-Term Evolution Experiment with Escherichia coli. *Journal of Molecular Evolution*.

Lenski, R. E., Ofria, C., Pennock, R. T. et Adami, C. (2003). The evolutionary origin of complex features. *Nature*, 423(6936):139–144.

Lenski, R. E., Rose, M. R., Simpson, S. C. et Tadler, S. C. (1991). Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2,000 generations. *The American Naturalist*, 138(6):1315–1341.

Lerat, E. (2004). - : Exploring the outer limits of bacterial pseudogenes. *Genome Research*, 14(11):2273–2278.

Lerat, E. (2005). Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Research*, 33(10):3125–3132.

Liard, V., Parsons, D. P., Rouzaud-Cornabas, J. et Beslon, G. (2020). The Complexity Ratchet : Stronger than Selection, Stronger than Evolvability, Weaker than Robustness. *Artificial Life*, 26(1):38–57.

Liu, L., Yu, L. et Edwards, S. V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, 10(1):302.

Liu, Q., Wang, H., Hu, H. et Zhang, H. (2015). Genome-wide identification and evolutionary analysis of positively selected miRNA genes in domesticated rice. *Molecular Genetics and Genomics*, 290(2):593–602.

Liu, Y., Harrison, P. M., Kunin, V. et Gerstein, M. (2004). [No title found]. *Genome Biology*, 5(9):R64.

Loewenthal, G., Wygoda, E., Nagar, N., Glick, L., Mayrose, I. et Pupko, T. (2022). The evolutionary dynamics that retain long neutral genomic sequences in face of indel deletion bias : A model and its application to human introns. *Open Biology*, 12(12):220223.

Long, H., Miller, S. F., Williams, E. et Lynch, M. (2018). Specificity of the DNA Mismatch Repair System (MMR) and Mutagenesis Bias in Bacteria. *Molecular Biology and Evolution*, 35(10):2414–2421.

Lower, S. S., Johnston, J. S., Stanger-Hall, K. F., Hjelmen, C. E., Hanrahan, S. J., Korunes, K. et Hall, D. (2017). Genome size in North American fireflies : Substantial variation likely driven by neutral processes. *Genome biology and evolution*, 9(6):1499–1512.

Łuksza, M. et Lässig, M. (2014). A predictive fitness model for influenza. *Nature*, 507(7490):57–61.

Lynch, M. (2006). Streamlining and Simplification of Microbial Genome Architecture. *Annual Review of Microbiology*, 60(1):327–349.

Lynch, M. (2007). The frailty of adaptive hypotheses for the origins of organismal complexity. *Proceedings of the National Academy of Sciences*, 104(Supplement 1):8597–8604.

Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics*, 26(8):345–352.

Lynch, M., Ackerman, M. S., Gout, J.-F., Long, H., Sung, W., Thomas, W. K. et Foster, P. L. (2016). Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, 17(11):704–714.

Lynch, M. et Conery, J. S. (2003). The Origins of Genome Complexity. *Science*, 302(5649):1401–1404.

Lynch, M. et Walsh, B. (2007). *The Origins of Genome Architecture*, volume 98. Sinauer Associates Sunderland, MA.

Mallo, D., De Oliveira Martins, L. et Posada, D. (2016). *SimPhy* : Phylogenomic Simulation of Gene, Locus, and Species Trees. *Systematic Biology*, 65(2):334–344.

McCutcheon, J. P. et Moran, N. A. (2012). Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology*, 10(1):13–26.

Mérot, C., Oomen, R. A., Tigano, A. et Wellenreuther, M. (2020). A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation. *Trends in Ecology & Evolution*, 35(7):561–572.

Mira, A., Ochman, H. et Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends in Genetics*, 17(10):589–596.

Mirarab, S., Bayzid, M. S. et Warnow, T. (2016). Evaluating Summary Methods for Multilocus Species Tree Estimation in the Presence of Incomplete Lineage Sorting. *Systematic Biology*, 65(3):366–380.

Misevic, D., Frénoy, A., Lindner, A. B. et Taddei, F. (2015). Shape matters : Lifecycle of cooperative patches promotes cooperation in bulky populations : SHAPE MATTERS. *Evolution*, 69(3):788–802.

Moran, N. A. et Plague, G. R. (2004). Genomic changes following host restriction in bacteria. *Current Opinion in Genetics & Development*, 14(6):627–633.

Mueller, R. L. et Jockusch, E. L. (2018). Jumping genomic gigantism. *Nature Ecology & Evolution*, 2(11):1687–1688.

Mukai, T. (1964). THE GENETIC STRUCTURE OF NATURAL POPULATIONS OF DROSOPHILA MELANOGASTER. I. SPONTANEOUS MUTATION RATE OF POLYGENES CONTROLLING VIABILITY. *Genetics*, 50(1):1–19.

Musumeci, O., Andreu, A. L., Shanske, S., Bresolin, N., Comi, G. P., Rothstein, R., Schon, E. A. et DiMauro, S. (2000). Intragenic inversion of mtDNA : A new type of pathogenic mutation in a patient with mitochondrial myopathy. *The American Journal of Human Genetics*, 66(6):1900–1904.

Nelson-Sathi, S., Sousa, F. L., Roettger, M., Lozada-Chávez, N., Thiergart, T., Janssen, A., Bryant, D., Landan, G., Schönheit, P., Siebers, B., McInerney, J. O. et Martin, W. F. (2015). Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature*, 517(7532):77–80.

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A. et Minh, B. Q. (2015). Iq-tree : a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, 32(1):268–274.

Noureen, M., Tada, I., Kawashima, T. et Arita, M. (2019). Rearrangement analysis of multiple bacterial genomes. *BMC Bioinformatics*, 20(S23):631.

Nuñez, P. A., Romero, H., Farber, M. D. et Rocha, E. P. (2013). Natural Selection for Operons Depends on Genome Size. *Genome Biology and Evolution*, 5(11):2242–2254.

Nyholm, S. V. et Graf, J. (2012). Knowing your friends : Invertebrate innate immunity fosters beneficial bacterial symbioses. *Nature Reviews Microbiology*, 10(12):815–827.

O'Neill, B. (2003). Digital Evolution. *PLoS Biology*, 1(1):e18.

Orgel, L. E. et Crick, F. H. (1980). Selfish DNA : The ultimate parasite. *Nature*, 284(5757):604.

Parsons, D. (2011). *Sélection indirecte en évolution Darwinienne : Mécanismes et implications*. Theses, INSA de Lyon.

Parsons, D. P., Knibbe, C. et Beslon, G. (2010). Importance of the rearrangement rates on the organization of transcription. *In Artificial Life 12 : Proceedings of the Fourteenth International Conference on the Synthesis and Simulation of Living Systems*, pages 479–486, Odense, Denmark.

Parsons, D. P., Knibbe, C. et Beslon, G. (2012). The paradoxical effects of allelic recombination on fitness. *In Proceedings of Artificial Life XIII*, pages 536–537.

Peng, B., Chen, H.-S., Mechanic, L. E., Racine, B., Clarke, J., Clarke, L., Gillanders, E. et Feuer, E. J. (2013). Genetic simulation resources : a website for the registration and discovery of genetic data simulators. *Bioinformatics*, 29(8):1101–1102.

Periwal, V. et Scaria, V. (2015). Insights into structural variations and genome rearrangements in prokaryotic genomes. *Bioinformatics*, 31(1):1–9.

Petrov, D. A. (2001). Evolution of genome size : New approaches to an old problem. *Trends in Genetics*, 17(1):23–28.

Petrov, D. A. (2002). Mutational Equilibrium Model of Genome Size Evolution. *Theoretical Population Biology*, 61(4):531–544.

Puigbò, P., Lobkovsky, A. E., Kristensen, D. M., Wolf, Y. I. et Koonin, E. V. (2014). Genomes in turmoil : Quantification of genome dynamics in prokaryote supergenomes. *BMC Biology*, 12(1):66.

Raeside, C., Gaffé, J., Deatherage, D. E., Tenaillon, O., Briska, A. M., Ptashkin, R. N., Cruveiller, S., Médigue, C., Lenski, R. E., Barrick, J. E. et Schneider, D. (2014). Large Chromosomal Rearrangements during a Long-Term Evolution Experiment with Escherichia coli. *mBio*, 5(5):e01377–14.

Randall, R. N., Radford, C. E., Roof, K. A., Natarajan, D. K. et Gaucher, E. A. (2016). An experimental phylogeny to benchmark ancestral sequence reconstruction. *Nature Communications*, 7(1):12847.

Rao, R. et Leibler, S. (2022). Evolutionary dynamics, evolutionary forces, and robustness : A nonequilibrium statistical mechanics perspective. *Proceedings of the National Academy of Sciences*, 119(13):e2112083119.

Robinson, M. D. et Vitek, O. (2019). Benchmarking comes of age. *Genome Biology*, 20(1):205, s13059–019–1846–5.

Rogozin, I. B. (2002). Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic Acids Research*, 30(19):4264–4271.

Romiguier, J., Ranwez, V., Douzery, E. et Galtier, N. (2013). Genomic Evidence for Large, Long-Lived Ancestors to Placental Mammals. *Molecular Biology and Evolution*, 30(1):5–13.

Rosenzweig, B. K., Pease, J. B., Besansky, N. J. et Hahn, M. W. (2016). Powerful methods for detecting introgressed regions from population genomic data. *Molecular Ecology*, 25(11):2387–2397.

Rutten, J. P., Hogeweg, P. et Beslon, G. (2019). Adapting the engine to the fuel : Mutator populations can reduce the mutational load by reorganizing their genome structure. *BMC Evolutionary Biology*, 19(1):191.

Sanjuán, R., Cuevas, J. M., Furió, V., Holmes, E. C. et Moya, A. (2007). Selection for Robustness in Mutagenized RNA Viruses. *PLoS Genetics*, 3(6):e93.

Scaduto, D. I., Brown, J. M., Haaland, W. C., Zwickl, D. J., Hillis, D. M. et Metzker, M. L. (2010). Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. *Proceedings of the National Academy of Sciences*, 107(50):21242–21247.

Schaack, S., Choi, E., Lynch, M. et Pritham, E. J. (2010). DNA transposons and the role of recombination in mutation accumulation in Daphnia pulex. *Genome biology*, 11(4):R46.

Schenk, M. F., Zwart, M. P., Hwang, S., Ruelens, P., Severing, E., Krug, J. et de Visser, J. A. G. M. (2022). Population size mediates the contribution of high-rate and large-benefit mutations to parallel evolution. *Nature Ecology & Evolution*, 6(4):439–447.

Seferbekova, Z., Zabelkin, A., Yakovleva, Y., Afasizhev, R., Dranenko, N. O., Alexeev, N., Gelfand, M. S. et Bochkareva, O. O. (2021). High Rates of Genome Rearrangements and Pathogenicity of Shigella spp. *Frontiers in Microbiology*, 12:628622.

Sela, I., Wolf, Y. I. et Koonin, E. V. (2016). Theory of prokaryotic genome evolution. *Proceedings of the National Academy of Sciences*, 113(41):11399–11407.

Senra, M. V. X., Sung, W., Ackerman, M., Miller, S. F., Lynch, M. et Soares, C. A. G. (2018). An Unbiased Genome-Wide View of the Mutation Rate and Spectrum of the Endosymbiotic Bacterium Teredinibacter turnerae. *Genome Biology and Evolution*, 10(3):723–730.

Sessegolo, C., Burlet, N. et Haudry, A. (2016). Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. *Biology letters*, 12(8):20160407.

Sjöstrand, J., Arvestad, L., Lagergren, J. et Sennblad, B. (2013). GenPhyloData : Realistic simulation of gene family evolution. *BMC Bioinformatics*, 14(1):209.

Slatkin, M. et Racimo, F. (2016). Ancient DNA and human history. *Proceedings of the National Academy of Sciences*, 113(23):6380–6387.

Song, S., Liu, L., Edwards, S. V. et Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences*, 109(37):14942–14947.

Spencer, H. (1864). *The Principles of Biology Vol. 1.* Hansebooks GmbH, Norderstedt, nachdruck der ausgabe von 1864 édition.

Starlinger, P. (1977). DNA REARRANGEMENTS IN PROCARYOTES. *Annual Review of Genetics*, 11(1):103–126.

Strope, C. L., Abel, K., Scott, S. D. et Moriyama, E. N. (2009). Biological Sequence Simulation for Testing Complex Evolutionary Hypotheses : Indel-Seq-Gen Version 2.0. *Molecular Biology and Evolution*, 26(11):2581–2593.

Swan, B. K., Tupper, B., Sczyrba, A., Lauro, F. M., Martinez-Garcia, M., González, J. M., Luo, H., Wright, J. J., Landry, Z. C., Hanson, N. W., Thompson, B. P., Poulton, N. J., Schwientek, P., Acinas, S. G., Giovannoni, S. J., Moran, M. A., Hallam, S. J., Cavicchioli, R., Woyke, T. et Stepanauskas, R. (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proceedings of the National Academy of Sciences*, 110(28):11463–11468.

Symons, J. (2008). Computational models of emergent properties. *Minds and Machines*, 18:475–491.

Syvanen, M. (2012). Evolutionary Implications of Horizontal Gene Transfer. *Annual Review of Genetics*, 46(1):341–358.

Szöllősi, G. J., Boussau, B., Abby, S. S., Tannier, E. et Daubin, V. (2012). Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proceedings of the National Academy of Sciences*, 109(43):17513–17518.

Tenaillon, O., Barrick, J. E., Ribeck, N., Deatherage, D. E., Blanchard, J. L., Dasgupta, A., Wu, G. C., Wielgoss, S., Cruveiller, S., Médigue, C. *et al.* (2016). Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature*, 536(7615):165–170.

Thrall, P. H., Oakeshott, J. G., Fitt, G., Southerton, S., Burdon, J. J., Sheppard, A., Russell, R. J., Zalucki, M., Heino, M. et Ford Denison, R. (2011). Evolution in agriculture : The application of evolutionary approaches to the management of biotic interactions in agro-ecosystems : Evolution in agriculture. *Evolutionary Applications*, 4(2):200–215.

Torres-Barceló, C., Cabot, G., Oliver, A., Buckling, A. et MacLean, R. C. (2013). A trade-off between oxidative stress resistance and DNA repair plays a role in the evolution of elevated mutation rates in bacteria. *Proceedings of the Royal Society B : Biological Sciences*, 280(1757):20130007.

Tristan, B., Aurelie, P. et Annabelle, H. (2019). Evidence for purifying selection on conserved noncoding elements in the genome of *Drosophila melanogaster*. Preprint, Evolutionary Biology.

Vadée-Le-Brun, Y., Rouzaud-Cornabas, J. et Beslon, G. (2015). Epigenetic inheritance speeds up evolution of artificial organisms. *In European Conference on Artificial Life*.

Vadée-Le-Brun, Y., Rouzaud-Cornabas, J. et Beslon, G. (2016). In silico experimental evolution suggests a complex intertwining of selection, robustness and drift in the evolution of genetic networks complexity. *In Artificial Life Conference Proceedings*, pages 174–174. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info ....

Vinogradov, A. E. (1997). Nucleotypic effect in homeotherms : Body-mass independent resting metabolic rate of passerine birds is related to genome size. *Evolution ; international journal of organic evolution*, 51(1):220–225.

Wagner, A. (2005). Robustness, evolvability, and neutrality. *FEBS Letters*, 579(8):1772–1778.

Wagner, A. (2007). *Robustness and Evolvability in Living Systems*. Princeton Studies in Complexity. Princeton University Press, Princeton, N.J. Oxford, 3. pr. and 1. paperback pr édition.

Waples, R. S. (2010). Spatial-temporal stratifications in natural populations and how they affect understanding and estimation of effective population size. *Molecular Ecology Resources*, 10(5):785–796.

Weigand, M. R., Peng, Y., Batra, D., Burroughs, M., Davis, J. K., Knipe, K., Loparev, V. N., Johnson, T., Juieng, P., Rowe, L. A., Sheth, M., Tang, K., Unoarumhi, Y., Williams, M. M. et Tondella, M. L. (2019). Conserved Patterns of Symmetric Inversion in the Genome Evolution of *Bordetella* Respiratory Pathogens. *mSystems*, 4(6):e00702–19.

Wells, J. N. et Feschotte, C. (2020). A Field Guide to Eukaryotic Transposable Elements. *Annual Review of Genetics*, 54(1):539–561.

Whitacre, J. M. (2012). Biological Robustness : Paradigms, Mechanisms, and Systems Principles. *Frontiers in Genetics*, 3.

Wilke, C. O. (2001). Selection for fitness versus selection for robustness in RNA secondary structure folding. *Evolution ; international journal of organic evolution*, 55(12):2412–2420.

Wilke, C. O. et Adami, C. (2003). Evolution of mutational robustness. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 522(1-2):3–11.

Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E. et Adami, C. (2001). Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–333.

Wolf, Y. I. et Koonin, E. V. (2013). Genome reduction as the dominant mode of evolution. *BioEssays*, 35(9):829–837.

Wright, S. (1931). EVOLUTION IN MENDELIAN POPULATIONS. *Genetics*, 16(2): 97–159.

Zhang, J. (2003). Evolution by gene duplication : An update. *Trends in Ecology & Evolution*, 18(6):292–298.

Zhang, Y., Tan, Z. et Krishnamachari, B. (2014). On the Meeting Time for Two Random Walks on a Regular Graph.

FOLIO ADMINISTRATIF

THESE DE L'INSA LYON, MEMBRE DE L'UNIVERSITE DE LYON

NOM : Foley

DATE de SOUTENANCE : 19/12/2023

Prénoms : Marco, Steven

TITRE : Dynamique des génomes bactériens : une étude expérimentale in silico avec la plateforme Aevol.

NATURE : Doctorat

Numéro d'ordre : 2023ISAL0130

Ecole doctorale : Infomaths (ED 512)

Spécialité : Informatique et application

RESUME :
Aevol est une plate-forme de simulation de l'évolution de populations d'organismes par variation et sélection. La conception du modèle est axée sur le réalisme de la structure du génome et des processus de mutations, permettant ainsi aux organismes simulés d'évoluer sur un fitness landscape comparable à celui d'organismes biologiques, avec des contraintes d'exploration similaires. Ces processus permettent l'émergence de comportements d'intérêt, pour l'étude de l'évolution de la structure des génomes, et pour produire des données de benchmarks pour tester les méthodes de phylogénie moléculaire. Les résultats obtenus jusqu'ici dans aevol concourent à suggérer que les éléments non-codants du génome sont soumis à sélection. Dans ce travail, nous avons utilisé Aevol pour mener une large campagne de simulation sur de très longues échelles de temps. Ces expériences nous permettent de montrer que la quantité de séquences non-codantes est finement régulée par deux forces contraires. La première est une force de sélection pour des génomes réduits car plus robustes face aux réarrangements chromosomiques. La seconde provient d'un biais mutationnel indirect favorisant les évènements de duplications neutres sur les délétions neutres menant à l'accumulation de non-codant par dérive génétique. Dans un deuxième temps, nous avons utilisé aevol comme outil de génération de benchmarks pour la phylogénie. En effet, Aevol ayant été développé indépendamment de la communauté de phylogénie moléculaire, il ne contient pas les a priori classiquement inclus dans les simulateurs de cette communauté, évitant ainsi la validation ad hoc des méthodes. Cependant, les séquences composant les génomes étant binaires dans Aevol, nous avons développé une version du simulateur utilisant des séquences génomiques quaternaire (ACTG). Cette nouvelle version a ensuite été utilisée pour générer des données de benchmarks afin de tester les reconstructions d'arbres phylogénétiques.

MOTS-CLÉS : évolution, taille des génomes, évolution expérimental in silico, benchmarks, robustesse, réarrangements chromosomiques

Laboratoire (s) de recherche :
Laboratoire d'informatique en image et systèmes d'information (LIRIS)

Directeurs de thèse:
Guillaume Beslon, Professeur des Universités, INSA de Lyon
Jonathan Rouzaud-Cornabas (co-encadrant), Maître de conférences, INSA de Lyon

Président de jury :

Composition du jury :

Lafontaine, Ingrid	Professeur des Universités	Sorbonne Université
Bredeche, Nicolas	Professeur des Universités	Sorbonne Université
Junier, Ivan	Directeur de recherche	Université Grenoble Alpes
Guillaume Beslon	Professeur des Universités	INSA de Lyon
Jonathan Rouzaud-Cornabas	Maître de conférences	INSA de Lyon