Structuration des génomes par sélection indirecte de la variabilité mutationnelle : une approche de modélisation et de simulation

Carole Knibbe

Numéro d'ordre 2006-ISAL-0093
Année 2006

Structuration des génomes par sélection indirecte de la variabilité mutationnelle, une approche de modélisation et de simulation

Thèse présentée par
Carole Knibbe, ingénieur INSA en Bioinformatique et Modélisation

Devant
L'Institut National des Sciences Appliquées de Lyon

Pour obtenir
Le grade de docteur

Formation doctorale
Evolution Ecosystèmes Microbiologie Modélisation (E2M2)

Spécialité
Approches mathématiques et informatiques du vivant

Soutenue le 4 décembre 2006 devant le jury composé de :

Guillaume Beslon	Maître de conférences, INSA Lyon, directeur de thèse
Laurent Duret	Directeur de recherche, CNRS, examinateur
Jean-Michel Fayard	Professeur, INSA Lyon, directeur de thèse
Michel Morvan	Professeur, ENS Lyon, rapporteur
Eduardo Rocha	Chargé de recherche HDR, CNRS, rapporteur
Marc Schoenauer	Directeur de recherche, INRIA, président du jury
François Taddéi	Chargé de recherche HDR, INSERM, membre invité

2005

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
	CHIMIE DE LYON Responsable : M. Denis SINOU	M. Denis SINOU Université Claude Bernard Lyon 1 Lab Synthèse Asymétrique UMR UCB/CNRS 5622 Bât 308 2ème étage 43 bd du 11 novembre 1918 69622 VILLEURBANNE Cedex Tél : 04.72.44.81.83 Fax : 04 78 89 89 14 sinou@univ-lyon1.fr
E2MC	ECONOMIE, ESPACE ET MODELISATION DES COMPORTEMENTS Responsable : M. Alain BONNAFOUS	M. Alain BONNAFOUS Université Lyon 2 14 avenue Berthelot MRASH M. Alain BONNAFOUS Laboratoire d'Economie des Transports 69363 LYON Cedex 07 Tél : 04.78.69.72.76 Alain.bonnafous@ish-lyon.cnrs.fr
E.E.A.	ELECTRONIQUE, ELECTROTECHNIQUE, AUTOMATIQUE M. Daniel BARBIER	M. Daniel BARBIER INSA DE LYON Laboratoire Physique de la Matière Bâtiment Blaise Pascal 69621 VILLEURBANNE Cedex Tél : 04.72.43.64.43 Fax 04 72 43 60 82 Daniel.Barbier@insa-lyon.fr
E2M2	EVOLUTION, ECOSYSTEME, MICROBIOLOGIE, MODELISATION http://biomserv.univ-lyon1.fr/E2M2 M. Jean-Pierre FLANDROIS	M. Jean-Pierre FLANDROIS UMR 5558 Biométrie et Biologie Evolutive Equipe Dynamique des Populations Bactériennes Faculté de Médecine Lyon-Sud Laboratoire de Bactériologie BP 1269600 OULLINS Tél : 04.78.86.31.50 Fax 04 72 43 13 88 e2m2@biomserv.univ-lyon1.fr
EDIIS	INFORMATIQUE ET INFORMATION POUR LA SOCIETE http://www.insa-lyon.fr/ediis M. Lionel BRUNIE	M. Lionel BRUNIE INSA DE LYON EDIIS Bâtiment Blaise Pascal 69621 VILLEURBANNE Cedex Tél : 04.72.43.60.55 Fax 04 72 43 60 71 ediis@insa-lyon.fr
EDISS	INTERDISCIPLINAIRE SCIENCES-SANTE http://www.ibcp.fr/ediss M. Alain Jean COZZONE	M. Alain Jean COZZONE IBCP (UCBL1) 7 passage du Vercors 69367 LYON Cedex 07 Tél : 04.72.72.26.75 Fax : 04 72 72 26 01 cozzone@ibcp.fr
	MATERIAUX DE LYON http://www.ec-lyon.fr/sites/edml M. Jacques JOSEPH	M. Jacques JOSEPH Ecole Centrale de Lyon Bât F7 Lab. Sciences et Techniques des Matériaux et des Surfaces 36 Avenue Guy de Collongue BP 163 69131 ECULLY Cedex Tél : 04.72.18.62.51 Fax 04 72 18 60 90 Jacques.Joseph@ec-lyon.fr
Math IF	**MATHEMATIQUES ET INFORMATIQUE FONDAMENTALE** http://www.ens-lyon.fr/MathIS M. Franck WAGNER	M. Franck WAGNER Université Claude Bernard Lyon1 Institut Girard Desargues UMR 5028 MATHEMATIQUES Bâtiment Doyen Jean Braconnier Bureau 101 Bis, 1er étage 69622 VILLEURBANNE Cedex Tél : 04.72.43.27.86 Fax : 04 72 43 16 87 wagner@desargues.univ-lyon1.fr
MEGA	**MECANIQUE, ENERGETIQUE, GENIE CIVIL, ACOUSTIQUE** http://www.lmfa.ec-lyon.fr/autres/MEGA/index.html M. François SIDOROFF	M. François SIDOROFF Ecole Centrale de Lyon Lab. Tribologie et Dynamique des Systêmes Bât G8 36 avenue Guy de Collongue BP 163 69131 ECULLY Cedex Tél :04.72.18.62.14 Fax : 04 72 18 65 37 Francois.Sidoroff@ec-lyon.fr

Remerciements

Je tiens à saluer ici les personnes qui, de près ou de loin, ont contribué à la concrétisation de ce travail de thèse. J'adresse ainsi mes plus sincères remerciements...

... à Michel Morvan et Eduardo Rocha, mes rapporteurs, qui ont accepté de lire en détail ce manuscrit et d'en donner une critique constructive,

... à Marc Schoenauer, Laurent Duret et François Taddéi, qui ont accepté de participer à mon jury,

... à Joël Favrel et Gérard Febvay, qui m'ont accueillie dans leurs laboratoires et qui m'ont donné les moyens de travailler dans de bonnes conditions,

... à la direction de la recherche de l'INSA et à la région Rhône-Alpes, qui soutiennent le groupe de recherche pluri-disciplinaire BSMC (Biologie des Systèmes et Modélisation Cellulaire) dans lequel ce travail s'inscrit,

... à Christian Gautier, Christian Biémont, Hans Geiselmann, Alain Mille, Jean Lobry et Hubert Charles, qui, en parrainant mon DEA ou en "pilotant" ma thèse, ont été parmi les premiers à s'intéresser à l'approche développée,

... à Sylvain Mousset, Vincent Daubin, Chris Adami et à tous les membres du groupe BSMC, pour leurs commentaires constructifs sur les résultats,

... à Olivier Mazet, pour son aide dans les calculs de probabilité,

... à Fabien Chaudier, qui s'est démené pour que les simulations puissent tourner sur le cluster de calcul,

... à Gaël Kaneko, Sylvain Koos, Vincent Bloch, Alexandra Dumitru, Mirabela Rusu, Jing-Jing Zhou, Victor Cianni, Christelle Gobet, Joao Martins et Jérémie Becker, ces étudiants qui se sont attaqués au code pour tester de nouvelles hypothèses,

... à Sandrine Charles, qui restera ma tutrice de cœur pour mon monitorat,

... à Hédi, Virginie, Yolanda et Lorraine, mes compagnons de fortune et d'infortune du deuxième étage, avec qui j'ai pu rire de nos galères de thésards et profiter des réconforts distribués par la machine à café,

... à Nadira, mon fil d'Ariane dans les labyrinthes administratifs, mais tellement plus aussi,

... à Jean-Michel, qui a toujours veillé à ce que je garde le cap,

... à Aurélie, confidente et coach sportif, à qui je dois ma santé mentale et un formidable coup de pouce,

... à mes parents et grands-parents, qui, par leur soutien moral et matériel tout au long de mes études, m'ont permis d'arriver jusque là,

... à ma sœur et ex-colocataire Marion, qui heureusement n'a pas déménagé trop loin au début de ma thèse – je n'oublierai pas les soirées passées à regarder Lilo et Stitch !

... à ma sœur et sage conseillère Julie, qui m'a écoutée de longues heures au téléphone,

... à Guillaume, qui m'a appris à marcher quand le chemin est escarpé,

... à Antoine enfin, mon compagnon, qui m'a remise debout dans les moments difficiles.

Résumé

À long terme, le succès évolutif d'une lignée ne dépend pas seulement de la valeur adaptative de ses fondateurs. Il dépend également de la capacité des descendants à transmettre le génotype ancestral sans mutation délétère, tout en découvrant parfois des mutations favorables. Une partie de la progéniture d'un organisme doit donc être sans mutation (ou uniquement avec des mutations neutres), tandis que l'autre partie, portant des mutations non neutres, explore de nouveaux phénotypes – souvent moins bien adaptés, parfois avantageux. Nous appelons *variabilité mutationnelle du phénotype* la façon dont les essais reproductifs sont ainsi répartis entre reproduction à l'identique et exploration de nouveaux phénotypes. À nombre d'essais reproductifs égal, cette variabilité mutationnelle détermine l'issue de la compétition évolutive : les lignées trop variables s'éteignent sous l'effet des mutations délétères, et les lignées trop peu variables sont tôt ou tard détrônées par celles qui ont pu découvrir des innovations avantageuses. Un niveau intermédiaire de variabilité mutationnelle est donc, de fait, indirectement sélectionné.

Le taux de mutation et la façon dont les mutations dans les gènes affectent le phénotype sont les deux composantes les plus étudiées de la variabilité mutationnelle du phénotype. Plusieurs travaux de modélisation ont ainsi montré que la sélection indirecte d'un niveau donné de variabilité peut déterminer l'évolution du taux de mutation et celles des mécanismes de "canalisation", limitant l'effet des mutations géniques. Dans ce travail, nous montrons par une approche de modélisation et de simulation que la structure du génome est une troisième composante tout aussi importante de la variabilité mutationnelle du phénotype. En effet, la quantité d'ADN non codant, le nombre d'éléments répétés et leur répartition sont autant de caractéristiques structurelles du génome susceptibles d'influencer le nombre moyen de gènes touchés par un réarrangement chromosomique.

En simulant, à l'aide d'un modèle individu-centré, l'évolution de génomes soumis à la fois à des mutations locales et à des réarrangements chromosomiques, nous montrons que la structure du génome peut faire l'objet de pressions de sélection indirecte. Le nombre de gènes et, de façon plus surprenante, la quantité de non codant s'ajustent ainsi spontanément en fonction du taux de mutation et de l'effet moyen des mutations géniques, sous l'effet de la sélection indirecte d'un niveau intermédiaire de variabilité mutationnelle. L'émergence de ces couplages surprenants suggère que les génomes ne sont pas seulement façonnés par les biais mutationnels et les coûts sélectifs directs, mais aussi, à plus long terme, par des pressions plus indirectes.

Liste des publications personnelles

Articles parus

- [article 1] **C. Knibbe**, O. Mazet, F. Chaudier, J.-M. Fayard et G. Beslon (2006). Evolutionary coupling between the deleteriousness of gene mutations and the amount of non-coding sequences. *J. Theor. Biol.*, sous presse. Lien DOI : `http://dx.doi.org/10.1016/j.jtbi.2006.09.005`
- [article 2] V. Lefort, **C. Knibbe**, G. Beslon et J. Favrel (2006). Simultaneous optimization of weights and structure of a RBF Neural Network. *Lect. Notes Comput. Sci.* 3871 :49-60.
- [article 3] **C. Knibbe**, G. Beslon, V. Lefort, F. Chaudier et J.-M. Fayard (2005). Self-adaptation of genome size in artificial organisms. *Lect. Notes Artif. Intell.* 3630 :423-432.
- [article 4] **C. Knibbe**, G. Beslon, F. Chaudier et J.-M. Fayard (2005). Designing artificial organisms to study the influence of gene pleiotropy on genome evolution. *Proceedings of the Systems Biology Workshop at the 8th European Conference on Artificial Life, ECAL 2005*, Canterbury, Royaume-Uni.
- [article 5] V. Lefort, **C. Knibbe**, G. Beslon et J. Favrel (2004). Introducing "proteins" into genetic algorithms, *Proceedings of CSIMTA'04, Int. Conf.*, Cherbourg, France, pp. 181-186.
- [article 6] V. Lefort, **C. Knibbe**, G. Beslon, J. Favrel (2004). A bio-inspired genetic algorithm with a self-organizing genome: The RBF-Gene model. *Lect. Notes Comput. Sci.* 3103 :406-407.
- [article 7] G. Beslon, **C. Knibbe**, H. Soula et J.-M. Fayard (2003). The RBF-Gene Model. In Pearson D.W. et al (Eds), *Artificial Neural Nets and Genetic Algorithms, Proceedings of the International Conference*, Roanne, France, Springer, pp. 187-192.

Articles soumis

- [article 8] **C. Knibbe**, A. Coulon, O. Mazet, J.-M. Fayard et G. Beslon. A long-term evolutionary pressure on the amount of non-coding DNA. En cours de révision pour la revue *Molecular Biology and Evolution.*
- [article 9] **C. Knibbe**, J.-M. Fayard et G. Beslon. The topology of the protein network influences the dynamics of gene order: From Systems Biology to a systemic understanding of evolution. En cours de révision pour la revue *Artificial Life.*

Résumés de conférences

- **C. Knibbe**, A. Coulon, O. Mazet, J.-M. Fayard et G. Beslon (2006). Amount of non-coding sequences depends on mutation rate. 10th Evolutionary Biology Meeting, Marseille, France. Poster et communication orale.
- **C. Knibbe**, G. Beslon et J.-M. Fayard (2006). Long-term needs for evolvability and robustness shape genome structure: A simulation study. JOBIM 2006, Bordeaux, France. Poster et communication orale.
- **C. Knibbe**, G. Beslon et J.-M. Fayard (2005). Evolutionary influence of the protein network topology on gene organisation in artificial organisms. European Conference on Complex Systems (ECCS 2005), Paris, France. Poster.
- G. Beslon et **C. Knibbe** (2005). Modelling evolution of prokaryotic genomes: an integrative approach. FEBS Advanced Lecture Course, Systems Biology: From Molecules and Modeling To Cells, Gosau, Autriche. Communication orale.
- **C. Knibbe**, V. Lefort et G. Beslon et J.-M. Fayard (2004). Evolutionary dynamics of overlapping genes in virtual organisms. Intl. Conf. on Integrative Post-Genomics, Lyon, France. Communication orale.
- **C. Knibbe**, G. Beslon et J.-M. Fayard (2003). Modeling evolutionary interactions between genomic and functional structures. 4th International Conference on Systems Biology (ICSB 2003), Saint-Louis, Missouri, USA. Poster.

Table des matières

Table des figures

Introduction

Je pense que la fascination exercée sur tant de gens par la théorie de l'évolution réside dans trois de ses caractéristiques. D'abord elle est, en l'état actuel de son développement, assez élaborée pour procurer un sentiment de satisfaction et de confiance, mais en même temps suffisamment peu avancée pour proposer moult mystères. En second lieu, elle est située au centre d'un continuum qui s'étend des sciences traitant de généralités intemporelles et quantitatives à celles qui touchent directement aux singularités de l'histoire. Elle offre donc asile aux chercheurs de tous styles et de toutes tendances, depuis ceux qui cherchent la pureté de l'abstraction (les lois de la croissance démographique et la structure de l'ADN) jusqu'à ceux qui se délectent dans le fatras des particularités irréductibles (que pouvait donc bien faire le tyrannosaure de ses deux pattes de devant si chétives, si jamais il en faisait quelque chose ?). Troisièmement, elle nous concerne tous dans notre vie ; car comment pouvons-nous être indifférents devant les grandes questions de la généalogie : d'où venons-nous et qu'est-ce que cela signifie ? Et puis, bien entendu, il y a tous ces organismes : plus d'un million d'espèces décrites, de la bactérie à la baleine bleue, avec une foultitude de bestioles entre les deux – chacune avec sa beauté propre, et chacune avec une histoire à raconter.

Stephen Jay Gould[1]

Presque 150 ans se sont écoulés depuis la parution de *L'Origine des Espèces*, l'ouvrage dans lequel Charles Darwin exposa les principes fondamentaux de la théorie de l'évolution. Les deux principes mis en avant par Darwin sont la variabilité inter-individuelle et la sélection naturelle : au sein d'une population, les individus présentant une caractéristique avantageuse dans l'environnement ont davantage de chances de survivre et de se reproduire ; si cette caractéristique est héréditaire, elle est transmise de génération en génération et devient de plus en plus commune dans la population (Darwin, 1859). Plusieurs

[1] In *Le Pouce du Panda*, Paris : Grasset, 1980 (1982 pour la traduction française), p. 10.

questions de taille restaient cependant en suspens à l'époque. En particulier, le support matériel de l'hérédité restait mystérieux, ainsi que l'origine des variations entre individus. La théorie de l'évolution est donc restée un champ de recherche très dynamique.

La question de la base matérielle de l'hérédité fut ainsi progressivement élucidée par la redécouverte des travaux de Mendel (Correns, 1900; De Vries, 1900; Von Tschermak, 1900), l'élaboration de la théorie chromosomique de l'hérédité (Morgan *et al.*, 1915), et l'identification de l'ADN (acide désoxyribonucléique) comme support biochimique de l'information génétique (Avery *et al.*, 1944). Pour confirmer le rôle de l'ADN dans l'hérédité, il fallait cependant montrer comment l'ADN pouvait être copié fidèlement de génération en génération. Un mécanisme possible, dit semi-conservatif, fut entrevu lors de la découverte par Watson et Crick (1953) de la structure en double hélice de l'ADN et du système d'appariement des bases complémentaires. Ce mécanisme fut confirmé expérimentalement cinq ans plus tard, par Meselson et Stahl (1958). Ces avancées permirent aussi de comprendre l'origine des variations héréditaires, condition *sine qua non* de l'évolution : des mutations, c'est-à-dire des modifications de la séquence des bases de l'ADN, peuvent occasionnellement se produire lors de la réplication ou sous l'effet d'agents mutagènes (acide nitreux, peroxydes, rayons ultra-violets...). Ces mutations étaient cependant supposées très rares dans les conditions naturelles, car l'ADN était vu comme intrinsèquement stable (Fox Keller, 2000).

Les progrès de la biologie moléculaire ont profondément bousculé cette vision, en montrant que la fidélité de la réplication de l'ADN n'est pas seulement liée à sa structure, mais qu'elle dépend aussi de l'action coordonnée d'enzymes capables de détecter et de réparer les erreurs. Selon le biologiste moléculaire Franklin Stahl, ces découvertes ont remis en cause "une croyance (tacite, mais confinant à la dévotion) répandue parmi les généticiens, selon laquelle les gènes sont si précieux qu'ils doivent d'une manière ou d'une autre être protégés des attaques biochimiques, peut-être en étant soigneusement enveloppés. La possibilité que les gènes fussent dynamiquement stables, soumis à un ballottement perpétuel entre les dommages causés et les tentatives maladroites (c'est-à-dire enzymatiques) destinées à réparer ces dommages, était inconcevable" (Fox Keller, 2000). L'idée nouvelle d'un ADN dynamique fut renforcée par la découverte des éléments transposables et des mécanismes de recombinaison ectopique[1]. L'existence de ces mécanismes implique en effet que ADN n'est pas seulement sujet à des modifications ponctuelles ; des segments entiers peuvent être excisés, insérés ou déplacés. Ainsi, le fait que l'information génétique soit stable ou au contraire variable d'une génération à l'autre dépend, au moins en partie, de l'action combinée de divers systèmes enzymatiques. Le degré de variabilité, matière première de l'évolution, est donc partiellement sous le contrôle de mécanismes qui sont eux-mêmes des produits de l'évolution.

Si le degré de variation peut ainsi être façonné par la sélection naturelle, c'est la faculté même d'évolution (*evolvability*[2]) qui peut évoluer. L'évolution est à même d'influencer

[1] La recombinaison ectopique désigne la recombinaison entre séquences répétées dispersées dans le génome.

[2] Ce terme évocateur, couramment utilisé aujourd'hui, souffre pourtant d'une absence de définition claire. Selon les auteurs, cette faculté d'évolution désigne tantôt la capacité à varier (dans quelque direction

son propre cours et elle apparaît donc comme un phénomène bien plus complexe que ce que l'on aurait pu soupçonner au début des années 50. Variation et sélection ne sont plus des processus indépendants, mais au contraire imbriqués dans une interaction subtile dont l'issue est difficile à prévoir.

Pour appréhender l'évolution dans sa complexité nouvelle, la modélisation s'est rapidement révélée indispensable. Les modèles peuvent en effet être conçus comme des représentations opératoires, sur lesquelles il est aisé d'agir et d'observer des effets à long terme : "Je n'ai jamais cru aux explications", disait Paul Valéry, "mais j'ai cru qu'il fallait chercher des représentations sur lesquelles on peut opérer, comme on travaille sur une carte, ou l'ingénieur sur une épure... et qui puissent servir à faire"[1]. Les modèles permettent la compréhension de phénomènes complexes parce qu'ils sont des dispositifs de médiation de l'action, comme, de façon plus incarnée, le développement cognitif de l'enfant passe par l'expérience de la manipulation d'objets[2]. Ils se sont avérés essentiels face au bouleversement conceptuel porté par l'idée d'évolution de la variabilité. Il n'est donc pas surprenant que le développement de cette idée soit intimement lié à la conception de nouveaux modèles.

Les premiers modèles visant à étudier l'évolution du taux de mutation sont ainsi apparus dans les années 70. Il s'agit de modèles mathématiques de génétique des populations qui mettent en jeu un gène dit "modifieur", contrôlant le taux de mutation des autres gènes (Leigh, 1970, 1973; Karlin et McGregor, 1974; Gillespie, 1981; Holsinger et Feldman, 1983). Ces travaux pionniers ont permis de formaliser l'action indirecte de la sélection sur un gène qui ne contribue pas directement à la valeur sélective (fitness) de l'organisme, mais qui détermine la variabilité de l'information génétique d'une génération à la suivante. Ainsi, dans une population asexuée, un allèle très mutagène disparaît de fait de la population s'il n'a généré que des mutations délétères au niveau des autres loci. Mais s'il génère une mutation favorable, il se propage en même temps qu'elle dans la population, un phénomène connu sous le nom d'"auto-stop" (hitch-hiking). Cela a pu être confirmé avec le développement de modèles informatiques dits "individu-centrés" (voir par exemple Taddei *et al.*, 1997), permettant de simuler explicitement la survie et la reproduction des individus d'une population pendant de nombreuses générations. Ce type de modélisation a aussi permis d'étudier des situations plus complexes que celles des premiers modèles mathématiques, en autorisant par exemple un grand nombre de valeurs pour le taux de mutation, plutôt que simplement deux valeurs extrêmes. Dans le domaine de l'évolution artificielle, certains travaux ont alors montré que le taux de mutation indirectement sélectionné est intermédiaire (Bedau et Packard, 2003) : la lignée qui se maintient sur le long terme est celle qui a pu générer suffisamment de mutations favorables sans pour autant disparaître sous l'effet des mutations délétères. Les différentes approches de modélisation ont ainsi permis de mieux comprendre les principes généraux susceptibles de gouverner

---

que ce soit), tantôt la capacité à générer des mutations favorables. Dans la suite de ce travail, nous préférerons donc éviter ce terme ambigu et parler de variabilité pour désigner la capacité à varier.

[1] In *Carnets*, Gallimard, Bibliothèque de la Pléiade, 1973.

[2] Voir Piaget, J., *La représentation du monde chez l'enfant*, Paris : Librairie Félix Alcan, 1926, ainsi que Thelen, E. et Smith, L. B., *A Dynamic Systems Approach to the Development of Cognition and Action*, Cambridge : MIT Press, 1996.

l'évolution des mécanismes de réplication et de réparation de l'ADN.

Parallèlement, la biologie moléculaire dévoilait la complexité de la relation entre l'ADN d'un organisme (son génotype, sujet aux mutations) et ses caractéristiques physiologiques et morphologiques (son phénotype, qui donne prise à la sélection naturelle). Une mutation dans un gène peut ainsi être sans conséquence phénotypique apparente ou bien au contraire avoir des effets dramatiques sur la survie, et la connaissance du génome complet de l'organisme ne suffit pas, en général, pour le prédire. Les effets d'une mutation au sein d'un gène se jouent à de multiples niveaux, depuis la dégénérescence du code génétique jusqu'aux processus de développement, en passant par les réseaux d'interactions entre protéines. La variabilité visible pour la sélection naturelle, c'est-à-dire la variabilité des traits phénotypiques, ne dépend donc pas seulement du taux de mutation au niveau de l'ADN, elle dépend aussi de la façon dont les mutations se répercutent sur le phénotype. Or il existe chez les êtres vivants des mécanismes qui peuvent moduler l'effet phénotypique des mutations. L'exemple le plus frappant est peut-être celui des protéines dites "chaperonnes", qui permettent à d'autres protéines de conserver un repliement correct malgré des changements dans leurs séquences d'acides aminés (Rutherford et Lindquist, 1998). L'effet des mutations dépend ainsi, au même titre que le taux de mutation, de mécanismes qui sont eux-mêmes le produit de l'évolution, comme le suggéraient Gerhart et Kirschner en 1997 : "Tout au long de [l'histoire de la génétique], l'organisme est resté une boîte noire, traduisant des changements aléatoires survenus dans ses gènes en des variations phénotypiques sur lesquelles devait agir la sélection. La biologie moderne est en train d'ouvrir rapidement cette boîte noire. On y découvre que les liens entre génotype et phénotype ont été façonnés par l'évolution pour collaborer avec l'évolution". L'idée était encore formulée de façon floue et quelque peu finaliste, mais là encore, diverses approches de modélisation mathématique et de simulation informatique ont permis de clarifier le phénomène. Elles ont montré comment des mécanismes de "canalisation" (neutralisation) des mutations peuvent être indirectement sélectionnés dans un environnement stable (Schuster et Swetina, 1988; Wagner *et al.*, 1997; Van Nimwegen *et al.*, 1999; Wilke *et al.*, 2001), et, inversement, comment des mécanismes amplifiant l'effet des mutations peuvent être indirectement sélectionnés dans un environnement variable (Huynen et Hogeweg, 1994; Ancel Meyers *et al.*, 2005).

Cependant, même dans ces développements récents, les mutations restent implicitement envisagées comme locales : chaque mutation n'affecte que quelques bases de l'ADN, ou, tout au plus, un gène entier – mais un seul. Or les programmes de séquençage et les études de génomique comparative ont souligné le rôle majeur des grands réarrangements chromosomiques (duplications, délétions, inversions...) dans l'évolution de l'information génétique (Hughes, 2000; Eichler et Sankoff, 2003; Sankoff, 2003; Coghlan *et al.*, 2005). Ces réarrangements peuvent affecter plusieurs gènes à la fois, voire le génome entier pour certaines duplications. Le nombre moyen de gènes touchés par mutation (au sens large, c'est-à-dire réarrangements compris) apparaît donc comme une troisième composante de la variabilité globale du phénotype, en plus du taux de mutation et de l'effet phénotypique de chaque modification de gène. La portée des mécanismes de recombinaison ectopique, mais aussi le nombre total de gènes, la quantité d'ADN répétitif, la répartition relative des séquences répétées et des gènes – en un mot, la structure du génome – influencent le

nombre de gènes touchés par un réarrangement aléatoire. La structure même du génome apparaît donc comme un levier supplémentaire pour l'évolution de la variabilité.

La question que nous abordons dans ce travail de thèse est donc la suivante : la structure du génome peut-elle faire l'objet de pressions de sélection indirecte, analogues à celles mises en évidence sur le taux de mutation et les mécanismes de canalisation ? Au vu de ses apports précédents, la modélisation semble particulièrement appropriée pour apporter des éléments de réponse à cette question. Nous verrons cependant que les modèles existants ne permettent pas de l'aborder, souvent parce qu'ils ne prennent pas explicitement en compte la structure génomique. Ce travail de thèse a donc d'abord eu pour objet la conception d'un nouveau modèle, permettant d'étudier les pressions indirectes susceptibles de s'exercer sur la structure d'un génome en raison de son rôle dans la variabilité mutationnelle globale. Les expériences *in silico* menées ensuite avec ce modèle ont montré que la sélection indirecte d'un certain niveau de variabilité mutationnelle peut effectivement se manifester au niveau de la structure du génome. Dans le modèle, ce phénomène affecte en particulier l'évolution de la quantité d'ADN non codant, du nombre de gènes et de l'ordre des gènes.

Ce travail est présenté dans le présent manuscrit à travers quatre chapitres. Dans le chapitre I, nous revenons plus en détail sur les trois composantes de la variabilité mutationnelle globale : le taux de mutation, l'effet d'une mutation dans un gène et – à travers le nombre de gènes touchés par mutation – la structure du génome. Nous présentons en particulier l'apport des approches de modélisation pour comprendre l'évolution de chacune de ces composantes. Dans le chapitre II, nous décrivons le modèle individu-centré *aevol* (pour *a*rtificial *evol*ution), que nous avons spécifiquement développé pour étudier l'évolution de la structure du génome. Les chapitres III et IV présentent les résultats des simulations menées avec ce modèle. Nous montrons ainsi au chapitre III que la quantité de non codant et le nombre de gènes dépendent du taux de mutation, et que cela reflète la sélection indirecte d'un niveau constant de variabilité mutationnelle globale. Au chapitre IV, nous montrons qu'à taux de mutation constant, la quantité de non codant et le nombre de gènes dépendent de l'impact phénotypique moyen d'une mutation dans un gène. Ce couplage, surprenant de prime abord, reflète lui aussi la sélection indirecte d'un certain niveau de variabilité mutationnelle. Nous montrons également dans ce chapitre que la stabilité de l'ordre des gènes dépend elle aussi de l'impact phénotypique moyen des mutations géniques. L'ensemble de ce travail suggère qu'en plus des éventuels biais mutationnels et coûts sélectifs directs, des pressions indirectes complexes peuvent également s'exercer sur la structure d'un génome.

Chapitre I

Variabilité mutationnelle : mécanismes et évolution

I say over and over again that Natural Selection can do nothing without variability, and that variability is subject to the most complex fixed laws...

C. Darwin à C. Lyell, 1er octobre 1862[1]

L'apparition de variations individuelles au sein d'une population est le préalable indispensable à l'action de la sélection naturelle, qui ne peut que "faire le tri" parmi différents phénotypes. Ces variations individuelles peuvent être dues à des mutations de l'ADN, mais aussi à des facteurs environnementaux, ou encore à des effets stochastiques dans l'expression des gènes et l'action des protéines. Ce sont cependant les variations dues aux mutations qui permettent l'évolution, car elles sont héritées lors de la transmission du génotype. C'est ce potentiel de variation phénotypique (due aux mutations) entre un individu et ses descendants que l'on appelle ici variabilité mutationnelle du phénotype. Il joue un rôle clé dans l'évolution.

Le degré de variabilité mutationnelle du phénotype dépend à la fois de la fréquence des mutations spontanées et de l'impact de ces mutations sur le phénotype. En effet, certaines mutations n'ont pas d'effet visible sur le phénotype, d'autres le modifient légèrement, et d'autres encore le bouleversent (anéantissant alors souvent les chances de survie de l'individu). Les progrès de la biologie moléculaire ont permis de révéler des mécanismes modulateurs de la variabilité mutationnelle, tant au niveau de la fréquence des mutations qu'au niveau de leur effet phénotypique. Comme nous l'avons déjà mentionné, il existe des enzymes capables de réparer certaines erreurs de réplication, réduisant ainsi le taux net de mutation entre un progéniteur et son descendant. L'effet phénotypique des mutations peut quant à lui être réduit par l'action de protéines telles que les "chaperonnes" évoquées

[1] In F. Darwin, éd., *Life and Letters of Charles Darwin*, New York, D. Appleton & Co., 1905, vol. 2.

en introduction : celles-ci permettraient à des protéines mutées de conserver un repliement correct et donc d'assurer normalement leurs fonctions.

Ces mécanismes modulateurs de la variabilité ont une base génétique et sont donc susceptibles d'évoluer. L'évolution serait alors à même d'influencer son propre cours (Fox Keller, 2000; Hermisson et Wagner, 2004). Ce phénomène intriguant est principalement étudié à travers deux sous-problématiques : (i) l'évolution des taux de mutation, et (ii) l'évolution de l'effet d'une mutation dans un gène.

L'objet de ce chapitre est tout d'abord de présenter les avancées principales dans ces deux domaines (sections 1 et 2). Dans les deux cas, il ne s'agira pas d'être exhaustif en ce qui concerne la description des mécanismes moléculaires, mais plutôt de fournir quelques exemples illustrant leur diversité. Nous verrons que pour inscrire ces mécanismes dans une perspective évolutive, la modélisation s'avère indispensable. Une large place est donc faite ici aux modèles développés pour comprendre l'évolution de tels mécanismes. Ces études permettent d'identifier les conditions qui conduisent à l'évolution du degré de variabilité mutationnelle, en termes d'intensité de sélection ou de taille de population par exemple.

Dans une troisième section, nous mettrons en évidence un chaînon manquant dans l'étude de l'effet d'une mutation : celui du nombre de gènes simultanément touchés. En effet, la majorité des modèles existants se focalisent sur les mutations qui touchent exactement un gène. Or on sait que d'une part, les génomes peuvent contenir de grandes quantités d'ADN non génique, et que d'autre part, les réarrangements de grands segments chromosomiques – pouvant toucher plusieurs gènes – jouent un rôle majeur dans l'évolution des génomes. La structure du génome, et en particulier la densité en gènes et en éléments répétés, est donc elle aussi susceptible de moduler l'effet moyen des mutations. Nous conclurons sur le besoin d'un nouveau type de modèle, permettant d'intégrer ce nouvel acteur de la variabilité mutationnelle.

1 Taux de mutation

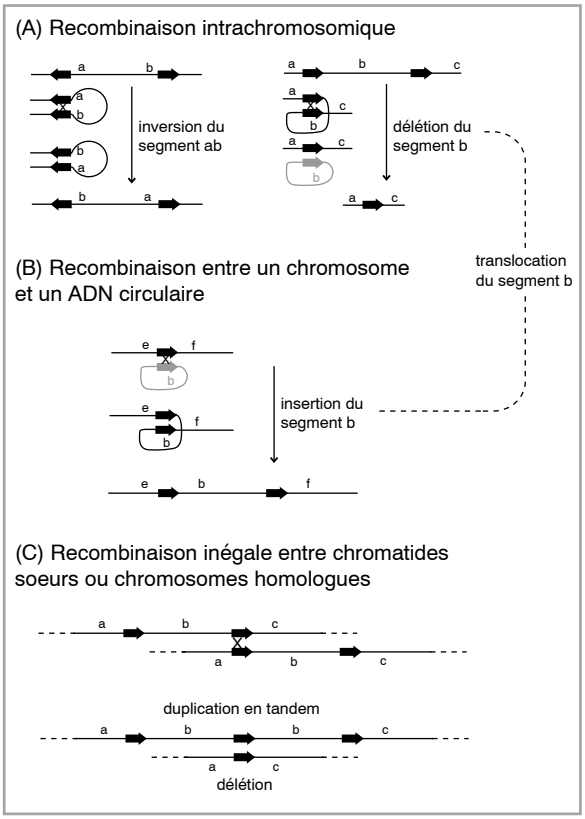
L'existence de mutations est la condition nécessaire de la variabilité mutationnelle du phénotype. Le taux de mutation est donc le point de départ logique de l'étude de cette variabilité. Le terme "mutation" est ici entendu au sens large : il peut s'agir de toute modification de l'information génétique, qu'elle concerne un nucléotide ou tout un segment chromosomique. Dans cette section, nous commencerons par décrire brièvement les mécanismes moléculaires qui peuvent moduler la fréquence des mutations. Le taux de mutation étant ainsi partiellement sous contrôle génétique, il peut lui-même évoluer par mutation et sélection, et nous présenterons donc les pressions sélectives qui peuvent s'exercer sur ce taux. Il est cependant difficile d'étudier expérimentalement ces pressions sélectives, et nous nous tournerons donc vers les approches de modélisation. Nous présenterons d'abord les modèles qui visent à déterminer *a priori* la valeur optimale du taux de mutation, puis ceux qui le font explicitement évoluer.

1.1 Mécanismes modulateurs du taux de mutation

À l'intérieur des cellules, l'ADN et les nucléotides libres sont sujets à des dommages physico-chimiques. Les bases de l'ADN peuvent par exemple être spontanément désaminées ou hydrolysées, ou encore être endommagées par la présence de radicaux libres. Ces dommages sont des sources de mutation car l'information nécessaire à la réplication de l'ADN est manquante ou erronée : le complexe de réplication, s'il n'est pas bloqué, risque d'incorporer au brin néosynthétisé une mauvaise base. Par ailleurs, des mutations ponctuelles et des petites insertions ou délétions peuvent être produites pendant la réplication même lorsque le brin à copier (brin matrice) est intact. Enfin, l'information génétique peut aussi être modifiée à plus grande échelle par divers mécanismes de transposition et de recombinaison, conduisant à la translocation (déplacement), l'inversion, la délétion ou la duplication de grands segments d'ADN (voir figure I.1).

Cependant, chez les procaryotes comme chez les eucaryotes, il existe des protéines dont l'action réduit la fréquence nette des mutations. En reprenant la typologie proposée par Radman *et al.* (1999), on peut distinguer cinq catégories de mécanismes moléculaires réduisant le taux de mutation (voir (Kornberg et Baker, 1992) et (Friedberg *et al.*, 1995) pour une description plus détaillée).

- **Préservation des brins matrices.** De nombreuses protéines dégradent les radicaux libres ou empêchent leur formation, ce qui réduit les dommages qu'ils peuvent causer à l'ADN. D'autres protéines peuvent réparer ces dommages, soit en effectuant la modification chimique qui restaure la base, soit en excisant la base endommagée et en plaçant une nouvelle à la place. Lorsque les deux brins sont cassés (double strand break), la cassure peut être réparée par recombinaison avec une molécule identique intacte (par exemple, entre la fin de la réplication et la division cellulaire, on dispose temporairement de deux copies de chaque chromosome dans la cellule).
- **Maintien d'un pool équilibré de nucléotides chimiquement sains.** Certaines enzymes permettent l'élimination des bases incorrectes. D'autres enzymes permettent que les concentrations des quatre nucléotides correspondent à la fréquence des différentes bases dans le génome.
- **Auto-correction réalisée par le complexe de réplication.** Lorsque le brin matrice et les nucléotides libres sont chimiquement de bonne qualité, l'ADN polymérase choisit les nucléotides à incorporer dans le brin néosynthétisé avec un taux d'erreur de l'ordre de 10^{-5}. Cependant, en cas d'erreur, le complexe de réplication est capable de revenir en arrière et d'éliminer la base incorrectement appariée. Cette activité exonucléasique du complexe de réplication ramène le taux d'erreur à 10^{-7}.
- **Correction des mésappariements post-réplicatifs.** Il existe un système d'enzymes qui détectent les mésappariements laissés par le complexe de réplication. Il peut s'agir de bases mal appariées (non complémentaires), ou bien de bases non appariées (sans vis-à-vis). Ce système est capable de distinguer le brin original du brin néosynthétisé. Il conserve alors l'original et excise la partie erronée de la copie, qui est ensuite resynthétisée. Chez *Escherichia coli*, ce système peut réparer les mutations ponctuelles et les insertions ou délétions de 1 à 4 nucléotides. Le taux d'erreur à l'issue de ces réparations descend alors à 10^{-10} (Radman et Wagner, 1986).



Réarrangements dus à la recombinaison homologue, illégitime et site-spécifique

Le schéma ci-contre montre comment la recombinaison homologue (aussi appelée recombinaison générale) entre séquences similaires peut conduire à des réarrangements. Chaque ligne représente une double hélice d'ADN. Les séquences répétées et leur orientation sont symbolisées par les flèches noires.

(A) La recombinaison entre séquences répétées d'un même chromosome conduit soit à l'inversion du segment qui les sépare, soit à sa délétion, selon que les répétitions sont inversées ou orientées dans le même sens. (B) Une séquence excisée peut être réinsérée dans le chromosome par le mécanisme inverse. Une délétion suivie d'une insertion peut donc aboutir à une translocation (déplacement). Notons que de l'ADN étranger circulaire peut par ce même mécanisme être intégré dans le chromosome s'il présente une zone de similarité avec une séquence chromosomique. Alternativement, un autre mécanisme de recombinaison, dit "site-spécifique", peut reconnaître des séquences cibles spécifiques dans les deux molécules et réaliser l'intégration comme sur le schéma. C'est par recombinaison site-spécifique que le phage λ s'intègre dans un chromosome bactérien. (C) Les séquences répétées peuvent aussi causer des décalages dans l'alignement de deux chromosomes homologues. La recombinaison est alors dite inégale : l'un des chromosomes porte une duplication en tandem et l'autre une délétion. Ce mécanisme existe aussi chez les organismes haploïdes car il peut également se produire entre deux copies d'un même chromosome, après la réplication.

Il faut enfin noter que la recombinaison dite illégitime entre séquences non homologues ou entre séquences homologues courtes peut conduire aux mêmes types de réarrangements. L'un des mécanismes sous-jacents est la réparation incorrecte des ruptures de la double hélice (double-strand breaks) par NHEJ (non homologous end-joining).

Réarrangements dus aux éléments transposables

Les éléments transposables, en tant que séquences répétitives dispersées dans le génome, peuvent médier les réarrangements par recombinaison homologue présentés dans les schémas A, B, C. Cependant, ils peuvent être à l'origine de réarrangements plus spécifiques.

En particulier, la présence d'éléments transposables (doubles flèches) de part et d'autre d'une séquence peut la rendre mobile (D). En effet, l'excision d'un transposon se fait en principe par rapprochement de ses deux "pieds" ( ), mais si une boucle plus grande est formée, ce sont les pieds de deux transposons différents qui peuvent se rapprocher, comme sur le schéma D. La séquence qui sépare les deux transposons est alors excisée en même temps qu'eux. L'intégration de ce segment excisé au niveau du site cible conduit alors à une translocation.

Lorsque ce ne sont pas les éléments eux-mêmes qui se déplacent, mais des copies (transposition réplicative), on obtient une duplication, comme sur le schéma E.

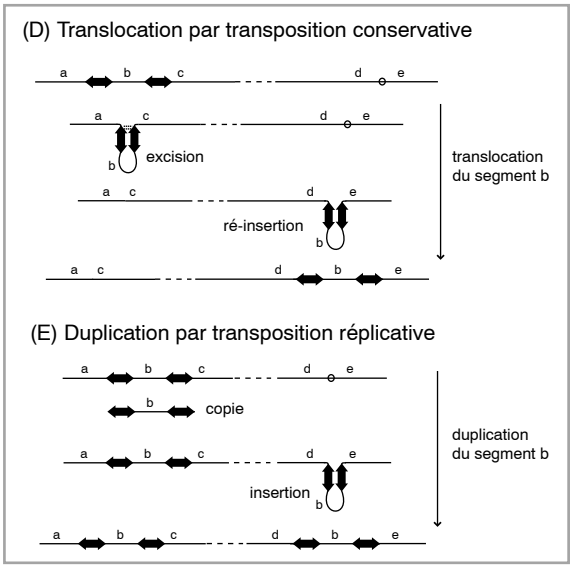


Fig. I.1: Mécanismes de réarrangements de l'ADN.

- **Maintien de la stabilité globale des chromosomes.** Ce même système de reconnaissance des mésappariements empêche la recombinaison excessive entre séquences non identiques et diminue donc la fréquence des réarrangements chromosomiques (Petit *et al.*, 1991). Chez les bactéries, ce système réduit aussi la fréquence des échanges génétiques entre espèces différentes (Rayssiguier *et al.*, 1989).

La perte ou l'altération de l'une de ces enzymes conduit à une augmentation parfois considérable (jusqu'à un facteur 10 000) du taux de mutation, et de tels hyper-mutants ne sont pas rares en laboratoire (Cox, 1976; Miller, 1996). Des souches de laboratoire présentant un taux de mutation très supérieur au type sauvage ont été décrites dès les années 40 chez la Drosophile (Plough, 1941) et quelques années plus tard chez les bactéries *Escherichia coli* (Treffers *et al.*, 1954) et *Salmonella typhimurium* (Miyake, 1960). On a découvert plus récemment que les souches hyper-mutantes étaient aussi présentes à une fréquence étonnamment élevée dans les populations bactériennes naturelles (LeClerc *et al.*, 1996; Oliver *et al.*, 2000; Denamur *et al.*, 2002; Richardson *et al.*, 2002). Inversement, des souches "anti-mutatrices", présentant par exemple des complexes de réplication plus fidèles, ont été isolées chez le bactériophage T4 et chez *Escherichia coli* (Schaaper, 1998). Cela montre que le taux de mutation est partiellement génétiquement contrôlé et qu'il peut être durablement augmenté ou diminué par des mutations dans les gènes impliqués dans la réplication et la réparation de l'ADN.

Le taux de mutation peut aussi être modifié transitoirement, par l'activation temporaire de certaines enzymes mutagènes. Parmi les plus étudiées, on trouve celles qui participent à la réponse "SOS" chez *Escherichia coli*. Leur expression est induite lorsque la réplication de l'ADN est gravement altérée, du fait de lésions multiples dans l'ADN, d'une inactivation de la machinerie de réplication ou de l'épuisement des précurseurs de l'ADN. Ces situations sont causées par des conditions environnementales particulières, représentant un stress pour les bactéries (irradiation par rayonnements UV, carence en nutriments...). Parmi les enzymes activées dans ces conditions, on trouve des ADN polymérases spéciales qui, paradoxalement, ne peuvent répliquer fidèlement l'ADN que s'il présente un certain type de dommage. Inversement, lorsqu'elles répliquent des portions d'ADN intact, ou présentant un type de dommage non reconnu, ces polymérases commettent un taux d'erreur de 100 à 10 000 fois plus grand que l'ADN polymérase exprimée en conditions normales (Friedberg *et al.*, 2002). Elles pourraient ainsi faire saturer le système de réparation des mésappariements (Rosenberg, 2001). L'activation de ces polymérases spéciales conduit vraisemblablement de nombreuses cellules à la mort en raison des mutations délétères produites, mais elle peut permettre à d'autres cellules de découvrir une mutation avantageuse et de survivre au stress (Rosenberg, 2001). Il existe des homologues de ces polymérases mutagènes chez les eucaryotes. Chez l'Homme en particulier, elles pourraient être impliquées dans l'hypermutation somatique des gènes d'immunoglobulines, qui permet l'expression d'un vaste répertoire d'anticorps (Friedberg *et al.*, 2002). Ainsi, l'existence dans un même génome de différentes polymérases plus ou moins mutagènes permet de moduler le taux de mutations ponctuelles dans l'espace et dans le temps.

La fréquence des grandes mutations peut aussi être transitoirement augmentée par l'activation de certaines enzymes. Par exemple, chez *Escherichia coli*, la réponse SOS semble également favoriser la mobilité des transposons (Kuan *et al.*, 1991), les grandes duplica-

tions (Dimpfl et Echols, 1989) et la recombinaison inter-spécifique (Matic *et al.*, 1995), en plus des mutations locales induites par les polymérases spéciales.

Ces différents mécanismes enzymatiques font augmenter ou diminuer le taux de mutation à l'échelle du génome entier. Des variations plus fines sont cependant possibles : l'hyper-mutabilité peut être limitée à certaines régions du génome. Des régions particulièrement sujettes aux petites insertions et délétions ont été par exemple identifiées dans les promoteurs ou les séquences codantes de certains gènes, chez des bactéries pathogènes comme *Haemophilus influenzae*, *Neisseria meningitidis* ou *Helicobacter pylori* (Moxon *et al.*, 2006). Ces régions, appelées "loci de contingence", sont constituées de répétitions de courtes séquences nucléotidiques. Lors de la réplication, ces répétitions peuvent induire la formation de boucles dans le brin matrice ou le brin néosynthétisé, ce qui cause alors l'ajout ou la suppression d'une ou plusieurs unités répétées (slipped-strand mispairing). On estime qu'à chaque réplication, un locus de contingence a, en fonction de sa longueur, une probabilité de 10^{-5} à 10^{-2} de subir ainsi une petite insertion ou une petite délétion (De Bolle *et al.*, 2000; Moxon *et al.*, 2006). Cette insertion/délétion peut affecter le niveau de transcription du gène ou bien causer un changement de cadre de lecture, ce qui entraîne généralement une interruption prématurée de la traduction. Ces régions hypermutables permettent donc des basculements génétiques de type "on/off", fréquents, stochastiques, héritables mais facilement réversibles. Chez les bactéries pathogènes mentionnées ci-dessus, cette hypermutabilité locale permet une variation antigénique considérable, car la majorité des gènes hypermutables correspondent soit à des molécules de surface, comme des adhésines ou des invasines (Moxon et Thaler, 1997), soit à des enzymes impliquées dans la synthèse du lipopolysaccharide (LPS)[1] (Van der Woude et Baumler, 2004).

Chez les eucaryotes, ces séquences répétitives, sujettes aux insertions/délétions par slipped-strand mispairing, sont plus connues sous le nom de microsatellites. Leur rôle biologique est moins bien connu et ils sont souvent utilisés comme des marqueurs neutres (Ellegren, 2004). Pourtant, si la plupart se trouvent dans des régions intergéniques ou introniques, il existe aussi des répétitions de triplets de nucléotides dans certaines régions codantes. D'autres microsatellites, situés en amont des gènes, sont des sites de fixation pour des protéines régulatrices. Le rôle de certains microsatellites dans l'expression et la fonction des gènes a pu être mis en évidence expérimentalement (Kashi *et al.*, 1997). Les génomes eucaryotes pourraient donc aussi contenir des gènes hypermutables.

Ainsi, différents mécanismes peuvent moduler le taux de mutation, localement ou à l'échelle du génome entier, et ces mécanismes dépendent en dernière analyse de la séquence génomique de l'organisme (présence de motifs répétés, gènes codant pour des polymérases ou des enzymes de réparation). Le taux de mutation local ou général peut donc évoluer, comme tout caractère partiellement sous contrôle génétique, par mutation, dérive et sélection. Mais la particularité de ce caractère est qu'il influence le cours de l'évolution, ce qui amène une question difficile : quelle est la dynamique d'un système

[1] Le LPS est un composant essentiel de la paroi bactérienne des bactéries à Gram négatif. Il constitue leur antigène O. Il s'agit d'un puissant stimulant de la réponse immunitaire chez l'Homme (Van der Woude et Baumler, 2004).

bouclé dans lequel le tempo de l'évolution est lui-même soumis à l'évolution ? Comme nous allons le voir, la modélisation s'est avérée particulièrement utile pour comprendre le fonctionnement d'un tel système.

1.2 Pressions sélectives sur le taux de mutation

Avant de présenter les différents modèles développés, leurs spécificités et leurs apports respectifs, nous donnons ici les grandes lignes du raisonnement qui leur est commun. Les modèles qui s'intéressent au taux de mutation et à son évolution distinguent généralement deux types de pressions sélectives potentielles. Les plus simples sont les pressions sélectives directes, qui s'appliquent lorsqu'un gène qui influence le taux de mutation influence aussi la fitness de l'organisme. C'est le cas par exemple si l'activité d'une enzyme de réparation a un coût énergétique. Le second type de pressions regroupe les pressions sélectives indirectes, qui propagent dans la population telle ou telle variante du gène modifieur en raison de son association avec des mutations avantageuses (ou, au contraire, qui empêchent sa propagation en raison de son association avec des mutations délétères). Considérons par exemple deux allèles M et m d'une ADN polymérase, M étant plus mutagène que m. Supposons que ces deux allèles n'ont pas d'effet direct sur la fitness, et donc qu'un individu M produit autant de descendants qu'un individu m. Cependant, la progéniture d'un individu M comporte plus de mutants. Autrement dit, les mutants produits à chaque génération sont majoritairement de type M. Lorsque les mutations sont délétères, les mutants sont contre-sélectionnés et l'allèle M l'est donc aussi. Le taux de mutation moyen dans la population diminue. Mais il peut aussi arriver qu'un des mutants porte une mutation avantageuse, auquel cas l'allèle M se propage dans la population par "auto-stop" (hitch-hiking) avec cette mutation avantageuse. Le taux de mutation moyen de la population augmente. On voit donc comment le taux de mutation peut évoluer sous l'effet de la sélection, même si l'on suppose qu'il n'a pas d'effet direct sur la fitness.

Ce mécanisme de sélection indirecte dépend cependant fortement de l'existence d'un déséquilibre de liaison[1] entre le gène qui modifie le taux de mutation et les mutations dont il est responsable (Sniegowski *et al.*, 2000; Chicurel, 2001). C'est le cas dans une population asexuée. Au contraire, s'il existe de la recombinaison allélique entre individus, l'allèle mutagène M peut être dissocié de la mutation avantageuse, ce qui réduit la probabilité de propagation par auto-stop. La force des pressions sélectives indirectes dépend donc de la fréquence des événements de recombinaison dans la population.

Intuitivement, il semble que le taux de mutation va s'équilibrer autour d'une valeur qui réalise un compromis optimal étant donné les forces relatives de trois pressions : le coût direct éventuel d'une réplication fidèle, la pression indirecte liée aux mutations délétères et la pression indirecte liée aux mutations avantageuses.

[1] Deux gènes sont en équilibre de liaison si leurs allèles sont associés au hasard dans la population. Ils sont en déséquilibre de liaison sinon. Pour un exemple extrême de déséquilibre, on peut imaginer une population où tous les individus sont soit $M-A$ soit $m-a$.

1.3 Quel est le taux de mutation optimal ?

Les premiers travaux théoriques concernant l'évolution du taux de mutation n'ont pas abordé la question dynamiquement mais ont cherché à déterminer *a priori* la valeur optimale vers laquelle il devait converger. Le raisonnement consiste à se donner un modèle de l'évolution et à choisir plus ou moins arbitrairement un indicateur mesurant le "succès" d'une population asexuée. Il peut par exemple s'agir de la fitness moyenne de la population à l'équilibre, ou bien de son fardeau génétique (genetic load)[1]. Le taux de mutation considéré comme optimal est alors celui qui optimise le critère choisi.

Certains travaux ont permis de délimiter la zone de recherche en identifiant un taux maximal. À l'aide du modèle des quasi-species (qui sera décrit dans la section suivante), Eigen (1971) a ainsi montré qu'au-delà d'un certain taux de mutation, une population infinie perd sa structure et se disperse aléatoirement dans l'espace des séquences : les génotypes les plus adaptés ne sont pas plus représentés que les autres, car ils sont plus vite perdus par mutation que reproduits par sélection. Le taux de mutation à partir duquel cet effet se produit est connu sous le nom d'"error threshold". Cependant, cet effet ne se produit que dans le cas où aucun génotype n'est létal (Wagner et Krall, 1993; Wilke, 2005). Si, au contraire, les mutations peuvent rendre les individus non viables, ceux-ci ne contribuent pas à la génération suivante. Deux cas sont alors possibles : (i) soit la taille de la population est constante, auquel cas elle peut rester structurée autour des génotypes adaptés malgré la forte pression mutationnelle (Wilke, 2005), (ii) soit la taille de la population peut diminuer, auquel cas elle risque de s'éteindre par "mutational meltdown" (Lynch *et al.*, 1993). L'existence d'un taux de mutation maximal dépend donc du modèle évolutif choisi, et lorsque ce seuil existe, il se manifeste par une "catastrophe" qui dépend elle aussi du modèle (dispersion ou extinction).

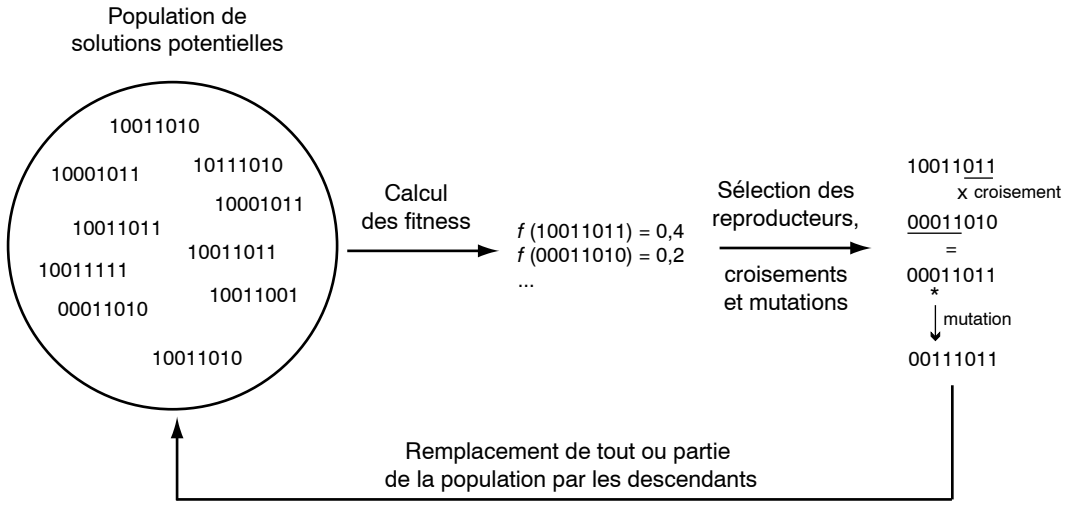
L'éventuel taux de mutation optimal doit se trouver entre 0 et ce taux de mutation catastrophique. L'existence d'un tel taux optimal est un phénomène bien connu en évolution artificielle. Lorsqu'on tente de résoudre un problème complexe avec un algorithme génétique (c'est-à-dire en employant métaphoriquement les principes de mutation et de sélection, voir figure I.2), on observe généralement que les meilleures performances sont obtenues pour un taux de mutation intermédiaire, réalisant ce qu'on appelle le "compromis exploitation-exploration" (Holland, 1975). La population doit en effet conserver d'une génération à l'autre les meilleures solutions obtenues jusqu'alors (exploitation) tout en explorant des solutions différentes, peut-être meilleures encore (exploration). Un taux de mutation trop élevé empêche l'exploitation et peut conduire à la perte des bonnes solutions, voire à la dispersion aléatoire de l'ensemble de la population si le taux de mutation catastrophique est atteint. Inversement, un taux de mutation trop faible ralentit l'exploration et la limite à des solutions très proches de celles existantes. La population peut alors stagner sur une solution suboptimale.

Ce type d'observations n'est pas limité aux algorithmes génétiques – l'existence d'un

[1] Le fardeau génétique d'une population est un nombre compris entre 0 et 1 qui mesure l'écart relatif entre la fitness moyenne de la population et la meilleure fitness possible.

Principe des algorithmes génétiques

Il existe différents types d'algorithmes dits "évolutionnaires", utilisant métaphoriquement les mécanismes de l'évolution pour résoudre des problèmes d'optimisation. Les algorithmes génétiques, proposés par John Holland en 1975, sont probablement les plus connus. Dans le cas le plus classique, l'objectif est d'optimiser une fonction à *n* paramètres. Chaque solution potentielle est donc un vecteur de *n* valeurs numériques, mais ce vecteur est codé sous forme d'une chaîne binaire : les *n* valeurs numériques sont exprimées en base 2 et concaténées. Comme le montre le schéma ci-dessous (avec *n*=4 paramètres codés chacun sur deux digits), l'algorithme travaille itérativement sur une population de solutions potentielles ("individus"). La première génération est initialisée avec des individus aléatoires. On évalue ensuite la fitness de chaque individu en décodant les valeurs des *n* paramètres et en calculant la valeur obtenue pour la fonction à optimiser. Les individus qui présentent les meilleurs scores de fitness sont sélectionnés comme reproducteurs. Les descendants sont construits par croisement ("crossover") et mutation ("switch") des chaînes binaires parentales. Les descendants remplacent tout ou partie de la population parentale et le processus est réitéré jusqu'à obtention d'une solution jugée satisfaisante.



D'après (Rennard, 2002)

Notons que dans la pratique, le codage binaire n'est pas toujours le plus adapté. Il est parfois plus efficace de travailler directement sur les valeurs réelles des paramètres, en utilisant des opérateurs de croisement et de mutation appropriés.

Fig. I.2: Principe des algorithmes génétiques.

taux de mutation optimal se généralise à d'autres systèmes d'évolution artificielle plus complexes, basés par exemple sur un partage explicite de ressources entre agents mobiles (Bedau et Packard, 2003). Cependant, pour aller au-delà de l'observation empirique et tenter de prédire analytiquement ce taux optimal, il faut analyser en détail l'interférence entre mutations délétères et avantageuses. Il faut alors faire appel à des modèles plus simples comme ceux utilisés classiquement en génétique des populations.

Kimura (1960, 1967) a ainsi tenté de déterminer le taux de mutation optimal en cherchant le taux de mutation qui minimise le fardeau génétique. À l'équilibre, Kimura distingue deux contributions au degré d'inadaptation de la population : les mutations délétères et les changements environnementaux. L'effet du taux de mutation sur le fardeau génétique est donc complexe : lorsque le taux de mutation est fort, la population s'adapte plus rapidement aux variations environnementales, mais c'est au prix d'une augmentation du nombre de mutations délétères. Kimura a montré qu'il existe un taux de mutation qui minimise le fardeau génétique en réalisant un compromis entre ces deux effets. Selon ce critère, le taux de mutation génomique optimal serait égal à la fréquence des changements environnementaux.

Cependant, la valeur optimale du taux de mutation dépend fortement des hypothèses relatives à la fréquence de fixation des mutations favorables. Dans le modèle de Kimura (1967), la population est supposée adaptée la plupart du temps, sauf lorsqu'un changement environnemental se produit. Les mutations ne peuvent donc être favorables que juste après un changement environnemental. Le nombre de mutations favorables fixées est ainsi supposé égal au nombre de changements environnementaux. Il est donc indépendant du taux de mutation.

Si l'on suppose au contraire que la population est en permanence loin de l'optimum, alors le nombre de mutations favorables fixées est cette fois déterminé par le nombre de mutations favorables spontanées, qui est lui-même proportionnel au nombre de mutations. Un taux de mutation élevé paraît donc avantageux à première vue. Deux mécanismes vont pourtant à l'encontre d'un taux de mutation trop élevé : "l'interférence clonale" (Gerrish et Lenski, 1998) et l'effet "ruby in a rubbish" (Peck, 1994). L'interférence clonale, aussi connue sous le nom d'effet Hill-Robertson, correspond à la compétition entre mutations avantageuses au sein d'une population asexuée. Une mutation avantageuse qui n'est pas encore fixée peut être supplantée par une autre mutation plus avantageuse. Augmenter le taux de mutation n'accélère donc pas nécessairement la vitesse d'adaptation (Gerrish et Lenski, 1998), ce qui a été confirmé expérimentalement (De Visser *et al.*, 1999). L'effet "ruby in a rubbish" empire la situation. Il s'agit de l'interférence des mutations avantageuses avec les mutations délétères. Lorsque le taux de mutation est élevé, une mutation avantageuse a de grandes chances de se produire dans un génome qui comporte déjà des mutations délétères, ce qui réduit ses chances de fixation. Dans ces conditions, le taux de mutation optimal doit réaliser un compromis entre une fréquence suffisante de mutations favorables et une proportion suffisante de génomes sans mutation délétère. Orr (2000) et Johnson et Barton (2002) ont montré que la valeur optimale en question est approximativement égale à $\max(s_b, s_d)$, où s_b (resp. s_d) est l'effet d'une mutation bénéfique (resp. délétère)

sur la fitness[1].

Une troisième possibilité consiste à supposer que la fréquence de fixation des mutations favorables est nulle, par exemple parce que l'optimum est atteint et qu'il ne change pas au cours du temps. Lorsqu'on néglige ainsi les mutations favorables, le taux de mutation génomique optimal est 0, sauf s'il existe un coût direct à la fidélité de la réplication. Dans ce cas, le taux de mutation optimal dépend de l'équilibre entre le coût d'un faible taux de mutation et les effets délétères associés à un fort taux de mutation (Dawson, 1998, 1999).

1.4 Le taux de mutation peut-il évoluer vers son optimum ?

L'existence d'un taux de mutation optimal – étant donnés un modèle de l'évolution et un critère de succès – n'est pas suffisante pour prédire que la dynamique du processus évolutif va effectivement amener le taux de mutation vers cet optimum. Si l'on met en compétition des sous-populations différant par leurs taux de mutation, on peut être tenté d'utiliser les résultats précédents et de conclure que la sous-population la plus représentée à long terme sera celle qui maximise le critère de succès (par exemple, celle qui a la meilleure fitness moyenne à l'équilibre). Le taux de mutation moyen dans l'ensemble de la population aura alors évolué vers son optimum. Le problème de ce raisonnement est qu'il se fonde sur la notion controversée de sélection de groupe et non sur une sélection inter-individus. Il n'est correct que si les sous-populations sont strictement indépendantes, ce qui suppose l'absence de recombinaison, mais aussi l'absence de mutations susceptibles de faire changer le taux de mutation (Johnson, 1999a). En effet, lorsqu'un individu subit une mutation qui change son taux de mutation, il change de fait de sous-population. Les sous-populations ne sont donc plus indépendantes.

Il est donc souhaitable prendre en compte explicitement la dynamique du processus évolutif et de faire réellement évoluer le taux de mutation, en même temps et de la même façon que les autres loci : par mutation et sélection. Cette idée fait partie des champs de recherche actuels en évolution artificielle (Eiben *et al.*, 1999; Meyer-Nieberg et Beyer, 2006). Cependant, l'objectif dans ce domaine reste le plus souvent l'optimisation des performances de l'algorithme, et dans la plupart des cas, on ne s'intéresse pas au taux de mutation obtenu en tant que tel. Peu d'études font la comparaison explicite entre le taux de mutation fixe optimal et le taux de mutation obtenu dynamiquement au cours de l'évolution. Cela a été fait pour le système multi-agents (asexués) étudié par Bedau et Packard (2003), en faisant subir au taux de mutation de chaque agent une variation aléatoire lorsque l'agent se reproduit. Le taux de mutation moyen de la population évolue alors spontanément vers la valeur qui serait optimale si le taux de mutation était fixe (il s'agit d'un taux intermédiaire, qui reflète le compromis exploitation-exploration).

Dans les modèles de génétique de populations, l'idée de faire explicitement évoluer le taux

[1] Dans les modèles simples présentés ici, les effets des mutations sont supposés indépendants du contexte génétique et constants au cours du temps. Nous verrons cependant dans la section 2 de ce chapitre que l'effet des mutations est lui aussi susceptible d'évoluer.

de mutation apparaît dans les années 70, avec l'ajout d'un locus "modifieur" qui contrôle le taux de mutation des autres loci (qui, eux, influencent la fitness) (Leigh, 1970, 1973; Karlin et McGregor, 1974; Gillespie, 1981; Holsinger et Feldman, 1983). Dans ces travaux pionniers, la situation est simplifiée dans la mesure où seuls quelques allèles (2 en général) pré-existent dans la population au niveau du locus modifieur, sans mutation possible de l'un à l'autre. Ces différents allèles peuvent alors être indirectement sélectionnés ou contre-sélectionnés, s'ils sont durablement associés à des mutations avantageuses ou délétères (voir paragraphe 1.2).

Ainsi, le coefficient de sélection d'un allèle mutagène dépend du niveau de recombinaison et des hypothèses faites sur la fréquence des mutations favorables. Considérons par exemple le modèle le plus simple, où les individus sont haploïdes et où l'on s'intéresse à seulement deux loci : un locus modifieur avec deux allèles M et m, et un locus directement sélectionné avec deux allèles A et a. La fitness d'un individu A est 1, et celle d'un individu a est $1-s_d$. Le taux de mutation au niveau du locus sous sélection vaut u si l'individu possède l'allèle m, et $u+\Delta u$ s'il possède l'allèle M. On appelle r le taux de recombinaison entre les deux loci. Le coefficient de sélection de M vaut $\frac{-s_d \Delta u}{s_d+r}$ si les mutations de a vers A sont négligées (Sniegowski *et al.*, 2000). Autrement dit, si les mutations avantageuses sont négligées, un allèle qui augmente le taux de mutation est contre-sélectionné, et un allèle qui le diminue est sélectionné. Cela signifie que si les mutations sont toutes délétères et s'il n'y a pas de coût direct à la fidélité, le taux de mutation ne peut que diminuer. La convergence vers le taux nul, optimal dans ces conditions, est d'autant plus rapide que la recombinaison est rare.

Si l'on suppose au contraire que des mutations favorables sont possibles, par exemple parce que l'environnement varie, alors les modèles les plus simples prédisent qu'un allèle mutagène peut être fixé dans une population asexuée (Leigh, 1970, 1973; Taddei *et al.*, 1997; Tenaillon *et al.*, 1999), ce qui a été confirmé expérimentalement (Chao et Cox, 1983; Mao *et al.*, 1997; Sniegowski *et al.*, 1997). Earl et Deem (2004) ont obtenu des résultats similaires avec un modèle plus complexe, où des séquences protéiques évoluent par mutations ponctuelles et par "domain swapping" (remplacement de toute une sous-séquence par une autre sous-séquence choisie aléatoirement dans un pool prédéfini) : le taux de domain swapping indirectement sélectionné est d'autant plus fort que l'environnement varie fréquemment. Pour les populations asexuées, il semble donc que les mutations favorables puissent faire augmenter le taux de mutation. De par l'action opposée des mutations favorables et des mutations délétères, le taux de mutation pourrait donc évoluer vers un optimum non nul, tel que celui prédit par Kimura (1967). Cependant, cette valeur optimale ne peut être atteinte que si des variations de plus en plus fines du taux de mutation sont possibles (Sniegowski *et al.*, 2000). Les mutations des mécanismes de réplication et de réparation connues à l'heure actuelle conduisent plutôt à de fortes variations du taux de mutation, mais cela n'exclut pas la possibilité de variations plus subtiles, plus difficiles à détecter.

Pour les populations recombinantes, le rôle des mutations favorables dans le maintien d'un fort taux de mutation est plus controversé. Les modèles de Gillespie (1981) et Ishii *et al.* (1989) suggèrent qu'un environnement variable peut faire augmenter le taux de mutation

dans une population recombinante. Au contraire, le modèle de Leigh (1973) suggère que l'effet des mutations favorables est négligeable dans ce cas. Les deux types de modèles diffèrent par l'existence ou non de mutations inconditionnellement délétères (Johnson, 1999b). En effet, dans le modèle de Leigh (1973), une certaine partie des loci ne peut subir que des mutations délétères, quel que soit l'environnement. Cela réduit de fait la proportion de mutations favorables.

Si les mutations favorables sont tellement rares que leur effet sur le taux de mutation est négligeable en présence de recombinaison, alors on doit supposer que c'est un coût direct à la fidélité de la réplication qui l'empêche de tendre vers zéro (Drake *et al.*, 1998; Sniegowski *et al.*, 2000). Ce point de vue, qui semble privilégié par une partie importante des biologistes moléculaires et des généticiens des populations, est quasi-orthogonal à l'idée du compromis exploitation-exploration largement répandue en évolution artificielle. La divergence porte fondamentalement sur un paramètre extrêmement difficile à estimer : le rapport entre la fréquence des mutations favorables et celle des mutations délétères. Selon Johnson (1999b), les quelques données empiriques disponibles suggèrent que pour les eucaryotes pluricellulaires, les mutations favorables sont trop rares pour affecter l'évolution du taux de mutation. Elles pourraient au contraire jouer un rôle important chez les micro-organismes qui font face à un environnement nouveau ou fluctuant. La fiabilité de ces données reste cependant incertaine, le taux de mutations favorables restant dans tous les cas un paramètre très difficile à estimer.

On retiendra que même si les différentes approches de modélisation peuvent diverger sur la valeur d'équilibre du taux de mutation, elles ont permis d'inscrire les mécanismes moléculaires de mutation et de réparation dans une perspective évolutive, en formalisant et en analysant les pressions indirectes auxquels ils peuvent être soumis. Avec la notion de sélection indirecte par association avec les mutations favorables, et, inversement, de contre-sélection indirecte par association avec les mutations délétères, les travaux relatifs à l'évolution du taux de mutation ont formellement mis en place les principes fondamentaux de l'évolution de la variabilité en général.

2 Effet des mutations géniques

Si le taux de mutation constitue une composante fondamentale de la variabilité mutationnelle du phénotype, il n'en est pas pour autant le seul acteur. La complexité de la relation entre génotype et phénotype fait qu'une petite modification génotypique ne se traduit pas mécaniquement par une petite modification du phénotype. L'effet d'une mutation ponctuelle, par exemple, peut être catastrophique ou au contraire indétectable. L'effet phénotypique des mutations est donc aussi une composante importante de la variabilité mutationnelle du phénotype. Dans cette section, nous nous intéresserons à l'aspect du problème le plus documenté à la fois du point de vue expérimental et du point de vue théorique, à savoir l'effet d'une mutation qui affecte un seul gène. Nous commencerons par montrer la multiplicité des niveaux auxquels l'effet d'une telle mutation peut se

jouer. Cette revue mettra en évidence des propriétés étonnantes de robustesse, qu'il serait tentant de considérer comme adaptatives. Nous verrons cependant qu'il est difficile de prouver expérimentalement que la robustesse mutationnelle des organismes est sélectionnée. C'est donc à travers les résultats des approches théoriques que nous présenterons les conditions de l'évolution adaptative de la robustesse mutationnelle. Avec les modèles de génétique quantitative, nous verrons dans quelles conditions un mécanisme de "canalisation" (neutralisation) des mutations peut être indirectement sélectionné. Avec les modèles de quasi-species, nous montrerons que l'évolution de la robustesse peut être étudiée sans inclure explicitement un gène canalisateur, en distinguant des mutations neutres et des mutations délétères. Avec les approches de vie artificielle, nous verrons dans quelles conditions la survie du plus robuste (ou du plus neutre) peut prédominer sur la survie du plus apte. Il nous faudra enfin constater que la plupart de ces approches ont ignoré les mutations favorables, et montrer que lorsqu'on les prend en compte, le phénotype peut au contraire devenir plus sensible aux mutations.

2.1 Mécanismes modulateurs de l'effet des mutations

De nombreux facteurs peuvent moduler l'effet d'une mutation dans un gène. Nous en citons quelques-uns ici pour illustrer leur diversité.

Effet des mutations sur la séquence protéique

L'un des premiers niveaux où se joue l'effet d'une mutation est celui du code génétique. Comme toute table de correspondance entre 64 triplets de nucléotides et 20 acides aminés (plus un signal d'arrêt de la traduction), le code génétique dit universel est dégénéré : certaines mutations ponctuelles dans la séquence codante sont sans effet sur la séquence de la protéine. Cependant, comparé à des codes aléatoires de même dimension, le code génétique universel présente une régularité qui lui confère des propriétés particulières. Il fait partie des codes qui confèrent aux séquences protéiques une grande robustesse, sans toutefois la maximiser (Maeshiro et Kimura, 1998; Judson et Haydon, 1999). Selon Maeshiro et Kimura (1998), ce compromis entre robustesse et variabilité des séquences serait un facteur essentiel à la survie et à l'évolution des organismes. De plus, la structure du code est telle que lorsque la séquence protéique varie, les propriétés physico-chimiques globales de la protéine sont souvent conservées. En effet, si l'on considère un usage non biaisé du code, une mutation ponctuelle qui change un acide aminé a de grandes chances de conduire à un acide aminé d'hydrophobicité similaire (Woese, 1965; Epstein, 1966; Goldberg et Wittes, 1966; Haig et Hurst, 1991; Maeshiro et Kimura, 1998). Cette robustesse des propriétés physico-chimiques des protéines vis-à-vis des mutations ponctuelles est encore plus forte si les biais mutationnels sont pris en compte[1] (Freeland et Hurst, 1998). Cependant, il est possible d'obtenir des codes encore plus conservatifs par sélection

[1] Toutes les mutations ponctuelles ne sont pas équiprobables. Les transitions (A $\leftrightarrow$ G, C $\leftrightarrow$ T) sont généralement plus fréquentes que les transversions (A $\leftrightarrow$ C, A $\leftrightarrow$ T, G $\leftrightarrow$ C, G $\leftrightarrow$ T).

artificielle *in silico*.

Ainsi, lorsqu'il est utilisé sans biais, le code génétique universel confère une robustesse importante aux protéines tout en laissant des possibilités d'exploration non négligeables, tant au niveau des séquences que des propriétés physico-chimiques globales (Judson et Haydon, 1999). Toutefois, comme nous le verrons plus loin (paragraphe 2.6), un usage biaisé du code peut faire pencher la balance d'un côté ou de l'autre, c'est-à-dire vers une plus grande robustesse ou au contraire vers une plus grande variabilité.

Effet des mutations sur la structure et la fonction de la protéine

La traduction n'est que la première étape dans l'évaluation de l'effet attendu d'une mutation ponctuelle dans une séquence codante. Après avoir identifié les changements d'acides aminés les plus probables grâce à la structure du code génétique, il faut évaluer l'impact moyen d'un changement d'acide aminé sur la structure tridimensionnelle et la fonction de la protéine. Il est possible de perturber expérimentalement une protéine en changeant un ou plusieurs de ses acides aminés et de regarder si sa structure ou sa fonction a changé. Ces études de mutagénèse suggèrent que la fonction de nombreuses protéines est insensible à la majorité des changements d'acides aminés possibles (Wagner, 2005a). Par exemple, Rennell *et al.* (1991) ont construit 2015 variantes de la lysozyme du bactériophage T4, une enzyme nécessaire à la lyse de la bactérie hôte et donc à la propagation du virus. Seules 16% de ces variantes affectaient sérieusement la capacité de lyse du virus. Sur les 163 sites de la protéine, 89 ont pu tolérer tous les acides aminés testés. Des études comparables sur une protéase du virus HIV (Loeb *et al.*, 1989), sur diverses protéines enzymatiques ou régulatrices d'*Escherichia coli* (Kleina et Miller, 1990; Huang *et al.*, 1996) et sur l'hémoglobine humaine (Weatherall et Clegg, 1976) ont montré que la fonction de ces protéines résiste à de très nombreux changements ponctuels dans leurs séquences d'acides aminés. La structure et la fonction de certaines protéines peuvent également résister à des événements de recombinaison, qui modifient potentiellement plusieurs acides aminés contigus (Leong *et al.*, 2003).

Les approches comparatives peuvent fournir un autre faisceau d'indices pour ce niveau de robustesse. On trouve chez des espèces distinctes des protéines de structure et de fonction similaires, mais de séquences très différentes (Aronson et Hendrickson, 1994; Thornton *et al.*, 1999). Il existe cependant des contre-exemples, notamment parmi les protéines structurales comme les histones, les actines ou les tubulines, qui peuvent présenter jusqu'à 98% de similarité de séquence dans des taxons aussi éloignés que les mammifères et les plantes (Doolittle, 1995). Il est donc difficile de dégager l'effet moyen d'un changement d'acide aminé en se basant sur un petit nombre protéines choisies plus ou moins arbitrairement. Chez l'Homme, une tentative de quantification plus systématique a été réalisée pour une cinquantaine de protéines (Eyre-Walker *et al.*, 2002). Les auteurs ont compté le nombre de changements d'acides aminés observés entre l'Homme et le Chimpanzé et l'ont comparé au nombre attendu si ces changements étaient neutres. Cette évaluation du niveau de contraintes sélectives a permis d'estimer à environ 30% la proportion de

changements d'acides aminés neutres ou quasi-neutres[1]. Cette estimation repose cependant sur un certain nombre d'hypothèses simplificatrices, et les critères qui ont présidé au choix des cinquante protéines ne sont pas clairs. Il faudra une étude plus complète pour confirmer cette estimation, mais elle donne déjà un ordre de grandeur du niveau de robustesse des protéines vis-à-vis des changements d'acides aminés.

Lorsque la fonction d'une protéine n'est pas intrinsèquement robuste vis-à-vis des changements d'acides aminés, elle peut parfois être rétablie par l'action de protéines dites "chaperonnes", qui stabilisent les protéines dénaturées, déstabilisées ou susceptibles de l'être. Elles les empêchent de s'agréger et leur permettent de reprendre leur configuration native. Ces protéines sont également appelées "protéines de choc thermique" (Heat Shock Proteins) car elles sont activées en réponse à un choc thermique. C'est ainsi qu'elles ont été découvertes dans les années 60, mais il a été montré par la suite qu'elles répondent également à d'autres types de stress (Feder et Hofmann, 1999). De nombreuses protéines chaperonnes sont également actives en conditions normales, chez les procaryotes comme chez les eucaryotes. L'une des plus connues est la protéine HSP90, qui stabilise de nombreuses protéines de transduction du signal. Certains changements d'acides aminés peuvent rendre ces protéines instables et donc, en principe, affecter leur fonctionnement, mais HSP90 les stabilise et les garde ainsi prêtes à être activées par le signal. Rutherford et Lindquist (1998) ont montré que chez la Drosophile, muter ou inactiver HSP90 produit des variations phénotypiques quasiment sur tous les traits morphologiques, depuis la forme des ailes jusqu'à la structure oculaire. Cela suggère que l'inactivation d'HSP90 permet à des mutations jusque-là "cryptiques" (Cossins, 1998) de s'exprimer. Le même phénomène a été observé suite à l'inactivation de HSP90 chez la plante *Arabidopsis thaliana* (Queitsch *et al.*, 2002). Chez *Escherichia coli*, la surexpression d'une autre chaperonne, GroEL, permet à des souches ayant accumulé des mutations délétères de retrouver un meilleur taux de croissance (Fares *et al.*, 2002). Ainsi, dans des conditions normales, de nombreuses mutations pourraient s'accumuler dans le génome de façon neutre, et s'exprimer lors de la perturbation du système de "tampon moléculaire" que constituent les chaperonnes.

Effet des mutations sur le profil d'expression

La contribution d'un gène à la survie de l'organisme ne dépend pas seulement de la structure de la protéine, mais aussi de son profil d'expression dans l'espace et dans le temps. Autrement dit, les mutations qui se produisent dans les régions régulatrices d'un gène sont tout aussi susceptibles d'affecter la survie de l'organisme que celles qui se produisent dans la séquence codante. Les sites de régulation sont en général courts et la disparition et l'apparition de sites peut donc être un phénomène très rapide. Par exemple, dans une population d'un million de drosophiles, il ne faut théoriquement que 24 ans (à raison de 10 générations par an) pour qu'un site de 6 paires de bases apparaisse par mutations

[1] Une mutation est quasi-neutre si son effet délétère est suffisamment faible pour que la sélection ne puisse pas l'éliminer : lorsque le coefficient de sélection d'une mutation est inférieur à $\frac{1}{N_e}$, où N_e est la taille efficace de la population, la mutation est sujette à la dérive génétique et se comporte de fait comme une mutation neutre.

ponctuelles aléatoires et soit fixé dans la population par dérive génétique (Stone et Wray, 2001).

Certaines régions régulatrices semblent avoir une organisation rigide, très conservée au cours de l'évolution. Cependant, on observe aussi un grand nombre de gènes dont les régions régulatrices sont organisées très différemment d'une espèce à l'autre, tout en présentant le même profil d'expression spatio-temporelle. Par exemple, si l'on compare les régions régulatrices dites "stripe2" du gène de développement *eve* chez deux espèces de Drosophiles, on remarque que cinq sites (sur un total de 12) ne sont présents que dans l'une des deux espèces, alors que le profil d'expression est identique (Ludwig *et al.*, 2000). De même, en analysant le gène *endo16* chez deux espèces d'oursins, on observe un même profil d'expression malgré une importante divergence des séquences régulatrices (Romano et Wray, 2003). Ces observations peuvent être dues à une certaine robustesse du profil d'expression vis-à-vis des mutations dans les régions régulatrices. Cependant, les indices qui vont dans ce sens restent encore anecdotiques. Il faudrait ici des études plus systématiques. Par ailleurs, le fait que deux espèces actuelles présentent le même profil d'expression n'implique pas nécessairement que ce profil soit robuste aux mutations. Il peut par exemple s'agir d'une convergence évolutive. C'est une des limites des approches comparatives en général : il manque la référence de l'état ancestral. Les méthodes de mutagénèse sont plus directes mais malheureusement peu nombreuses dans ce cas.

Effet des mutations sur le fonctionnement d'un réseau

Lorsque l'activité d'une protéine est réduite ou annulée par une mutation (dans la séquence codante ou dans la région régulatrice), il reste encore un niveau qui peut moduler l'effet de la mutation sur la fitness de l'organisme – celui du réseau d'interactions avec les autres produits géniques. Nous prenons ici le terme d'interaction au sens large, dans une perspective fonctionnelle : interactions physiques avec d'autres protéines, régulation de l'expression d'autres gènes, participation à une voie métabolique ou à une cascade de signalisation, etc. Du fait de ces interactions, une mutation génique peut être pléiotrope, c'est-à-dire affecter plusieurs caractères phénotypiques. On peut donc s'attendre à ce que l'effet phénotypique global d'une mutation soit lié à sa position dans le réseau d'interactions fonctionnelles. Il semble en effet que chez la Levure *Saccharomyces cerevisiae*, la probabilité qu'une perte de gène soit létale soit corrélée avec le nombre d'interactions physiques de la protéine (Jeong *et al.*, 2001). Mais si l'organisation en réseau démultiplie ainsi l'effet des pertes de protéines centrales, elle peut en même temps "diluer" l'effet de nombreuses autres mutations. Ainsi, dans un réseau, la modification de la pondération d'une connexion ou de l'activité d'un noeud peut n'avoir aucun effet sur les points de stabilité globaux.

Dans le cas simple d'une voie métabolique linéaire, Kacser et Burns (1981) ont ainsi montré que le flux dans la voie est théoriquement relativement insensible aux variations de l'activité d'une enzyme individuelle, et ce d'autant plus que le nombre d'enzymes dans la voie est grand. Dès lors que la voie comporte plus d'une dizaine d'enzymes, on

peut notamment diminuer l'activité d'une enzyme de 50% sans affecter significativement le flux. Dykhuizen *et al.* (1987) ont obtenu des résultats expérimentaux analogues chez *Escherichia coli* : une réduction de 75% de l'activité de la β-galactosidase ne réduit la fitness que de 1,3%, alors que cette enzyme est centrale dans le métabolisme du lactose et que le lactose était, dans cette expérience, la seule source de carbone disponible. Notons que les mécanismes de régulation des enzymes par les métabolites, qui permettent en général de garder une sortie constante lorsque les concentrations en substrats changent, peuvent corrélativement renforcer la robustesse de la voie vis-à-vis des mutations (Fell, 1997).

Dans le cas de la perte complète d'une enzyme métabolique, causée par exemple par la délétion du gène, la voie métabolique correspondante se trouvera bloquée, mais pas nécessairement le réseau métabolique dans son ensemble (sauf si l'enzyme en question était un point central du réseau). L'intégration des voies métaboliques dans un réseau complexe permet en effet l'existence de chemins alternatifs et donc un niveau supplémentaire de robustesse. Par exemple, l'analyse *in silico* du métabolisme central d'*Escherichia coli* suggère que sur 48 réactions éliminées, seules 7 seraient essentielles à la survie et à la croissance aérobie en milieu minimal (Edwards et Palsson, 2000). Pour la majorité des 41 restantes, la perte de la réaction ne réduirait le taux de croissance que de 5%.

Les réseaux de régulation présentent également des propriétés étonnantes de robustesse. En effet, en simulant des mutations dans un modèle simple de réseau de régulation, Wagner (1996) a montré que près de 60% des changements d'interactions sont sans effet sur l'état d'activation des gènes à l'équilibre. Cette forte proportion de mutations neutres est indépendante de la taille du réseau et de sa connectivité. Von Dassow *et al.* (2000) ont obtenu des résultats comparables en simulant cette fois le fonctionnement d'un réseau réel, impliqué dans la morphogénèse chez la Drosophile. Seule la topologie de ce réseau est connue, les paramètres (force des activations et des inhibitions, taux de dégradation des protéines, ...) ne le sont pas. Les auteurs ont montré qu'en les choisissant au hasard, on obtient quasiment systématiquement un réseau fonctionnel, et que ce réseau est en général étonnamment robuste vis-à-vis des variations des paramètres.

Robustesse distribuée ou redondance ?

Les mécanismes de robustesse présentés jusqu'ici reposent sur le fait que les systèmes biologiques sont constitués d'un grand nombre de composants, dont l'interaction peut masquer ou compenser la variation d'un composant individuel. On peut penser par exemple à la structure tridimensionnelle d'une protéine, résultat des interactions entre plusieurs centaines d'acides aminés, ou encore à un réseau métabolique, mettant en jeu des centaines d'enzymes. En reprenant la terminologie proposée par Wagner (2005a), on peut qualifier ces mécanismes de robustesse *distribuée*. Cependant, un autre type de robustesse est souvent invoqué : la robustesse par *redondance*. Il s'agit alors d'expliquer la neutralité d'une mutation dans un gène par la présence d'un gène dupliqué intact.

Il est difficile de quantifier la contribution relative de ces deux types de robustesse. Chez la Levure, Gu *et al.* (2003) ont montré qu'une mutation dans un gène qui possède une copie a 20% de chances de plus d'avoir un faible effet sur la fitness qu'une mutation dans un gène non dupliqué. Mais lorsqu'un gène est dupliqué, la copie peut avoir divergé fonctionnellement et n'est donc pas nécessairement responsable du faible effet de la mutation. De plus, environ 40% des gènes mutables sans effet sur la fitness ne sont pas dupliqués (Wagner, 2000; Gu *et al.*, 2003). La robustesse distribuée est vraisemblablement responsable dans ces cas.

2.2 Origine évolutive de la robustesse : les difficultés expérimentales

Quelle est l'origine évolutive des propriétés de robustesse mutationnelle d'un trait phénotypique ? Ont-elles été indirectement sélectionnées en raison de l'avantage qu'elles peuvent conférer à long terme (scénario adaptatif), ou bien ne sont-elles qu'un effet secondaire de la sélection d'une autre propriété, corrélée à la robustesse (scénario intrinsèque) (De Visser *et al.*, 2003) ? Par exemple, l'action des protéines chaperonnes masque l'effet de certaines mutations, mais cela n'est peut-être qu'une conséquence fortuite de leurs autres rôles essentiels à la survie de la cellule (De Visser *et al.*, 2003). De même, la forte proportion de mutations neutres dans un réseau de régulation n'est peut-être pas due à la sélection d'une topologie spécifiquement robuste, mais au simple fait qu'une organisation en réseau, quelle qu'elle soit, est déjà intrinsèquement robuste (voir à ce sujet les travaux de Wagner (1996) et Von Dassow *et al.* (2000)).

La robustesse pourrait donc être une propriété intrinsèque de certains traits phénotypiques. Ainsi, montrer qu'un trait est robuste ne suffit pas pour affirmer que cette robustesse est une propriété sélectionnée. Il faut, entre autres, montrer que le même caractère peut aussi être présent sous des formes moins robustes (Wagner, 2005a). Un tel polymorphisme du niveau de robustesse est en effet une condition nécessaire à l'évolution de la robustesse par sélection indirecte (figure I.3).

Dès les années 50, Waddington (1957) a mis au point des expériences visant à tester le scénario adaptatif. Ces expériences et celles qui en furent inspirées (voir la revue de Scharloo (1991)) montrent que des organismes portant une mutation majeure ou ayant été exposés à un fort stress environnemental présentent une plus grande diversité phénotypique que le type sauvage. Une grande partie de cette diversité peut être artificiellement sélectionnée et est donc d'origine génétique. En effectuant divers croisements, on peut montrer que cette variation génétique était déjà présente dans le type sauvage, mais qu'elle ne s'exprimait pas. Ces expériences sont généralement interprétées de la façon suivante. Dans la souche sauvage, le développement du phénotype est génétiquement "canalisé", c'est-à-dire insensible aux mutations. Des mutations peuvent alors s'accumuler dans le génotype sans modifier le phénotype : elles sont maintenues à l'état neutre (ou "cryptique") par le ou les mécanismes de canalisation (Rutherford et Lindquist, 1998; Cossins, 1998). Le stress extrême ou la mutation majeure détruisent ces mécanismes, et les mutations accumulées

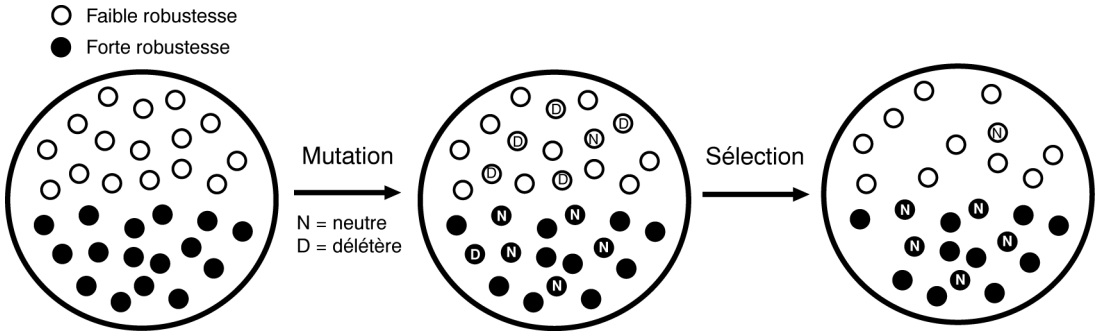


Fig. I.3: Evolution de la robustesse par sélection indirecte, d'après (Wagner, 2005a). Le grand cercle représente une population et les petits cercles les organismes. On considère ici deux types d'organismes identiques en tous points, excepté leur niveau de robustesse : les deux types ont la même fitness, donc la sélection ne peut agir directement. Mais comme les mutations sont majoritairement délétères dans le type variable alors qu'elles sont majoritairement neutres dans le type robuste, le type robuste est finalement le plus représenté. La robustesse est donc indirectement sélectionnée. Notons que le scénario considéré par Wagner (2005a) suppose que les mutations non neutres sont toutes délétères. Si au contraire des mutations favorables sont possibles, elles seront plus probables chez le type variable et pourront éventuellement le sauver.

s'expriment soudainement au niveau du phénotype. Le phénotype serait donc moins robuste dans la souche modifiée que dans la souche sauvage, ce qui démontrerait l'existence du polymorphisme de la robustesse, condition nécessaire du scénario adaptatif.

Cette interprétation n'est cependant pas la seule possible (Stearns, 2002). Par exemple, la révélation de variations génétiques peut aussi être due à la présence de loci conditionnellement neutres, c'est-à-dire neutres seulement dans un certain contexte génétique ou environnemental (Hermisson et Wagner, 2004). Dans ce cas, les loci neutres ne sont pas les mêmes dans la souche sauvage et dans la souche modifiée, mais leur nombre est identique : l'effet moyen des mutations reste le même. Ainsi, le changement de contexte peut révéler des variations génétiques même si la souche sauvage n'est globalement pas plus robuste que la souche modifiée. Un autre facteur susceptible d'influencer la variance mutationnelle d'un trait est le nombre de gènes impliqués dans ce trait. Houle (1998) et Hermisson et Wagner (2005) ont ainsi souligné qu'une plus faible variance mutationnelle peut ne pas être due à un mécanisme de canalisation des mutations mais simplement découler d'une "cible mutationnelle" plus petite (moins de loci impliqués). Autrement dit, les expériences inspirées des travaux de Waddington ne démontrent pas nécessairement l'existence de formes plus ou moins canalisées d'un même phénotype.

Une autre approche expérimentale, plus directe, consiste à introduire les mêmes mutations dans deux souches différentes. On peut alors comparer leurs effets dans les deux contextes génétiques. Elena et Lenski (2001) ont ainsi comparé les effets de 12 insertions d'éléments transposables dans deux souches d'*Escherichia coli*, l'une dite ancestrale et l'autre obtenue après 10000 générations d'évolution *in vitro* en milieu minimal. Cette dernière souche est supposée adaptée à l'environnement de laboratoire. En d'autres termes, on suppose que les mutations dans cette souche sont, dans leur majorité, soit neutres soit délétères. Acquérir

un mécanisme de canalisation semble alors avantageux. Pourtant, sur les 12 mutations, seules 3 ont eu un effet significativement différent dans les deux souches, et sur ces 3 mutations, 2 ont eu un effet plus grand dans la souche adaptée. L'expérience réalisée n'a donc pas pu confirmer que la souche adaptée était plus robuste que la souche ancestrale – mais cela tient peut-être au faible nombre de mutations testées.

Toujours est-il qu'à l'heure actuelle, les études expérimentales réalisées ne permettent pas d'affirmer avec certitude que la robustesse des phénotypes a été sélectionnée ou peut l'être en laboratoire. Pourtant, il semble intuitivement avantageux de réduire la sensibilité du phénotype aux mutations lorsque celles-ci sont fréquentes (fort taux de mutation) et quasi-systématiquement délétères (sélection stabilisatrice). C'est par des approches de modélisation et de simulation que cette intuition a pu être confirmée. Par exemple, dans l'étude pionnière de Wagner (1996) sur l'évolution *in silico* de réseaux de gènes, une sélection stabilisatrice sur l'état d'activation des gènes conduit à des réseaux de plus en plus robustes aux mutations : au cours des quelques centaines de générations d'évolution simulées, la proportion de mutations neutres passe de 60% (robustesse intrinsèque) à environ 90% (robustesse adaptative). Dans les paragraphes suivants, nous montrons comment le scénario adaptatif a été testé dans d'autres modèles issus de traditions différentes, de la génétique quantitative à la vie artificielle en passant par la théorie des "quasi-species".

2.3 Modèles de génétique quantitative : propagation d'un mécanisme de canalisation des mutations délétères

Les modèles de génétique quantitative visent à étudier l'évolution d'un trait phénotypique quantitatif déterminé par plusieurs gènes, comme la taille d'un animal ou le nombre d'épis sur un plant de maïs. Wagner *et al.* (1997) ont utilisé ce cadre théorique pour calculer le coefficient de sélection d'un allèle "canalisateur", qui réduit l'effet des variations des autres gènes sur le trait considéré. Les auteurs considèrent une population d'organismes diploïdes, sexués, qui s'apparient au hasard. Dans leur modèle, le trait sous sélection est noté X. L'action combinée de n gènes détermine la valeur d'une variable physiologique intermédiaire $Y = \sum_{i=1}^{n} y_i$. Les mutations font directement varier les y_i et donc Y. Un locus supplémentaire, dit "modifieur" ou "canalisateur", détermine la sensibilité de X vis-à-vis des variations de Y : on pose $X = cY$, où $c \in]0, 1]$ dépend des allèles présents au locus canalisateur. Si une mutation dans l'un des gènes provoque une variation ΔY, alors la variation de X vaudra $c\Delta Y$. On suppose que deux allèles B et b sont possibles. L'allèle B n'a pas d'effet canalisateur ($c_{BB} = 1$), tandis que l'allèle b réduit la sensibilité de X vis-à-vis des variations de Y ($0 < c_{bb} \leq c_{bB} < 1$). Il est important de noter que dans ce modèle, les mutations ont forcément un effet phénotypique, même s'il peut être modulé.

Modéliser la robustesse avec ce coefficient de proportionnalité est pratique d'un point de vue analytique, mais conceptuellement problématique. En effet, le locus canalisateur a un effet systématique sur la valeur du trait et peut donc être sélectionné (ou contre-sélectionné) en raison de cet effet direct et non de son effet sur la robustesse. Pour contourner ce problème, les auteurs fixent la valeur optimale de X à 0 : toute déviation de X

par rapport à 0 est délétère, la sélection est stabilisatrice. Ainsi, on peut avoir dans la population un polymorphisme de la robustesse (différentes valeurs de c) à fitness égale (si $Y = 0$, c n'a pas d'effet sur la valeur de X). Pour étudier l'évolution de la robustesse, on s'intéresse alors à la propagation de l'allèle canalisateur (b) dans la population.

L'étude analytique et les simulations réalisées par les auteurs montrent que dans ce contexte, le coefficient de sélection de l'allèle canalisateur est positif : les individus qui le possèdent ont en moyenne une plus grande fitness que ceux qui ne le possèdent pas. Ainsi, tandis que la valeur moyenne de X dans la population reste à 0 au cours du temps, la valeur moyenne de c augmente au fur et à mesure que l'allèle canalisateur se propage.

Le coefficient de sélection de l'allèle canalisateur dépend de son efficacité, de l'intensité de la sélection stabilisatrice, mais aussi de la variation génétique présente dans la population. Tous les facteurs qui augmentent cette variation génétique vont favoriser la propagation de l'allèle canalisateur : grande taille de la population, fort taux de mutation par locus, grand nombre de gènes contribuant au trait phénotypique. L'influence de l'intensité de la sélection stabilisatrice est plus complexe : une sélection plus intense, correspondant à des pertes de fitness plus lourdes lorsque le trait varie, augmente directement l'avantage procuré par l'allèle canalisateur, mais réduit en même temps la variation génétique présente dans la population. Lorsque la sélection reste faible, le premier effet prédomine. Mais lorsqu'elle est très intense, les deux effets se compensent et le coefficient de sélection du canalisateur est finalement indépendant de la force de la sélection stabilisatrice. Cela signifie qu'un mécanisme de robustesse ne se propagera pas forcément plus rapidement si le trait est plus contraint.

Il faut noter qu'une sélection stabilisatrice n'est pas le seul régime capable d'induire l'évolution de la robustesse. En reprenant le modèle de Wagner *et al.* (1997), Kawecki (2000) a montré qu'une sélection directionnelle[1] avec un optimum fluctuant fréquemment peut aussi favoriser la propagation du canalisateur. En effet, lorsque la valeur optimale du trait varie à chaque génération autour d'une valeur moyenne, il est plus avantageux à long terme de rester sur la valeur moyenne. Or le canalisateur empêche les mutations de faire varier le trait. Ainsi, si les variations environnementales sont suffisamment fréquentes, il peut aller jusqu'à la fixation. Comme précédemment, un fort taux de mutation accélère le phénomène. Des fluctuations environnementales très rapides autour d'une moyenne fixe produisent ainsi les mêmes effets qu'une sélection stabilisatrice.

Cependant, dans ce modèle, l'existence d'un seul "gène-maître" modifiant l'effet des mutations à tous les loci est peu réaliste, car la robustesse d'un trait quantitatif est généralement multifactorielle (Scharloo, 1991). Wagner *et al.* (1997) ont donc proposé un modèle plus complexe, où chaque gène peut moduler l'effet des mutations à d'autres loci, en plus de contribuer directement au trait sous sélection. Une mutation peut alors faire varier à la fois la valeur du trait et l'effet de certaines mutations ultérieures. Les simulations réalisées montrent que sous une sélection stabilisatrice, l'effet moyen des mutations tend à diminuer au fur et à mesure de l'évolution de la population. Là encore, cette augmentation de la

[1] Une sélection directionnelle signifie que le caractère peut être amélioré par sélection naturelle et que les mutations peuvent générer des variants de plus grande fitness (Wagner, 2005a).

robustesse est d'autant plus rapide que le taux de mutation est élevé. Ainsi, l'existence d'un "gène-maître" n'est pas nécessaire pour l'évolution de la robustesse. Celle-ci peut résulter d'interactions entre petits groupes de gènes.

2.4 Modèles de "quasi-species" : évolution de la proportion de mutations neutres

Il est également possible d'étudier l'évolution de la robustesse à l'aide des modèles de "quasi-species". La théorie des quasi-species décrit l'évolution d'une population supposée infinie d'autoréplicateurs soumis à de forts taux de mutation (Eigen *et al.*, 1988). Elle est souvent utilisée pour décrire l'évolution de macromolécules d'ARN, de virus à ARN, ou d'organismes haploïdes asexués. Bien qu'ayant été développée indépendamment de la génétique des populations classique, elle lui est mathématiquement équivalente lorsque la population est infinie, haploïde et asexuée (Wilke, 2005). Cependant, les deux écoles se sont focalisées sur des cas particuliers différents. Alors que la génétique des populations s'est longtemps limitée à des modèles à un ou deux loci, tout en intégrant les effets stochastiques liés à la taille finie des populations, la théorie des quasi-species traite d'emblée un nombre arbitraire de loci, mais souvent dans le cadre déterministe d'une population infinie.

Le modèle des quasi-species considère que les entités autoréplicantes peuvent être représentées par des séquences constituées d'un petit nombre de symboles (A, C, G, U par exemple). De nouvelles séquences ne peuvent entrer dans le système que par réplication, correcte ou erronée, de séquences déjà présentes. La sélection est due au fait que les séquences différentes peuvent avoir des taux de réplication (fitness) différents. Si on a S séquences possibles, la dynamique du système en temps discret est décrite par S équations de la forme :

$$x_i(t+1) = \frac{1}{E(t)} \sum_{j=1}^{S} W_j Q_{ij} x_j(t)$$

où $x_i(t)$ est la concentration de la séquence i, W_i est son taux de réplication (fitness), Q_{ij} est la probabilité de muter de la séquence j vers la séquence i, et $E(t) = \sum_k \sum_j W_j Q_{kj} x_j(t)$ est la production totale de nouvelles séquences (Demetrius, 1983). Dans le cas des séquences d'ARN, la fitness W_i est souvent attribuée en fonction de la structure secondaire de la molécule.

Comment étudier l'évolution de la robustesse avec un tel modèle, qui n'inclut pas explicitement de locus canalisateur ? Considérons le voisinage mutationnel d'une séquence i. Il est constitué des séquences j atteignables après une réplication, qui sont telles que $Q_{ij} \neq 0$. Par exemple, lorsque le taux de mutation est faible, il s'agira des séquences qui ne diffèrent de i qu'en une position. L'effet des mutations dépend des fitness W_j des séquences appartenant à ce voisinage mutationnel. Si les W_j sont identiques à la fitness de départ (W_i), la fitness de la séquence i est robuste aux mutations.

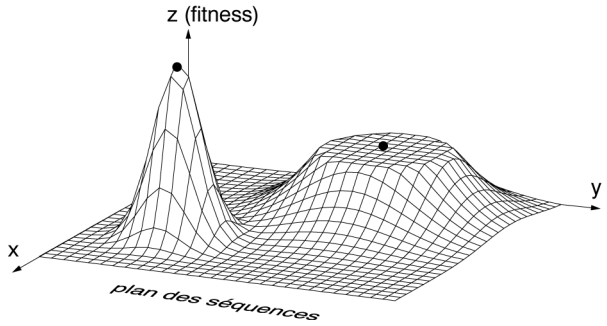


Fig. I.4: Métaphore du paysage adaptatif. Dans cet exemple simple, le pic de gauche correspond à une séquence dont la fitness est très élevée mais peu robuste aux mutations. A l'inverse, au centre du plateau de droite se trouve une séquence dont la fitness est moyenne mais très robuste aux mutations.

Pour illustrer cela, on peut utiliser la métaphore du "paysage adaptatif" (fitness landscape). Telle qu'elle fut introduite par Wright (1932), il s'agit d'une illustration informelle et non d'une représentation mathématique rigoureuse (Skipper, 2004). Chaque séquence i est vue comme un point dans un plan (x, y). Sur ce plan, la distance entre deux séquences dépend du nombre de mutations qui les séparent. On ajoute ensuite une troisième dimension z en associant à chaque séquence sa fitness W_i (figure I.4). Les pics du paysage ainsi obtenu correspondent aux séquences de fitness élevée. Ces pics sont plus ou moins pointus, en fonction de la fitness des séquences voisines. Au centre des plateaux, on trouve donc les séquences dont la fitness est robuste : de petits déplacements dans le plan (x, y) ne provoquent pas de variation sur l'axe z, ce qui signifie que les mutations sont neutres.

Ainsi, dans les modèles basés sur la théorie des quasi-species, la robustesse n'est pas "codée" explicitement dans un locus supplémentaire canalisateur. Mais comme les mutations s'appliquent sur la séquence et non directement sur la valeur du trait sous sélection, certaines mutations peuvent être neutres. Certains génotypes sont alors intrinsèquement plus robustes que d'autres. Etudier l'évolution de la robustesse revient alors à caractériser la répartition de la population sur les S séquences : va-t-elle se concentrer sur les séquences où la majorité des mutations sont neutres ?

Dans le cas où le paysage adaptatif présente deux pics de même hauteur, les simulations réalisées par Schuster et Swetina (1988) montrent que la population se concentre effectivement sur le pic le plus plat, c'est-à-dire le plus robuste aux mutations. La même observation a été faite avec un paysage plus complexe, mais toujours fixe dans le temps. La population se concentre sur les séquences qui non seulement ont elles-mêmes une forte fitness, mais qui ont aussi dans leur voisinage mutationnel des séquences de forte fitness (Huynen et Hogeweg, 1994).

Van Nimwegen *et al.* (1999) ont étudié analytiquement ce phénomène en considérant un paysage très simple, constitué d'un seul plateau appelé "neutral network" (réseau de séquences connectées par des mutations neutres (Schuster *et al.*, 1994)). Toutes les séquences en dehors de ce plateau ont une fitness négligeable. Les séquences sont de taille fixe L et le taux de mutation par position est supposé suffisamment faible pour qu'à

chaque réplication, une seule mutation ponctuelle puisse se produire au maximum. Pour les séquences de forte fitness, certaines mutations sont neutres et permettent de rester sur le plateau, alors que d'autres mutations peuvent faire tomber le descendant du plateau. La proportion relative des deux types de mutations est variable selon les séquences. Les séquences dont la fitness est la plus robuste se trouvent au centre du plateau. La fitness des séquences qui se trouvent en périphérie du plateau est plus fragile.

L'étude analytique de ce système démontre qu'à l'équilibre, une population infinie n'est pas répartie uniformément sur le plateau. Au contraire, elle est concentrée sur les séquences dont la fitness est robuste aux mutations. Cela signifie que l'évolution d'une population sur un plateau de fitness n'est pas une marche aléatoire. Dans cette configuration comparable à une sélection stabilisatrice, on a évolution de la robustesse. Par des simulations stochastiques, les auteurs montrent que ce résultat peut également s'appliquer à une population de taille finie, à condition que $NU >> 1$, où N est la taille de la population et U le taux de mutation par réplication. Dans la composition de la population à l'équilibre, la fréquence des séquences les plus robustes augmente alors avec le produit NU. Inversement, si $NU << 1$, toute la population se trouve la plupart du temps concentrée sur une même séquence. Il n'y a plus de polymorphisme du niveau de robustesse et donc plus de sélection de la robustesse.

L'analyse mathématique de ce système a été généralisée par Wilke (2001a) pour un taux de mutation quelconque, autorisant ainsi plusieurs mutations par réplication. La population converge alors vers les génotypes qui maximisent la probabilité qu'un descendant conserve la fitness de son progéniteur. Cette probabilité est appelée "fraction of neutral offspring", c'est-à-dire proportion de descendants neutres. Pour une séquence donnée i, elle s'écrit ici

$$F_\nu = (1-u)^L \left(1 + \sum_{k=1}^{L} \left(\frac{u}{(1-u)(A-1)} \right)^k d_i^{(k)} \right) \tag{I.1}$$

où u est le taux de mutation par position, L la taille de la séquence, A le nombre de symboles dans l'alphabet, et $d_i^{(k)}$ le nombre de génotypes distants de k mutations et ayant la même fitness que la séquence i. Lorsque le taux de mutation est faible, seuls les $d_i^{(1)}$ voisins immédiats (distants de 1 mutation) influencent cette proportion. On retrouve alors un critère déjà obtenu par Van Nimwegen *et al.* (1999) : la population converge vers les séquences qui maximisent $d^{(1)}$. Si au contraire le taux de mutation est fort, ces voisins immédiats perdent de leur importance, car il est peu probable que la réplication les produise. Dans ce cas, la réplication produit des génotypes distants de plusieurs mutations, et c'est la fitness de ces génotypes distants qui compte pour la survie du descendant.

Dans le cas général, nous préfèrerons tenir compte du taux de mutations, et donc utiliser la proportion F_ν de descendants neutres comme mesure de la robustesse d'un génotype, plutôt que $d_i^{(1)}$. Cependant, si la formule de F_ν donnée ci-dessus présente l'avantage d'être exacte, elle n'est que très rarement utilisable en pratique, car il faut connaître la fitness de toutes les séquences possibles pour calculer les $d_i^{(k)}$. Nous utiliserons au chapitre 3 une formulation approchée de F_ν, nécessitant moins d'informations.

2.5 Modèles de vie artificielle : survie du plus apte ou survie du plus robuste ?

Les résultats précédents montrent qu'à fitness égale, la population se concentre sur les zones les plus plates du paysage adaptatif, c'est-à-dire sur les zones où la fitness est robuste aux mutations. Qu'en est-il lorsque le paysage comporte plusieurs pics de largeurs *et de hauteurs* différentes, comme sur la figure I.4 ? Lorsque la robustesse a un coût, la population se concentre-t-elle sur les séquences où la fitness est la plus élevée, ou sur celles où la fitness est la plus robuste ?

Ce problème a été abordé pour la première fois à l'aide d'une plate-forme de vie artificielle appelée *Avida* (Adami, 2006). Cette plate-forme simule l'évolution d'une population d'organismes artificiels haploïdes et asexués. Ces "organismes" sont en fait des programmes informatiques : leurs génotypes sont des séquences d'instructions (comparables à des instructions assembleur) et leurs phénotypes sont obtenus par l'exécution de ces instructions. Certaines instructions permettent de réaliser des opérations logiques récompensées par de l'énergie. D'autres permettent à l'organisme de se reproduire en recopiant son code, avec éventuellement des mutations : des instructions peuvent être ajoutées, enlevées, ou encore remplacées par d'autres choisies aléatoirement. La vitesse de l'exécution du code, et donc la vitesse de réplication (fitness) de l'organisme, dépendent de l'énergie acquise.

Wilke *et al.* (2001) ont utilisé cette plate-forme pour obtenir deux types d'organismes : les organismes de type A ont une fitness élevée mais sont peu robustes aux mutations, alors que les organismes de type B ont une fitness moindre mais sont plus robustes. Pour obtenir un organisme robuste (B), les auteurs ont fait évoluer une population sous un fort taux de mutation ($U_B = 2$ mutations par génome répliqué) pendant 1000 générations. A l'inverse, un taux de mutation plus faible ($U_A = 0{,}5$) est utilisé pour obtenir un organisme moins robuste, mais dont la fitness peut être de 1,5 à 12 fois plus forte.

Après avoir vérifié que les mutations ont effectivement moins d'effet dans le type B, les auteurs ont mis les deux types en compétition : une population mixte, initialisée avec 50% de type A et de 50% de type B, évolue pendant 50 générations sous un taux de mutation U donné, allant de 0,5 à 3 mutations par génome répliqué. Ces expériences de compétition ont été répétées pour 12 paires A et B différentes. Dans tous les cas, le type A domine la population si le taux de mutation est faible. Au contraire, c'est le type B qui l'emporte si le taux de mutation est fort, même s'il se reproduit 12 fois moins vite. En effet, lorsque le taux de mutation est fort, le type A produit de nombreux descendants, mais la plupart d'entre eux portent des mutations très délétères. Le type B produit moins de descendants, mais ceux-ci portent des mutations neutres et vont donc pouvoir se reproduire. Le type B peut donc être finalement le plus représenté (figure I.5). Un taux critique sépare donc ici deux modes de sélection : la survie du plus apte lorsque le taux de mutation est faible, et la survie du plus robuste lorsque le taux de mutation est fort. Le taux de mutation critique observé dans l'étude de Wilke *et al.* (2001) est de l'ordre de 1 mutation par génome répliqué, un taux rencontré chez les virus à ARN et chez plusieurs eucaryotes (Drake, 1993; Drake *et al.*, 1998; Eyre-Walker et Keightley, 1999; Lynch *et al.*, 1999).

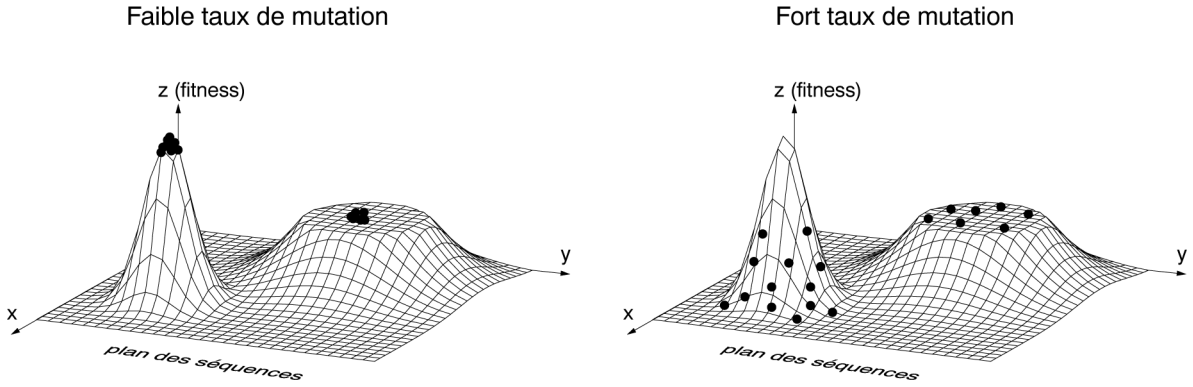


Fig. I.5: Évolution de la robustesse et taux de mutation. Considérons les descendants de deux individus différents, l'un (au sommet du pic) qui présente une fitness élevée mais sensible aux mutations, et l'autre (au centre du plateau) qui présente une fitness plus faible mais plus robuste. Lorsque le taux de mutation est faible, la grande majorité des descendants sont identiques à leur progéniteur, et le type de fitness élevée finira par gagner la compétition. Lorsqu'au contraire le taux de mutation est très élevé, une grande partie des descendants d'un individu porte des mutations. Si celles-ci sont délétères, ils ne se reproduiront pas ou peu, et la descendance effective de l'individu est de fait très réduite. Si au contraire ces mutations sont neutres, tous les descendants, même les mutants, pourront se reproduire. À terme, c'est le type le plus robuste qui domine la population, même s'il présente un taux brut de reproduction plus faible.

Notons qu'à taux de mutation égal, c'est la taille de la population qui départage ces deux modes. Krakauer et Plotkin (2002) ont en effet montré, à l'aide d'un modèle de quasi-species à taille de population finie, que les grandes populations se concentrent sur une séquence optimale mais sensible aux mutations, alors que les petites populations restent sur des séquences suboptimales mais robustes.

Selon Wilke (2001b), ces deux modes – la survie du plus apte ou la survie du plus robuste – sont en fait les deux facettes d'un même critère de sélection : dans les deux cas, il s'agit de maximiser le nombre de descendants *phénotypiquement identiques* au progéniteur, c'est-à-dire, avec les notations précédentes (page 45), le produit $F_\nu W$. Il faut toutefois noter que ce critère correspond à une situation où les mutations ne peuvent pas être avantageuses. En effet, dans l'expérience de compétition de Wilke *et al.* (2001), les mutations normalement avantageuses étaient ignorées dans le calcul de la fitness. Plus généralement, la plupart des résultats concernant l'évolution de la robustesse ont été obtenus en considérant les mutations avantageuses impossibles. La généralisation de ces résultats lorsque la sélection est directionnelle plutôt que stabilisatrice n'est donc pas assurée.

2.6 Evolution de l'anti-robustesse

Lorsqu'un caractère présente une forte robustesse mutationnelle, une grande partie des variations génétiques présentes dans la population est invisible pour la sélection. À court terme au moins, cela réduit la capacité du caractère à répondre à une sélection direction-

nelle (Stearns, 1993). La robustesse pourrait donc être contre-sélectionnée lorsque l'état du caractère n'est pas optimal (Layzer, 1980). Ainsi, Kawecki (2000) a montré, à l'aide du modèle de génétique quantitative de Wagner *et al.* (1997), qu'un allèle canalisateur est contre-sélectionné dès lors qu'une sélection directionnelle est maintenue pendant plusieurs générations.

Une certaine sensibilité du caractère aux mutations est en effet nécessaire, au moins transitoirement, pour passer d'un état du caractère à un autre (Fontana et Schuster, 1998; Ancel et Fontana, 2000; Ancel Meyers *et al.*, 2005). On peut reprendre la métaphore du paysage adaptatif pour illustrer cela. Imaginons que la population se trouve initialement au centre d'un plateau, c'est-à-dire sur une séquence dont la fitness est robuste. À proximité de ce plateau se trouve un autre plateau plus élevé. Seules les séquences situées au bord du premier plateau permettent d'accéder au second. Par des mutations neutres, une partie de la population peut s'éloigner du centre et atteindre ces séquences peu robustes. Ces séquences sont dites à fort potentiel génétique (Ancel Meyers *et al.*, 2005), car les mutations qui s'y produisent peuvent permettre l'accès à une meilleure fitness[1].

Cette "anti-robustesse", ou cette sensibilité aux mutations, peut-elle être une condition stable ? Si le plateau nouvellement atteint est le plus élevé de tout le paysage, ou si les plateaux plus élevés sont inatteignables, la sélection redevient stabilisatrice, et la population converge à nouveau vers des séquences où la fitness est robuste. Dans les modèles où le génotype est de longueur fixée, comme c'est le cas pour le modèle des quasi-species, le nombre de génotypes est fini et l'un ou l'autre de ces deux cas finit par se produire (si le paysage adaptatif est fixe au cours du temps). Mais les mutations favorables sont-elles impossibles en réalité ? Les possibilités génétiques sont bien plus vastes et la fitness pourrait être en permanence sous sélection directionnelle, même si l'environnement est relativement stable. Si de plus l'environnement varie au cours de l'évolution, le potentiel génétique, c'est-à-dire l'existence de mutations non neutres, est d'autant plus nécessaire.

Les travaux théoriques de Huynen et Hogeweg (1994) et d'Ancel Meyers *et al.* (2005) confirment qu'une population évoluant dans un environnement changeant tend à se concentrer sur les séquences où les mutations ont un grand effet sur le phénotype. Les indices empiriques qui vont dans ce sens sont pour l'instant limités. On peut citer l'usage du code génétique chez deux bactéries pathogènes pour l'Homme, *Mycobacterium tuberculosis* et *Plasmodium falciparum* (Plotkin *et al.*, 2004). Leurs protéines de surface, qui sont des cibles pour le système immunitaire, sont codées par des codons peu robustes aux mutations ponctuelles (c'est-à-dire des codons pour lesquels une mutation ponctuelle a de fortes chances de changer l'acide aminé). Parallèlement, Tan *et al.* (2004) ont pris en compte à la fois l'usage du code génétique et les biais mutationnels et ont montré que chez les anticorps murins, les mutations se produisent plus fréquemment au niveau des acides aminés polaires. Selon les auteurs, cela favorise la variation de l'affinité des anticorps pour les antigènes.

[1] Notons que la robustesse des séquences initiales permet d'atteindre ces séquences facilement, sans perte de fitness : les mutations neutres mettent en place le contexte nécessaire aux mutations favorables. En ce sens, la présence de neutralité dans le paysage adaptatif n'est donc pas nécessairement antagoniste à l'innovation (Huynen, 1996; Huynen *et al.*, 1996; Fontana et Schuster, 1998; Wagner, 2005b).

À ce stade de notre passage en revue des différentes composantes de la variabilité mutationnelle, il ressort qu'à bien des égards, l'évolution de l'effet des mutations géniques et l'évolution du taux de mutation sont des problématiques sœurs. Dans les deux cas, les approches de modélisation ont permis de mettre en évidence la nature indirecte de la sélection mise en jeu, en restant à chaque fois partagées sur la direction de cette sélection – favorise-t-elle la stabilité ou la variabilité ? La réponse semble dépendre, dans les deux cas, d'un paramètre difficile à estimer : la fréquence des mutations favorables.

3 Rôle de la structure du génome

La grande majorité des modèles présentés jusqu'à présent se focalisent sur deux composantes de la variabilité mutationnelle du phénotype : le taux de mutation et l'effet d'une mutation touchant un locus. Le nombre de gènes touchés par une mutation constitue cependant une troisième composante tout aussi importante, car ce nombre n'est pas toujours égal à 1. Une mutation peut ne toucher aucun gène ou au contraire en affecter plusieurs simultanément. Cela dépend à la fois des mécanismes mutationnels mis en jeu (mutations ponctuelles, grands réarrangements, ...) et de la structure du génome (quantité d'ADN non codant, gènes chevauchants, densité et répartition des éléments répétés, ...). La structure du génome joue donc aussi un rôle dans l'effet phénotypique moyen des mutations, en influençant le nombre de gènes qu'elles affectent simultanément.

Comme nous l'avons vu dans les sections précédentes, le taux de mutation et les mécanismes de canalisation peuvent faire l'objet d'une sélection indirecte. Est-ce aussi le cas pour la structure du génome ? En d'autres termes, une structure de génome permettant une variabilité optimale peut-elle être indirectement sélectionnée ? Les travaux qui ont abordé cette question sont rares, probablement car (i) le mécanisme de sélection indirecte est plus spontanément associé à l'évolution du taux de mutation ou de la canalisation (voir sections 1 et 2), et (ii) la structure des génomes est souvent présentée comme le résultat de biais mutationnels et/ou de pressions sélectives directes. Par exemple, la quantité d'ADN non fonctionnel maintenue dans un génome est en général expliquée par un équilibre entre les différents mécanismes d'insertions et de délétions (petites insertions et délétions, délétions larges par recombinaison, prolifération des éléments transposables, rétroposition des ARN messagers...), auquel vient éventuellement s'ajouter un coût sélectif direct à la taille du génome (volume occupé par l'ADN, vitesse et coût énergétique de la réplication...).

Dans cette section, nous commencerons par montrer que les mutations ne touchent pas toujours exactement un gène, qu'il s'agisse de mutations locales ou de réarrangements génomiques. Nous présenterons ensuite trois caractéristiques structurelles du génome susceptibles d'influencer le nombre de gènes touchés par une mutation : la présence de gènes chevauchants, la quantité d'ADN non fonctionnel et la répartition relative des gènes et des éléments répétés. Nous détaillerons les pressions évolutives les plus fréquemment évoquées pour expliquer ces caractéristiques, et nous verrons que la sélection indirecte d'un certain niveau de variabilité n'en fait généralement pas partie. Nous nous tournerons alors vers

les approches de modélisation mettant en jeu une structure de génome flexible, avec un nombre variable de gènes ou des segments non codants par exemple. Nous verrons qu'en évolution artificielle, cette flexibilité du génome est malheureusement associée à des mécanismes de mutation ou à des transitions génotype-phénotype peu réalistes, causant des comportements artefactuels (relativement à notre problématique). Nous montrerons ensuite que les modèles développés en biologie pour rendre compte de l'évolution du nombre de gènes, du nombre d'éléments transposables ou de la taille du génome sont radicalement différents, dans la mesure où la séquence génomique et/ou le niveau populationnel ne sont pas explicitement inclus. Il s'agit plutôt de modèles mathématiques qui, en général, ne sont pas adaptés à l'étude de pressions sélectives indirectes. Il nous faudra donc développer un nouveau type de modèle, combinant les atouts des différentes approches.

3.1 Nombre de gènes touchés par une mutation

Une mutation locale (c'est-à-dire une mutation ponctuelle ou une insertion/délétion de quelques bases) ne touche pas nécessairement exactement un gène. Elle peut en toucher plusieurs simultanément – cela dépend de la présence de gènes chevauchants –, ou au contraire n'en affecter aucun – cela dépend de la quantité d'ADN non fonctionnel contenue dans le génome.

Le nombre de gènes simultanément touchés est encore plus variable si l'on considère les réarrangements chromosomiques (au sens large : duplications, délétions, inversions, translocations). Bien qu'ils soient difficiles à prendre en compte dans la plupart des modèles décrits ci-avant, on sait depuis longtemps qu'ils sont fréquents (Starlinger, 1977) et qu'ils jouent un rôle évolutif important. La duplication de segments chromosomiques, voire même du génome complet, est un mode très répandu d'acquisition de gènes (Ohno, 1970; Venter *et al.*, 2001; Betran et Long, 2002; Eichler et Sankoff, 2003; Gevers *et al.*, 2004; Cannon *et al.*, 2004; Dujon *et al.*, 2004). Selon Teichmann et Babu (2004), au moins 50% des gènes procaryotes et plus de 90% des gènes eucaryotes proviennent de duplications de gènes. Inversement, les grandes délétions peuvent conduire à la perte simultanée de plusieurs gènes. Des expériences d'évolution expérimentale menées sur la bactérie *Salmonella enterica* ont montré que de tels événements se produisent fréquemment et qu'ils peuvent toucher plusieurs dizaines de gènes à la fois (Nilsson *et al.*, 2005; Ochman, 2005). De telles délétions ont par exemple pu jouer un rôle crucial dans la réduction des génomes des bactéries endosymbiotiques (Ochman, 2005). La comparaison des génomes d'espèces plus ou moins proches, ou de différentes souches d'une même espèce, fournit un autre faisceau d'indices de la fréquence et de la diversité des réarrangements. Des délétions de plusieurs gènes ont ainsi été fixées dans certains isolats naturels d'*Escherichia coli* (Ochman et Jones, 2000). De même, plusieurs grandes délétions ont été identifiées dans différentes souches de *Mycobacterium tuberculosis* (Fang *et al.*, 1999; Kato-Maeda *et al.*, 2001). De nombreuses inversions et translocations ont également été mises en évidence par des comparaisons inter- et intra-spécifiques de génomes bactériens (Hughes, 2000). Ces événements modifient l'ordre des gènes : chez les bactéries, l'ordre des gènes est de moins en moins conservé lorsque les espèces sont phylogénétiquement distantes (Tamames,

2001). Les réarrangements sont également communs dans les génomes eucaryotes, où ils peuvent parfois affecter des centaines de gènes (Eichler et Sankoff, 2003; Sankoff, 2003; Coghlan *et al.*, 2005). La comparaison de différentes espèces de levures révèle ainsi des réorganisations génomiques massives (Dujon *et al.*, 2004). De même, le génome humain et celui du chimpanzé diffèrent par un certain nombre de grandes insertions, grandes délétions, translocations et duplications[1] (Wooding et Jorde, 2006). Le phénomène est encore plus frappant si l'on compare des espèces plus distantes, comme l'Homme et la Souris (International Human Genome Sequencing Consortium, 2001; Eichler et Sankoff, 2003).

Les réarrangements constituant une proportion non négligeable des variations génétiques, il est important de prendre en compte le nombre de gènes touchés lorsqu'on s'intéresse à l'effet phénotypique des mutations. Si un réseau de régulation ou un réseau métabolique peut résister à la perte d'un gène (voir paragraphe 2.1, p. 34), il n'est pas certain qu'il puisse aussi résister à la délétion d'une dizaine de gènes par exemple. En comparaison, les duplications sont souvent considérées comme relativement bénignes, mais elles peuvent tout de même conduire à un doublement des niveaux d'expression de tous les gènes dupliqués. À nouveau, le fonctionnement d'un réseau peut être robuste vis-à-vis d'une plus forte expression d'un seul gène, mais une variation simultanée de plusieurs gènes constitue une perturbation plus difficile à compenser. Plus précisément, Wagner (1994) a montré à l'aide d'un modèle mathématique que les duplications qui causent le moins de dommages dans un réseau de régulation sont soit les duplications d'un seul gène, soit les duplications du réseau complet. Au niveau du génome, deux types d'organisation seraient donc indirectement plus favorables pour les gènes du réseau : un espacement maximal (qui permet de dupliquer un seul gène à la fois) ou au contraire une agrégation maximale (qui permet de dupliquer tous les gènes simultanément) (Wagner, 1994).

Les inversions et les translocations peuvent sembler plus inoffensives, dans la mesure où elles se comportent en première approximation comme des paires ou des triplets de mutations locales, au niveau des points de rupture. Pour une inversion par exemple, seuls les gènes qui se trouveraient aux extrémités d'un segment inversé seraient touchés, les gènes complètement inclus dans le segment (régions régulatrices comprises) restant intacts. La réalité peut cependant être plus complexe. Dans les génomes bactériens par exemple, il semble qu'il existe des pressions sélectives directes sur l'orientation des gènes : les gènes essentiels ont tendance à se trouver sur le brin direct, ce qui éviterait les collisions frontales entre les machineries de réplication et de transcription (Rocha et Danchin, 2003a,b). L'inversion d'un gène peut augmenter la fréquence de ces collisions et ainsi causer des arrêts prématurés de la transcription. Selon ce point de vue, les effets d'une inversion ne sont pas nécessairement limités à ses points de rupture ; elle peut aussi affecter les produits géniques codés à l'intérieur du segment inversé.

En somme, les réarrangements génomiques affectent potentiellement les gènes situés au niveau des points de rupture ainsi que les gènes contenus dans le segment réarrangé, à des degrés divers selon le type de réarrangement. Certaines caractéristiques structurelles du génome vont donc influencer l'impact des réarrangements. Comme pour les mutations

[1] Le célèbre pourcentage d'identité de 98.9% entre les séquences humaines et les séquences de chimpanzé n'est en fait correct que pour les régions communes aux deux génomes (Wooding et Jorde, 2006).

ponctuelles, la quantité d'ADN non fonctionnel et la présence de gènes chevauchants vont influencer l'effet phénotypique au niveau des points de rupture. Le degré d'agrégation des gènes et la distribution des éléments répétés vis-à-vis des gènes vont aussi jouer un rôle important, en influençant le nombre moyen de gènes contenus dans un segment réarrangé.

3.2 Caractéristiques structurelles du génome influençant la variabilité mutationnelle

Les gènes chevauchants

Les gènes chevauchants sont très courants chez les virus à ADN ou à ARN (Barrell *et al.*, 1976; Normark *et al.*, 1983; Samuel, 1989; Lamb et Horvath, 1991; Pavesi *et al.*, 1997) et il en existe aussi de nombreux exemples chez les bactéries (Normark *et al.*, 1983; Rogozin *et al.*, 2002b; Fukuda *et al.*, 2003). Jusqu'à récemment, on pensait qu'ils étaient rares chez les eucaryotes, mais des études récentes ont montré que les génomes de vertébrés peuvent contenir des centaines voire des milliers de chevauchements (Veeramachaneni *et al.*, 2004; Makalowska *et al.*, 2005).

Plusieurs hypothèses ont été avancées pour expliquer l'apparition et l'évolution de ces chevauchements. Chez les virus, le génome doit tenir dans une petite capside protéique pour se transmettre, ce qui crée vraisemblablement une pression sélective directe en faveur d'une compression du génome (Normark *et al.*, 1983; Krakauer, 2000). Chez les bactéries, les chevauchements pourraient résulter d'un biais mutationnel en faveur des délétions, qui tend à réduire les séquences intergéniques (Clark *et al.*, 2001). Les chevauchements pourraient aussi avoir des fonctions régulatrices, comme un couplage traductionnel (Normark *et al.*, 1983). Cependant, comme les chevauchements tendent à augmenter l'effet des mutations ponctuelles, on peut également faire l'hypothèse que la sélection indirecte d'un certain niveau de variabilité peut agir sur leur fréquence dans le génome (Krakauer et Plotkin, 2002).

L'ADN non fonctionnel

La quantité d'ADN non fonctionnel contenue dans un génome est un paramètre difficile à estimer car, en toute rigueur, il est impossible de prouver l'absence de fonction. Elle est vraisemblablement liée à la quantité d'ADN intergénique, qui est plus facile à mesurer mais qui inclut aussi des séquences fonctionnelles, jouant par exemple un rôle dans la régulation de l'expression des gènes ou dans la conformation spatiale du chromosome (Gheldof et Dekker, 2004). La quantité d'ADN intergénique est très variable d'un domaine à l'autre, et au sein d'un même domaine, d'une espèce à l'autre. Chez les virus et les procaryotes, la distance intergénique moyenne varie de quelques bases à quelques centaines de bases (Mira *et al.*, 2001; Miller *et al.*, 2003; Giovannoni *et al.*, 2005; Claverie *et al.*, 2006). Chez les eucaryotes pluricellulaires, elle se compte plutôt en kilobases, avec par exemple environ

3 kb chez *Caenorhabditis elegans* et *Drosophila melanogaster* (Nelson *et al.*, 2004), contre environ 60 kb chez l'Homme (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001).

Des pressions sélectives directes limitant la taille du génome ont été évoquées pour expliquer la compacité des génomes viraux et procaryotes. Pour les virus, on retrouve la contrainte liée à la taille de la capside. Pour les cellules procaryotes, le coût sélectif d'un grand chromosome serait lié à la durée de sa réplication, qui limiterait la vitesse de reproduction (Maniloff, 1996; Silva *et al.*, 2001). Pourtant, dans les isolats naturels d'*Escherichia coli*, le taux de croissance ne dépend pas de la longueur du chromosome (Bergthorsson et Ochman, 1998) – il dépend surtout de l'efficacité de la traduction et de l'abondance des ARN de transfert (Mikkola et Kurland, 1991). De plus, la durée de la réplication ne limite pas nécessairement le taux de croissance, car plusieurs fourches de réplication peuvent être imbriquées et permettre ainsi d'initier la réplication plusieurs divisions cellulaires à l'avance (Lawrence *et al.*, 2001). Pour les génomes bactériens, l'hypothèse qui est semble-t-il en passe de supplanter l'idée du coût sélectif direct est celle d'un biais mutationnel : les petites délétions seraient spontanément plus fréquentes que les petites insertions (Mira *et al.*, 2001), ce qui conduirait à une érosion des séquences non fonctionnelles.

Si ce biais a également été observé dans certains génomes eucaryotes (Graur *et al.*, 1989; Ophir et Graur, 1997; Petrov *et al.*, 2000), il a été récemment remis en question par Denver *et al.* (2004). En utilisant une méthode plus directe chez *Caenorhabditis elegans*, ils ont mis en évidence un biais inverse, en faveur des petites insertions. Par ailleurs, les taux de petites délétions et insertions ne peuvent à eux seuls expliquer la taille des génomes eucaryotes (Gregory, 2004). Celle-ci peut varier beaucoup plus rapidement par l'activité des éléments transposables (Kidwell, 2002), par l'insertion d'ARN messagers rétro-transcrits (rétroposition) (Maestre *et al.*, 1995; Esnault *et al.*, 2000), par des grandes duplications ou délétions.

Comeron (2001) défend une vision plus sélectionniste : l'ADN intergénique, en diminuant le déséquilibre de liaison entre les loci sous sélection, limiterait l'effet "ruby in a rubbish" (lorsque la fixation d'une mutation favorable est gênée par des mutations délétères ailleurs dans le génome) et l'effet Hill-Robertson (lorsque la fixation d'une mutation favorable est gênée par la présence d'une mutation encore plus favorable dans la population).

À ces différentes pressions susceptibles d'expliquer la quantité d'ADN non fonctionnel présente dans un génome, nous pouvons en ajouter une autre, liée au rôle des séquences intergéniques et des séquences répétitives dans la variabilité mutationnelle du phénotype. Il s'agit d'un rôle complexe. L'ADN intergénique semble promouvoir la robustesse car il permet à certaines mutations locales et à certains réarrangements de n'affecter aucun gène existant. Mais paradoxalement, il favorise aussi la variabilité mutationnelle du phénotype. Tout d'abord, les mutations locales qui se produisent dans de l'ADN non fonctionnel peuvent le rendre fonctionnel, en le transformant en un signal de régulation (Hahn *et al.*, 2003; Lynch, 2006b) ou même en un gène : dans le génome humain, certains anciens éléments transposables et certains rétroposons (ARN messagers rétrotranscrits et

insérés dans un chromosome) ont ainsi donné naissance à des gènes fonctionnels (Smit, 1999; International Human Genome Sequencing Consortium, 2001; Brosius, 2003). Ensuite, l'ADN non fonctionnel situé à l'intérieur d'un gène peut augmenter la variabilité mutationnelle de la protéine : par exemple, les mutations qui se produisent dans la partie transcrite non traduite en amont d'une séquence codante (5'UTR) peuvent créer un nouveau signal d'initiation et ainsi augmenter la longueur de la protéine, ou bien causer un décalage du cadre de lecture. Les éléments transposables sont aussi particulièrement mutagènes dans la mesure où ils sont susceptibles de s'insérer au sein des séquences géniques et donc d'affecter le phénotype. Enfin, l'ADN non fonctionnel peut contenir de nombreuses séquences répétitives[1] susceptibles de médier des grands réarrangements, qui peuvent affecter plusieurs gènes simultanément. Il semble donc que l'ADN non fonctionnel représente un réservoir important de mutations non neutres.

La sélection indirecte d'un certain niveau de variabilité pourrait alors se traduire par une modulation de la quantité d'ADN non fonctionnel. Si la sélection est directionnelle, c'est-à-dire si des mutations favorables sont possibles, les génomes avec le plus grand "réservoir d'innovations" (Brosius et Gould, 1992) les trouveront peut-être plus rapidement, et une grande quantité d'ADN non fonctionnel pourrait ainsi être indirectement sélectionnée. Si au contraire la sélection est stabilisatrice, alors la variabilité causée par l'ADN non fonctionnel n'est pas avantageuse et les génomes les plus longs devraient être indirectement contre-sélectionnés. Lynch (2006b) parle ainsi de l'ADN non fonctionnel comme d'un "fardeau mutationnel" qui devrait être éliminé par la sélection naturelle. Mais selon Lynch et Conery (2003), lorsque la taille efficace de la population est faible (comme pour les vertébrés), la sélection n'est pas suffisamment efficace pour cela et les insertions d'ADN non fonctionnel, même indirectement délétères, peuvent être fixées par dérive génétique. Les deux points de vue permettraient d'expliquer, au moins partiellement, les variations de taille de génome entre phyla. Tous deux sont fondés sur les effets mutagènes de l'ADN non fonctionnel et la sélection indirecte qu'ils peuvent occasionner, et c'est encore une fois la possibilité ou non des mutations favorables qui détermine la direction de cette sélection indirecte.

Répartition relative des éléments répétés et des gènes

Nous avons rapidement évoqué dans le paragraphe précédent le rôle de l'ADN répétitif dans les réarrangements génomiques et donc dans la variabilité mutationnelle du phénotype. L'ADN n'étant constitué que de 4 bases différentes, il contient nécessairement des répétitions dites fortuites, apparaissant simplement par mutations ponctuelles (Achaz, 2002), et ces répétitions fortuites sont d'autant plus nombreuses que la molécule est longue. Cependant, les mécanismes de duplication et de transposition génèrent des répétitions supplémentaires, qui peuvent être plus longues que les répétitions fortuites et ainsi constituer un meilleur substrat pour les mécanismes de recombinaison. Le nombre de répétitions pouvant initier des événements de recombinaison joue ainsi un rôle important

[1] Dans le génome humain par exemple, une grande partie de l'ADN intergénique dérive vraisemblablement d'anciens éléments transposables (International Human Genome Sequencing Consortium, 2001).

dans le nombre de réarrangements et donc dans la stabilité d'un génome (Achaz *et al.*, 2003; Rocha, 2003a,b). Mais à nombre égal, la façon dont ces répétitions sont distribuées le long du génome est aussi un facteur important, dans la mesure où elle influence le nombre moyen de gènes affectés et donc l'impact moyen de chaque événement.

Chez *Escherichia coli*, presque deux tiers des répétitions longues se trouvent dans des régions intergéniques, alors que chez *Bacillus subtilis*, 60% des répétitions se trouvent dans des séquences codantes (Rocha *et al.*, 1999a). Les réarrangements ont donc vraisemblablement des effets différents chez les deux espèces. Chez *Mycoplasma genitalium* et *Mycoplasma pneumoniae*, les répétitions sont agrégées : il existe des "clusters" de répétitions séparés par de longues séquences sans répétitions (Rocha *et al.*, 1999a). Ces répétitions correspondent à des gènes codant pour des protéines de surface, qui sont hautement répétés dans les génomes des deux espèces. Les réarrangements qui se produisent entre ces répétitions génèrent une grande diversité antigénique (Rocha *et al.*, 1999b), et peuvent ainsi permettre à ces bactéries pathogènes d'échapper au système immunitaire de l'hôte. Cette stratégie n'est cependant pas généralisable à toutes les bactéries pathogènes, puisque les génomes de *Chlamydia trachomatis* et *Rickettsia prowazekii* présentent très peu de répétitions (Rocha *et al.*, 1999b).

Dans le génome humain, la majeure partie des répétitions est due aux éléments transposables. Ceux-ci ne sont pas répartis uniformément le long du génome ; il existe des régions plus riches que d'autres en ADN répétitif (International Human Genome Sequencing Consortium, 2001). Une région de 525 kb du chromosome X présente par exemple une densité en éléments transposables de 89%. Au contraire, certaines régions comme celles des gènes homéotiques sont quasiment dépourvues de répétitions. Un examen plus fin des différents types de répétitions montre qu'ils se distribuent différemment le long du génome. Les séquences LINEs[1] se trouvent principalement dans les régions riches en A et T, alors que les SINEs[2] se trouvent surtout dans les régions riches en G et C. Ces régions riches en G+C et en SINE sont aussi les régions les plus riches en gènes (Mouchiroud *et al.*, 1991; International Human Genome Sequencing Consortium, 2001).

Achaz *et al.* (2001) ont examiné la répartition des longues répétitions autres que les éléments transposables, dans le génome humain et dans d'autres génomes eucaryotes comme ceux de la Levure *Saccharomyces cerevisiae*, de la Drosophile *Drosophila melanogaster* et de l'Arabette *Arabidopsis thaliana*. Ces différents génomes présentent une caractéristique commune : les répétitions directes, c'est-à-dire orientées dans le même sens, sont en général plus proches l'une de l'autre que ne le sont les répétitions inversées. Un examen plus détaillé mené chez *Saccharomyces cerevisiae* a montré que les répétitions directes sont plus longues et beaucoup plus proches que dans un génome aléatoire, alors que les répétitions

[1]Long INterspersed Elements. Ces rétrotransposons d'environ 6 kb de long contiennent un promoteur reconnu par l'ADN polymérase II et deux séquences codantes, dont l'une a une activité reverse-transcriptase (copie d'un ARN en ADN) et intégrase. Grâce à cette activité, l'ARN messager produit par l'ADN polymérase II est rétro-transcrit en ADN et réinséré à un autre endroit dans le génome. Les LINEs constituent environ 21% du génome humain.

[2]Short INterspersed Elements. Ces courtes séquences d'ADN (de 100 à 400 bp) contiennent un promoteur reconnu par l'ADN polymérase III mais aucune séquence codante. Ils utilisent vraisemblablement la machinerie enzymatique des LINEs pour se propager. Ils représentent environ 11% du génome humain.

inversées sont, elles, distribuées comme dans un génome aléatoire (Achaz *et al.*, 2000). De plus, les répétitions directes proches sont principalement situées à l'intérieur d'une même séquence codante, alors que les répétitions directes éloignées et les répétitions inversées sont beaucoup moins fréquemment situées dans des régions codantes (Achaz *et al.*, 2000). Les répétitions directes proches sont sujettes à la délétion par recombinaison, et ce processus consomme l'une des deux copies. Cela expliquerait qu'elles aient disparu dans les zones non fonctionnelles. Les contraintes fonctionnelles qui pèsent sur les séquences codantes protègeraient les répétitions qui s'y trouvent. Cependant, si la délétion est systématiquement délétère, on peut imaginer que des mutations qui réduiraient la similarité entre les séquences (et donc leur taux de recombinaison) sans affecter la fonction de la protéine soient indirectement sélectionnées, car elles rendraient le phénotype plus robuste. Le maintien d'une forte similarité sur une grande durée évolutive pourrait, à l'inverse, refléter la sélection indirecte d'un phénotype variable.

Ainsi, dès lors que l'on prend en compte l'existence de grands réarrangements, la structure du génome apparaît comme un levier supplémentaire pour la variabilité mutationnelle du phénotype, un levier aux multiples facettes permettant potentiellement des ajustements plus ou moins fins.

3.3 Evolution de la structure du génome dans les algorithmes évolutionnaires

Si la structure du génome joue un rôle important dans la variabilité mutationnelle du phénotype, il est pertinent de se demander si elle peut évoluer par sélection indirecte, au même titre que le taux de mutation ou les mécanismes de canalisation. C'est une question difficile à aborder expérimentalement et les approches de modélisation sont donc particulièrement utiles. Comme nous allons le voir, l'idée de laisser la structure du génome évoluer est présente dans certains algorithmes d'évolution artificielle. Cependant, la motivation restant souvent l'amélioration des performances, la structure génomique obtenue et la façon dont elle a évolué ne constituent pas l'objet principal de ces études, ce qui induit parfois des choix et des comportements biologiquement peu réalistes.

Duplications, délétions, inversions

Les algorithmes génétiques traditionnels travaillent explicitement sur une séquence de taille fixe (voir figure I.2, p. 29). Cependant, l'idée de permettre au nombre de gènes d'évoluer était déjà présente dans les textes fondateurs du domaine (Schutz, 1995). Holland (1975) proposait ainsi d'utiliser les concepts de délétion et de duplication génique pour augmenter la puissance des algorithmes. Avec un nombre de gènes variable, il est en effet envisageable de chercher la solution d'un problème complexe à partir de solutions plus simples, répondant partiellement au problème (Goldberg *et al.*, 1989). Selon Conrad et Ebeling (1992), introduire des dimensions supplémentaires dans le paysage adaptatif est

aussi un moyen de réduire la probabilité d'être bloqué sur un optimum local, car le paysage contient moins de pics isolés et davantage de points-selles. Yu *et al.* (2003) ont aussi montré qu'un algorithme où le nombre de gènes est variable réalise de meilleures performances qu'un algorithme traditionnel lorsque le problème (l'environnement) varie dans le temps. Enfin, pour certains problèmes, le nombre de dimensions nécessaires n'est pas connu à l'avance et il paraît donc plus pertinent de le laisser évoluer que de le fixer arbitrairement (Harvey, 1992). C'est le cas par exemple en vie artificielle lorsqu'on cherche un ensemble de règles pour régir le comportement des agents : le nombre de règles nécessaires pour obtenir le comportement souhaité n'est pas toujours connu à l'avance (Schutz, 1995).

Cependant, dans le cadre de l'application la plus répandue des algorithmes génétiques, à savoir l'optimisation d'une fonction à n paramètres (n étant donné et fixe dans le temps), permettre la duplication et la délétion de gènes pose un certain nombre de problèmes. En particulier, il n'est plus possible d'utiliser la position du gène dans le génome pour interpréter sa fonction, c'est-à-dire pour savoir à quel paramètre il correspond. Ce problème est en général résolu en "marquant" chaque bit avec sa signification (quel digit de quel paramètre ?). Dans le "Messy GA" de Goldberg *et al.* (1993), le génotype d'un individu est ainsi une suite de paires $(\mathit{paramètre}_{\mathit{digit}}, \mathit{valeur})$, les mutations ponctuelles s'appliquant sur la *valeur*. Par exemple, dans le cadre de l'optimisation d'une fonction à $n = 4$ paramètres (a, b, c, d) codés chacun sur trois digits, un individu pourrait être de la forme

$$(a_1, 0), (a_2, 1), (a_3, 0), (b_1, 1), (b_2, 1), (b_3, 1), (c_1, 1), (c_1, 0), (c_2, 0), (c_3, 0), (d_1, 1), (d_2, 1), (d_3, 0)$$

le bit c_1 ayant été dupliqué puis l'une des copies mutée. Dans ce cas, c_1 est dit sur-spécifié. Au moment de calculer la fitness, sa valeur est souvent déterminée en choisissant arbitrairement l'une des valeurs possibles. Ce codage permet aussi à l'ordre des bits d'évoluer, par inversions par exemple (Bagley, 1967; Goldberg *et al.*, 1993; Harik, 1997). Cependant, avec ce type de codage, les paramètres peuvent se retrouver sous-spécifiés suite à des délétions. Dans l'exemple précédent, la délétion de la paire $(a_1, 0)$ conduirait à l'indétermination du premier digit du paramètre a. Il faut alors définir une heuristique pour remplir les digits manquants et évaluer la fitness, mais l'heuristique en question n'a en général pas de sens biologique. Ainsi, introduire des opérateurs naturels comme la duplication et la délétion dans un algorithme évolutionnaire ne le rend pas nécessairement plus proche de la biologie. Puisque la transition du génotype au phénotype est radicalement différente dans un algorithme évolutionnaire et dans un système vivant, l'utilisation d'opérateurs bio-inspirés sur le génotype ne produit pas nécessairement des effets réalistes au niveau du phénotype, et le système artificiel peut évoluer très différemment d'un système naturel (Fogel, 1995).

Segments non codants de taille fixe

Parallèlement aux recherches sur les opérateurs de duplication, délétion et inversion, un certain nombre de travaux ont été menés sur l'intérêt des segments non codants. Les performances d'un algorithme génétique peuvent en effet être améliorées en ajoutant un nombre fixe de positions non codantes entre les blocs de positions qui spécifient les para-

mètres (Levenick, 1991; Wu et Lindsay, 1995). Ces positions non codantes sont ignorées dans le calcul de la fitness. Par exemple, pour optimiser une fonction à 4 paramètres codés chacun sur 3 bits, les individus seront de la forme $a_1a_2a_3{}^{***}b_1b_2b_3{}^{***}c_1c_2c_3{}^{***}d_1d_2d_3$, plutôt que $a_1a_2a_3b_1b_2b_3c_1c_2c_3d_1d_2d_3$ comme dans la version traditionnelle. L'amélioration apportée est liée à la façon dont la recombinaison est effectuée entre les individus. Dans un algorithme génétique traditionnel, on fixe une probabilité de "crossover", c'est-à-dire la probabilité de produire un descendant en combinant les génotypes de deux individus. Dans ce cas, une position de rupture est choisie uniformément le long du chromosome. Le génotype du descendant est construit en recopiant l'information de l'individu 1 à gauche de cette position, puis l'information de l'individu 2 à droite de cette position. L'effet protecteur du non codant est lié au fait qu'on ne réalise qu'un point de rupture, quelle que soit la longueur du génome. En ajoutant des segments non codants entre les blocs de positions formant une bloc fonctionnelle, on réduit la probabilité que ce point de rupture soit situé à l'intérieur d'un bloc. Ainsi, le crossover ne modifie pas les briques de base – c'est davantage le rôle de la mutation – mais en crée de nouvelles combinaisons[1].

Segments non codants de taille variable

Ce rôle protecteur joué par les segments non codants vis-à-vis des effets délétères du crossover est vraisemblablement l'une des raisons du "code bloat" (littéralement, gonflement du code) observé en programmation génétique. Cette branche de l'évolution artificielle vise à trouver automatiquement le code d'un programme informatique qui réalise une tâche donnée, en faisant évoluer une population de programmes par mutation, recombinaison et sélection (voir (Koza, 1992) pour une introduction). Les solutions potentielles du problème sont typiquement de taille inconnue, et la programmation génétique a donc fait d'emblée appel à des représentations de taille variable et à des opérateurs comme la duplication et la délétion (Koza et Andre, 1995). Selon le type de langage utilisé, les instructions et les données peuvent être représentées linéairement, ou bien dans une structure arborescente, particulièrement adaptée au langage LISP par exemple. Dans ce cas, le "génotype" n'est pas aisément comparable à un chromosome. Cependant, comme certaines parties d'un code évolué par programmation génétique peuvent être enlevées sans modifier le comportement du programme, les chercheurs du domaine ont pris l'habitude de distinguer d'une part les parties fonctionnelles et d'autre part les "introns". Ces introns sont partiellement responsables d'un problème récurrent en programmation génétique, à savoir la croissance illimitée des génotypes au fur et à mesure de l'évolution, sans augmentation de fitness. Blickle et Thiele (1994), Nordin et Banzhaf (1995) et Nordin *et al.* (1997) ont suggéré que les introns protègent les parties fonctionnelles des effets délétères du crossover (Luke, 2000). Des génotypes de plus en plus longs sont donc indirectement sélectionnés. Encore une fois, ce comportement est très lié au fait que le nombre de points de rupture causés

[1] Ce résultat peut sembler comparable à l'argument de Comeron (2001), qui met en avant le rôle de l'ADN intergénique dans la recombinaison pour expliquer sa présence dans les génomes réels (voir paragraphe 3.2, p. 3.2). Si les segments non codants augmentent effectivement l'efficacité du crossover à la fois dans les algorithmes génétiques et dans les génomes réels, ce n'est cependant pas par le même mécanisme. Le nombre de points de rupture dans un génome réel augmente vraisemblablement avec sa taille, ce qui annule l'effet protecteur du non codant décrit par Levenick (1991).

par le crossover (en l'occurrence 1 en général) est indépendant de la taille du génome. On n'observe pas l'effet de code bloat dans la plate-forme de vie artificielle Avida (Ofria *et al.*, 2003), qui est pourtant relativement proche des algorithmes de programmation génétique (voir paragraphe 2.5, p. 46), et l'une des raisons est vraisemblablement qu'il n'y a pas de crossover.

D'autres hypothèses ont cependant été proposées pour expliquer le phénomène du code bloat. Certaines sont plus spécifiques aux représentations arborescentes (Soule et Foster, 1998; Luke, 2003), alors que d'autres sont plus générales et vont même au-delà du domaine de la programmation génétique. Parmi ces dernières, on trouve notamment l'hypothèse de Langdon et Poli, "Fitness Causes Bloat" (Langdon et Poli, 1997; Langdon, 1998). Selon cette hypothèse, les génotypes longs de fitness élevée sont plus nombreux que les génotypes courts de fitness élevée, tout simplement parce qu'il existe plus de génotypes longs. Plus rigoureusement, il existe plus de représentations longues d'une solution donnée que de représentations courtes. Si l'exploration se faisait uniformément dans l'espace de recherche, la taille moyenne des solutions serait constante. Comme la population est en général initialisée avec des solutions courtes, la croissance des solutions pourrait ne refléter que la transition vers l'équilibre. Une autre hypothèse relativement générale a été proposée par Luke (2005) : la croissance des génomes serait due au fait que les duplications sont moins délétères que les délétions et ont donc plus de chances d'être fixées. L'auteur propose un modèle simple pour démontrer cet effet, en prenant en compte des duplications, des délétions et des translocations. Le problème de ces simulations est que le nombre d'événements subis par un génome est indépendant de sa taille, et que la taille des segments dupliqués, excisés ou déplacés est aussi indépendante de la taille du génome. On retrouve ainsi, sous une forme légèrement différente, le problème du crossover 1-point : de très longs segments non codants protègent les gènes des effets délétères des mutations. Cette hypothèse, qui a également figuré pendant un temps dans la littérature biologique (Hsu, 1975, *in* (Petrov, 2001; Bolzer *et al.*, 2005)), n'est cependant valable que pour des mutations dont le nombre n'augmente pas avec la taille du génome, comme peut-être l'influx de séquences virales.

Structure bio-inspirée du génome

Il faut noter que si les segments non codants de taille variable sont naturels en programmation génétique, on peut aussi les autoriser dans un algorithme génétique, et laisser alors l'évolution déterminer le nombre de bits non codants et leur répartition (Wu et Lindsay, 1996). Plusieurs méthodes sont possibles. On peut utiliser le système de paires $(paramètre_{digit}, valeur)$ en ajoutant un pseudo-paramètre ignoré dans le calcul de la fitness. De façon plus bio-inspirée, on peut travailler sur des chaînes de bits et choisir des signaux de "start" et "stop" qui définiront les bornes des gènes, c'est-à-dire ici les blocs de bits qui seront utilisés dans le calcul de la fitness. Cette seconde option permet plus de souplesse au niveau de la structure du génome : le nombre de gènes, mais aussi la longueur des gènes et la longueur des séquences intergéniques sont variables. Des mutations ponctuelles peuvent faire disparaître les signaux de start et de stop et les faire réapparaître à d'autres endroits. En principe, des duplications, délétions ou inversions peuvent

aussi faire varier l'organisation du génome. Ce type de codage nécessite cependant une transition du génotype au phénotype qui puisse s'accommoder de la flexibilité totale de la structure génomique. En d'autres termes, la fitness doit être calculable pour tout nombre de gènes, pour toutes longueurs de gènes.

Un des exemples les plus connus est le "Virtual Virus" de Burke *et al.* (1998). Dans ce modèle, le génotype est une chaîne de "nucléotides" (A, C, G, T), de longueur variable. Un code génétique artificiel met en relation chaque triplet de nucléotides avec une lettre de l'alphabet, ou bien avec un signal start ou stop. Le phénotype est déterminé en lisant le génotype de gauche à droite dans les 3 cadres de lecture : lorsqu'un signal start est rencontré, les triplets suivants sont interprétés à l'aide du code génétique artificiel, jusqu'à rencontrer un signal stop. On obtient ainsi les séquences "protéiques". Le critère de sélection repose directement sur ces séquences protéiques : le phénotype cible est l'ensemble de séquences {COREPROTEIN, POLYMERASE, ENVELOPE}. Un individu qui possède les séquences {COREPRODEI, JOLYMERISE, ENVEKLOPE} se reproduit donc plus qu'un individu qui possède les séquences {CPGJIN, PPLKOJMEFHOSE, NTLPA}. Comme dans un algorithme génétique classique, les mutations sont ponctuelles, et une certaine fraction des individus de chaque génération est créée par crossover 1-point entre deux parents. La différence est que le point de rupture ne se trouve pas nécessairement au même endroit sur les deux chromosomes : une région est choisie au hasard dans l'un des génomes, puis on recherche dans le second une région de séquence similaire. Les deux régions sont placées face à face et l'échange est réalisé. C'est ce mécanisme qui fait varier la taille du génome. Une simplification lourde de conséquences est faite dès que le génome contient plus de trois gènes : seuls les trois qui correspondent le mieux aux séquences cibles sont pris en compte dans le calcul de la fitness. Sans coût sélectif à la taille du génome, celle-ci augmente donc indéfiniment, car les génomes les plus longs ont à la fois leurs gènes fonctionnels protégés du crossover (voir plus haut) et à la fois une plus grande probabilité de trouver de meilleurs gènes. Avec un coût à la taille du génome et une taille maximale de 3000 nucléotides, la population converge vers une taille d'équilibre, qui augmente avec le taux de mutation par position (Burke *et al.*, 1998; Ramsey *et al.*, 1998). Il semble que ce résultat soit à nouveau lié au fait que seuls 3 gènes sont pris en compte dans le calcul de la fitness. En effet, pour conserver la fitness, il suffit que l'une des copies de chaque gène soit exempte de mutations. Plus le génome contient de copies de chaque gène, plus on a de chances que cela soit le cas. Les génomes longs ont donc ici, en plus de leur avantage exploratoire, une plus grande robustesse. Ils sont donc d'autant plus favorisés que la pression mutationnelle est forte.

Ainsi, les modèles d'évolution artificielle existants permettent d'étudier certaines pressions indirectes s'exerçant sur la structure d'un génome, et en particulier sur le nombre de gènes et la quantité d'ADN non fonctionnel. Cependant, ces pressions indirectes sont souvent liées à des artefacts du modèle, au niveau des mécanismes de recombinaison ou au niveau de la transition du génotype au phénotype.

3.4 Evolution de la structure du génome dans les modèles biologiques

L'évolution de la structure des génomes a également fait l'objet d'études de modélisation dans la communauté biologique. La méthode est cependant radicalement différente dans la mesure où les modèles font souvent abstraction de la notion de séquence et/ou du niveau populationnel.

Des modèles de génétique des populations ont par exemple été développés pour étudier l'évolution du nombre de transposons contenus dans un génome (Charlesworth et Charlesworth, 1983; Basten et Moody, 1991; Deceliere *et al.*, 2005). Cependant, ce type de modèle ne prend pas explicitement en compte la séquence génomique, avec ses gènes, ses régions intergéniques et ses éléments répétés. Dans le modèle de Basten et Moody (1991) par exemple, la fitness d'un individu dépend directement du nombre de transposons qu'il contient. Le nombre de transposons à l'équilibre dépend alors de leur taux de transposition, du taux de délétion et de ce coût sélectif direct. Un tel coût direct peut refléter, par exemple, les coûts énergétiques de la rétrotransposition ou les effets délétères de l'activité des transposases, qui créent des cassures dans les chromosomes (Nuzhdin, 1999). Un tel modèle n'est pas adapté pour isoler les effets indirects des transposons : un transposon peut en effet ne pas être pénalisant pour un individu donné mais représenter un danger pour sa descendance, car ils peuvent s'insérer dans des gènes ou médier des réarrangements chromosomiques.

On se heurte à des problèmes similaires avec certains modèles de l'évolution du nombre de gènes. Bengtsson (2004) a par exemple proposé un modèle mathématique qui distingue d'une part les gènes dédiés au métabolisme, et d'autre part les gènes de réparation contrôlant le taux de mutation. Là encore, la séquence génomique n'est pas explicitement prise en compte. La fitness d'un organisme est directement fonction du nombre de gènes dédiés au métabolisme, et, de façon plus problématique, du taux de mutation et donc du nombre de gènes qui le contrôlent. L'idée est que la probabilité que le génome subisse au moins une mutation délétère lors d'une réplication augmente avec le nombre de gènes qu'il contient. Il a en effet été suggéré que comme la majorité des mutations sont délétères, il existe un nombre maximal de gènes pour un taux de mutation par locus donné (Eigen, 1971; Maynard Smith, 1983; Hurst, 1995; Ofria et Adami, 1999; Pal et Hurst, 2000; Ofria *et al.*, 2003). Dans le modèle de Bengtsson (2004), la probabilité que le génome subisse au moins une mutation diminue directement la fitness de l'organisme. Selon l'auteur, il s'agit donc d'une "fitness à long terme". Il existerait un nombre de gènes optimal selon ce critère. Cela suggère que, du fait de l'effet délétère des mutations, le génome ne va pas croître indéfiniment mais se stabiliser sur un nombre de gènes optimal, et ce, même en présence de gènes de réparation et en l'absence d'un coût direct du nombre de gènes sur le taux de croissance. Plus le taux de mutation basal est grand, plus le nombre optimal de gènes est faible. Il faut cependant noter que ce modèle n'inclut pas la dynamique d'une population et du processus évolutif. Le raisonnement suppose implicitement qu'une sélection interclones va sélectionner la sous-population qui présente la plus grande fitness à long terme. Comme l'avait déjà souligné Johnson (1999a) dans le contexte de l'évolution du taux de

mutation, un tel raisonnement n'est correct que si les sous-populations sont strictement indépendantes. Cela supposerait ici l'absence de mutations susceptibles de faire varier le nombre de gènes, ce qui est pour le moins problématique lorsqu'on s'intéresse à l'évolution du nombre de gènes. Le rôle des mutations délétères dans l'évolution du nombre de gènes reste une hypothèse intéressante, mais pour la tester, il paraît préférable d'utiliser un modèle incluant explicitement les mutations et leurs effets phénotypiques, le niveau populationnel et une sélection inter-individus.

Cependant, si l'on souhaite travailler sur des données génomiques réelles pour obtenir des indicateurs quantitatifs de la dynamique structurelle du génome, il est impossible de calculer le phénotype et la fitness. Il faut alors se placer à un niveau encore plus haut que les précédents : en travaillant directement sur les taux d'événements fixés, on peut s'affranchir du niveau populationnel et d'une définition explicite de la fitness.

L'évolution du nombre de gènes dupliqués peut ainsi être modélisée à l'aide d'équations mettant en jeu un taux de duplications fixées et d'un taux d'inactivations fixées (fixation d'une mutation conduisant à l'inactivation du gène) (Lynch et Conery, 2000; Maere *et al.*, 2005). En simulant la dynamique d'un tel modèle, on peut caractériser la distribution de l'âge des gènes fonctionnels qui est attendue pour un jeu de paramètres donné. Il est ensuite possible de trouver les taux de duplications et d'inactivations qui donnent la distribution théorique la plus proche de la distribution observée. Cela a permis de montrer que dans des génomes eucaryotes allant de la Levure à l'Homme, la grande majorité des gènes dupliqués deviennent non fonctionnels en quelques millions d'années (Lynch et Conery, 2000). Chez *Arabidopsis thaliana*, ce type d'approche a permis de montrer que le taux d'inactivation dépend à la fois du type de duplication dont provient le gène (duplication d'un seul gène ou du génome complet) et de sa catégorie fonctionnelle (Maere *et al.*, 2005).

Il est également possible de définir un modèle mathématique de l'évolution de la taille du génome à partir du taux de délétions fixées et du taux d'insertions fixées. Petrov (2002) a ainsi proposé un modèle d'équilibre mutationnel pour la taille d'un génome, reposant sur un équilibre entre la fréquence de fixation des petites délétions et celle des grandes insertions. L'auteur fait en effet l'hypothèse que (i) les petites délétions sont spontanément plus fréquentes que les petites insertions, et que (ii) les grandes délétions ont une contribution négligeable dans l'évolution de la taille d'un génome dans la mesure où elles ont de grandes chances d'affecter au moins un gène et donc d'être contre-sélectionnées. Les grandes insertions ont plus de chances d'être conservées car elles peuvent s'insérer sans dommages dans les régions non fonctionnelles. Selon ce modèle, la taille du génome évolue jusqu'à atteindre un équilibre, où les pertes d'ADN par petites délétions sont compensées par les grandes insertions. La quantité d'ADN non fonctionnel ne fait ici l'objet d'aucune forme de sélection, ni directe ni indirecte. Cette approche comporte cependant un certain nombre de raccourcis. Tout d'abord, lorsque le génome contient de grandes régions non fonctionnelles, de grandes délétions peuvent être fixées. Le taux de délétions fixées n'est donc pas indépendant de la taille du génome. Ensuite, il faut supposer que le taux d'insertions fixées augmente très lentement (puissance 1/4) avec la taille du génome pour qu'il compense exactement le taux de pertes par petites délétions (Gregory, 2004). Or si

l'on suit les hypothèses initiales, de plus grandes régions non fonctionnelles fournissent plus de points d'insertions possibles, ce qui devrait faire augmenter rapidement le nombre d'insertions fixées. Le principal problème du modèle de Petrov (2002) est que l'idée d'un équilibre stable est postulée a priori, sans justification, et le taux de fixation des grandes insertions est déduit de ce postulat plutôt que de données observées.

Cet exemple montre les limites des approches qui travaillent directement sur les taux d'événements fixés. Le comportement de ces modèles dépend fortement des hypothèses simplificatrices faites sur les mécanismes de mutation-sélection sous-jacents. Comme le niveau populationnel n'est pas explicitement inclus, ces simplifications sont implicites et ne sont donc pas toujours faciles à déceler. Par ailleurs, pour étudier les pressions sélectives indirectes qui peuvent s'exercer sur la structure du génome, il est impératif de prendre en compte tous les événements mutationnels, y compris ceux qui ne seront pas fixés. Ce sont en effet la fréquence et l'effet des événements spontanés qui déterminent la variabilité mutationnelle du phénotype, cible potentielle d'une sélection indirecte.

4 Conclusion

Le niveau de variabilité mutationnelle du phénotype d'un organisme est partiellement sous contrôle génétique et est donc susceptible d'évoluer. Les approches de modélisation ont permis de montrer que le taux de mutation et les mécanismes de canalisation peuvent évoluer par sélection indirecte. Cependant, les différents modèles sont parfois en désaccord sur le degré optimal de variabilité mutationnelle : faut-il être plutôt robuste ou plutôt variable ? Le fait que la sélection soit directionnelle ou stabilisatrice – autrement dit, la possibilité ou non des mutations favorables – apparaît comme un facteur crucial dans ce contexte. Les mutations favorables étant en principe possibles mais rares, on peut se demander si la clé du succès évolutif ne serait pas un compromis entre robustesse et variabilité, c'est-à-dire une variabilité mutationnelle intermédiaire.

Par ailleurs, ce tour d'horizon des modèles dédiés à l'étude de la variabilité mutationnelle montre que la mutation y est souvent envisagée comme ponctuelle, ou restreinte à un locus. Pourtant, dans un génome réel, la présence de gènes chevauchants, de séquences intergéniques, et d'éléments répétés pouvant médier de grands réarrangements chromosomiques, font que le nombre de gènes touchés par une mutation est variable. La structure du génome, en influençant le nombre moyen de gènes affectés par une mutation, joue donc potentiellement un rôle dans la variabilité mutationnelle du phénotype et serait donc susceptible d'évoluer par sélection indirecte, au même titre que le taux de mutation ou les mécanismes de canalisation. Si cette hypothèse est correcte, l'organisation d'un génome ne résulterait pas seulement de biais mutationnels et de pressions sélectives directes. Elle pourrait également refléter la sélection indirecte d'un niveau adéquat de variabilité mutationnelle. Cette hypothèse est cependant difficile à tester expérimentalement, car la sélection d'un certain niveau de variabilité ne peut s'exercer qu'à long terme, et est difficile à isoler des autres pressions sélectives qui pèsent sur le génome. Dans ce contexte,

la modélisation est un recours possible, qui doit permettre de tester si (i) des formes de sélection indirecte émergent effectivement dans un système en évolution, et (ii) ces formes de sélection indirecte agissent sur la structure d'un génome. Il s'avère que les modèles existants mettant en jeu l'évolution de la structure génomique ne permettent pas, tels quels, de répondre à cette double question. Les modèles développés pour quantifier la dynamique des génomes réels se heurtent à l'impossibilité de simuler explicitement leur évolution, du fait de la complexité de la transition du génotype au phénotype. En travaillant directement sur les événements fixés, ils ne peuvent pas capturer la compétition entre organismes plus ou moins variables induite par les mutations spontanées. Les modèles développés en évolution artificielle simulent explicitement l'évolution d'une population, et certains le font avec des génomes structurellement flexibles. Mais le manque de réalisme biologique de ces modèles, tant au niveau des mécanismes mutationnels qu'au niveau de la transition génotype-phénotype, rend leurs comportements difficilement transférables vers la biologie. Pour étudier la structuration des génomes par sélection indirecte, il faut donc un nouveau type de modèle, combinant autant que possible les forces des approches existantes. Dans le chapitre suivant, nous présentons le modèle et la plate-forme de simulation individu-centrée *aevol* (pour *a*rtificial *evol*ution), développée pour étudier l'évolution de l'organisation fonctionnelle des génomes, avec le souci d'intégrer les principaux niveaux auxquels se jouent la variabilité mutationnelle du phénotype.

Chapitre II

Le modèle *aevol*

La connaissance-projet se produit – et se représente – par ***conception*** *de modèles (...) et non plus par* ***analyse****. Le modèle alors, qu'il soit iconique ou symbolique, devient source de connaissance et non plus résultat. Il ne décrit plus, ex-post, une connaissance-objet tenue pour ex-ante ; il représente a priori une connaissance-projet qui n'existe que par lui (...). Etonnante et intelligible conjonction d'actions, la représentation théâtrale est bien riche métaphore décrivant la modélisation d'une connaissance-projet : auteur, acteur, spectateur, chacun cherche – et souvent trouve – ce qui n'existe pas encore et qui n'existera peut-être que pour lui.*

Jean-Louis Le Moigne[1]

Les modèles – schémas, diagrammes, équations...– peuvent permettre de synthétiser les connaissances acquises en analysant un objet, un génome par exemple. Mais la modélisation et la simulation permettent aussi d'explorer de nouvelles hypothèses, parfois inattendues. La construction du modèle ne passe pas, dès lors, par l'analyse et la synthèse exhaustive de la masse de données accumulées sur un objet. Le modèle est, au contraire, imaginé, *conçu*, à partir d'un nombre limité d'axiomes, de connaissances considérées vraies, formant un socle sur lesquelles de nouvelles connaissances pourront être construites. Le niveau de détail avec lequel ces axiomes sont représentés et implémentés doit réaliser un compromis entre réalisme, généricité et pragmatisme. C'est la problématique, la question que l'on veut poser au modèle, qui va déterminer le niveau de détail adéquat.

Voici brièvement introduits les principes qui ont guidé la construction du modèle *aevol*. L'objectif de ce modèle est de permettre l'exploration de nouvelles hypothèses relatives à

[1] "Qu'est-ce qu'un modèle ?", in *Les modèles expérimentaux et la clinique*, numéro spécial de la revue *Confrontations psychiatriques*, 1987.

l'évolution structurelle des génomes, en répondant aux deux questions laissées en suspens à la fin du chapitre précédent : des formes de sélection indirecte, liées à la variabilité mutationnelle du phénotype, émergent-elles dans un système en évolution ? Et si oui, ces formes de sélection indirecte peuvent-elles agir sur l'organisation fonctionnelle de l'information génétique ? Il s'agit ici de simuler l'évolution d'un système générique en s'appuyant sur quelques postulats de base (par exemple, les mutations s'appliquent sur le génotype, la sélection s'applique sur le phénotype, et la relation génotype-phénotype permet des degrés de liberté au niveau de la structure du génome), pour tenter d'*expliquer* l'évolution des structures génomiques. Il ne s'agit donc pas d'analyser des génomes réels donnés pour déterminer les équations qui prédiraient leur évolution. En cela, notre modèle est plus "source de connaissance [que] résultat".

D'un point de vue épistémologique, notre démarche est proche de celle de la vie artificielle. Ce champ a pris forme à la fin des années 80 lorsqu'une centaine de biologistes, chimistes, physiciens et informaticiens, travaillant indépendamment sur des sujets proches, se sont retrouvés au premier "atelier interdisciplinaire sur la synthèse et la simulation des systèmes vivants", sous l'impulsion de Chris Langdon, avec l'aide du Centre d'études des phénomènes non linéaires de Los Alamos et du Santa Fe Institute. La vie artificielle vise à inférer des principes universels sous-jacents à tout système vivant, en créant des systèmes artificiels qui capturent (partiellement) la complexité du vivant pour la rendre accessible à de nouvelles formes d'expérimentation. Ces systèmes artificiels permettent en effet de contrôler précisément les paramètres, de répliquer facilement les expériences, et d'accéder à toutes les données pertinentes pour analyser les résultats (Miller, 1995). L'évolution est une thématique centrale dans ce contexte : Bedau *et al.* (2000), dans un article intitulé "Open problems in artificial life", fixent l'objectif de "déterminer ce qui est inévitable" dans l'évolution d'un système vivant, c'est-à-dire de dégager ce qui est général, reproductible, de ce qui est contingent. Il s'agit de "réaliser les expériences que la méthode scientifique nous dicte, mais que nous ne pouvons pas faire avec les échelles de temps et d'espace de structures matérielles comme les cellules elles-mêmes" (Forbes, 2004). De telles expériences ne sont pas, à notre avis[1], des démonstrations au sens strict, mais permettent de "stimuler l'intuition" (Rennard, 2002) et de dégager des mécanismes potentiels qu'il faudra ensuite confronter au réel.

D'un point de vue technique, notre modèle s'apparente aux modèles d'évolution artificielle présentés dans le chapitre précédent (quasi-species, algorithmes évolutionnaires, plate-forme Avida...). Nous avons vu que certains de ces modèles se sont révélés utiles pour étudier la sélection indirecte d'un certain taux de mutation ou d'un certain niveau de neutralité des mutations. En revanche, en ce qui concerne la sélection indirecte de la structure du génome, on se heurte rapidement à des problèmes méthodologiques. Comme nous l'avons vu au chapitre précédent, aucun de ces modèles n'a été conçu pour étudier ce phénomène et ils ne présentent donc pas le réalisme minimal nécessaire au niveau de la transition génotype-phénotype et au niveau du génome lui-même, de son organisation fonctionnelle et des mécanismes mutationnels responsables de sa dynamique. Le modèle

[1] Dans la communauté de la vie artificielle, le statut de ces expériences *in silico* est très débattu. Certains y voient des nouvelles formes de vie, alors que d'autres les considèrent comme des simulations du réel.

aevol et la suite de logiciels associée ont donc été conçus pour prendre en compte explicitement ces aspects, et ainsi permettre d'étudier leur rôle dans la variabilité mutationnelle du phénotype.

Dans ce chapitre, nous commençons par replacer le modèle dans le contexte de la modélisation individu-centrée, en précisant le type d'individu qui doit être implémenté (section 1). Nous détaillerons ensuite chaque composante du modèle, du génome au phénotype en passant par le protéome, sans oublier les processus de mutation et de sélection (section 2). Dans une troisième section, nous décrirons le comportement typique du modèle pour les plages de paramètres utilisées dans ce travail, en y distinguant le pertinent de l'artefactuel, et en fournissant les principaux éléments d'analyse de sensibilité.

1 Généralités

1.1 Modélisation individu-centrée, émergence et immergence

Contrairement aux modèles mathématiques classiques exprimant directement des dynamiques sur des variables globales (moyennes sur l'ensemble de la population), les modèles dits "individu-centrés" se fondent sur une représentation explicite de l'ensemble des individus d'un système, ainsi que de leurs interactions éventuelles. Ces modèles sont très utilisés en vie artificielle, mais aussi dans de nombreuses autres disciplines, les individus pouvant représenter des plantes ou des animaux d'un écosystème (Hogeweg et Hesper, 1990; Grimm, 1999), des insectes d'une colonie (Corbara *et al.*, 1993), des poissons formant un banc (Reynolds, 1987), des voitures dans un réseau routier (Byrne *et al.*, 1982), ou encore des personnes formant des réseaux sociaux (Gilbert, 2004). Les modèles individu-centrés sont typiquement à vocation explicative. Il s'agit souvent de caractériser le comportement collectif, global, de la population, et de comprendre comment ce comportement peut *émerger* des propriétés des individus et de leurs interactions locales. On peut par exemple montrer que le déplacement cohérent d'un banc de poissons ne reflète pas nécessairement une hiérarchie dans le groupe (présence d'un "leader"), mais peut aussi spontanément émerger de comportements locaux (éviter les collisions avec les voisins immédiats).

Le concept d'émergence, en général résumé par la maxime "le tout est plus que la somme des parties", est un concept central des approches individu-centrées et de la vie artificielle. Cependant, lorsque ces approches sont utilisées pour modéliser l'évolution, ce n'est pas tant le tout, le collectif, qui est l'objet d'intérêt. Ce sont davantage les parties, c'est-à-dire les individus et leurs propriétés après plusieurs milliers de générations. Plus exactement, il s'agit de comprendre comment le fait d'appartenir à une population qui évolue (et donc, de subir la sélection naturelle) façonne les individus. On s'intéresse donc à la façon dont le tout agit sur les parties, en renversant le concept d'émergence. Si les parties acquièrent des propriétés inattendues dans le contexte du tout – des propriétés "immergentes", "micro-émergentes" ou "localement émergentes" selon les auteurs –, alors les parties sont plus que simplement des parties. Comme celui de l'émergence, le concept d'immergence requiert

la présence explicite du niveau micro en plus du niveau macro. Les approches individu-centrées sont donc particulièrement adaptées à l'étude de ce type de phénomènes.

Ici, nous utilisons ce type d'approche pour interroger la capacité du processus évolutif le plus fondamental – mutation au niveau du génotype, sélection au niveau du phénotype – à faire apparaître des structures génomiques particulières. Si la structure génomique ne fait l'objet d'aucune pression sélective directe, peut-elle tout de même spontanément acquérir des propriétés particulières, du simple fait que chaque individu est en compétition avec d'autres individus, plus ou moins adaptés et subissant plus ou moins de mutations ?

1.2 Quel type d'individu ?

Comment modéliser les individus dans ce contexte ? La structure génomique, notre principal objet d'étude, doit être suffisamment réaliste et suffisamment flexible pour pouvoir évoluer, ce qui exclut une représentation génotypique analogue à celle des algorithmes génétiques classiques. Plus précisément, comme pour un génome réel, la structure génomique (par exemple l'ordre des gènes) doit pouvoir être modifiée sans nécessairement entraîner une modification du phénotype et de la fitness. Le génome ne peut donc être une suite d'instructions formant un programme informatique, comme en programmation génétique (Koza, 1992) ou dans la plate-forme Avida (Adami, 2006). L'organisation génomique doit avoir de réels degrés de liberté.

De plus, nous avons vu au chapitre précédent, avec notamment le modèle "Virtual Virus" (Burke *et al.*, 1998), que donner des degrés de liberté à la structure génomique ne suffit pas pour obtenir une évolution structurelle réaliste. Il faut aussi que la transition du génotype au phénotype ne soit pas biologiquement aberrante. Cela implique d'abord de pouvoir prendre en compte tous les gènes dans le calcul du phénotype, et non un nombre pré-déterminé comme dans le Virtual Virus (Burke *et al.*, 1998) ou le Messy GA (Goldberg *et al.*, 1993). En d'autres termes, la complexité du phénotype ne doit pas être arbitrairement fixée une fois pour toutes, mais doit au contraire pouvoir évoluer, parallèlement à l'évolution du génome. Par ailleurs, dès lors que l'on s'intéresse à l'effet des mutations, il est important que la correspondance gènes-fonctions reflète deux propriétés rencontrées chez les organismes réels : la pléiotropie (un gène peut contribuer à plusieurs fonctions) et la polygénie (une fonction peut résulter de l'action de plusieurs gènes).

À ce nécessaire réalisme, nous devons cependant opposer généricité et pragmatisme. Généricité, parce que notre objectif est de découvrir des principes d'organisation, et non d'étudier une structure génétique donnée chez un organisme particulier. Pragmatisme, car le modèle doit rester calculable et analysable : "ce qui est simple est toujours faux, ce qui ne l'est pas est inutilisable", disait Paul Valéry[1]. L'une des simplifications majeures réalisées ici est l'omission de l'échelle de temps de l'individu : dans le modèle, un génotype donne un seul phénotype, supposé invariable au cours de la vie de l'individu. De plus, ce phénotype est exprimé uniquement en termes de capacités fonctionnelles (métaboliques

[1] *Mauvaises pensées et autres*, in *Oeuvres*, vol. II, Gallimard, Bibliothèque de la Pléiade, 1942.

par exemple) et l'aspect morphologique n'est pas pris en compte. Nous ne prenons donc pas en compte les processus de développement.

Par ailleurs, même en se restreignant aux aspects métaboliques, il est impossible de concevoir un modèle compréhensible contenant toutes les réactions biochimiques à l'œuvre dans un organisme réel. En outre, cela nuirait à la généricité du modèle, car à ce niveau de détail, il faudrait se restreindre à une espèce particulière. Nous avons donc utilisé une description plus abstraite des fonctions réalisées par une protéine : chaque protéine réalise ou inhibe un certain nombre de "processus biologiques" parmi un ensemble abstrait de processus réalisables. La section suivante détaille ce formalisme ainsi que tous les autres aspects du modèle.

2 Description du modèle *aevol*

La description d'un modèle tel qu'*aevol* est nécessairement imparfaite, oscillant entre description formelle des choix axiomatiques, justification de ces choix et apport de précisions sur leur implémentation. Afin de permettre différents niveaux de lecture, nous avons choisi de centrer le texte sur la description formelle et de fournir les principales justifications en notes de bas de page. Nous espérons que l'alourdissement évident que cela entraîne sera compensé par la possibilité offerte au lecteur de choisir le niveau de détail auquel il souhaite accéder. Les précisions relatives à l'implémentation ne seront fournies que ponctuellement, lorsqu'il s'agit de points sensibles, mais le code de la suite de logiciels associée au modèle est libre et disponible sur Internet : `http://bsmc.insa-lyon.fr/~cknibbe/aevol.html`. Il s'agit d'un code C++ qui a été testé sur plusieurs distributions de Linux (Debian, Mandrake, Fedora).

2.1 Vue générale

Le modèle *aevol* simule l'évolution d'une population d'individus, qui modélisent le plus simplement possible des organismes haploïdes tout en présentant les propriétés nécessaires à l'étude de la structuration du génome. Chacun de ces individus possède un chromosome circulaire[1] double-brin[2], constitué de "nucléotides" binaires, 0 étant complémentaire de 1 et réciproquement. Ce chromosome contient des séquences codantes séparées par des régions non codantes. Chaque séquence codante définit une "protéine" capable d'activer ou d'inhiber une certaine gamme de processus biologiques (abstraits). La combinaison de toutes les protéines codées dans le génome donne l'ensemble des processus biologiques que l'individu est globalement capable de réaliser. Ces capacités fonctionnelles globales

[1] La circularité ne vise pas à imiter un chromosome bactérien mais à éviter les effets de bord, ce qui est important étant donné que l'on travaille dans le modèle avec des génomes bien plus courts que les génomes réels.

[2] Cela permet d'effectuer des inversions, qui jouent potentiellement un rôle important dans l'évolution de l'ordre des gènes.

définissent ici le phénotype de l'individu (la figure II.1 illustre ce passage du génotype au phénotype, qui est détaillé dans le paragraphe suivant). On mesure ensuite l'adaptation d'un individu à l'environnement en comparant ses capacités fonctionnelles avec un ensemble – arbitrairement choisi – de fonctions à réaliser pour survivre dans cet environnement. Les individus les plus adaptés ont de plus grandes chances de se reproduire. Dans tout ce travail, la reproduction est asexuée. Lorsqu'il est répliqué, le chromosome peut subir des mutations ponctuelles et de petites insertions ou délétions, mais aussi des réarrangements intrachromosomiques pouvant affecter plusieurs séquences codantes à la fois. Il peut s'agir de translocations, d'inversions, de duplications ou encore de grandes délétions.

Après avoir initialisé la population avec des génomes aléatoires, les étapes de calcul du phénotype, sélection et reproduction avec mutations sont répétées pendant plusieurs dizaines de milliers de générations, selon la boucle générationnelle présentée à la figure II.1.

2.2 Du génotype au phénotype

Transcription

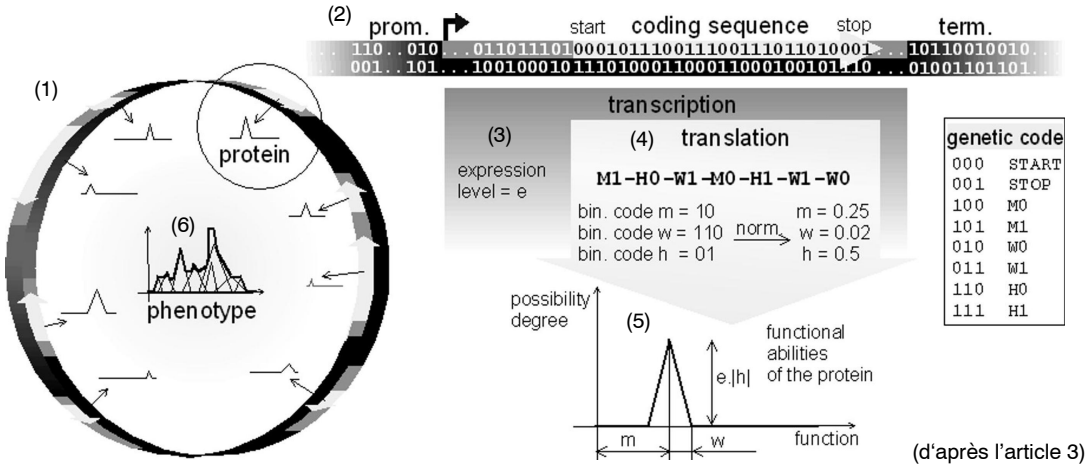
À partir de la séquence génomique, le calcul du phénotype passe d'abord par la détection de signaux d'initiation et de terminaison de la transcription, sur chacun des deux brins. Ces signaux, appelés respectivement séquences promotrices et séquences terminatrices, délimitent les régions transcrites. C'est à l'intérieur de ces régions transcrites que les signaux de traduction (et donc les séquences codantes) seront recherchés.

Les séquences promotrices sont celles qui présentent une distance de Hamming $d \leq d_{\max}$ avec une séquence consensus définie au début de la simulation. Dans toutes les expériences présentées dans ce travail, la séquence consensus[1] est 01010110011110010010110 (22 bp) et on autorise jusqu'à $d_{\max} = 4$ différences (sans insertion ni délétion). Les séquences terminatrices sont celles qui seraient capables de former une structure en tige-boucle de taille donnée, comme le font les terminateurs ρ-indépendants des bactéries[2]. Par exemple,

[1] Cette séquence est suffisamment longue pour que la probabilité d'apparition d'un gène dans une région non codante ne soit pas exagérément élevée. Elle est équilibrée en 0/1, n'est pas palindromique et ne contient pas de terminateur potentiel.

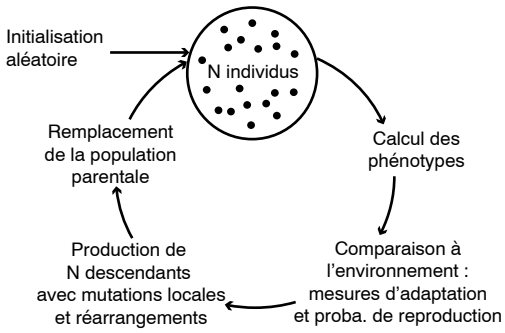
[2] Dans les premières implémentations du modèle, les séquences terminatrices étaient définies de façon analogue aux promoteurs, mais cela posait un certain nombre de problèmes. En effet, des séquences terminatrices fréquentes – donc courtes si l'on se base sur la recherche d'un consensus – sont nécessaires pour éviter que les chevauchements de régions transcrites (et donc de séquences codantes) ne soient trop abondants. Mais si une séquence courte est choisie, par exemple 11111, alors aucune séquence codante ne peut contenir ce motif, ce qui contraint significativement l'évolution. Il faut donc que les séquences terminatrices soient à la fois longues *et fréquentes*. C'est le cas si on se base sur une structure secondaire en tige-boucle, à la manière des terminateurs ρ-indépendants des bactéries. Il est tentant de spéculer sur la raison d'être de ce type de terminateurs dans les génomes réels : auraient-ils pu être indirectement sélectionnés en raison de cette propriété ?

Du génotype au phénotype



(1) Le génome est une chaîne (double-brin et circulaire) de "nucléotides" binaires, où 0 est complémentaire de 1 et réciproquement. Chaque brin contient des régions transcrites et des régions non transcrites. Chaque région transcrite peut contenir une ou plusieurs séquences codantes. (2) Les régions transcrites sont délimitées par des séquences signal, dites promotrices et terminatrices. (3) La similarité de la séquence promotrice avec un consensus détermine le niveau d'expression *e* de la région. (4) Au sein d'une région transcrite, des codons START et STOP délimitent les bornes des séquences codantes. Chaque séquence codante est lue codon par codon et traduite en 3 nombres réels, *m*, *w* et *h* à l'aide d'un code génétique artificiel. (5) Ces trois nombres réels caractérisent le sous-ensemble de "processus biologiques" (abstraits) que la protéine active (si $h>0$) ou inhibe (si $h<0$). La protéine intervient dans les processus $]m\text{-}w,m\text{+}w[$ avec une efficacité variable, comprise entre 0 et 1, appelée degré de possibilité. La fonction qui associe le degré de possibilité à chaque processus biologique est appelée distribution de possibilité. Le degré de possibilité maximal est atteint pour le processus *m*. Il dépend à la fois de la séquence de la protéine (qui détermine *h*) et de son niveau d'expression *e*. (6) Toutes les protéines sont combinées pour obtenir les capacités fonctionnelles de l'individu, exprimées sous forme d'une distribution de possibilité. Cette distribution caractérise le degré avec lequel l'individu peut réaliser chaque processus biologique. C'est cette distribution que nous appelons phénotype et qui fait donc l'objet de la sélection. Voir le texte principal pour plus de détails.

Boucle générationnelle



Les N individus de la première génération sont initialisés avec un même génome aléatoire contenant au moins une séquence codante avantageuse. A chaque génération, on calcule la distribution de possibilité phénotypique de chaque individu, et on la compare à une distribution arbitrairement choisie, représentant l'environnement. Plus les deux distributions sont proches, plus la probabilité de reproduction est grande. On tire le nombre de descendants de chaque individu en fonction de sa probabilité de reproduction et sachant qu'il faut produire au total N descendants. Pendant le processus de reproduction, divers types de mutations peuvent se produire : mutations ponctuelles, insertions, délétions, duplications, inversions... Les N descendants produits remplacent l'ensemble de la population parentale et on réitère le processus.

Fig. II.1: Vue générale du modèle *aevol*.

pour une tige de longueur 4 et une boucle de longueur 3 (valeurs utilisées dans tout ce travail), la séquence doit être de la forme $abcd***\bar{d}\bar{c}\bar{b}\bar{a}$, où $\bar{a}$ est la base binaire complémentaire de a.

L'algorithme de transcription est le suivant. On recherche d'abord toutes les séquences promotrices sur l'un des brins[1]. Puis, pour chaque séquence promotrice, on avance sur le même brin jusqu'à trouver une séquence terminatrice. Le niveau d'expression e de chaque région transcrite ainsi délimitée (et donc de toutes les séquences codantes qu'elle contiendra) dépend de la similarité de son promoteur avec le consensus[2] : $e = 1 - \frac{d}{d_{\max}+1}$. On recommence ensuite ces étapes pour l'autre brin. Notons que plusieurs promoteurs peuvent avoir le même terminateur, ce qui conduit alors à des régions transcrites chevauchantes.

Traduction

Après avoir localisé toutes les régions transcrites, on recherche dans chacune d'entre elles des signaux d'initiation et de terminaison de la traduction. Chaque fois qu'un signal d'initiation est trouvé, les bases suivantes sont lues trois par trois (codon par codon) jusqu'à trouver un signal de terminaison (si aucun signal de terminaison n'est trouvé avant la fin de la région transcrite, aucune protéine n'est produite). Plusieurs séquences codantes, chevauchantes ou non, peuvent se trouver dans une même région transcrite, ce qui permet l'apparition d'opérons.

Le signal d'initiation est constitué d'une courte séquence modélisant la séquence de Shine-Dalgarno des bactéries, qui doit précéder un codon START. Dans toutes les expériences décrites ici, le signal d'initiation complet est le motif $011011***000$ (où 000 est le codon START), et le signal de terminaison est simplement le codon STOP, 001. On remarque que pour la traduction comme pour la transcription, le signal d'initiation est globalement plus rare que le signal de terminaison. Cela a pour effet de diminuer la probabilité de création *ex nihilo* d'un gène dans une région intergénique. En effet, si l'on souhaite étudier comment la quantité d'ADN non codant évolue en raison de son rôle dans la variabilité mutationnelle du phénotype, il est impératif qu'il joue dans le modèle un rôle comparable à celui qu'il joue en réalité. Si, dans un génome réel, une mutation ponctuelle dans une séquence aléatoire a peu de chances de créer un gène, alors cela doit aussi être le cas dans le modèle.

[1] En pratique, le chromosome n'est entièrement parcouru qu'à la première génération. La liste des promoteurs est gardée en mémoire et mise à jour à chaque mutation, ce qui permet de gagner en temps d'exécution.

[2] Cette forme très simple de modulation du niveau d'expression a initialement été introduite pour permettre à des gènes dupliqués de réduire temporairement leur contribution phénotypique (par mutations ponctuelles dans le promoteur) et de dériver vers d'autres fonctions. Cependant, en pratique, il n'est pas certain que cette étape de "réduction au silence" soit indispensable pour que les individus acquièrent de nouvelles fonctions. Mais la modulation de l'expression en fonction du promoteur peut se révéler utile pour l'étude de l'organisation des gènes en opérons. Elle crée en effet un lien de co-régulation entre les séquences codantes d'une même région transcrite. L'existence d'opérons de gènes liés fonctionnellement, pouvant être activés simultanément lorsque l'environnement change, peut alors être avantageuse. C'est une hypothèse qui pourrait être explorée avec le modèle.

Il faut ensuite exprimer la contribution phénotypique de chaque séquence codante détectée, à travers les capacités fonctionnelles de la protéine. Pour cela, on utilise un formalisme répandu en intelligence artificielle, la théorie des possibilités et des ensembles flous (Zadeh, 1978; Dubois et Prade, 1980). On considère un ensemble abstrait de "processus biologiques". Cet ensemble est appelé Ω. Chaque protéine peut réaliser (ou inhiber) une partie des processus de Ω, mais avec un *degré de possibilité* variable selon le processus (le degré de possibilité est un réel compris entre 0 et 1). Une protéine donnée peut ainsi contribuer à certains processus et pas à d'autres – mais parmi les processus possibles, certains le sont plus que d'autres. À chaque protéine correspond ainsi ce qu'on appelle un *sous-ensemble flou* de Ω.

Pour garder le modèle simple, nous avons choisi de travailler dans un espace unidimensionnel[1] : l'ensemble Ω est un intervalle $[a, b]$ de $\mathbb{R}$ ($[0, 1]$ dans les expériences présentées ici), ce qui signifie que dans le modèle, un processus biologique est simplement un nombre réel. Notons que comme $\mathbb{R}$ est un ensemble ordonné, certains processus sont de fait plus proches que d'autres. Illustrons cela de la manière suivante : le processus 0,10 est plus proche du processus 0,11 que du processus 0,20, de même que de façon très informelle, le métabolisme du glucose peut être vu comme plus proche du métabolisme du lactose que des mécanismes de chemotaxie. Cette modélisation de l'espace fonctionnel est bien entendu très grossière (d'autant qu'en réalité, la notion de fonction est très subjective). Il faut cependant garder à l'esprit que l'objectif du modèle n'est pas de modéliser le fonctionnement d'un organisme vivant, mais de mettre en évidence les forces de sélection indirectes qui peuvent s'exercer sur l'organisation de l'information génétique. Des simplifications sont nécessaires pour garder le modèle compréhensible et donc utile pour cette question. L'inconvénient de ces hypothèses simplificatrices réside bien sûr dans l'interprétation des résultats, car il faut alors identifier ceux qui en sont dépendants et donc artefactuels.

Une fois munis de l'espace de processus biologiques $\Omega = [a, b]$, on peut exprimer l'action d'une protéine par sa *distribution de possibilité* f. Cette fonction de $\Omega = [a, b]$ vers $[0, 1]$ exprime, pour chaque processus x, le degré de possibilité $f(x)$ avec lequel la protéine est impliquée dans le processus. Nous avons choisi des distributions linéaires par morceaux, en forme de triangles isocèles[2] comme sur la figure II.1 (p. 71). Trois paramètres sont nécessaires pour caractériser une telle distribution :

- la position m ("mean") du triangle sur l'axe, qui correspond au processus principal de la protéine,
- la hauteur maximale H ("height"), qui détermine le degré de possibilité maximal

[1] D'un point de vue technique, l'utilisation d'un espace multi-dimensionnel ne poserait pas de problème majeur. Il suffirait de décrire chaque protéine avec plus de paramètres, ce qui nécessiterait soit des codons plus longs (de taille 4 ou 5 par exemple) soit plus de types de nucléotides (0, 1, 2, 3 au lieu de 0, 1 par exemple). Cependant, il deviendrait rapidement impossible de représenter graphiquement les distributions de possibilité des protéines, ce qui nuirait à la compréhension de l'évolution du phénotype.

[2] On pourrait choisir des formes plus complexes, par exemple avec plusieurs lobes. Tout comme le choix d'un espace Ω multi-dimensionnel, cela nécessiterait plus de paramètres pour décrire l'action de chaque protéine. Cela peut être obtenu par des codons plus longs ou un alphabet plus riche. Mais cela augmenterait la complexité du modèle, et son comportement serait plus difficile à analyser. Il est plus sage de commencer par comprendre le fonctionnement du modèle dans sa version la plus simple.

avec lequel la protéine peut contribuer à ce processus,

- et enfin la demi-largeur w ("width") du triangle, qui représente l'étendue de la gamme de processus à laquelle la protéine peut contribuer, et qui est donc une façon de modéliser sa pléiotropie[1].

La protéine peut ainsi intervenir dans les processus allant de $m - w$ à $m + w$, avec une possibilité maximale H pour le processus m. Le sous-ensemble flou de processus associé à la protéine est donc l'intervalle $]m - w, m + w[\subset \Omega$. Les paramètres m et w sont spécifiés entièrement par la séquence codante, alors que H est un paramètre composite, qui dépend à la fois de la séquence de la protéine (aspect qualitatif) et de son niveau d'expression (aspect quantitatif) : $H = e|h|$, où e est le niveau d'expression de la région et h est spécifié par la séquence codante.

La séquence codante est lue codon par codon et un code génétique artificiel est utilisé pour la traduire en trois nombres réels, correspondant à m, w et h. Comme le montre le tableau II.1, deux codons sont affectés à chaque paramètre[2]. Par exemple, le paramètre w sera calculé à l'aide des codons 010 et 011, appelés respectivement W_0 et W_1. Tous les codons W rencontrés en parcourant la séquence codante seront utilisés pour former un code binaire Gray[3] du paramètre w. Le premier digit du code Gray de w sera ainsi un 0 (resp. un 1) si le premier codon W rencontré est un codon W_0 (resp. W_1), et ainsi de suite. Dans l'exemple de la figure II.1, la séquence codante contient 3 codons W : $W_1 \ldots W_1 W_0$. Le code Gray de w sera donc 110, ce qui équivaut à 100 en binaire traditionnel, c'est-à-dire 4 en base 10. Avec ce type de codage, on obtient donc une valeur entière comprise entre 0 et $2^n - 1$ si la séquence codante contient n codons W. Une normalisation est ensuite effectuée pour ramener le paramètre dans une plage de valeurs définie. Le paramètre w, qui, rappelons-le, représente la demi-largeur du "triangle", est normalisé entre 0 et $w_{\max}$, où $w_{\max}$ est fixé au début de la simulation : la valeur entière (4 dans notre exemple) est multipliée par $\frac{w_{\max}}{2^n - 1}$. Par défaut, $w_{\max} = \frac{1}{30}$, ce qui donne pour notre exemple $w = 4 * \frac{1}{30} \frac{1}{2^3 - 1} \simeq 0{,}02$. Il est cependant possible de choisir d'autres valeurs de $w_{\max}$, ce qui permet d'explorer l'effet d'une pléiotropie plus ou moins grande (ce sera l'objet du chapitre IV).

[1] Il existe cependant dans le modèle une nécessaire proximité entre ces processus, ce qui constitue une nouvelle simplification : en réalité, une protéine peut aussi être impliquée dans des processus très différents. Ce type de pléiotropie pourrait être modélisé par une distribution multimodale, mais comme nous l'avons déjà souligné, cela ne devrait être envisagé qu'après avoir bien compris le fonctionnement du modèle avec des distributions simples.

[2] Avec en plus un codon START et un codon STOP, notre code génétique doit donc comporter au total huit codons différents. Cela implique, avec des nucléotides binaires, que les codons soient de taille 3.

[3] Le codage Gray est une variante du codage binaire traditionnel (codage de position). Il garantit que seul un digit sera modifié pour passer de i à $i + 1$, ce qui n'est pas toujours le cas avec le codage binaire usuel. Par exemple, pour passer de 3 à 4 en binaire traditionnel, il faut changer 3 digits : 011 à 100. Ainsi, dans notre cas, pour que la valeur de w (avant normalisation) passe de 3 à 4, il faudrait que la séquence codante passe de $W_0 W_1 W_1$ à $W_1 W_0 W_0$, soit trois mutations ponctuelles simultanées. Si les trois mutations se produisent séquentiellement, les individus intermédiaires présentent des valeurs éloignées de la valeur initiale et risquent d'être contre-sélectionnés. Le codage Gray évite ces "falaises de Hamming" : 3 et 4 s'écrivent respectivement 010 et 110 en Gray, donc seule une mutation sur le premier digit est nécessaire pour passer de l'un à l'autre. Ce codage facilite ainsi l'évolution des capacités fonctionnelles des protéines. La conversion du Gray vers le binaire s'effectue selon la règle suivante : chaque digit binaire est le XOR (OU exclusif) des digits Gray de poids supérieur ou égal.

Codon	Nom	Paramètre concerné	Digit ajouté à son code Gray
000	START	-	-
001	STOP	-	-
100	M_0	m	0
101	M_1	m	1
010	W_0	w	0
011	W_1	w	1
110	H_0	h	0
111	H_1	h	1

Tab. II.1: Code génétique artificiel utilisé dans tout ce travail.

La position du triangle sur l'axe (m) et sa hauteur (h) sont obtenues de façon analogue, m étant normalisée entre a et b, et h entre -1 et 1. Si h est positive (resp. négative), la protéine réalise (resp. inhibe) les processus $]m-w, m+w[$[1]. Ainsi, le signe de h détermine la nature activatrice ou inhibitrice de la protéine, et sa valeur absolue est utilisée pour calculer le degré de possibilité maximal $H = e|h|$, avec lequel la protéine réalise ou inhibe le processus m.

Interactions fonctionnelles entre protéines et calcul du phénotype

Avec le formalisme décrit ci-dessus, les sous-ensembles flous de plusieurs protéines (ou, graphiquement parlant, leurs "triangles") peuvent se recouvrir partiellement ou totalement, comme sur la figure II.2. Cela signifie que plusieurs protéines peuvent contribuer à un même processus. Nous dirons que ces protéines se trouvent en interaction fonctionnelle[2]. Ainsi, pour connaître le degré de possibilité avec lequel l'individu peut réaliser un processus donné, il faut prendre en compte toutes les protéines qui y contribuent et combiner leurs distributions de possibilité élémentaires. Nos caractères phénotypiques – les processus biologiques réalisés – sont donc potentiellement polygéniques, comme ceux des organismes réels. Il est relativement aisé ici de combiner l'action de différentes protéines, car la théorie des ensembles flous fournit une panoplie d'opérateurs pour l'union et l'intersection de tels ensembles. Pour déterminer les capacités fonctionnelles globales d'un individu, il suffit de trouver le sous-ensemble (flou) de processus biologiques qu'il peut réaliser, c'est-à-dire les processus qui sont activés ET NON inhibés, en prenant en compte toutes les protéines activatrices et toutes les protéines inhibitrices codées dans le génome. Un processus est considéré activé (resp. inhibé) s'il l'est par la protéine 1 OU par la protéine 2 OU par la protéine 3, etc. Plus formellement, si A_i est le sous-ensemble de

[1] Dans le modèle, une protéine ne peut donc pas en même temps activer certains processus et en inhiber d'autres. Il s'agit à nouveau d'une simplification qui pourrait être levée en choisissant des distributions de possibilités plus complexes et donc spécifiées par plus de trois paramètres.

[2] Le terme est ici à prendre au sens large. Il ne s'agit pas nécessairement d'une interaction physique : cela représente plus généralement la participation à une même voie métabolique, à une même cascade de transduction du signal...

la i-ème protéine activatrice et I_j celui de la j-ème protéine inhibitrice, alors l'ensemble global des processus réalisables est l'ensemble $P = (\cup A_i) \cap \overline{(\cup I_j)}$. C'est la distribution de possibilité de cet ensemble P que nous appelons ici phénotype. Cette distribution est notée f_P. Nous utilisons les opérateurs dits de Lukasiewicz[1] pour effectuer ce calcul :

$$\begin{cases} \text{NON :} & f_{\bar{A}_1}(x) & = & 1 - f_{A_1}(x) \\ \text{OU :} & f_{A_1 \cup A_2}(x) & = & \min(f_{A_1}(x) + f_{A_2}(x), 1) \\ \text{ET :} & f_{A_1 \cap A_2}(x) & = & \max(f_{A_1}(x) + f_{A_2}(x) - 1, 0) \end{cases} \quad \text{(II.1)}$$

Le calcul du phénotype revient donc intuitivement à sommer d'un côté les distributions de possibilités des protéines activatrices, à faire de même avec celles des inhibitrices, puis à soustraire la seconde somme à la première, en veillant à rester dans l'intervalle $[0, 1]$ à chaque étape. Notons que ces seuils, 0 et 1, créent des effets non linéaires. Par exemple, l'interaction d'une protéine faiblement activatrice avec une forte inhibitrice ne se traduit pas par un degré de possibilité négatif, mais par un degré nul. Si la protéine activatrice gagne progressivement en efficacité (par des mutations dans la séquence codante qui augmentent h), le degré de possibilité global reste nul dans un premier temps, puis devient brusquement positif lorsque l'efficacité de l'inhibitrice est dépassée. De même, toujours du fait du seuillage, l'efficacité conjointe de deux protéines n'est pas toujours égale à la somme de leurs efficacités élémentaires.

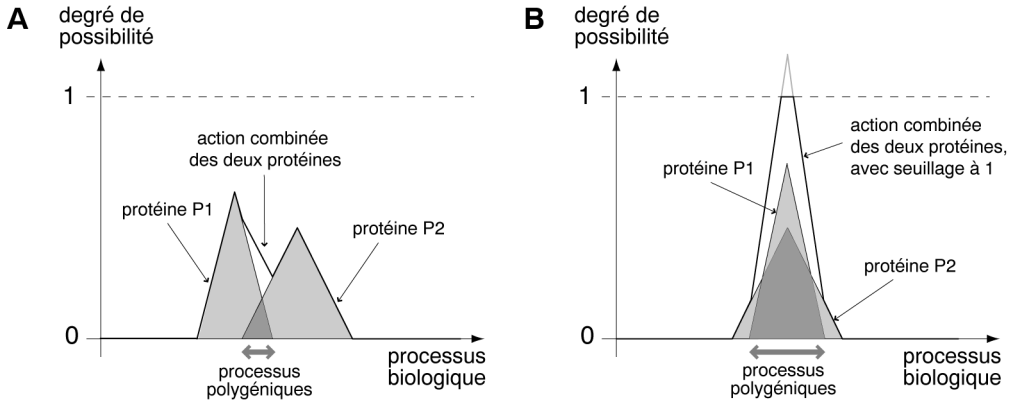


Fig. II.2: Deux exemples d'interactions fonctionnelles entre protéines. Lorsque les sous-ensembles flous de deux protéines se recouvrent, cela signifie que ces protéines interviennent ensemble dans certains processus biologiques et sont donc en interaction fonctionnelle (au sens large). Au niveau des processus polygéniques, il faut combiner les distributions de possibilité des protéines pour connaître le degré de possibilité résultant. Avec l'opérateur OU de Lukasiewicz, cela revient à sommer les deux distributions de possibilité, puis à seuiller à 1 si nécessaire, comme dans le cas B.

[1] D'autres jeux d'opérateurs flous pourraient être utilisés, mais ceux de Lukasiewicz sont préférables pour modéliser une interaction fonctionnelle. En effet, avec ces opérateurs, le degré de possibilité d'un processus polygénique ne peut pas se ramener au degré de possibilité de l'une de ses composantes : $f_{A_1 \cup A_2}(x)$ n'est ni égal à $f_{A_1}(x)$ ni à $f_{A_2}(x)$. Les opérateurs min/max souvent utilisés en logique floue reviennent au contraire à choisir comme résultante globale l'une ou l'autre des composantes, ce qui se prête mal à la modélisation d'une coopération.

2.3 Environnement, adaptation et sélection

L'environnement dans lequel la population évolue est également modélisé à travers un sous-ensemble flou $E \subset \Omega$, qui peut être vu comme l'ensemble des processus biologiques à réaliser pour survivre. La distribution de possibilité de E, notée f_E, est spécifiée au début de la simulation et peut éventuellement varier aléatoirement au cours du temps. L'adaptation de chaque individu est mesurée par l'écart entre ses capacités fonctionnelles et celles requises dans l'environnement, c'est-à-dire par l'écart entre f_P (distribution de possibilité phénotypique), et f_E (distribution de possibilité environnementale). Cet écart est noté g ("gap") et est calculé selon la formule $g = \int_\Omega |f_E(x) - f_P(x)|dx = \int_a^b |f_E(x) - f_P(x)|dx$. Comme le montre la figure II.3, cette mesure pénalise à la fois les processus "trop peu" réalisés et les processus "trop" réalisés.

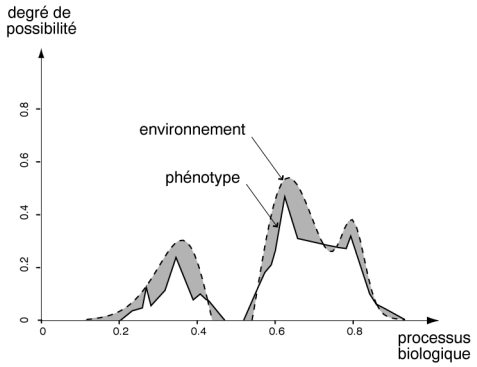


Fig. II.3: Mesure de l'adaptation d'un individu. Dans cet exemple, la distribution de possibilité de l'environnement, f_E, est tracée en pointillé. La ligne pleine correspond au phénotype de l'individu : il s'agit de sa distribution de possibilité f_P. L'adaptation est mesurée par l'écart entre les deux courbes, c'est-à-dire l'écart entre ce que l'individu peut réaliser et ce qu'il faut réaliser dans l'environnement. Graphiquement, cet écart g correspond à l'aire grisée.

La population est asexuée et gérée très simplement : elle est de taille fixe[1], N, au cours du temps et est complètement renouvelée à chaque pas de temps (génération). Il s'agit donc d'affecter une probabilité de reproduction à chaque individu, en fonction de son adaptation g, et de tirer le nombre de reproductions effectives par un tirage multinomial. Comme pour un algorithme génétique, différentes méthodes peuvent être utilisées pour calculer la probabilité de reproduction d'un individu, connaissant son adaptation g. Les trois méthodes utilisées dans ce travail sont brièvement décrites ci-dessous. On peut se reporter à (Blickle et Thiele, 1996) pour plus de détails sur les méthodes existantes.

- **"Fitness-proportionate".** La probabilité de reproduction est directement fonction de la valeur numérique de g. Ici, une probabilité proportionnelle à $1/g$ donnerait une sélection bien trop faible, et la population évoluerait quasi-exclusivement par dérive génétique. Dans ce travail, nous utilisons donc plutôt une probabilité proportionnelle à $\exp(-kg)$, où k (de l'ordre de 100 à 1 000 ici) influence l'intensité de la sélection.

[1] Une variante du modèle avec une population de taille variable et un partage explicite de ressources est actuellement à l'étude ; un prototype est en cours de développement.

Un facteur de proportionnalité normalise les probabilités de sorte à ce que la somme des probabilités de reproduction de tous les individus de la population soit égale à 1. Si cette méthode est la plus intuitive, elle ne permet pas de conserver une pression de sélection constante tout au long de la simulation. Il est donc difficile d'explorer l'effet de l'intensité de la sélection avec une telle méthode. Les méthodes basées sur le rang (voir ci-dessous), bien que moins réalistes biologiquement, permettent un meilleur contrôle (Whitley, 1989).

- **"Linear ranking".** La probabilité de reproduction dépend linéairement du rang de l'individu dans la population, après avoir trié les individus par g décroissant. Un individu de rang r aura donc une probabilité de reproduction de $\frac{1}{N}\left(\eta^- + (\eta^+ - \eta^-)\frac{r-1}{N-1}\right)$, où $\frac{\eta^+}{N}$ et $\frac{\eta^-}{N}$ sont des constantes représentant respectivement la probabilité de reproduction de l'individu le mieux adapté et celle du moins adapté. Comme la population est de taille constante, on doit vérifier $\eta^- = 2 - \eta^+$, et seule η^+ doit ainsi être spécifiée. Comme η^- doit être positive et comme la probabilité de reproduction doit augmenter avec le rang r, on doit choisir η^+ dans l'intervalle $[1, 2]$. Plus η^+ est proche de 2, plus la sélection est intense. Cependant, l'intensité maximale reste relativement faible, puisqu'avec $\eta^+ = 2$, le meilleur individu n'obtient en moyenne que deux reproductions.
- **"Exponential ranking".** Le principe est le même que pour le linear ranking, mais la relation entre la probabilité de reproduction et le rang est exponentielle, ce qui permet d'obtenir une sélection plus intense qu'avec une relation linéaire. La probabilité de reproduction est ainsi de la forme $\frac{c-1}{c^N-1}c^{N-r}$, où $c \in]0, 1[$ est une constante qui règle l'intensité de la sélection. Plus c est proche de 0, plus la sélection est intense. En pratique, pour $N = 1\,000$, il faut choisir $c \simeq 0{,}998$ pour obtenir une sélection comparable à celle obtenue en régime linéaire lorsque η^+ est proche de 2.

2.4 Mutations

Chaque fois qu'un individu est reproduit, son génome peut subir des mutations. Il peut s'agir de mutations dites locales, concernant quelques nucléotides, mais aussi de grands réarrangements chromosomiques. Au total, sept types de mutations, reprenant les grandes catégories de mutations observées dans les génomes réels (voir figure I.1, p. 24), peuvent se produire. Considérons un chromosome circulaire de L positions, numérotées de 0 à $L-1$.

- **Mutation ponctuelle.** Une position p est choisie uniformément entre 0 et $L-1$. Si cette position porte un 1, celui-ci est changé en 0. Inversement, si c'est un 0, il est changé en 1.
- **Petite insertion.** Une position p est choisie comme ci-dessus. Une courte séquence aléatoire (de 1 à 6 nucléotides dans les expériences décrites ici) est insérée dans le chromosome, entre les positions $p-1$ et p[1].
- **Petite délétion.** Une position p est choisie comme précédemment. On tire au hasard le nombre n de nucléotides à exciser (entre 1 et 6 à nouveau) et on supprime les nucléotides $\{p, p+1, ..., p+n-1\}$.

[1] Comme le chromosome est circulaire, toutes les positions s'entendent modulo L.

- **Grande délétion.** Deux positions p_1 et p_2 sont choisies sur le chromosome, puis le segment chromosomique allant de p_1 à p_2 (incluses) dans le sens horaire est excisé. Comme le chromosome est circulaire, p_2 peut être inférieure à p_1, et dans ce cas, il faut en fait exciser les segments $\{p_1, ..., L-1\}$ et $\{0, ..., p_2\}$.
- **Inversion.** Comme la grande délétion, mais le segment $\{p_1, ..., p_2\}$ est inversé et non excisé. La longueur du chromosome est inchangée.
- **Duplication.** Deux positions p_1 et p_2 sont choisies sur le chromosome, puis le segment chromosomique allant de p_1 à p_2 (incluses) dans le sens horaire est copié. Une position p_3 est choisie sur le chromosome, et enfin le segment copié est inséré entre les positions $p_3 - 1$ et p_3. L'orientation du segment est conservée.
- **Translocation.** Comme la duplication, mais le segment $\{p_1, ..., p_2\}$ est excisé (et non copié) avant d'être ré-inséré entre les positions $p_3 - 1$ et p_3. La longueur du chromosome est finalement inchangée. Si la duplication est un "copier-coller", alors la translocation est le "couper-coller".

Pour choisir les points de rupture des réarrangements (positions p_1 et p_2), le plus réaliste serait de rechercher les paires de séquences similaires. Cependant, cela serait très coûteux en temps de calcul. Nous faisons donc l'hypothèse simplificatrice que toute position peut servir de point de rupture. Nous avons considéré deux lois pour la distance l_{seg} entre p_1 et p_2 :

- une loi uniforme entre 1 et L. Dans ce cas, p_1 et p_2 sont toutes deux tirées uniformément entre 0 et $L-1$. Ce mécanisme modélise l'action de la recombinaison homologue, qui semble pouvoir agir sur de grandes distances (Rocha, 2003a).
- une loi quasi-géométrique de paramètre q, où $\text{prob}(l_{\text{seg}} = k) = \frac{1}{1-(1-q)^L} q(1-q)^{k-1}$. Comme avec une loi géométrique, la probabilité que le segment réarrangé soit de taille k décroît exponentiellement avec k, la vitesse de décroissance dépendant de q. La différence réside dans le facteur de normalisation $\frac{1}{1-(1-q)^L}$, qui permet de borner la longueur du segment entre 1 et L. Dans ce cas, on tire d'abord p_1 uniformément entre 0 et $L-1$, puis on tire l_{seg} selon la loi géométrique normalisée pour obtenir p_2. Ce mécanisme modélise l'action de la recombinaison illégitime, qui agit à un niveau plus local que la recombinaison homologue (Rocha, 2003a).

Notons que dans un génome réel, la distribution de la taille des segments réarrangés est certainement plus complexe que l'une ou l'autre de ces lois. Plusieurs mécanismes – recombinaison homologue, recombinaison illégitime, transposition, duplication complète du génome, amplification de gènes par rétroposition.... – agissent simultanément et à des échelles différentes, ce qui rend vraisemblablement la distribution multimodale. Il est cependant difficile de caractériser cette distribution : on n'a bien souvent accès qu'aux événements fixés, alors que c'est la distribution de taille des réarrangements spontanés dont nous avons besoin ici.

Les taux de mutation sont fixés par nucléotide pour chacun des sept types de mutation, et les sept nombres de mutations à réaliser pendant la réplication sont tirés selon des lois binomiales à L essais. Par exemple, si u_{ponct} est la probabilité qu'un nucléotide subisse une mutation ponctuelle, alors on réalisera $\mathcal{B}(L, u_{\text{ponct}})$ mutations ponctuelles. De même, on réalisera $\mathcal{B}(L, u_{\text{inv}})$ inversions, etc. Ainsi, pour tous les types de mutation, le nombre

d'événements moyen augmente avec la taille du génome. Pour les réarrangements, cela permet de modéliser simplement le fait que plus le génome est grand, plus on a de chances de trouver des séquences répétées susceptibles de les médier.

L'algorithme est le suivant : lorsqu'on reproduit un individu, on tire les quatre nombres d'événements à réaliser pour les quatre types de réarrangements. On effectue ensuite tous les réarrangements dans un ordre aléatoire. Pendant cette étape, la longueur du génome peut varier. Ainsi, si l'on veut effectuer une inversion après une duplication qui aurait fait passer la taille de L à L', on tire les positions p_1 et p_2 entre 0 et $L'-1$. Les réarrangements successifs ne sont donc pas indépendants. Une fois tous les réarrangements effectués, on tire les trois nombres de mutations locales (mutations ponctuelles, petites insertions et petites délétions) et on effectue toutes les mutations locales, toujours dans un ordre aléatoire.

2.5 Échange de matériel génétique entre individus

L'échange de matériel génétique entre individus, appelé "crossover" en évolution artificielle, est techniquement possible : il est tout à fait envisageable de construire un nouveau chromosome à partir de morceaux de deux chromosomes parentaux. Cependant, comme le nombre et l'ordre des gènes sont variables, il faudrait effectuer une recherche de similarité pour réaliser des échanges d'allèles pertinents, ce qui est très coûteux en temps de calcul. Une alternative serait de modéliser les échanges de matériel génétique non pas comme des crossing-overs, mais comme des insertions, à la manière de la transduction bactérienne[1]. Le problème est que cela crée un biais vers la croissance du génome. Nous avons préféré étudier dans un premier temps l'évolution des génomes sans biais, en nous limitant à des populations strictement asexuées (ce sera le cas dans tout ce travail). Une fois le comportement du modèle connu dans ce cas simple, on pourra ensuite étudier spécifiquement l'effet d'un mécanisme de transduction.

3 Paramétrage et comportements typiques

3.1 Paramétrage

Le paramétrage d'un modèle individu-centré est en général une étape délicate. Ici, le nombre de paramètres en jeu est tel qu'une exploration exhaustive de l'espace des paramètres n'est pas envisageable. Certaines valeurs sont exclues d'emblée, car elles conduiraient à des comportements artefactuels. Par exemple, comme nous l'avons souligné précédemment, les signaux d'initiation de la transcription et de la traduction sont volontairement longs, de sorte à ce que la probabilité de création d'un gène *ex nihilo* ne soit pas

[1] La transduction est le processus par lequel de l'ADN est transféré d'une bactérie à une autre par un virus.

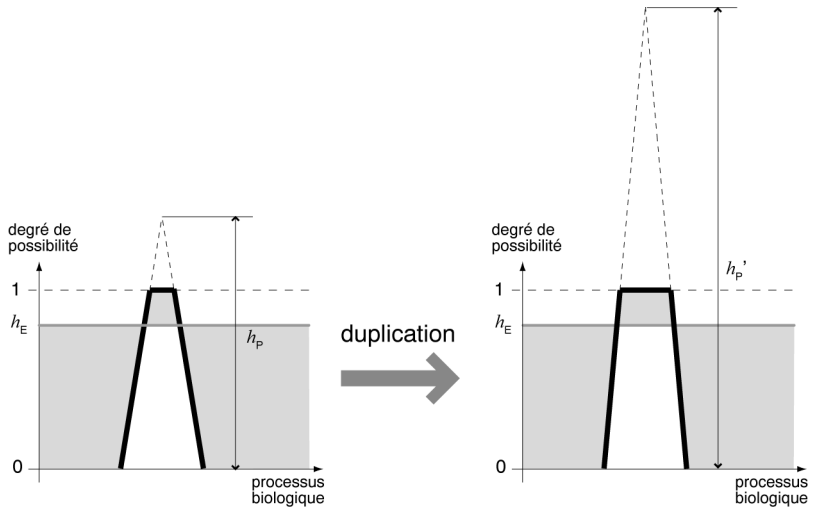


Fig. II.4: Avantage artefactuel à la duplication de gènes lorsque la distribution de possibilité de l'environnement excède 0,6. Considérons par exemple une distribution environnementale f_E constante, égale à h_E pour tout processus biologique (représentée en gris). Considérons également un individu qui possèderait n protéines activatrices identiques, dont la somme h_P dépasse 1. Du fait du seuillage à 1, le phénotype résultant serait en forme de trapèze (ligne noire). L'écart g entre le phénotype et l'optimum environnemental correspond à l'aire en gris clair. Après duplication d'un ou plusieurs gènes, la hauteur du phénotype avant seuillage devient $h_P\prime > h_P$. On peut montrer géométriquement que si $h_E > \frac{\sqrt{2}}{2}$, alors l'écart g diminue lorsque h_P augmente. Il y a donc un avantage systématique à la duplication, ce qui induit une explosion difficilement réversible de la taille du génome. En pratique, avec un protéome plus complexe, comprenant aussi des protéines inhibitrices, l'effet se produit dès que $h_E > 0{,}6$. Cet effet est un artefact induit par le seuillage à 1, qui répond davantage aux contraintes imposées par la théorie des possibilités qu'à une réalité biologique. Nous l'évitons donc en choisissant des distributions environnementales qui n'excèdent pas 0,6.

exagérément forte. Il faut également veiller à ce que la distribution de possibilité choisie pour l'environnement, f_E, n'induise pas de phénomènes irréalistes. Le formalisme des ensembles flous implique en effet un seuillage de la distribution phénotypique f_P à 1, ce qui induit une explosion artefactuelle du génome si f_E est supérieure à 0,6 (voir figure II.4). Pour éviter cet effet, nous nous limitons donc dans ce travail à des distributions environnementales qui n'excèdent pas 0,6.

Pour d'autres paramètres, ce sont des raisons pratiques, liées au temps de calcul et à l'espace mémoire notamment, qui nous empêchent d'utiliser les valeurs observées dans les populations naturelles. C'est le cas par exemple de la taille de la population, qui doit rester de l'ordre du millier pour que les temps de simulations restent raisonnables. De même, il n'est pas envisageable d'utiliser des tailles de génomes réelles. La taille des génomes que l'ont fait évoluer ici varie de quelques centaines de paires de bases à quelques dizaines de milliers de paires de bases. Le temps de simulation est aussi très lié aux taux de mutations : il faut des taux relativement élevés – de l'ordre de 10^{-6} à 10^{-4} par paire de base – pour que l'évolution du phénotype soit significative après quelques jours de calcul. La comparaison quantitative avec des données réelles est donc exclue, mais il faut garder à l'esprit qu'une telle comparaison n'aurait que peu de sens étant donné les simplifications

effectuées au niveau fonctionnel. L'objectif de ce modèle réside davantage dans l'étude qualitative des principes d'organisation des génomes.

3.2 Comportement général : une simulation typique

Dans les plages de paramètres pertinentes et calculables, le comportement du modèle présente un certain nombre de régularités. Pour les illustrer, nous détaillons ici une simulation ("run") représentative, dont les paramètres figurent dans le tableau II.2.

Paramètre	Symbole	Valeur
Taille de la population	N	1 000 individus
Taille initiale des génomes	L_{init}	5 000 bp
Séquence promotrice	-	01010110011100100101 10, $d_{\max}$ = 4 différences autorisées
Séquence terminatrice	-	De la forme $abcd***\bar{d}\bar{c}\bar{b}\bar{a}$
Signal d'initiation de la traduction	-	011011***000
Signal de terminaison de la traduction	-	001
Code génétique	-	Voir tableau II.1
Ensemble des processus biologiques	Ω	$[0, 1]$
Pléiotropie maximale des protéines	$w_{\max}$	$\frac{1}{30}$
Distribution de possibilité de l'environnement	f_E	Trimodale, comme sur la figure II.3, et fixe au cours du temps
Méthode de sélection	-	"Linear ranking"
Intensité de la sélection	η^+	1.998
Taux de mutation ponctuelle	u_{ponct}	10^{-5} par bp
Taux de petite insertion	u_{ins}	10^{-5} par bp
Taux de petite délétion	u_{del}	10^{-5} par bp
Taux de grande délétion	u_{gdel}	10^{-5} par bp
Taux d'inversion	u_{inv}	10^{-5} par bp
Taux de duplication	u_{dup}	10^{-5} par bp
Taux de translocation	u_{transloc}	10^{-5} par bp
Loi de la longueur des petits indels	-	Loi uniforme entre 1 et 6 bp
Loi de la longueur des segments réarrangés	l_{seg}	Loi uniforme entre 1 et L

Tab. II.2: Paramètres utilisés pour la simulation détaillée dans cette section. Toutes les simulations décrites dans ce travail utilisent les mêmes paramètres de transcription et de traduction.

Suivi au cours de la simulation

Avec ce jeu de paramètres, la simulation de 20 000 générations dure environ trois jours sur un processeur cadencé à 2 GHz. La figure II.5 montre concrètement les sorties graphiques

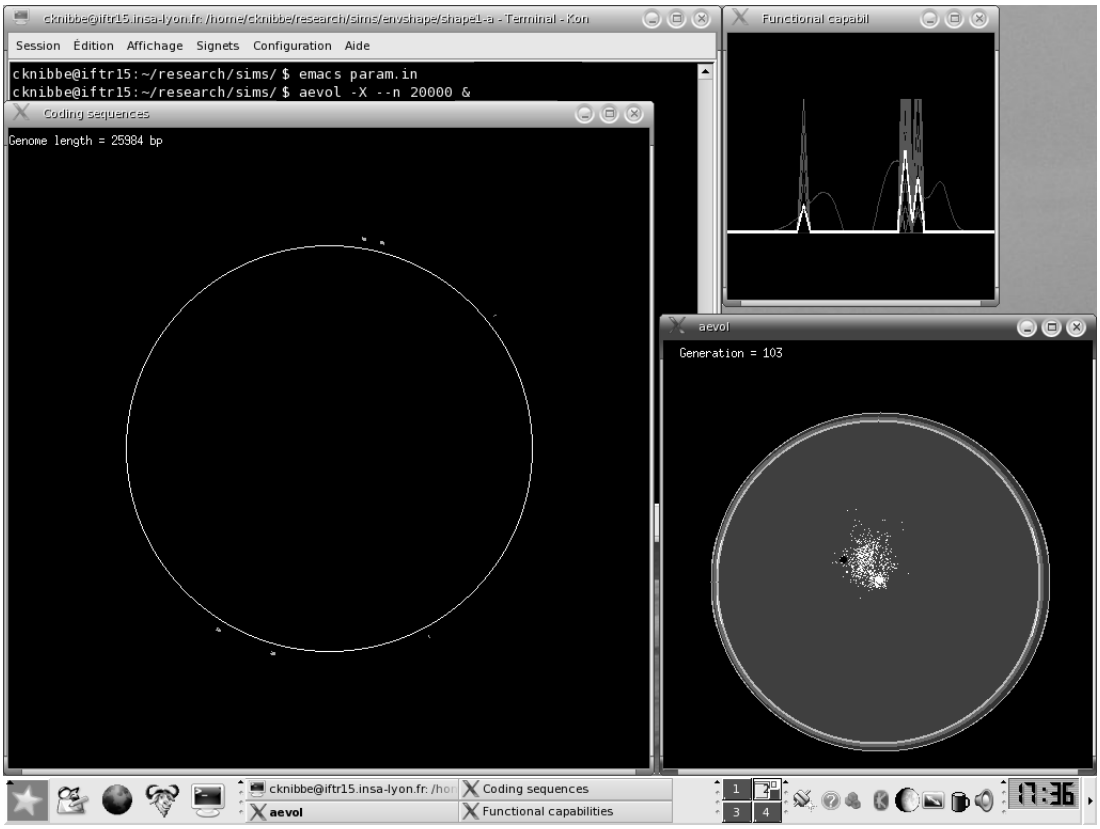


Fig. II.5: Capture d'écran au cours d'une simulation. La simulation est lancée en ligne de commande. La fenêtre principale, en bas à droite, affiche le compteur des générations et montre une représentation abstraite de la population qui permet de distinguer visuellement les éventuelles sous-populations (chaque individu est un point blanc, sauf le meilleur qui est en noir ; lorsqu'on reproduit un individu, on positionne le descendant à l'endroit où se trouvait son progéniteur, sauf s'il a subi des mutations, auquel cas on l'éloigne d'une distance proportionnelle au nombre de mutations). La fenêtre en haut à droite affiche la superposition des phénotypes des individus de la génération courante (le meilleur étant en blanc), ainsi que l'environnement. Enfin, la grande fenêtre à gauche représente le génome du meilleur organisme de la génération courante : le cercle blanc représente le chromosome et les petits arcs de cercle les séquences codantes. On pourra cependant se référer aux figures suivantes pour des représentations plus lisibles des génomes obtenus.

accessibles pendant que la simulation se déroule.

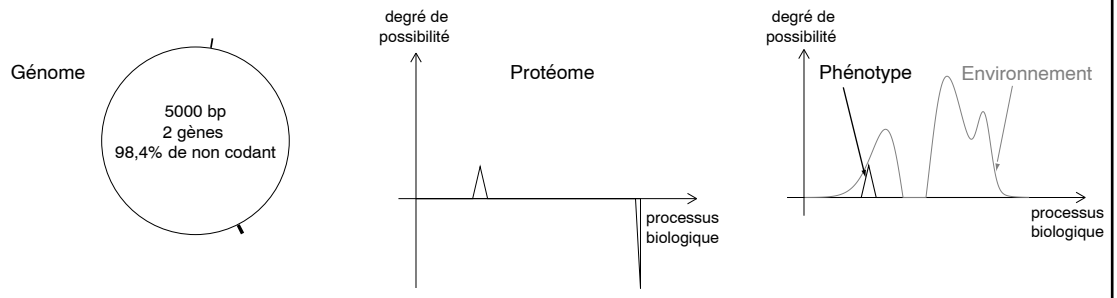
La figure II.6 montre le génome aléatoire utilisé pour initialiser l'ensemble de la population. La grande majorité de ce génome est non codante. Il ne comporte que deux courtes séquences codantes, l'une codant pour une protéine activatrice et l'autre pour une protéine inhibitrice. Les capacités fonctionnelles de l'individu sont donc très limitées. Cependant, au cours des générations suivantes, les séquences codantes initiales – ainsi que les régions non codantes adjacentes – sont rapidement dupliquées, ce qui cause une explosion de la taille du génome (figure II.7A). Les séquences codantes copiées, en subissant des mutations ponctuelles et des petites insertions ou délétions, codent pour des protéines avec de nouvelles capacités fonctionnelles : par exemple, lorsqu'un codon M est muté, la protéine

se translate le long de l'axe des processus biologiques, réalisant ainsi de nouveaux processus. De même, son efficacité peut être affinée en modifiant les codons H ou le promoteur. On retrouve donc un mécanisme d'acquisition de gènes par duplication puis divergence fonctionnelle, similaire à celui qui se produit vraisemblablement dans les génomes réels (Ohno, 1970; Venter *et al.*, 2001; Betran et Long, 2002; Eichler et Sankoff, 2003; Cannon *et al.*, 2004; Dujon *et al.*, 2004; Gevers *et al.*, 2004). Ce mécanisme permet aux individus de devenir rapidement mieux adaptés, comme le montre la figure II.8.

L'explosion de la taille du génome est cependant temporaire : des séquences codantes (figure II.7C), mais aussi et surtout des séquences non codantes (figure II.7B), sont perdues. Après environ 6 000 générations, la taille du génome se stabilise autour d'une valeur d'équilibre, de l'ordre de 10 000 paires de bases ici. Cette valeur d'équilibre est indépendante de la taille choisie pour l'initialisation. Il est important de noter que cette taille d'équilibre ne correspond pas à l'élimination de tout le non codant. Bien au contraire, dans cet exemple, il subsiste environ 80% de non codant à la fin de la simulation (figure II.6). Comme il n'y a pas de biais mutationnel vers la production de non codant, ni d'avantage sélectif direct à sa présence, le maintien d'une telle quantité de non codant est intriguant. Comme nous l'avons vu dans le chapitre précédent, la cause de ce maintien peut être recherchée dans les pressions de sélection indirecte qui s'exercent sur la variabilité mutationnelle du phénotype – ce sera l'objet des chapitres suivants.

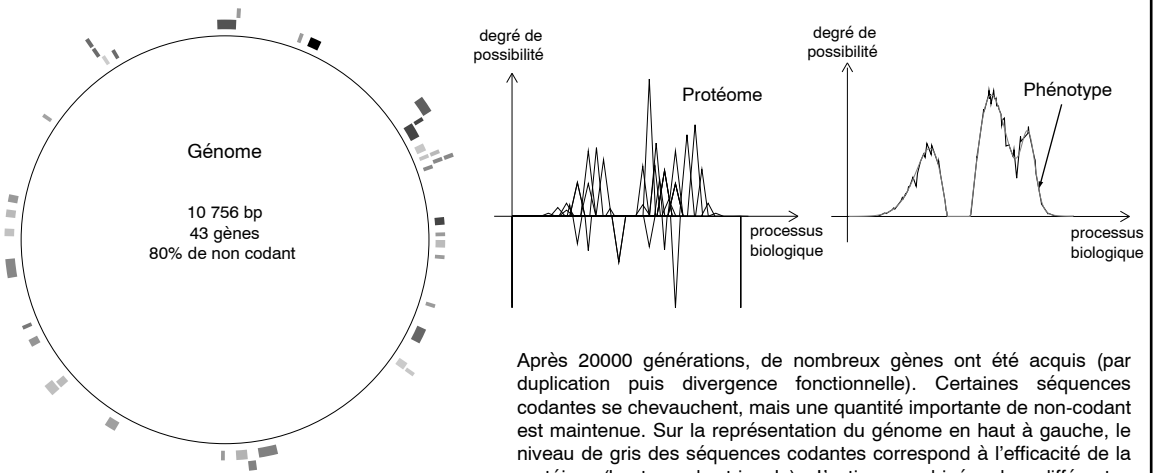
Une fois la quantité de non codant stabilisée, l'adaptation des individus continue de progresser, mais de plus en plus lentement (figure II.8). De nouveaux gènes apparaissent par duplication-divergence, mais de façon beaucoup moins massive qu'initialement (figure II.7C) : la plupart des duplications deviennent en effet délétères lorsque le phénotype approche de l'optimum environnemental. Certaines séquences codantes apparaissent aussi par mutations locales dans des régions transcrites, ce qui peut causer des chevauchements (figure II.6). Les gènes existants sont améliorés par des mutations locales, et notamment par de petites insertions : comme le montre la figure II.7D, la longueur moyenne des gènes augmente progressivement. Cela est dû au fait que la précision des paramètres m, w et h des protéines est directement corrélée au nombre de codons dans la séquence codante. L'augmentation de la taille des gènes ne peut donc être considérée comme un résultat exportable aux génomes réels ; il s'agit d'un artefact lié aux simplifications réalisées dans la transition du génotype au phénotype.

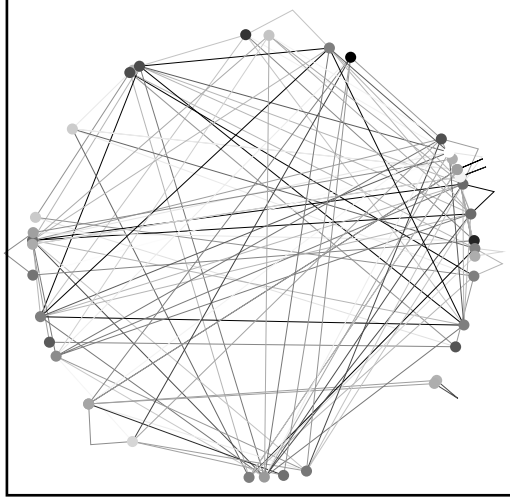
Génération 0 - Ancêtre commun



Le génome initial est une séquence aléatoire de 5000 paires de bases, de phénotype non nul. En général, il ne contient que très peu de gènes, comme sur cet exemple. Ici, sur les deux séquences codantes (bâtonnets situés autour du cercle), l'une correspond à une protéine activatrice, et l'autre à une protéine inhibitrice (représentée ici par un triangle de hauteur négative). Le reste du génome est non codant. Les capacités fonctionnelles globales de l'individu (phénotype) sont très éloignées de l'optimum environnemental.

Génération 20000 - Meilleur individu final





Après 20000 générations, de nombreux gènes ont été acquis (par duplication puis divergence fonctionnelle). Certaines séquences codantes se chevauchent, mais une quantité importante de non-codant est maintenue. Sur la représentation du génome en haut à gauche, le niveau de gris des séquences codantes correspond à l'efficacité de la protéine (hauteur du triangle). L'action combinée des différentes protéines confère à l'individu des capacités fonctionnelles globales qui sont en bien meilleure adéquation avec l'environnement qu'à t=0. Notons la présence d'une protéine inhibitrice au niveau de la zone nulle de l'environnement : cette protéine n'a aucun effet direct sur le phénotype, sauf si une mutation fait apparaître une protéine activatrice dans cette zone. Cela pourrait donc refléter la sélection indirecte d'un mécanisme de robustesse mutationnelle.

Dès que le génome contient des protéines en interaction fonctionnelle (recouvrement des triangles), on peut, comme ci-contre, projeter le réseau d'interactions sur le chromosome. Sur cette représentation, chaque point correspond à une séquence codante et est coloré en fonction de l'aire du triangle associé. Deux points sont reliés par un trait si les deux protéines sont en interaction fonctionnelle. Plus le trait est foncé, plus l'interaction est forte. Cette représentation permet donc de voir si les gènes qui interagissent tendent à se rapprocher les uns des autres, ou, au contraire, à se disperser régulièrement sur le chromosome. On peut également voir comment se répartissent les gènes qui ont un grand poids dans le phénotype (ci-dessus : triangles de grande aire, ci-contre : points foncés).

Fig. II.6: Individu initial et individu final dans un run typique (paramètres : voir tableau II.2).

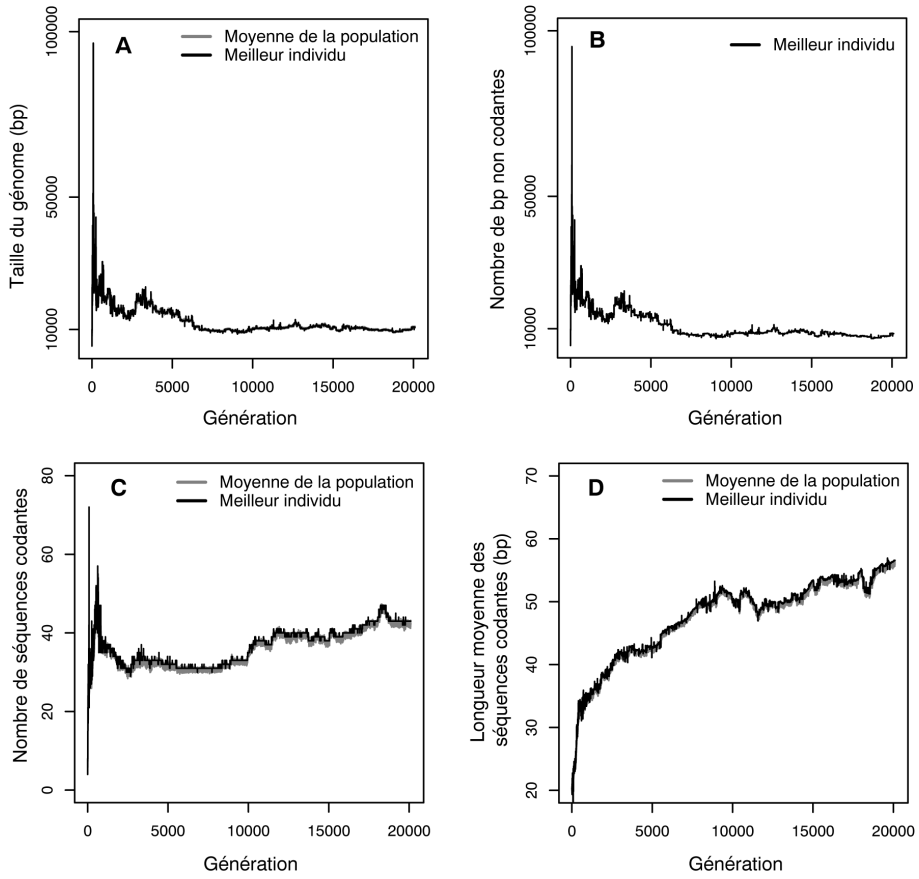


Fig. II.7: Évolution de quelques caractéristiques génomiques dans un run typique. Voir les paramètres dans le tableau II.2.

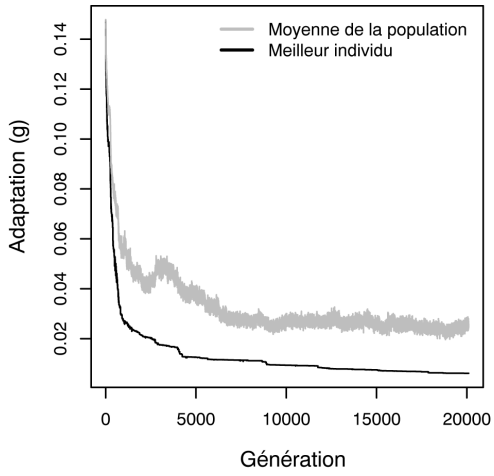


Fig. II.8: Évolution de l'adaptation (mesurée par l'écart g entre les distributions de possibilité phénotypique et environnementale) dans un run typique. Voir les paramètres dans le tableau II.2.

Analyses post-simulation

L'un des intérêts de l'approche mise en œuvre réside dans la connaissance exhaustive des ancêtres, des relations de parenté et des mutations qui se sont produites. Une fois la simulation achevée, on peut donc identifier la lignée ancestrale du meilleur individu final, ainsi que les mutations qui se sont produites sur cette lignée (figure II.9). Ces mutations (hormis celles qui se sont produites pendant les toutes dernières générations), sont celles qui ont été fixées dans la population. On peut alors "rejouer le film de l'évolution" en montrant la succession des ancêtres et les mutations qui les séparent (figure II.10).

Nous pouvons donc comparer la succession des ancêtres avec celle des meilleurs individus. La figure II.11 montre que le t-ième ancêtre des individus finaux n'est pas nécessairement le meilleur individu de la génération t, et que certaines mutations favorables sont perdues. Cela peut être simplement dû aux effets stochastiques de l'échantillonnage des reproducteurs (dérive génétique) : même le meilleur individu court le risque de ne pas se reproduire. Mais cela peut aussi être dû au fait que tous les descendants produits par cet individu ont subi des mutations délétères : dans ce cas, il est possible qu'une autre lignée, plus robuste, soit finalement la plus représentée au pas de temps suivant. Les différences entre la courbe des meilleurs individus et celle des ancêtres sont donc potentiellement les premiers indices d'un effet de sélection indirecte du niveau de variabilité mutationnelle. Elles suggèrent que le niveau d'adaptation n'est pas la seule clé du succès évolutif à long terme.

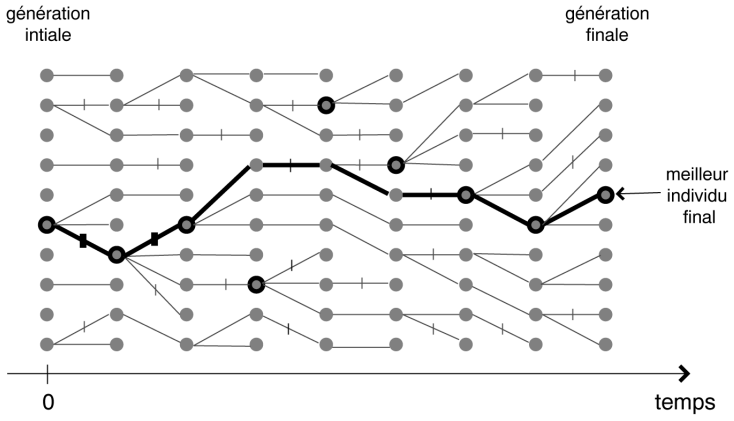


Fig. II.9: Identification de la lignée ancestrale et des mutations fixées. Tout au long de la simulation, on garde en mémoire les relations de parenté (traits horizontaux et obliques) entre les individus (points), ainsi que les mutations (traits verticaux) qui se produisent lors des reproductions. Une fois la simulation terminée, on peut retracer la lignée ancestrale du meilleur individu final (traits forts). Les meilleurs individus à un moment donné (points entourés par un cercle noir) ne sont pas toujours dans cette lignée, ce qui signifie que la valeur immédiate de l'adaptation n'est pas la seule clé du succès évolutif à long terme. Les mutations fixées, c'est-à-dire portées par tous les individus finaux, se trouvent sur la lignée ancestrale (traits verticaux forts). Notons que les toutes dernières mutations sur la lignée ancestrale ne sont pas encore fixées.

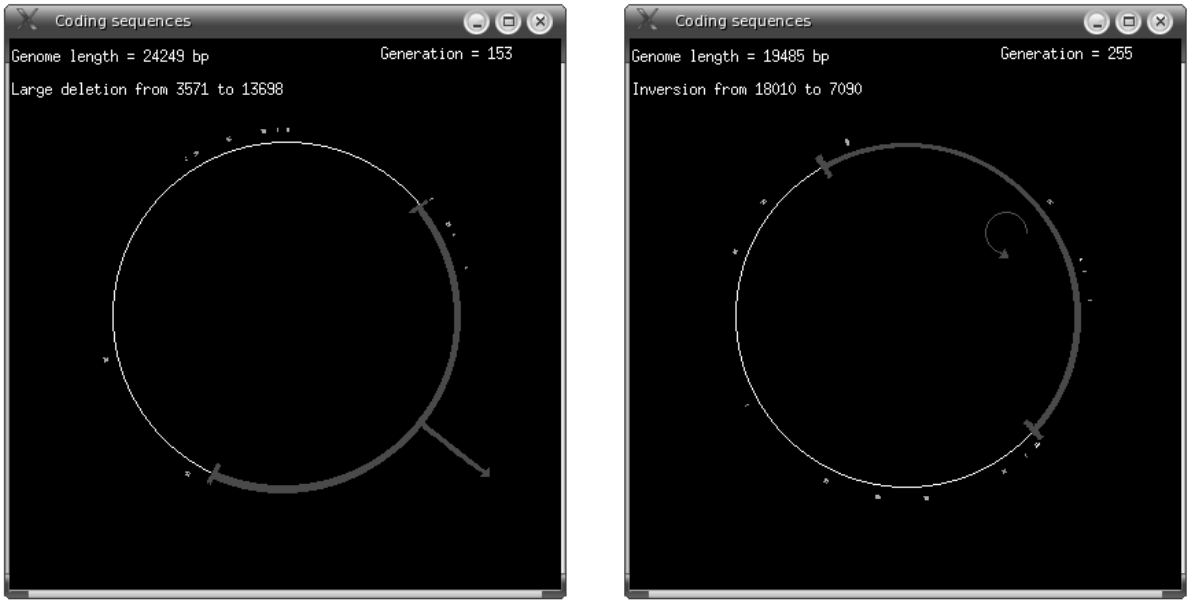


Fig. II.10: Captures d'écran pendant le "film" de l'évolution. Dans cet exemple, on voit que le génome a subi une grande délétion à la génération 153 et que cette délétion a affecté trois séquences codantes. On voit également qu'un bloc de cinq séquences codantes a été inversé à la génération 255.

La figure II.12 montre qu'il existe aussi des différences entre les caractéristiques des meilleurs et celles des ancêtres au niveau génomique. De manière générale, pour caractériser l'évolution du génome, nous préférerons nous fonder sur la lignée ancestrale des individus finaux plutôt que sur les données capturées au cours de la simulation (caractéristiques du meilleur individu par exemple). En effet, toutes les variations observées sur la lignée ancestrale sont attribuables à des mutations fixées, alors que les variations sur la séquence des meilleurs peuvent aussi être dues à des changements de lignée.

Les mutations qui se produisent dans la lignée ancestrale peuvent être analysées pour quantifier la dynamique de l'organisation du génome. Comme le montre la figure II.13, la majorité des événements fixés sont des mutations locales (mutations locales, petites insertions et délétions). Viennent ensuite les inversions et les translocations, qui modifient potentiellement l'ordre des gènes mais pas la taille du génome. Enfin, les duplications et les grandes délétions fixées sont peu nombreuses et majoritairement groupées dans la phase initiale d'acquisition de gènes puis de réduction du génome. Cela reflète la stabilisation de la taille du génome observée après quelques milliers de générations. Rappelons que tous les types d'événements se produisent spontanément avec la même fréquence (10^{-5} par bp en l'occurrence). Nous observons donc bien une contre-sélection plus prononcée des duplications et des grandes délétions. Cette contre-sélection peut simplement être due à des effets délétères directs (doublement du niveau d'expression, perte de gènes...), mais nous nous intéresserons davantage au cas où les effets délétères sont indirects, liés à la variabilité mutationnelle du phénotype.

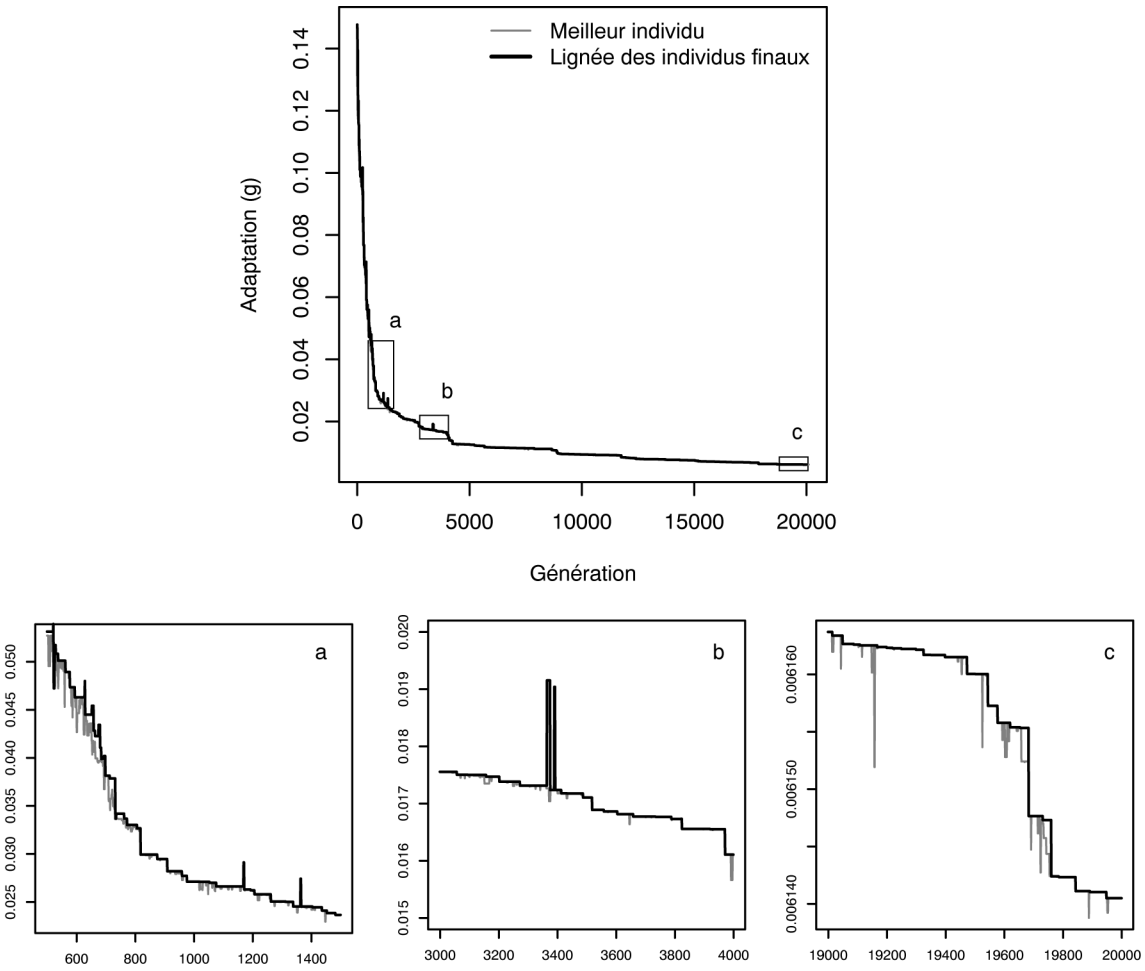


Fig. II.11: Évolution de la mesure d'adaptation le long de la lignée ancestrale du meilleur individu final (en noir), comparée à la succession des adaptations des meilleurs individus (en gris). Les encarts a, b, c sont des zooms de la courbe principale. On voit que les deux courbes ne sont pas toujours confondues, ce qui signifie que l'individu qui sera finalement l'ancêtre de la génération finale n'est pas toujours le meilleur de sa génération.

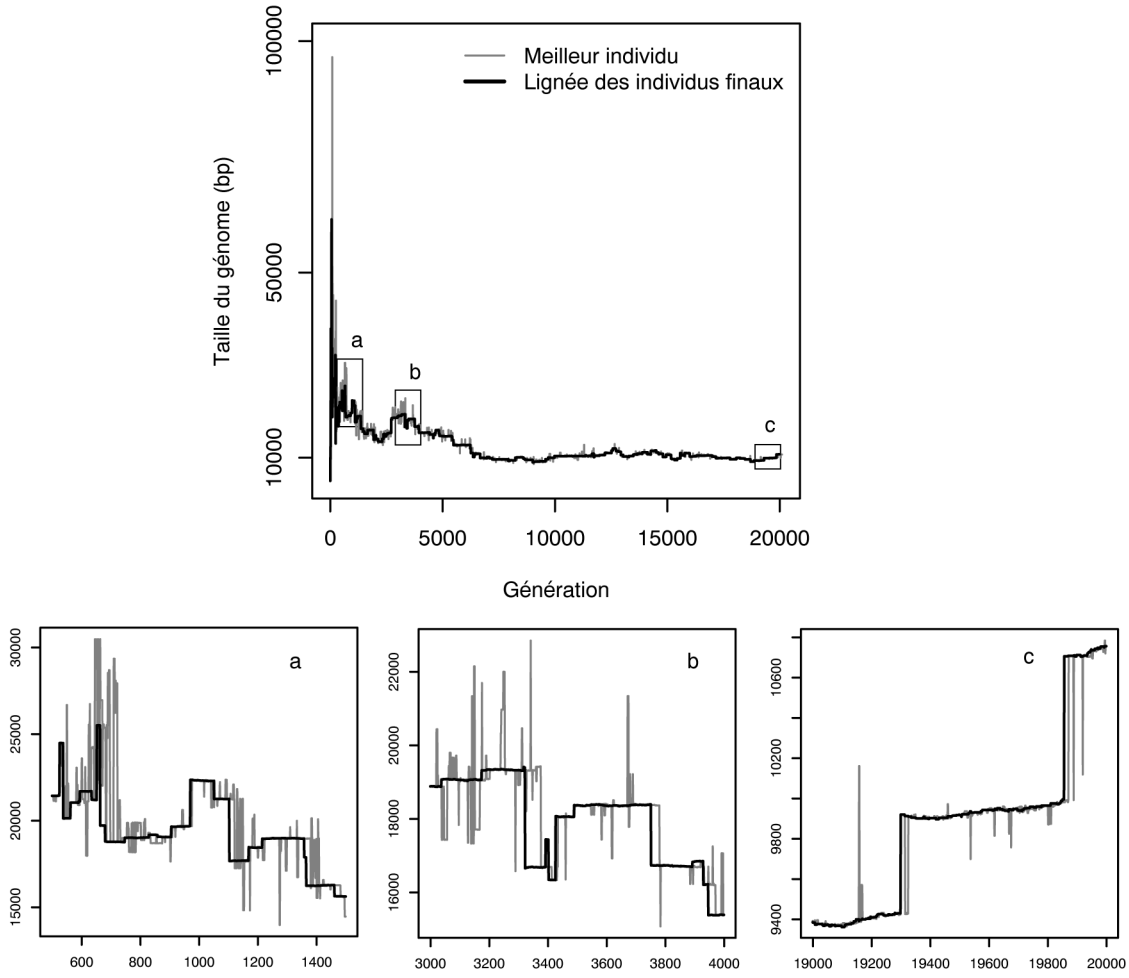


Fig. II.12: Évolution de la taille du génome le long de la lignée ancestrale du meilleur individu final (en noir), comparée à la succession des tailles des génomes des meilleurs individus (en gris). Là encore, les deux courbes ne sont pas confondues. La courbe noire, celle de la lignée ancestrale, est globalement plus stable que la courbe grise, qui reflète la compétition entre différentes lignées de taille de génome différentes : le meilleur individu peut ainsi se trouver dans une lignée donnée à la génération t, puis dans une autre la génération $t+1$. Ces changements de lignée causent des variations dans la courbe grise qui ne sont pas nécessairement dues à des mutations affectant la taille du génome.

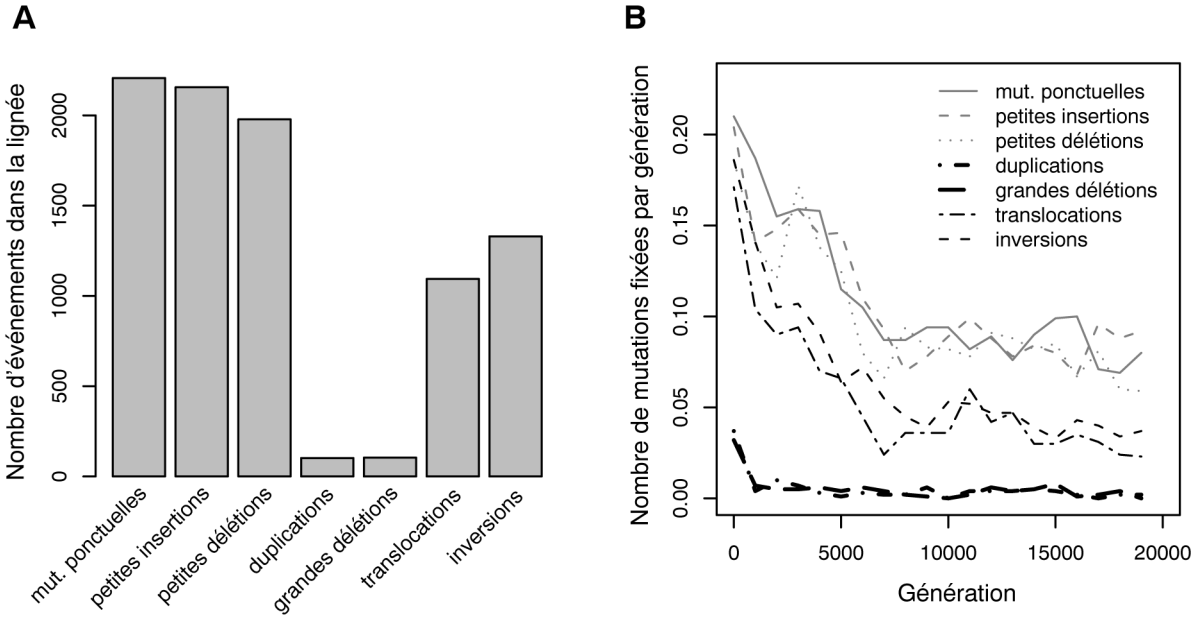


Fig. II.13: Caractérisation des mutations fixées, c'est-à-dire des mutations qui se sont produites sur la lignée ancestrale du meilleur individu final. Le graphe A montre le nombre total d'événements de chaque type, sur l'ensemble des 20 000 pas de temps. Le graphe B représente l'évolution au cours du temps du taux de fixation de chaque type de mutation. Ce graphe est obtenu en calculant le nombre d'événements dans chaque fenêtre de 1 000 générations et en le divisant par 1 000.

3.3 Éléments d'analyse de sensibilité

D'un point de vue qualitatif, les comportements décrits – acquisition de gènes par duplication, amélioration de l'adaptation, stabilisation de la quantité de non-codant, augmentation de la longueur des gènes, apparition de gènes chevauchants... – dépendent peu des valeurs choisies pour les paramètres. Celles-ci peuvent en revanche jouer sur la vitesse et l'ampleur des phénomènes. C'est l'influence des paramètres biologiquement pertinents, comme les taux de mutation (chapitre III) ou l'effet moyen des mutations (chapitre IV), qui nous intéresse en premier lieu. Cependant, il convient d'abord d'évaluer la sensibilité du modèle à des paramètres plus arbitraires, pour savoir ensuite distinguer ce qui est biologiquement pertinent de ce qui est artefactuel. Comme nous souhaitons garder le volume de ce manuscrit raisonnable, nous nous limitons ici aux deux paramètres qui, à notre avis, mettent en jeu le plus d'arbitraire : la forme de la distribution de possibilité de l'environnement et la méthode de sélection.

Effet de la forme de l'environnement

Pour tester l'effet de la forme de la distribution de possibilité de l'environnement, nous avons choisi six allures de f_E différentes (de même aire) et simulé l'évolution de 3 populations indépendantes dans chaque cas. Les six allures choisies sont représentées sur la figure II.14, les autres paramètres étant identiques à ceux qui figurent dans le tableau II.2.

Comme le montre la figure II.14, on retrouve dans tous les cas le phénomène de stabilisation de la taille du génome. Pour 17 des 18 simulations, cette stabilisation fait suite à un pic initial dû à des duplications massives. La hauteur et la largeur de ce pic sont variables d'un environnement à l'autre, et, pour un environnement donné, d'une répétition à l'autre. La variabilité inter-environnement de cette phase initiale est vraisemblablement due au fait que les six fonctions choisies sont de difficultés différentes[1], alors que la variabilité entre répétitions reflète plutôt l'effet fondateur des premières générations. Cette phase initiale, durant laquelle "tout est à faire", est qualitativement différente de la phase d'améliorations fines qui lui succède : le rapport entre les mutations favorables et les mutations délétères est exagérément élevé tant que la majorité de l'axe des processus n'est pas couverte. Ici, c'est plutôt la seconde phase qui nous intéresse, car notre problématique se centre sur les mécanismes qui pourraient être à l'œuvre dans des populations actuelles plutôt que sur l'origine de la vie. Dans la mesure où nous fonderons nos conclusions sur la phase stable, la sensibilité quantitative de la phase initiale vis-à-vis de la forme de l'environnement ne constitue pas un obstacle majeur.

Quelle est l'influence de la forme de l'environnement sur l'allure du génome final ? Comme le montre la figure II.15, la taille du génome, le nombre de gènes et la proportion de non codant sont du même ordre de grandeur, malgré les différences importantes imposées au

[1] L'environnement E_5, par exemple, est discontinu et donc impossible à réaliser exactement avec une somme de fonctions continues.

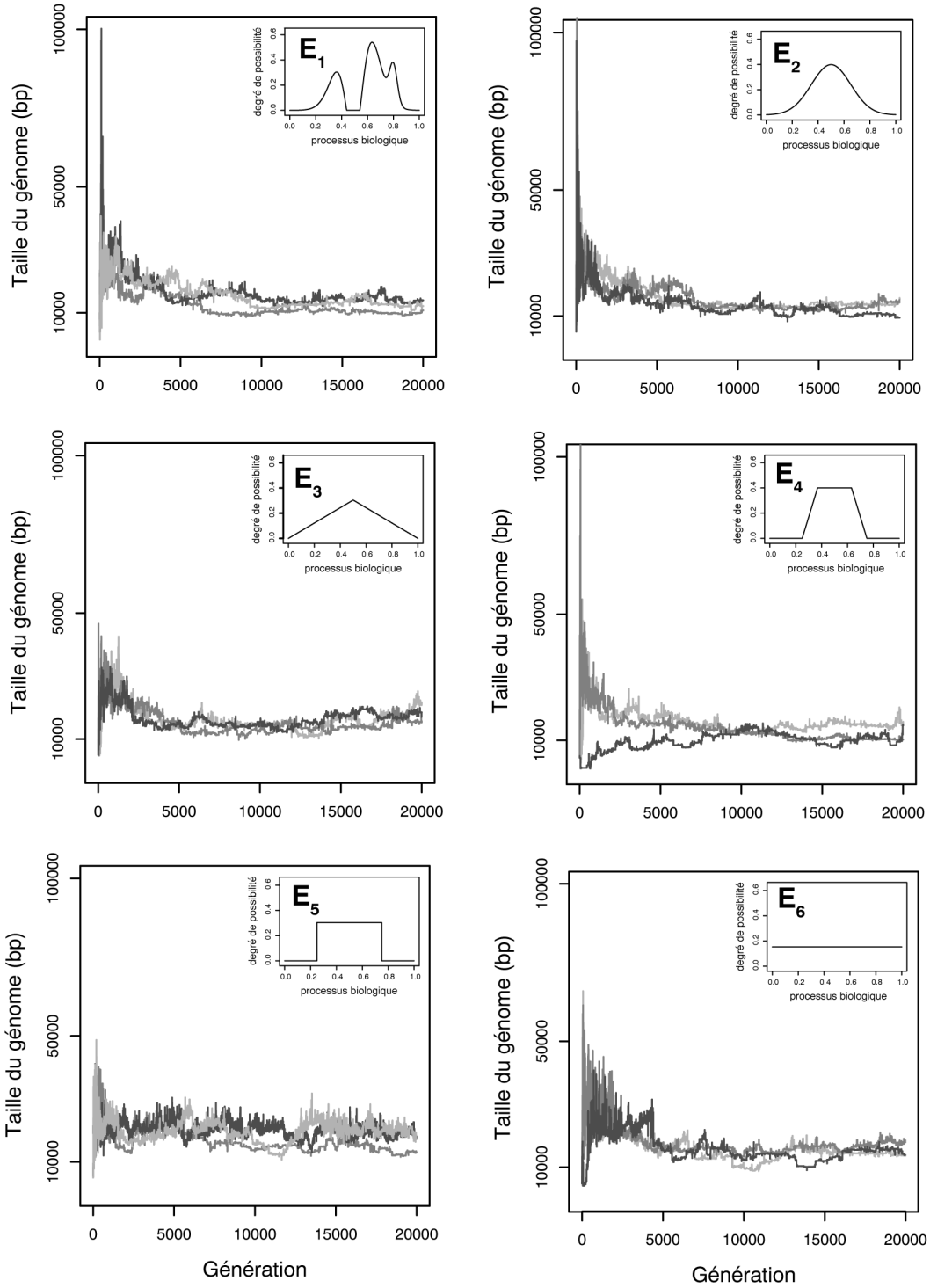


Fig. II.14: Évolution de la taille du génome du meilleur individu, pour six formes d'environnement différentes. Dans chacun des six graphes, l'encart en haut à droite représente la distribution de possibilité f_E avec laquelle les simulations ont été réalisées (celle-ci étant invariable au cours du temps). Les niveaux de gris correspondent aux trois répétitions.

niveau fonctionnel par la forme de l'environnement (répartition des triangles sur l'axe des processus, hauteur des triangles, ...). La figure II.16 montre que les variations de la taille du génome ou du nombre de gènes d'un environnement à l'autre sont sensiblement du même ordre que les variations inter-répétitions : l'effet de la forme de l'environnement sur ces caractéristiques génomiques n'est donc pas statistiquement significatif (au seuil de 5%). Son effet sur le nombre absolu de positions non codantes et sur la proportion codante du génome, bien que statistiquement significatif, reste faible, et une forte variabilité entre répétitions subsiste. Pour conclure sur ce point, il semble que dans le modèle, à aire constante, l'effet de la forme de l'environnement sur l'allure du génome existe mais qu'il ne soit pas démesurément plus grand que l'effet fondateur des premières générations et le hasard des mutations. Pour isoler l'effet d'un paramètre biologiquement pertinent comme le taux de mutation, il sera cependant préférable de se placer dans un même environnement. Toutes les expériences qui sont décrites dans la suite de ce manuscrit se sont ainsi déroulées dans l'environnement E_1. Cette forme a été choisie pour quatre raisons : (i) elle comporte une zone nulle, ce qui permet de tester si des gènes inhibiteurs peuvent y apparaître par sélection indirecte (voir figure II.6, p. 85), (ii) elle ne comporte pas de plateau (zone nulle exceptée), donc le glissement d'une protéine sur l'axe des processus n'est en général pas neutre, (iii) elle est continue mais pas linéaire par morceaux[1], ce qui correspond à un niveau de difficulté intermédiaire, et (iv) elle comporte trois lobes qui permettent d'identifier des modules fonctionnels.

[1]Dans l'implémentation, la fonction est de fait approximée par une fonction linéaire par morceaux, mais de résolution bien plus fine que celles des triangles des protéines.

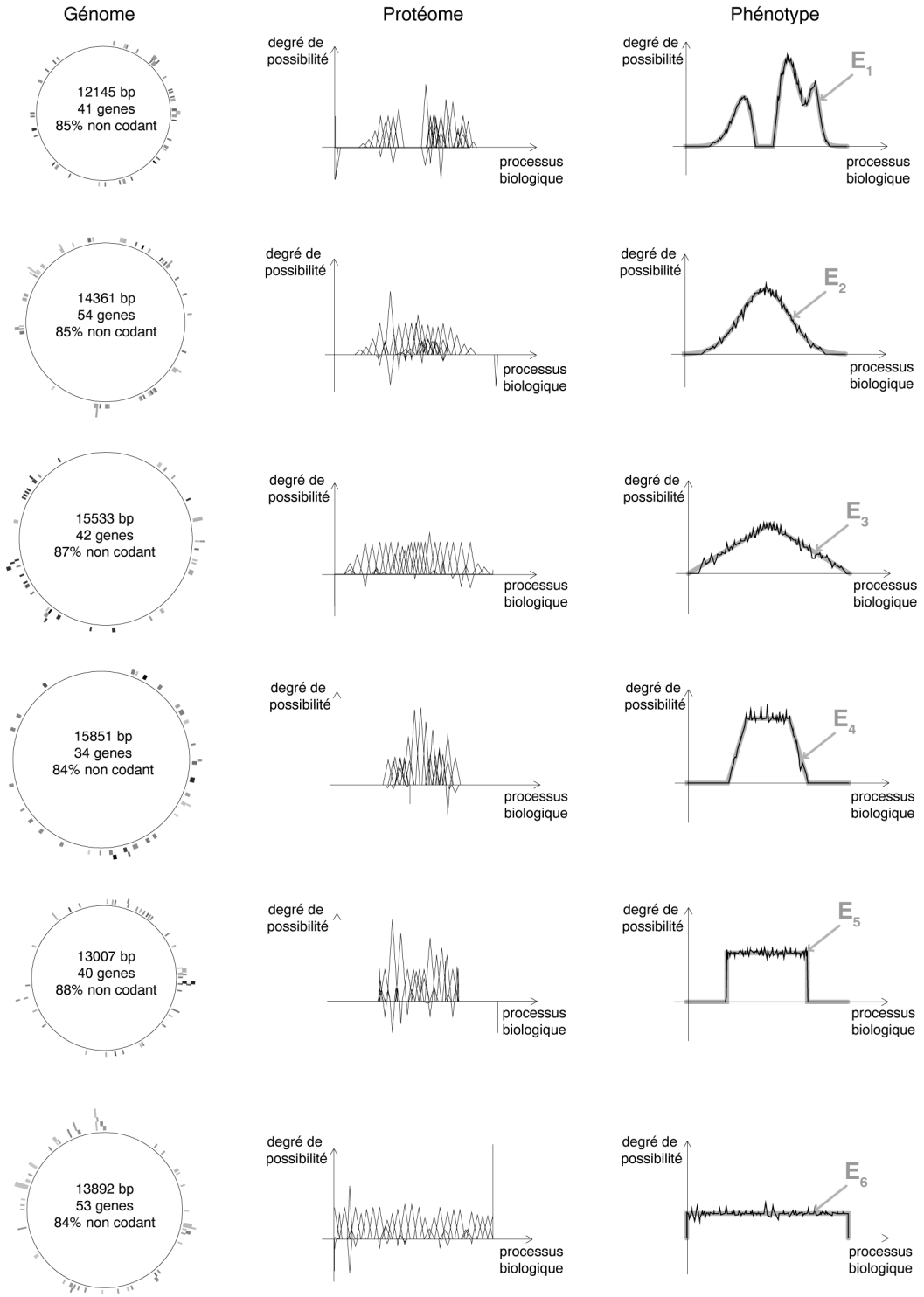


Fig. II.15: Effet de la forme de l'environnement sur l'allure des organismes après 20 000 générations. Pour chaque environnement testé, le meilleur individu final de l'une des trois répétitions est représenté, avec les mêmes conventions que pour la figure II.6.

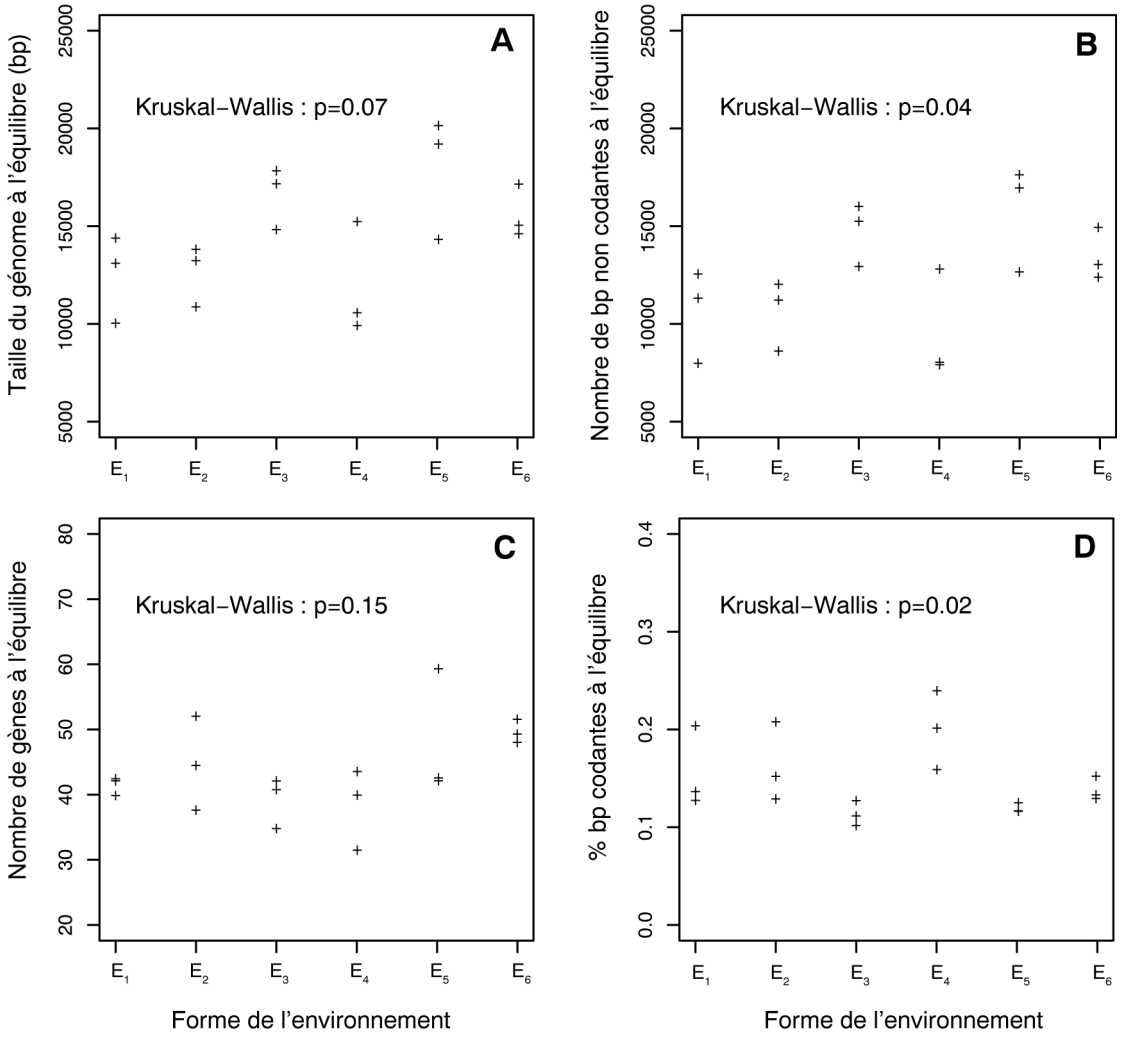


Fig. II.16: Effet de la forme de l'environnement sur la taille (A), le nombre absolu de positions non codantes (B), le nombre de gènes (C) et la proportion de positions codantes (D) du génome à l'équilibre. Chaque point correspond à une simulation et représente la moyenne du meilleur individu sur les 5 000 dernières générations. Étant donné la petite taille des échantillons (et le caractère non normal des données, au moins pour la proportion codante), c'est l'équivalent non paramétrique de l'analyse de variance, c'est-à-dire le test de Kruskal et Wallis, qui est utilisé pour tester l'effet de la forme de l'environnement. Si l'on se fixe un seuil de 5%, cet effet est statistiquement significatif pour le nombre de positions non codantes et pour la proportion codante du génome. Cependant, la variabilité entre répétitions reste élevée et il est possible que l'effet disparaisse si l'on ajoute des points supplémentaires.

Effet de la méthode de sélection

Les utilisateurs des algorithmes génétiques connaissent bien l'influence que la méthode de sélection peut avoir sur la trajectoire évolutive de la population. La méthode basée sur les valeurs brutes d'adaptation ("fitness-proportionate") est par exemple connue pour causer des convergences prématurées sur des optima locaux. En effet, au départ de la simulation, elle tend à sur-sélectionner les champions et donc à créer une perte importante de diversité : seule une voie est explorée, menant souvent à un optimum local. Par la suite, lorsque les différences d'adaptation sont trop faibles pour causer une reproduction différentielle, les individus les moins aptes sont sélectionnés à un rythme très proche des meilleurs, ce qui fait en général stagner la population sur l'optimum local. Les méthodes de sélection basées sur le rang (ranking) sont beaucoup moins sensibles à ce phénomène de convergence prématurée, car elles sont indépendantes des valeurs absolues des mesures d'adaptation comme de l'ampleur des écarts entre individus. La méthode de sélection a donc clairement une influence potentielle sur l'adaptation des individus finaux. Mais ici, nous ne souhaitons pas résoudre un problème d'optimisation et ce n'est pas tant l'adaptation finale qui nous intéresse – c'est davantage le génome et sa structure. La méthode de sélection a-t-elle donc aussi une influence sur l'organisation du génome ?

Pour répondre à cette question, nous avons considéré les trois méthodes de sélection décrites précédemment ("fitness-proportionate", "linear ranking" et "exponential ranking") et nous avons simulé l'évolution de trois populations indépendantes dans chaque cas, les autres paramètres étant identiques à ceux du tableau II.2. La comparaison directe des trois méthodes est délicate, car chacune d'entre elles a sa propre façon de régler l'intensité de la sélection. Nous avons choisi pour le "linear ranking" l'intensité de sélection η^+ la plus couramment utilisée, qui donne au meilleur individu N fois plus de chances de reproduction qu'au moins bien adapté, soit ici $\eta^+ = 1{,}998$. Pour l'"exponential ranking", nous avons choisi $c = 0{,}998$, qui fournit une relation rang - probabilité de reproduction relativement proche de celle du régime linéaire avec $\eta^+ = 1{,}998$. La méthode "fitness-proportionate" ne peut pas être directement comparée avec les deux précédentes, car les probabilités de reproduction relatives du meilleur et du moins bon individu dépendent à la fois du paramètre k et des valeurs brutes d'adaptation dans la population (donc du temps). Le choix de k est donc plus arbitraire. Nous avons choisi $k = 250$, qui semble donner des probabilités de reproduction relatives comparables à celles du "linear ranking" lorsque la plupart des gènes sont acquis. Cela reste cependant très approximatif : l'intensité de la sélection en mode "fitness-proportionate" n'est égale à celle du "ranking" que pour une période donnée de la simulation.

Quel que soit le mode de sélection, l'évolution de la taille du génome suit qualitativement les phases décrites précédemment, avec un pic initial correspondant à des duplications massives, puis un retour à une taille plus faible par élimination d'une partie des régions non codantes. Cependant, comme le montre la figure II.17, l'allure de ce pic initial varie d'un mode de sélection à l'autre et d'une répétition à l'autre. Pour l'une des répétitions du mode "fitness-proportionate", ce pic initial semble dédoublé. Ce n'est qu'après environ 5 000 générations que la trajectoire devient plus reproductible au sein d'un même mode

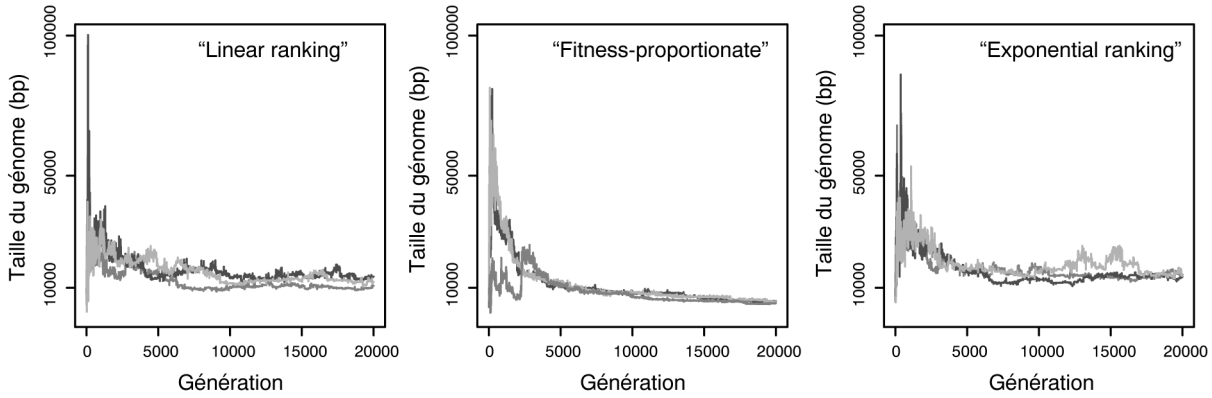


Fig. II.17: Évolution de la taille du génome du meilleur individu, pour les différents modes de sélection. Les niveaux de gris correspondent aux trois répétitions.

de sélection. Pour les deux méthodes basées sur le rang, la taille du génome se stabilise. Pour la méthode basée sur les valeurs brutes d'adaptation, le génome tend à se raccourcir progressivement au cours du temps. Les figures II.18 et II.19 montrent que c'est au niveau de la quantité de non-codant, et non au niveau du nombre de gènes, que les génomes finaux diffèrent le plus. Il semble donc que la quantité de non-codant maintenue dans le génome soit liée à l'intensité de la sélection, qui est constante en "ranking" mais décroissante au cours du temps en "fitness-proportionate". Ce phénomène peut s'expliquer par le rôle du non-codant dans les réarrangements génomiques et donc dans la variabilité mutationnelle du phénotype (ce rôle sera mis en évidence au chapitre suivant). En effet, au fur et à mesure de l'évolution, les gains d'adaptation dus aux mutations favorables deviennent de plus en plus petits, alors que les pertes d'adaptation dues aux mutations délétères peuvent rester conséquentes. Avec un mode de sélection basé sur les valeurs brutes d'adaptation, la conséquence est que les mutations favorables ne sont plus sélectionnées alors que les mutations délétères sont, elles, contre-sélectionnées. On passe ainsi progressivement d'une sélection directionnelle à une sélection stabilisatrice. Comme nous l'avons vu au chapitre I, cela implique que le niveau de variabilité mutationnelle indirectement sélectionné diminue. Il devient alors indirectement avantageux de réduire les régions non codantes stimulant les réarrangements.

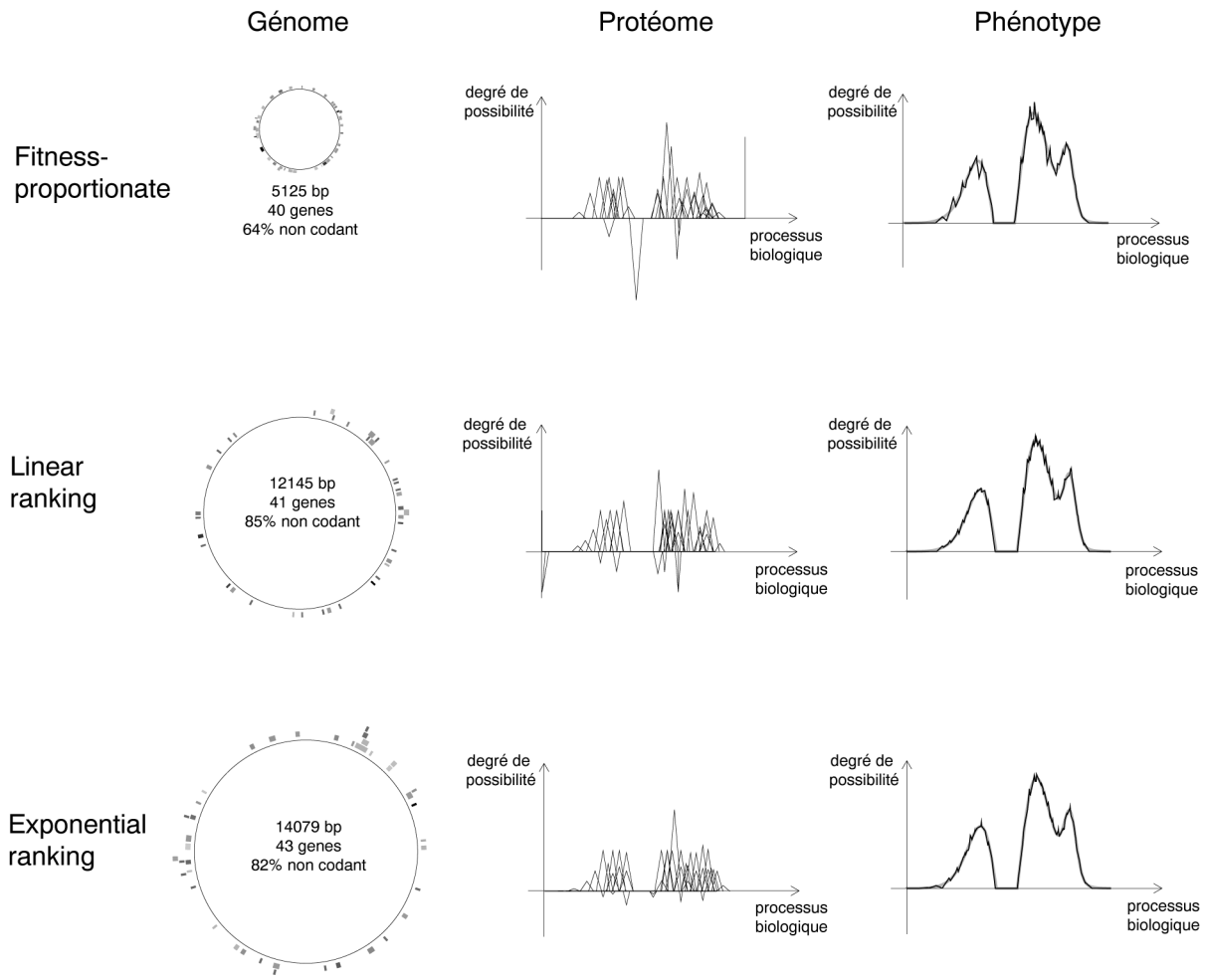


Fig. II.18: Effet de la méthode de sélection sur l'allure des organismes après 20 000 générations. Pour chaque mode de sélection, le meilleur individu final de l'une des trois répétitions est représenté, avec les mêmes conventions que pour la figure II.6.

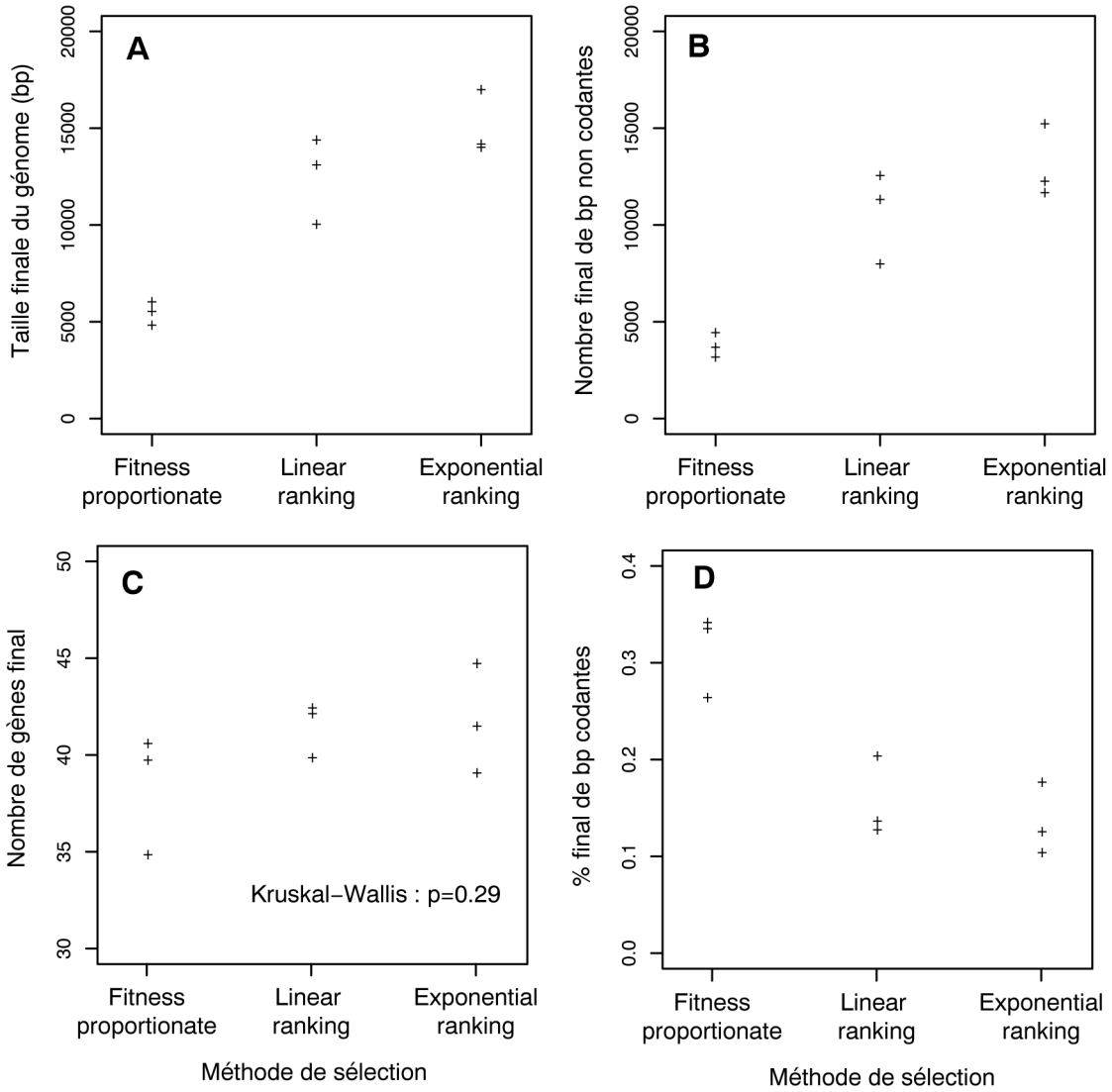


Fig. II.19: Effet de la méthode de sélection sur la taille (A), le nombre absolu de positions non codantes (B), le nombre de gènes (C) et la proportion de positions codantes (D) du génome. Chaque point correspond à une simulation et représente la moyenne du meilleur individu sur les 5 000 dernières générations. Il y a clairement un effet du mode de sélection sur la quantité absolue ou relative de non codant, alors que l'effet sur le nombre de gènes n'est pas statistiquement significatif au seuil de 5%. Il faut cependant être prudent dans l'interprétation de ces résultats : comme on ne peut pas garantir que l'intensité de la sélection soit identique dans les trois modes (voir texte principal), on voit en fait ici un mélange de l'effet du mode de sélection et de l'effet de l'intensité de la sélection.

4 Conclusion

Le modèle *aevol* a été développé avec l'objectif de dégager des principes généraux susceptibles de gouverner la structuration des génomes. Le choix d'une modélisation individu-centrée permet de simuler explicitement la compétition entre individus de génomes différents, ainsi que l'apparition de mutations et leur propagation dans la population. Pour gérer cette population, les méthodes traditionnellement employées en évolution artificielle ont pu être réutilisées. En revanche, au niveau de la représentation du génome et de la transition génotype-phénotype, une modélisation plus proche du réel a été mise en œuvre, afin de donner aux génomes artificiels des degrés de liberté comparables à ceux des génomes réels. L'utilisation de séquences signal pour détecter les gènes et la représentation des protéines comme intermédiaires élémentaires entre le génotype et le phénotype sont des éléments-clés qui permettent aux mutations de faire varier non seulement l'action des protéines, mais aussi le nombre de gènes, l'ordre des gènes ou encore la quantité de non codant. Les mutations ne sont donc pas limitées aux événements ponctuels mais peuvent aussi consister en de grands réarrangements chromosomiques.

L'adaptation des individus ne dépendant que de l'adéquation de leur phénotype avec l'environnement, aucune pression sélective directe n'est exercée sur le génome : la taille du génome, par exemple, n'est pas pénalisante. Il est alors possible de tester si une structuration spontanée du génome est possible, c'est-à-dire une structuration qui serait une propriété immergente de la compétition entre individus lorsque ceux-ci se reproduisent avec des mutations. En effet, si ces mécanismes fondamentaux peuvent induire la sélection indirecte d'un certain niveau de variabilité mutationnelle du phénotype, et si la structure du génome joue un rôle dans cette variabilité (en influençant notamment le nombre moyen de gènes touchés par un réarrangement chromosomique), alors on peut s'attendre à ce que des structures génomiques particulières émergent sur le long terme. Pour tester cette hypothèse, nous allons utiliser le fait que le modèle intègre les trois composantes de la variabilité mutationnelle du phénotype : le taux de mutation (directement paramétrable), l'effet phénotypique des mutations dans les gènes (indirectement paramétrable à travers la largeur maximale des "triangles" protéiques) et la structure du génome (libre d'évoluer). En faisant varier le taux de mutation (chapitre III) ou l'effet des mutations géniques (chapitre IV), nous serons en mesure de tester si la structure du génome s'y ajuste spontanément. Dans l'interprétation de ces résultats, il conviendra cependant de tenir compte des hypothèses simplificatrices effectuées au cours du processus de modélisation et des artefacts qu'elles peuvent engendrer.

Chapitre III

Structuration du génome en fonction du taux de mutation

La nature est remplie d'une infinité de raisons
dont l'expérience n'a jamais vu la trace.

Léonard de Vinci[1]

Bien que l'organisation des génomes soit généralement présentée comme le résultat de biais mutationnels ou de pressions sélectives directes, nous avons vu au chapitre I que certaines caractéristiques structurelles du génome pouvaient contribuer à la variabilité mutationnelle du phénotype. Elles pourraient alors, au même titre que le taux de mutation et les mécanismes de canalisation, faire l'objet d'une sélection indirecte, ce qui constituerait une force supplémentaire agissant sur le génome. Tester cette hypothèse requiert cependant des outils spécifiques et c'est dans ce cadre que le modèle *aevol* a été développé.

Parmi les caractéristiques structurelles qui peuvent être étudiées avec ce modèle, la quantité d'ADN non codant aiguise particulièrement la curiosité : comment expliquer que certains virus n'ont presque aucune base non codante, alors que dans le génome humain, les gènes sont séparés par environ 60 kilobases en moyenne (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001) ? Certaines séquences intergéniques portent l'empreinte de la sélection naturelle et sont donc vraisemblablement fonctionnelles (Duret *et al.*, 1993; Frazer *et al.*, 2001; Margulies *et al.*, 2003; Bejerano *et al.*, 2004; Andolfatto, 2005; Dermitzakis *et al.*, 2005; Keightley *et al.*, 2005; Lunter *et al.*, 2006), mais il n'est pas certain que cela suffise à expliquer la quantité totale d'ADN non codant contenue dans un génome comme le génome humain. Il subsisterait donc de l'ADN "non fonctionnel", ou plus précisément de l'ADN dont la séquence précise n'est pas directement contrainte. Comme nous l'avons déjà mentionné, cette quantité d'ADN non fonctionnel

[1] In *Les carnets de Léonard de Vinci*, Gallimard, 1994.

est difficile voire impossible à estimer précisément, mais elle est vraisemblablement liée à la quantité d'ADN non codant et donc variable selon les phyla. Nous voici donc ramenés à notre questionnement initial, quoique sous une forme légèrement différente : comment expliquer que certains virus ne possèdent apparemment que de l'ADN fonctionnel, alors que le génome humain semble contenir une quantité importante d'ADN dont la séquence précise n'est pas directement contrainte ?

Rappelons brièvement les deux grandes catégories d'hypothèses évoquées pour répondre à cette question (pour plus de détails, on pourra se rapporter au chapitre I, p. 52). La première est celle des pressions de sélection directes s'exerçant sur la quantité d'ADN totale, indépendamment de son contenu informationnel. La taille d'un génome viral serait ainsi directement limitée par le volume de la capside. Chez les procaryotes, on évoque plus fréquemment la durée de la réplication du chromosome comme facteur limitant pour la taille du génome, mais nous avons vu que cela ne semblait pas confirmé expérimentalement. Enfin, pour les eucaryotes, certains auteurs ont proposé l'existence d'une quantité optimale d'ADN, liée au volume de la cellule (Cavalier-Smith, 1985; Gregory, 2001). La seconde catégorie d'hypothèses adopte un point de vue opposé et voit dans la quantité d'ADN non fonctionnel le résultat passif de processus mutationnels biaisés en faveur des délétions ou au contraire des insertions (Graur *et al.*, 1989; Ophir et Graur, 1997; Petrov *et al.*, 2000; Mira *et al.*, 2001; Denver *et al.*, 2004). Mais comme nous l'avons souligné au chapitre I, on peut également envisager un troisième type de pression évolutive, lié au rôle (quelque peu paradoxal de prime abord) de l'ADN non fonctionnel dans l'apparition de mutations non neutres, et donc de nouveaux phénotypes. La sélection indirecte d'un certain niveau de variabilité mutationnelle du phénotype, élément clé du succès évolutif à long terme, pourrait alors se traduire par une modulation de la quantité d'ADN non fonctionnel.

Ce type de mécanisme évolutif est cependant difficile à mettre en évidence chez des organismes réels, et ce pour différentes raisons. La première est la nature indirecte du mécanisme, qui agit donc à long terme. Un grand nombre de générations est *a priori* nécessaire pour que son effet soit détectable. Un second obstacle, peut-être plus fondamental, réside dans la difficulté à *isoler* cet effet des autres pressions évolutives qui s'exercent sur la taille du génome, comme les biais mutationnels et les éventuelles pressions sélectives directes. Enfin, la sélection indirecte d'un degré donné de variabilité mutationnelle peut simultanément agir à d'autres niveaux que la structure du génome, et faire aussi varier le taux de mutation, l'action canalisatrice des protéines chaperonnes ou encore la robustesse des réseaux de régulation. L'expérimentation *in silico* est donc particulièrement utile dans ce contexte, car il est possible de maîtriser précisément (voire de supprimer) les biais mutationnels et les pressions sélectives directes, et de garder le taux de mutation fixe au cours de l'évolution. Avec le modèle *aevol*, on peut également contrôler l'effet moyen d'une mutation dans un gène par l'intermédiaire de la pléiotropie maximale des protéines (voir chapitre II). Le taux de mutation et l'effet moyen des mutations géniques étant fixes au cours du temps, la structure du génome devient théoriquement le levier principal par lequel la variabilité mutationnelle peut évoluer. On peut alors tester la capacité d'un processus de sélection indirecte à structurer le génome.

Dans ce chapitre, nous explorons ce mécanisme en simulant, à l'aide du modèle *aevol*, l'évolution de génomes sous différents taux de mutations. L'idée qui motive cette expérience est la suivante : si le taux de mutation et la quantité d'ADN non fonctionnel contribuent tous deux à la variabilité mutationnelle du phénotype, on peut s'attendre à ce que la sélection indirecte d'une variabilité donnée se traduise par l'ajustement spontané de la quantité d'ADN non fonctionnel en fonction du taux de mutation. Après avoir présenté plus en détail le plan d'expérience (section 1), nous montrerons que la compacité des génomes obtenus dépend en effet du taux de mutation, et plus particulièrement du taux de duplications et de grandes délétions (section 2) : le génome est d'autant plus compact que ces taux sont élevés. Nous confirmerons ensuite que c'est bien la sélection d'un niveau donné de variabilité mutationnelle qui détermine la quantité de non codant et qui est donc sous-jacente à sa dépendance au taux de mutation (section 3). Nous vérifierons alors que ces résultats sont robustes vis-à-vis de l'environnement choisi et du mode de sélection (section 4). Enfin, nous discuterons le sens de ces résultats pour les génomes réels (section 5). Ce travail fait l'objet des articles 3 et 8, selon la numérotation établie au début du manuscrit.

1 Plan d'expérience

Afin de garder l'expérience simple, tous les taux de mutation ont été fixés à la même valeur, notée u. Cela permet de ne pas introduire de biais mutationnel vers la croissance ou le rétrécissement du génome, et de ne pas donner *a priori* plus d'importance aux mutations locales qu'aux réarrangements, ou inversement. Six valeurs différentes ont été testées pour u : 5.10^{-6}, 10^{-5}, 2.10^{-5}, 5.10^{-5}, 10^{-4} et 2.10^{-4} par bp. Comme nous l'avons mentionné au chapitre II, ces taux ne se veulent pas quantitativement réalistes, en raison du temps de calcul que des taux réels entraîneraient, mais aussi parce que les simplifications du modèle au niveau fonctionnel sont telles qu'une comparaison quantitative avec les données réelles n'aurait pas de sens.

Comme l'analyse de sensibilité présentée au chapitre II a suggéré une dépendance de la quantité de non codant vis-à-vis de l'intensité de la sélection, nous avons répété les tests pour quatre intensités de sélection différentes. La méthode de sélection choisie est l'"exponential ranking", parce qu'elle fournit une pression de sélection constante au cours du temps et que cette pression peut être choisie forte. Nous avons ainsi successivement fixé l'intensité c de la sélection à 0,9995 (sélection faible), 0,9980, 0,9950 et 0,9900 (sélection forte).

Pour chaque taux et chaque intensité de sélection, nous avons fait évoluer indépendamment trois populations pendant 20 000 générations. Au total, nous avons donc simulé l'évolution de $4 \times 6 \times 3 = 72$ populations. L'ensemble des paramètres utilisés figure dans le tableau III.1.

Paramètre	Symbole	Valeur
Taille de la population	N	1 000 individus
Taille initiale des génomes	L_{init}	5 000 bp
Séquence promotrice	-	01010110011100100101 10, $d_{\max}$ = 4 différences autorisées
Séquence terminatrice	-	De la forme $abcd***\bar{d}\bar{c}\bar{b}\bar{a}$
Signal d'initiation de la traduction	-	011011***000
Signal de terminaison de la traduction	-	001
Code génétique	-	Voir tableau II.1, p. 75
Ensemble des processus biologiques	Ω	$[0, 1]$
Pléiotropie maximale des protéines	$w_{\max}$	$\frac{1}{30}$
Distribution de possibilité de l'environnement	f_E	Trimodale (voir la figure II.3), et fixe au cours du temps
Méthode de sélection	-	"Exponential ranking"
Intensité de la sélection	c	De 0,9900 à 0,9995
Taux de mutation ponctuelle	u	De 5.10^{-6} à 2.10^{-4} par bp
Taux de petite insertion	u	De 5.10^{-6} à 2.10^{-4} par bp
Taux de petite délétion	u	De 5.10^{-6} à 2.10^{-4} par bp
Taux de grande délétion	u	De 5.10^{-6} à 2.10^{-4} par bp
Taux d'inversion	u	De 5.10^{-6} à 2.10^{-4} par bp
Taux de duplication	u	De 5.10^{-6} à 2.10^{-4} par bp
Taux de translocation	u	De 5.10^{-6} à 2.10^{-4} par bp
Loi de la longueur des petits indels	-	Loi uniforme entre 1 et 6 bp
Loi de la longueur des segments réarrangés	l_{seg}	Loi uniforme entre 1 et L

Tab. III.1: Paramètres utilisés pour tester l'influence du taux de mutation et de l'intensité de la sélection sur la compacité du génome.

2 Relation entre le taux de mutation et la quantité de non codant

Les 1 000 individus de chaque population sont initialisés avec un même génome de 5 000 bp contenant au moins un gène. Ce génome est choisi aléatoirement et est différent d'une répétition à l'autre, mais l'évolution du génome suit qualitativement les mêmes étapes dans tous les cas : acquisition de la plupart des gènes par duplication-divergence, puis, après environ 5 000 générations, stabilisation de la taille du génome (figure III.1). D'un point de vue quantitatif, les valeurs d'équilibre du nombre de gènes et de la quantité de non codant sont assez bien reproductibles d'une répétition à l'autre, mais elles sont au contraire très variables d'un taux de mutation à l'autre. Comme le montre la figure III.2, plus le taux de mutation est élevé, moins le génome contient de gènes, et moins il contient de non codant. Ces deux tendances sont observées pour les quatre intensités de sélection testées, les génomes étant globalement plus longs lorsque la sélection est plus intense.

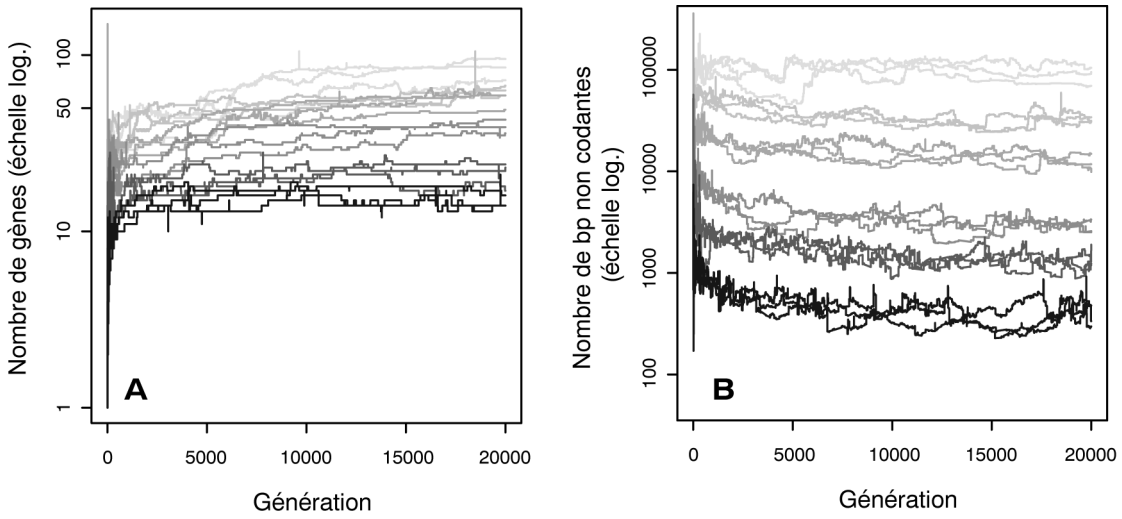


Fig. III.1: Évolution du nombre de séquences codantes (A) et de la quantité de non codant (B) sur la lignée ancestrale du meilleur individu final, pour un taux de mutation u variant de 5.10^{-6} (gris clair) à 2.10^{-4} (gris foncé) par bp. Ces courbes correspondent à l'intensité de sélection $c = 0{,}9950$. L'allure des courbes pour les autres valeurs de c est similaire.

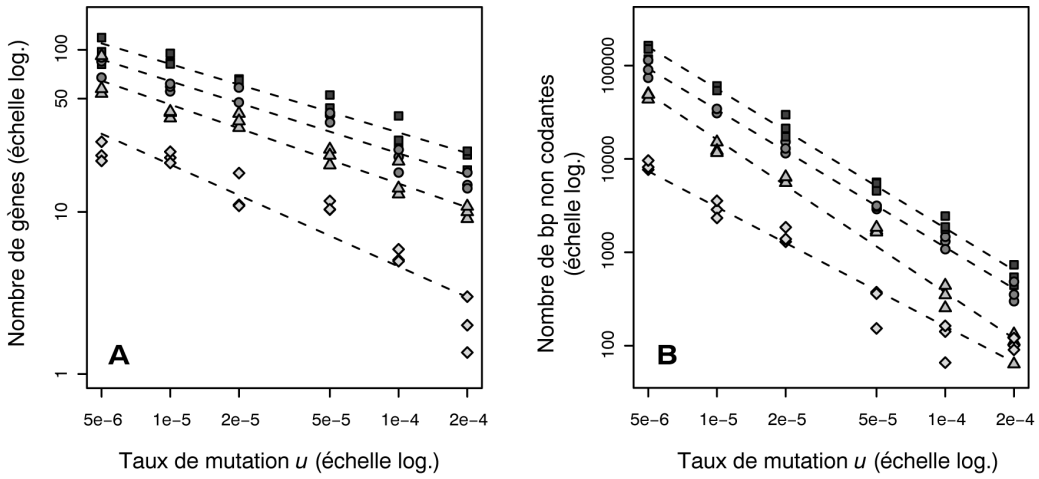


Fig. III.2: Influence du taux de mutation u sur le nombre de séquences codantes (A) et sur la quantité de non codant (B). Les différents symboles représentent les différentes intensités de sélection : carrés pour $c = 0,9900$, cercles pour $c = 0,9950$, triangles pour $c = 0,9980$ et losanges pour $c = 0,9995$. La valeur à l'équilibre est estimée par la moyenne des 5 000 dernières générations, c'est-à-dire des 5 000 derniers ancêtres du meilleur individu final. Chaque régression linéaire de $\log$(nombre de gènes) en fonction de $\log(u)$ est statistiquement significative : les p-values sont respectivement de 5.10^{-11}, 2.10^{-10}, 5.10^{-11}, 2.10^{-8} pour c =0,9900, 0,9950, 0,9980, 0,9995. Les coefficients de détermination r^2 sont respectivement de 0,94, 0,93, 0,94, 0,87. De même, pour $\log$(quantité de non codant) en fonction de $\log(u)$, les p-values sont respectivement de 2.10^{-16}, 2.10^{-16}, 2.10^{-15}, 10^{-11} et les coefficients r^2 sont de 0,99, 0,99, 0,98, 0,95.

La relation entre le taux de mutation et le nombre de gènes reflète un phénomène déjà connu, en général présenté comme suit : comme la plupart des mutations géniques sont

délétères, il doit exister un nombre maximal de gènes mutés tolérable par reproduction, et donc, pour un taux de mutation donné, un nombre maximal de gènes dans le génome (Eigen, 1971; Maynard Smith, 1983; Hurst, 1995; Pal et Hurst, 2000; Ofria *et al.*, 2003). En revanche, la relation entre le taux de mutation et la quantité de non codant est plus inattendue. Elle est frappante si l'on compare visuellement, comme sur la figure III.3, les génomes obtenus pour les taux de mutation extrêmes. Lorsque le taux de mutation est élevé, le génome obtenu ressemble à un génome viral : il contient peu de gènes[1] et ceux-ci sont soit chevauchants, soit séparés par des séquences intergéniques très courtes. Si au contraire le taux de mutation est faible, le génome contient plus de gènes et ceux-ci sont séparés par de grandes distances intergéniques. Sur l'exemple de la figure III.3, le génome contient ainsi environ 95% de non codant. Cela montre que lorsque le taux de mutation est faible, une grande quantité de non codant peut être maintenue malgré l'absence d'éléments génétiques "égoïstes" proliférants ou de biais en faveur des insertions.

[1] Corrélativement, le phénotype obtenu après 20 000 générations est assez éloigné de l'optimum environnemental lorsque le taux de mutation est élevé (voir figure III.3). Il semble donc que ce soit un taux de mutation *trop élevé* qui empêche la progression de l'adaptation. Cela peut sembler étonnant car dans un algorithme génétique classique, les mutations sont généralement vues comme ce qui permet de sortir des optima locaux et donc d'éviter la stagnation de la fitness. Comme nous allons le voir, cela vient du fait que dans notre modèle, les individus qui sont de fait sélectionnés sur le long terme ne sont pas les plus adaptés mais ceux qui réalisent le meilleur compromis entre adaptation, stabilité et variabilité. Les phénomènes de sélection indirecte peuvent donc entraver l'optimisation de l'adaptation et engendrer des effets inattendus lorsqu'on fait varier le taux de mutation. Ces phénomènes sont dus aux degrés de liberté donnés à la structure du génome : ici, comme dans un génome réel, les mutations peuvent faire varier non seulement le phénotype, mais aussi la structure du génome et donc la variabilité du phénotype (le degré d'exploration) pour les générations suivantes. Ainsi, si l'on veut utiliser l'évolution artificielle pour optimiser une fonction, il ne semble pas toujours souhaitable d'imiter au plus près le vivant.

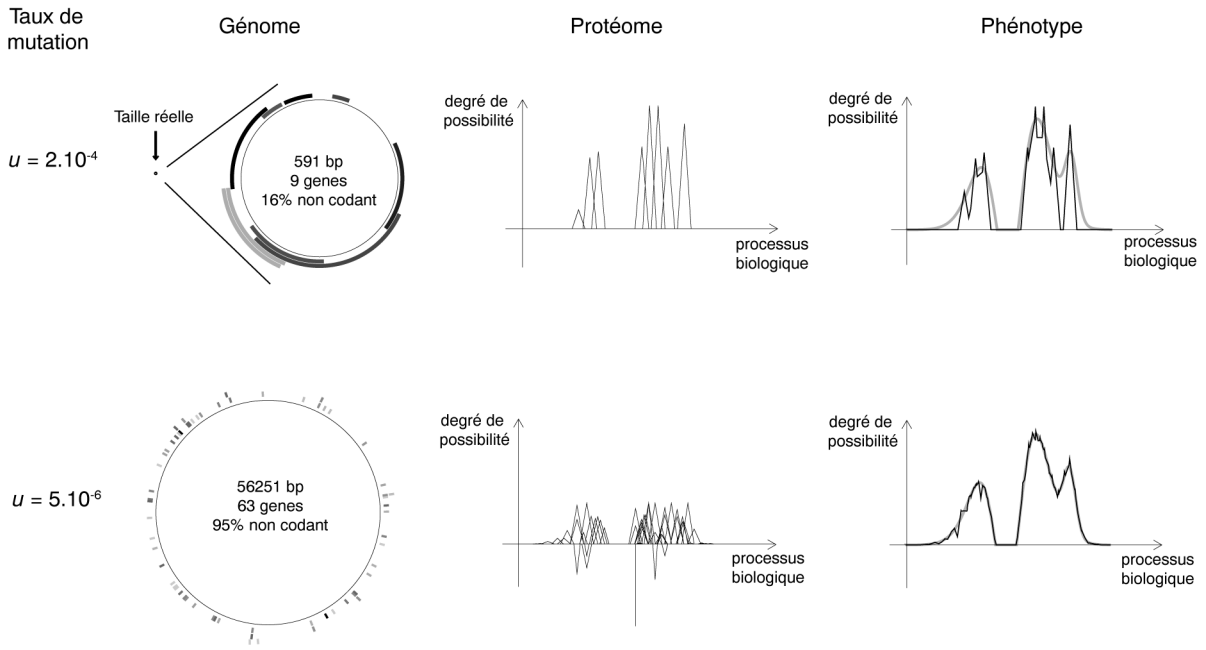


Fig. III.3: Effet du taux de mutation sur l'allure des organismes après 20 000 générations. La figure présente le meilleur individu final de l'une des trois répétitions pour le taux de mutation le plus fort ($u = 2.10^{-4}$, en haut) et pour le taux le plus faible ($u = 5.10^{-6}$, en bas), dans le cas où l'intensité de la sélection vaut $c = 0,9980$. Les conventions de représentation sont identiques à celles de la figure II.6, p. 85. Cependant, la différence de taille entre les deux génomes est telle que nous avons dû agrandir 70 fois le génome du haut (fort taux de mutation) pour que ses séquences codantes soient visibles.

3 Sélection indirecte du niveau de variabilité mutationnelle

Pour tester si la sélection indirecte d'un niveau constant de variabilité mutationnelle est bien le facteur qui détermine la quantité de non codant et qui sous-tend sa relation avec le taux de mutation, nous avons mesuré la variabilité mutationnelle du meilleur individu final de chaque simulation, qui constitue le niveau de variabilité qui a de fait été indirectement sélectionné. Pour cela, nous avons estimé, pour chaque meilleur individu final, la probabilité F_ν que ses descendants conservent son adaptation g. Cette probabilité représente la proportion attendue de "descendants neutres", c'est-à-dire de descendants qui n'ont subi aucune mutation ou bien uniquement des mutations neutres, sans effet sur l'adaptation (Wilke, 2001a; Ofria *et al.*, 2003). Comme le montre la figure III.4, il s'agit d'un indicateur partiel de la variabilité mutationnelle (de l'adaptation). Un indicateur plus complet serait la *distribution* de l'adaptation dans les descendants, mais il est plus commode de travailler avec un seul indicateur numérique, caractérisant partiellement cette distribution. Notons que dans notre modèle, variabilité mutationnelle de l'adaptation et variabilité mutationnelle du phénotype se confondent presque exactement, car il est impossible de faire varier g sans faire varier le phénotype, et, réciproquement, il est quasiment impossible de faire

varier le phénotype sans faire varier g [1].

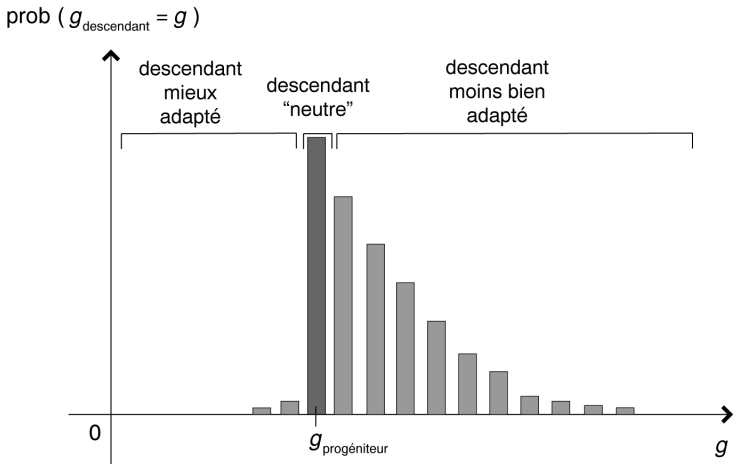


Fig. III.4: F_ν comme indicateur partiel de la variabilité mutationnelle de l'adaptation, c'est-à-dire de la propension de l'adaptation à varier entre un progéniteur et son descendant du fait des mutations. La figure présente une distribution de probabilité hypothétique pour l'adaptation g d'un descendant, connaissant l'adaptation $g_{\text{progéniteur}}$ de son progéniteur. Rappelons que g mesure l'écart entre le phénotype et l'environnement, et que les mutations favorables sont donc celles qui le font diminuer. C'est en principe l'ensemble de cette distribution de probabilité qui permet de caractériser la variabilité mutationnelle de l'adaptation. Elle dépend à la fois du taux de mutation et de l'effet des mutations. Il est cependant plus aisé de travailler sur des indicateurs partiels comme la probabilité F_ν que le descendant soit "neutre" (sans mutation ou avec uniquement des mutations neutres, hauteur du bâton gris foncé) ou l'espérance de la perte d'adaptation.

Dans notre modèle, il est possible d'obtenir une estimation théorique de F_ν à partir de la séquence génomique d'un individu, en faisant l'hypothèse simplificatrice qu'une mutation est neutre si elle n'affecte aucune "région fonctionnelle", une région fonctionnelle étant définie comme une région transcrite (promoteur et terminateur inclus) contenant au moins une séquence codante[2]. Pour cette estimation théorique de F_ν, nous négligeons donc (i) la probabilité qu'une mutation en dehors de toute région fonctionnelle crée un nouveau gène, (ii) la probabilité qu'une mutation dans une région fonctionnelle soit neutre[3], et (iii) la probabilité que plusieurs mutations dans une même région fonctionnelle se compensent exactement. On peut alors estimer F_ν par la probabilité $\tilde{F}_\nu$ qu'aucune région fonctionnelle

[1] En effet, lorsqu'une mutation modifie la distribution de possibilité phénotypique, l'intégrale de sa différence avec la distribution environnementale est également modifiée, à moins que pendant la même réplication, une autre mutation compense exactement l'effet de la première.

[2] Si deux régions transcrites se chevauchent, elles sont ici considérées comme une seule région fonctionnelle.

[3] Cette simplification est la plus forte, en comparaison à des organismes réels (voir chapitre I). Elle reflète les simplifications faites au niveau de la transition génotype-phénotype, qui suppriment des sources potentielles de robustesse, comme la redondance du code génétique, la robustesse du repliement des protéines ou encore l'existence de voies métaboliques alternatives. Dans notre modèle, l'effet d'une mutation génique peut être plus ou moins grand en fonction du codon concerné et de la pléiotropie de la protéine (largeur w du "triangle"), mais il peut difficilement être nul.

ne soit mutée pendant la réplication :

$$\tilde{F}_\nu = \prod_j (1 - u_j(1 - \tilde{\nu}_j))^L \tag{III.1}$$

où j désigne le type de mutation (ponctuelle, insertion, délétion...), u_j le taux par bp du type de mutation j, L la longueur du génome et $\tilde{\nu}_j$ la probabilité qu'une mutation de type j faite au hasard n'affecte aucune région fonctionnelle. Lorsque les mutations locales sont distribuées uniformément le long du génome et que la longueur des réarrangements suit une loi uniforme entre 1 et L, les $\tilde{\nu}_j$ peuvent s'exprimer comme suit :

$$\begin{cases} \tilde{\nu}_{\text{ponct}} &= \tilde{\nu}_{\text{ins}} = \tilde{\nu}_{\text{del}} = 1 - \frac{l}{L} \\ \tilde{\nu}_{\text{inv}} &= \left(1 - \frac{l}{L}\right)^2 \\ \tilde{\nu}_{\text{transloc}} &= \left(1 - \frac{l}{L}\right)^3 \\ \tilde{\nu}_{\text{gdel}} &= \frac{1}{2L^2} \sum_{i=1}^{N_G} \lambda_i \left(\lambda_i + 1\right) \\ \tilde{\nu}_{\text{dup}} &= \frac{1}{2L^2} \left(1 - \frac{l}{L}\right) \sum_{i=1}^{N_G} \lambda_i \left(\lambda_i + 1\right) \end{cases} \tag{III.2}$$

où N_G est le nombre de régions fonctionnelles, l leur longueur totale, et λ_i la distance en bp entre la fin de la région i et le début de la région $i+1$ (on a donc $\sum_{i=1}^{N_G} \lambda_i = L - l$). Les détails de ces calculs figurent en annexe. Ces équations montrent que pour un taux de mutation donné, des séquences non fonctionnelles plus longues réduisent la proportion de descendants neutres (figure III.5) et augmentent donc la variabilité mutationnelle. Il y a deux raisons à cela. La première est que quand le génome est plus long, il subit en moyenne plus d'événements mutationnels (voir paragraphe II.2.4, p. 78). La seconde est que contrairement aux événements locaux, les grandes délétions et les grandes duplications n'ont pas plus de chances d'être neutres lorsque les distances intergéniques augmentent. En effet, la longueur moyenne des segments affectés augmente avec celle du génome. Par conséquent, comme le montre la figure III.5, des séquences non fonctionnelles plus longues n'augmentent pas les probabilités $\tilde{\nu}_{\text{gdel}}$ et $\tilde{\nu}_{\text{dup}}$. Par exemple, la probabilité qu'une grande délétion soit neutre, $\tilde{\nu}_{\text{gdel}}$, plafonne à $1/(2N_G)$ (pour une quantité infinie d'ADN intergénique) si les gènes sont régulièrement distribués sur le chromosome, ou à $1/2$ s'ils forment un cluster unique. Notons que les inversions et les translocations, dont les effets délétères sont concentrés au niveau des points de rupture, se comportent de fait comme des paires ou des triplets d'événements locaux, et ont donc une chance accrue d'être neutres lorsque les distances intergéniques augmentent. En somme, allonger les séquences non fonctionnelles augmente ici la variabilité mutationnelle du phénotype car le génome subit en moyenne plus de grandes délétions et de duplications et que celles-ci n'ont pas plus de chances d'être neutres[1].

Ce rôle du non codant dans la variabilité mutationnelle du phénotype suggère qu'une élongation des séquences intergéniques peut compenser une baisse du taux de mutation,

[1] Il est important de noter qu'en l'absence de duplications et de grandes délétions, la quantité de non codant n'influencerait aucunement la variabilité mutationnelle du phénotype : si tous les événements mutationnels étaient ponctuels, l'augmentation de la proportion de mutations neutres compenserait exactement l'augmentation du nombre de mutations.

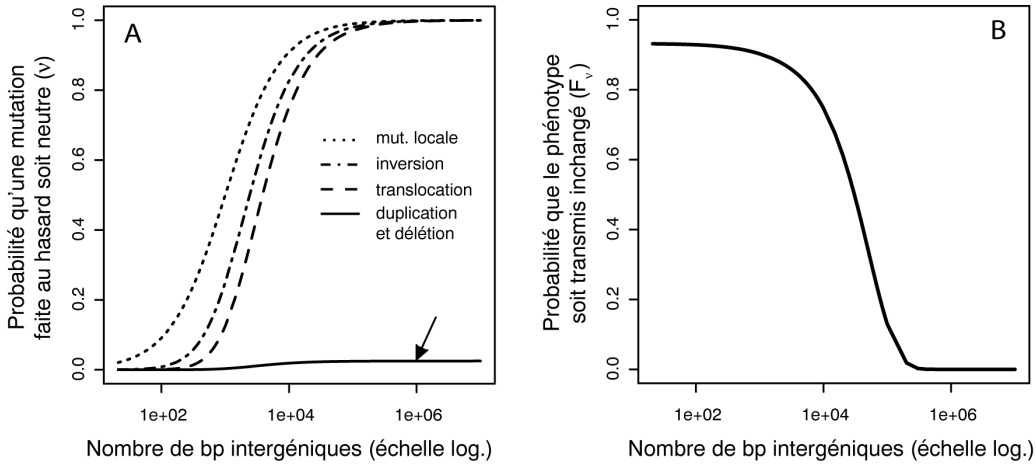


Fig. III.5: Influence de la quantité de non codant sur F_ν. Nous considérons ici le cas théorique d'un génome de 20 gènes (ou, plus précisément, de 20 "régions fonctionnelles", voir texte principal), tous de longueur égale à 50 bp, répartis régulièrement le long du chromosome. **A :** Si l'on ajoute des bases non codantes à ce génome, les mutations locales ainsi que les inversions et les translocations ont de plus en plus de chances d'être neutres, mais ce n'est pas le cas pour les grandes délétions et les duplications. Les courbes sont calculées d'après l'équation III.2. **B :** Comme par ailleurs le nombre d'événements subis par réplication augmente avec la taille du génome, un génome plus long risque davantage de subir au moins une duplication ou une grande délétion non neutre : à nombre de gènes égal, la probabilité F_ν que le descendant soit neutre (sans mutation ou avec uniquement des mutations neutres) décroît donc avec la quantité de non codant. Cette courbe est calculée d'après l'équation III.1 avec $u = 10^{-5}$.

ou, inversement, qu'un rétrécissement de ces séquences peut compenser une augmentation du taux de mutation : les caractéristiques structurelles du génome apparaissant aux côtés de u dans l'équation de F_ν (équations III.1 et III.2), elles peuvent constituer un levier d'ajustement et permettre de garder F_ν constante lorsque u est modifié. Comme le montre la figure III.6, c'est en effet ce qui se produit dans la lignée qui gagne la compétition évolutive. Pour une intensité de sélection donnée, les génomes obtenus après 20 000 générations d'évolution présentent tous la même proportion F_ν de descendants neutres, quel que soit le taux de mutation. La compacité du génome est donc bien déterminée par la sélection indirecte d'un niveau donné de variabilité mutationnelle entre un progéniteur et ses descendants. Cela signifie que dans la compétition évolutive, les individus qui présentent (par hasard) la compacité génomique adéquate étant donné le taux de mutation sont de fait ceux dont la descendance se maintient sur le long terme. À fitness égale, ceux qui ont un génome trop long ont une trop grande variabilité mutationnelle et ne peuvent donc transmettre fidèlement leur phénotype, et ceux qui ont un génome trop court produisent une descendance qui n'explore pas de nouveaux phénotypes et qui risque donc de disparaître lorsque d'autres, plus variables, découvriront une meilleure solution. C'est en cela que l'on peut dire qu'une pression de sélection indirecte s'exerce sur la structure du génome, et que les caractéristiques structurelles du génome sont des propriétés qui émergent – ou plutôt, ici, qui *immergent* – spontanément du contexte populationnel de compétition.

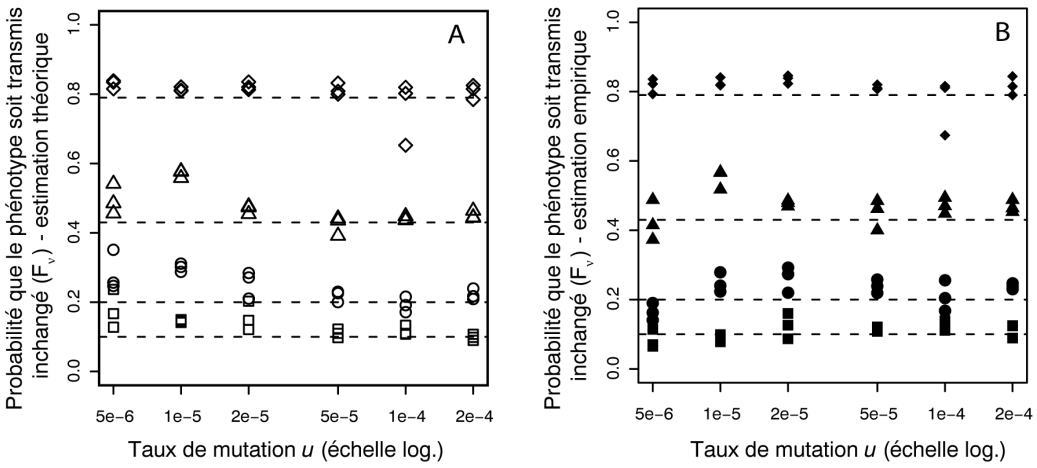


Fig. III.6: Valeurs de F_ν indirectement sélectionnées, en fonction du taux de mutation, pour les quatre intensités de sélection testées (carrés : $c = 0,9900$; cercles : $c = 0,9950$; triangles : $c = 0,9980$; losanges : $c = 0,9995$). La figure présente pour chaque "run" les estimations théoriques (A) et empiriques (B) de la proportion F_ν de descendants neutres du meilleur individu final. Les estimations théoriques ont été obtenues grâce aux équations III.1 et III.2. Les estimations empiriques ont été obtenues en simulant 1 000 réplications indépendantes pour chaque individu final, avec le même taux de mutation u que pendant l'évolution, et en comptant le nombre de descendants qui ont conservé la même valeur d'adaptation g. On note la bonne adéquation entre les deux méthodes, qui valide *a posteriori* les simplifications effectuées pour obtenir les estimations théoriques. Chaque ligne horizontale correspond à la valeur théorique de F_ν qui permettrait au meilleur individu (rang N) de produire en moyenne 1 descendant neutre (voir texte principal).

Bien que, par souci de simplicité, le raisonnement ait été présenté à fitness égale, il faut noter que la pression de sélection indirecte sur la structure du génome peut aussi s'opposer à la pression de sélection directe sur l'amélioration du phénotype : la figure III.3 (p. 109) montre clairement la différence d'adaptation entre le meilleur individu obtenu pour $u = 2.10^{-4}$ et celui obtenu pour $u = 5.10^{-6}$. L'amélioration de l'adaptation doit souvent passer par l'acquisition de nouveaux gènes (activateurs ou inhibiteurs), qui elle-même passe par la duplication d'un gène existant et, bien souvent, des séquences non fonctionnelles adjacentes. Les mutations favorables ont donc de grandes chances de se trouver dans les génomes les plus longs. Lorsque le taux de mutation est fort, il est quasiment impossible à ces génomes longs de ne pas subir de mutations non neutres lorsqu'ils se répliquent. La mutation *a priori* favorable a alors de fortes chances d'être liée à des mutations délétères, ce qui empêche sa fixation dans la population. On retrouve ici l'effet "ruby in a rubbish". Ainsi, lorsque le taux de mutation est élevé, le meilleur phénotype d'une génération donnée peut être perdu après quelques pas de temps, et c'est finalement un phénotype moins adapté, comme celui de la figure III.3, qui se maintient sur le long terme. On retrouve ici une idée déjà partiellement formulée par Wilke (2001b) : en pratique, l'évolution n'optimise pas l'adaptation à court terme, mais le succès à long terme de la descendance, qui dépend à la fois de l'adaptation et du niveau de variabilité mutationnelle du phénotype. La variabilité optimale (celle qui est indirectement sélectionnée) étant elle-même un compromis entre "exploitation" des solutions existantes et "exploration" de

nouvelles solutions, il s'agit finalement de concilier adaptation, fidélité de la transmission du phénotype et capacité à innover.

Si l'idée du compromis exploration-exploitation permet de comprendre pourquoi le niveau de variabilité optimal est intermédiaire (le F_ν des individus finaux n'étant ni 0 ni 1), il reste à comprendre plus finement pourquoi il prend spécifiquement telle ou telle valeur, et en particulier pourquoi cette valeur dépend de l'intensité de la sélection. Il s'agit d'une question difficile et ce qui suit constitue davantage une hypothèse à approfondir qu'une démonstration. Considérons les quelques individus les plus adaptés de la population. Le paramètre c, que nous appelons l'intensité de la sélection, contrôle leur probabilité de reproduction relative vis-à-vis des individus les moins adaptés. Lorsque c est proche de 1 (sélection faible), les meilleurs individus n'ont pas beaucoup plus d'essais reproductifs que les autres. Avec seulement $W = 1{,}3$ essais en moyenne si c=0,9995, il leur faut une forte proportion de descendants neutres, de l'ordre de $F_\nu = 0{,}79$, pour obtenir en moyenne 1 descendant neutre, condition *sine qua non* pour la propagation du phénotype. Si la sélection est plus efficace (c plus faible), les meilleurs individus sont assurés d'obtenir un grand nombre d'essais reproductifs et peuvent donc se permettre une plus faible F_ν : si c =0,9900, les meilleurs individus reçoivent environ W =10 essais reproductifs, et alors une F_ν de l'ordre 0,10 seulement suffit pour produire en moyenne 1 descendant neutre. Comme le montre la figure III.6, la valeur de F_ν qui est indirectement sélectionnée pour un c donné est justement très proche de $1/W$, valeur qui permet d'assurer en moyenne 1 descendant neutre : en réalité, les valeurs obtenues permettent d'assurer 1,1 descendant neutre en moyenne. On peut s'interroger sur la généralité de ce critère. Il est clair que produire au moins un descendant neutre est une condition nécessaire à la propagation d'un phénotype d'une génération à l'autre, et que $1/W$ constitue donc la valeur minimale admissible pour F_ν. Il est plus délicat d'expliquer pourquoi la valeur indirectement sélectionnée frôle cette valeur minimale. Cela pourrait "simplement" refléter la sélection directe sur l'adaptation (diminuer g requiert souvent d'augmenter la taille du génome et donc de baisser F_ν), mais cela pourrait aussi refléter la sélection indirecte d'une certaine capacité d'exploration, d'innovation, requise pour gagner la compétition évolutive. Ce dernier effet pourrait être favorisé par la méthode de sélection employée, qui se fonde sur les rangs et crée donc un effet permanent de "course aux armements" analogue à l'effet dit de la Reine Rouge (Van Valen, 1973).

En somme, la chaîne de causalité qui semble se dégager serait la suivante. L'adaptation g d'un individu et l'intensité de la sélection déterminent son nombre d'essais reproductifs, W. La descendance de cet individu se perpétue dans la population si F_ν est de l'ordre de $1/W$: c'est la sélection indirecte d'un niveau de variabilité mutationnelle intermédiaire, qui concilierait fidélité de transmission et capacité d'innovation. Le taux de mutation étant ici fixe dans le temps, c'est en variant la structure du génome, et en particulier la quantité d'ADN non fonctionnel, que la valeur optimale de F_ν peut être atteinte. La sélection indirecte d'un certain niveau de variabilité se traduit alors par la sélection indirecte d'une structure génomique particulière.

4 Robustesse des résultats

Nous avons souligné au chapitre précédent la nécessité de distinguer autant que possible les phénomènes biologiquement pertinents des éventuels artefacts, c'est-à-dire des comportements spécifiquement liés aux composantes arbitraires du modèle ou à certaines valeurs des paramètres. Nous avons donc testé si la relation obtenue entre le taux de mutation et la compacité du génome est généralisable à d'autres paramètres que ceux du tableau III.1 (p. 106).

4.1 Influence de la forme de l'environnement

Nous avons tout d'abord testé la robustesse de cette relation vis-à-vis de la distribution de possibilité de l'environnement, en répétant l'expérience dans un environnement de forme et d'aire différentes. Ce nouvel environnement est représenté en encart dans la figure III.7 : il est unimodal et son aire est de 0,2, contre 0,15 précédemment. Trois simulations ont été effectuées pour chaque taux de mutation, sous une intensité de sélection intermédiaire (c =0,9980). Comme le montre la figure III.7, on retrouve qualitativement et quantitativement l'ensemble des résultats obtenus précédemment. La quantité de non codant se stabilise à une valeur qui dépend fortement du taux de mutation et qui correspond à une valeur constante de F_ν, la proportion de descendants neutres. La valeur de F_ν indirectement sélectionnée est, comme précédemment, très légèrement supérieure à la valeur qui assure à un individu bien adapté une moyenne d'un descendant neutre. Nos conclusions sont donc robustes vis-à-vis de la forme et de l'aire de l'environnement.

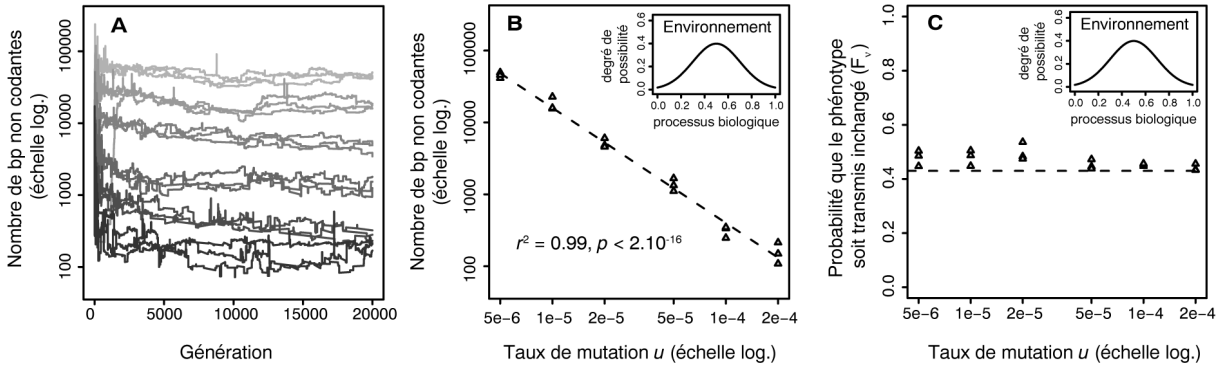


Fig. III.7: Robustesse des résultats vis-à-vis de la forme de l'environnement. **A** : Évolution de la quantité de non codant sur la lignée ancestrale du meilleur individu final, pour un taux de mutation u variant de 5.10^{-6} (gris clair) à 2.10^{-4} (gris foncé) par bp. **B** : Influence du taux de mutation sur la quantité de non codant à l'équilibre, celle-ci étant estimée par la moyenne sur les 5 000 dernières générations. **C** : Comme précédemment, la proportion F_ν de descendants neutres des meilleurs individus finaux est juste au-dessus de la valeur qui assure une moyenne d'un descendant neutre au meilleur individu (ligne horizontale).

4.2 Influence de la méthode de sélection

Dans un second temps, nous avons testé l'effet de la méthode de sélection, en réalisant à nouveau les simulations sous une sélection "fitness-proportionate". Deux valeurs de k ont été testées : $k = 250$ (sélection faible) et $k = 1\,000$ (sélection forte). Pour chaque valeur de k, trois simulations ont été effectuées pour chaque taux de mutation, dans l'environnement initial (voir tableau III.1).

Comme le montre la figure III.8, l'évolution au cours de temps de la quantité de non codant diffère légèrement de celle observée avec un mode de sélection basé sur le rang : après le traditionnel pic initial, la quantité d'ADN non codant décroît au lieu de se stabiliser rapidement. La stabilisation est plus progressive, mais on retrouve le fait que la quantité finale de non codant dépend du taux de mutation, avec des génomes beaucoup plus compacts lorsque le taux de mutation est élevé que lorsqu'il est faible. La figure III.9 suggère que la proportion F_ν de descendants neutres des individus finaux est, comme précédemment, de l'ordre de $1/W$ (W étant le nombre d'essais reproductifs des individus les plus adaptés), ce qui signifie que ces individus produisent en moyenne un descendant neutre à chaque génération. Cependant, avec une sélection "fitness-proportionate", W n'est plus constant, même pour un k fixé. Il dépend en effet de l'adaptation moyenne des autres individus de la population et peut donc varier en fonction du temps et du taux de mutation. W étant alors un paramètre libre, au même titre que F_ν, il est délicat d'interpréter la relation entre les deux quantités : dans quel sens va la causalité ? On peut penser que comme précédemment, le nombre d'essais reproductifs détermine – par sélection indirecte – le F_ν optimal, permettant de produire un nombre de descendants neutres qui concilie exploitation et exploration. Mais en sélection "fitness-proportionate", le F_ν du génotype le plus répandu pourrait aussi déterminer mécaniquement le nombre d'essais reproductifs de ses représentants. F_ν et W seraient alors en interaction mutuelle, ce qui complique sensiblement la chaîne de causalité qui se dégageait précédemment.

Ainsi, avec un mode de sélection directement basé sur les valeurs absolues de l'adaptation, on observe bien une modulation de la compacité du génome en fonction du taux de mutation, reflétant la sélection indirecte des individus présentant un niveau approprié de variabilité mutationnelle. Il est cependant plus difficile que précédemment de déterminer le ou les facteurs qui fixent ce niveau approprié, et une étude spécifique serait nécessaire pour éclaircir ce point. Il faudrait notamment chercher à caractériser la distribution complète de l'adaptation des descendants plutôt que de se limiter à l'indicateur simplifié que constitue F_ν.

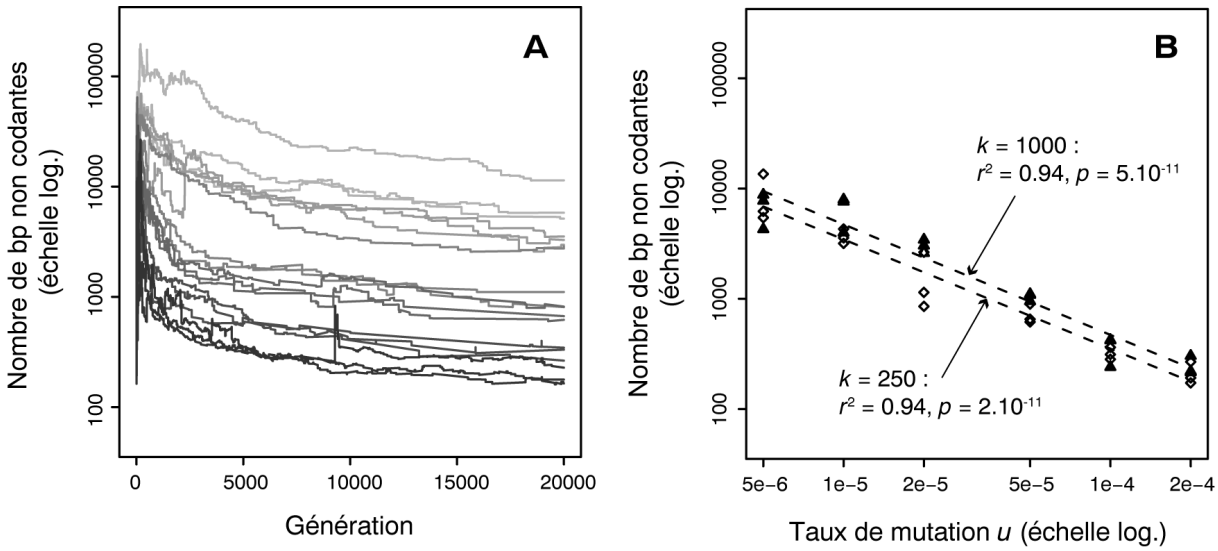


Fig. III.8: Influence du mode de sélection sur la relation entre le taux de mutation et la quantité de non codant. **A :** Évolution de la quantité de non codant sur la lignée ancestrale du meilleur individu final, pour un taux de mutation u variant de 5.10^{-6} (gris clair) à 2.10^{-4} (gris foncé) par bp, dans le cas où $k = 250$. **B :** Influence du taux de mutation sur la quantité finale de non codant, celle-ci étant estimée par la moyenne sur les 5 000 dernières générations. Les losanges correspondent à $k = 250$ et les triangles à $k = 1\ 000$.

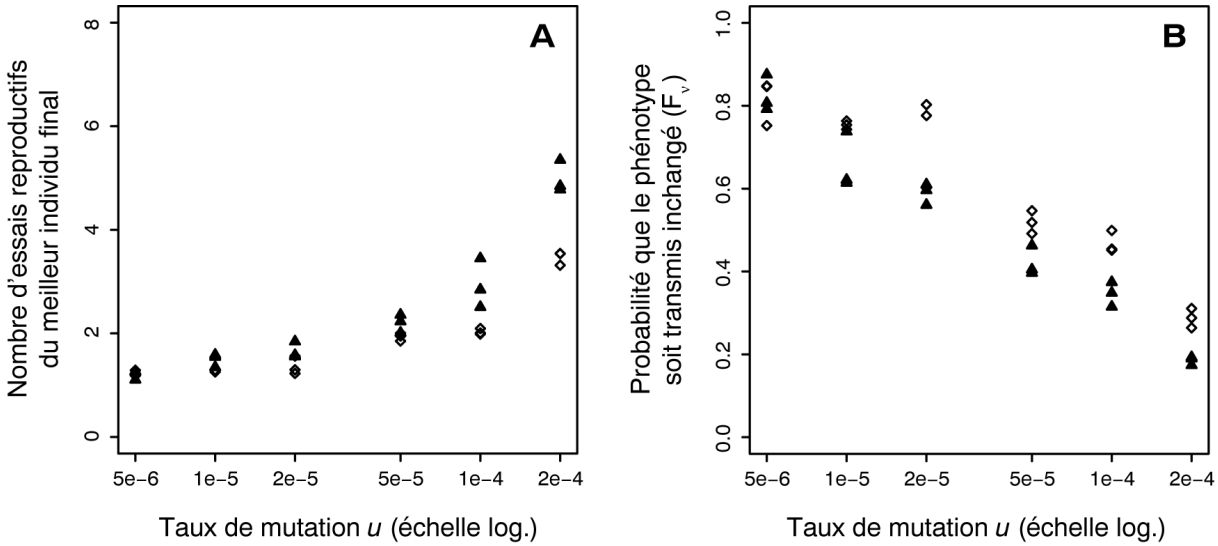


Fig. III.9: Influence du mode de sélection sur le nombre d'essais reproductifs du meilleur individu à $t = 20\ 000$ (A), et la proportion F_ν de descendants neutres de ce meilleur individu (B). Les losanges correspondent à $k = 250$ et les triangles à $k = 1\ 000$.

4.3 Influence de la distribution de la taille des réarrangements

La façon dont la sélection indirecte d'un certain F_ν se traduit au niveau du génome dépend bien sûr de la façon dont ses caractéristiques structurelles – taille, nombre de gènes, distances intergéniques – contribuent à F_ν. Cette contribution dépend de la façon dont les réarrangements sont effectués, et en particulier de la distribution de probabilité de la taille des segments réarrangés. Nous avons déjà établi cette relation dans le cas où la taille des réarrangements suit une loi uniforme entre 1 et L (équations III.1 et III.2). Dans ce cas, F_ν décroît avec la quantité de non codant, ce qui explique la contre-sélection indirecte des génomes longs lorsque le taux de mutation est fort. Mais on peut imaginer qu'avec d'autres patterns mutationnels, et donc avec une autre relation entre F_ν et la quantité de non codant, la sélection d'un certain F_ν se manifeste différemment au niveau du génome. Il semble donc pertinent de s'interroger sur l'effet de la distribution de la taille des réarrangements sur l'évolution structurelle du génome. Si, pour les raisons techniques que nous allons voir, l'étude systématique reste à réaliser, nous avons d'ores et déjà quelques résultats préliminaires dans le cas d'une loi quasi-géométrique de paramètre q (voir section II.2.4, p. 78).

Dans le cas d'une loi quasi-géométrique, nous avons retrouvé la décroissance de F_ν avec la quantité de non codant obtenue sous une loi uniforme, mais cette fois, la valeur asymptotique de F_ν est non nulle (et elle est d'autant plus grande que q est élevé, c'est-à-dire que les réarrangements sont courts). Cela signifie que des génomes même très longs ont une chance non nulle de se transmettre sans mutation majeure, et qu'au-delà d'un certain seuil, cette chance ne diminue plus significativement lorsque la quantité de non codant augmente. Ce seuil est d'autant plus bas que q est élevé. S'il est par hasard franchi au cours du pic initial de duplications massives, alors il n'y a plus d'avantage indirect à raccourcir le génome, puisque cela ne fera pas augmenter F_ν. Plus précisément, seule une diminution suffisamment drastique pour ramener en une fois la quantité de non codant sous le seuil peut être indirectement sélectionnée. Les premiers tests réalisés confirment ces prédictions. Lorsque q est élevé, on n'observe pas toujours la stabilisation de la taille du génome : le hasard des mutations des premières générations fait que d'une répétition à l'autre, la taille du génome peut se stabiliser comme précédemment, ou au contraire franchir la taille critique puis augmenter jusqu'à atteindre la capacité mémoire maximale de la machine. Dans ce dernier cas, la simulation s'arrête après seulement quelques centaines de générations, ce qui pose des problèmes évidents pour réaliser un plan de test complet.

Ces expériences préliminaires montrent la puissance de la notion de variabilité mutationnelle pour prédire le comportement du modèle, et soulignent le rôle clé des réarrangements agissant à grande échelle. La pression indirecte sur la quantité de non codant est en effet levée si q est fort[1], c'est-à-dire si la taille moyenne des réarrangements augmente très peu avec celle du génome : dans ce cas, lorsque le génome est très long, on peut quasiment considérer que tous les événements mutationnels sont locaux. Au contraire, si q est faible (ou si la loi est uniforme, ce qui correspondrait à $q = 0$), la taille moyenne des seg-

[1] En revanche, la pression indirecte sur le nombre de gènes subsiste. L'existence de pressions sélectives indirectes sur l'organisation du génome n'est donc pas complètement remise en cause.

ments réarrangés augmente de façon non négligeable avec celle du génome, les pressions de sélection indirectes peuvent s'exercer sur la compacité du génome.

5 Discussion

Les expériences réalisées avec le modèle *aevol* montrent comment le nombre de gènes et la quantité d'ADN non fonctionnel maintenus dans un génome peuvent être modulés en fonction du taux de mutation, par la sélection indirecte d'une variabilité mutationnelle intermédiaire, conciliant fidélité de transmission et capacité d'innovation. Peut-on s'attendre à ce qu'un tel phénomène se produise dans des génomes réels ? En d'autres termes, la sélection indirecte d'un niveau intermédiaire de variabilité mutationnelle peut-elle aussi contribuer, avec les biais mutationnels et les pressions sélectives directes, à déterminer la compacité d'un génome réel ?

Il est clair que le modèle comporte un certain nombre d'hypothèses simplificatrices. Celles-ci nous ont permis de décrire par de simples équations le lien entre la structure du génome et la variabilité mutationnelle du phénotype, mais ce lien peut être moins direct en réalité.

Tout d'abord, nous avons vu que ce lien dépend de la façon dont se produisent les réarrangements. Deux propriétés apparaissent comme essentielles pour observer une stabilisation de la taille du génome ainsi que sa dépendance au taux de mutation : (i) le nombre d'événements par réplication doit augmenter avec la taille du génome, et (ii) la taille moyenne des réarrangements doit également augmenter significativement avec la taille du génome. Or dans les génomes réels, le nombre et la taille des réarrangements ne dépendent pas aussi directement de la taille du génome que dans le modèle. Ils dépendent du nombre d'éléments répétés, de leur longueur, de leur répartition, ainsi que du rayon d'action des différents mécanismes de recombinaison ectopique. Il semble cependant que chez les procaryotes comme chez les eucaryotes, le nombre d'éléments répétés augmente avec la taille du génome (Achaz *et al.*, 2001, 2002; Frank *et al.*, 2002), ce qui remplirait la première condition. En ce qui concerne la distribution de la taille des réarrangements spontanés, nous avons déjà mentionné au chapitre II qu'elle était difficile à caractériser : d'une part, plusieurs mécanismes agissent simultanément à des échelles différentes, et d'autre part, on n'a souvent accès qu'aux événements fixés. On ne peut donc affirmer avec certitude que la longueur moyenne des segments réarrangés augmente avec la taille du génome, même si cela semble plausible, au moins tant que le génome reste relativement court.

Ensuite, si le modèle prend en compte (de façon simplifiée) le rôle des réarrangements et des mutations locales dans la variabilité mutationnelle, il ignore deux autres sources de variabilité mutationnelle du phénotype, à savoir l'auto-réplication des éléments transposables et la recombinaison allélique. L'auto-réplication des éléments transposables crée un biais mutationnel vers la croissance du génome, mais si les séquences intergéniques sont principalement constituées d'éléments transposables, alors elles deviennent particulièrement mutagènes – ce qui peut causer leur contre-sélection indirecte (Lynch, 2006b). En

effet, d'une part, les éléments transposables peuvent s'insérer dans des gènes, et d'autre part, ils peuvent constituer de meilleurs substrats que les répétitions fortuites pour les mécanismes de recombinaison intrachromosomiques (Achaz, 2002). Dans une population recombinante, la recombinaison allélique est également susceptible de générer de nouveaux phénotypes en créant de nouvelles combinaisons d'allèles. La recombinaison peut notamment rassembler deux mutations favorables qui étaient jusqu'alors en compétition, et contrer ainsi l'effet Hill-Robertson (parfois appelé interférence clonale). De grandes distances intergéniques, augmentant le taux de recombinaison entre les loci sous sélection, pourraient être alors indirectement sélectionnées (Comeron, 2001).

Enfin, nous avons supposé ici que (i) le taux de mutation par base est fixe dans le temps, et (ii) seules les mutations qui n'affectent pas les gènes sont neutres. Or, comme nous l'avons vu au chapitre I, le taux de mutation est susceptible d'évoluer chez les organismes réels, et ceux-ci présentent aussi une multitude de niveaux de robustesse dans la transition génotype-phénotype, qui permettent à certaines mutations géniques d'être neutres. La sélection indirecte d'un niveau donné de variabilité pourrait donc se manifester sur le taux de mutation et sur les mécanismes de canalisation en même temps que sur la structure du génome[1].

Pour toutes ces raisons, il semble que l'effet que nous avons pu isoler à l'aide du modèle – une pression indirecte sur le niveau de variabilité qui se manifeste par une relation entre la compacité du génome et le taux de mutation – soit susceptible d'exister dans les génomes réels, mais qu'il soit difficile à révéler par comparaison directe d'espèces différentes. Une telle relation a cependant été obtenue expérimentalement par Drake (1991) : pour différentes espèces microscopiques comprenant des bactériophages, une bactérie, une levure et un champignon, Drake a obtenu une relation linéaire entre le logarithme de la taille du génome et le logarithme du taux de mutation par base. Ces données pourraient donc refléter, comme dans nos simulations, la sélection indirecte de la taille du génome qui permet le meilleur compromis entre robustesse et capacité d'innovation, étant donné le taux de mutation. En effet, si les espèces testées partagent approximativement un mode de reproduction asexué, la même intensité de sélection, les mêmes patterns mutationnels au niveau des réarrangements, les mêmes niveaux de robustesse dans la transition génotype-phénotype, etc., alors on peut effectivement s'attendre à ce qu'elles s'alignent sur les graphes $\log(\text{nombre de gènes}) = f(u)$ et $\log(\text{non codant}) = f(u)$, comme dans nos simulations. Si, au contraire, on compare des espèces qui diffèrent vraisemblablement en termes d'intensité de sélection, de patterns mutationnels et de robustesse dans la transition du génotype au phénotype (Lynch, 2006a), alors ces deux relations ne peuvent être visibles.

[1] On peut cependant penser que l'évolution de ces trois composantes ne se fait pas nécessairement au même rythme : s'il est relativement facile de baisser la variabilité mutationnelle en excisant des séquences non fonctionnelles répétitives, il semble plus difficile de le faire en augmentant la fidélité de la réplication. Le taux de mutation pourrait donc évoluer plus lentement que la compacité du génome, et celle-ci pourrait donc s'ajuster à tout moment au taux de mutation courant.

6 Conclusion

Les expériences décrites dans ce chapitre illustrent bien le potentiel du modèle *aevol* en tant que générateur d'hypothèses : les simulations ont révélé l'existence d'une pression sélective indirecte sur le niveau de variabilité mutationnelle, qui se traduit au niveau de la structure du génome. En analysant ce phénomène dans un système simple comme celui-ci, on est mieux armé pour en rechercher les traces dans les génomes réels. Nous avons pu par exemple montrer que la pression sélective indirecte mise en évidence se traduit, tout étant égal par ailleurs, par une influence du taux de mutation sur le nombre de gènes et sur la quantité de non codant. Les données expérimentales bien connues de Drake (1991) ont ainsi pu être interprétées comme une signature possible du phénomène. On peut envisager de tester si ce type de pression indirecte peut également expliquer, au moins en partie, l'évolution réductive des génomes des bactéries endocytobiotes (Andersson et Andersson, 1999; Charles *et al.*, 1999; Moran et Wernegreen, 2000; Moran et Mira, 2001; Ochman et Moran, 2001; Silva *et al.*, 2001; Gil *et al.*, 2002), en étudiant l'effet d'un taux de mutation ponctuel plus élevé (Moran et Mira, 2001; Itoh *et al.*, 2002) dans le contexte d'une réduction de la taille efficace de la population (Moran et Wernegreen, 2000) et d'une stabilisation et/ou simplification de l'environnement (Moran et Mira, 2001).

L'apport potentiel du modèle ne se limite pas pour autant à l'étude de la taille des génomes. On peut en effet s'interroger sur l'effet indirect du taux de mutation sur d'autres caractéristiques structurelles du génome, comme la répartition des gènes sur le chromosome : sont-ils régulièrement répartis ou forment-ils des "clusters" ? Les équations III.1 et III.2 montrent en effet que la variance des distances intergéniques ($\mathrm{Var}(\lambda_i)$) influence aussi la proportion de descendants neutres, et donc la variabilité mutationnelle du phénotype. Une redistribution des gènes à taille de chromosome constante peut donc permettre un ajustement fin de la variabilité mutationnelle. Dans ce cas, il serait pertinent de tester si un changement du taux de mutation s'accompagne d'un changement de répartition des gènes.

Enfin, à taux de mutation constant, il est également possible d'étudier l'effet sur le génome d'une autre composante de la variabilité mutationnelle : l'impact moyen d'une mutation génique. La sélection indirecte d'un niveau donné de variabilité peut-elle là aussi induire un couplage inattendu, entre cet impact moyen et la structure du génome ? C'est ce que nous abordons dans le chapitre suivant.

Chapitre IV

Structuration du génome en fonction de l'effet des mutations

Donc toutes choses étant causées et causantes, aidées et aidantes, médiates et immédiates et toutes s'entretenant par un lien naturel et insensible qui lie les plus éloignées et les plus différentes, je tiens impossible de connaître les parties sans connaître le tout, non plus que de connaître le tout sans connaître particulièrement les parties.

Blaise Pascal[1]

Les simulations présentées au chapitre précédent ont montré que, tout étant égal par ailleurs, la structure du génome s'ajuste spontanément en fonction du taux de mutation par base. Ce couplage inattendu provient du fait que le taux de mutation et la structure du génome sont deux composantes d'une seule et même propriété indirectement sélectionnée : la variabilité mutationnelle globale du phénotype entre progéniteur et descendance. Or, comme nous l'avons vu au chapitre I, cette variabilité globale dépend aussi d'une troisième composante, à savoir la façon dont le phénotype varie lorsqu'un gène est modifié ou perdu. On peut donc se demander si, à taux de mutation égal, la structure du génome peut spontanément s'ajuster en fonction de l'impact phénotypique des mutations géniques. Par exemple, si modifier ou perdre un gène peut causer un fort impact sur le phénotype, va-t-on indirectement sélectionner une structure de génome qui permet de muter peu de gènes à chaque reproduction ?

Nous avons vu au chapitre I que chez les êtres vivants, l'impact phénotypique d'une mutation dans un gène résulte d'une multitude de paramètres, comme la dégénérescence du

[1] *Pensées* in Oeuvres Complètes, p. 527, Seuil, 1963.

code génétique, la robustesse de la fonction de la protéine vis-à-vis d'un changement de séquence, ou encore la position de la protéine dans le réseau d'interactions fonctionnelles. Ce réseau d'interactions fonctionnelles englobe les interactions physiques, mais aussi les réseaux de régulation, les réseaux métaboliques et les cascades de signalisation. Même pour les espèces modèles, ce "réseau de réseaux" (Barabasi et Oltvai, 2004) est loin d'être complètement connu, et sa complexité est telle qu'il est extrêmement difficile de prédire tous les effets qu'une mutation peut avoir sur le phénotype d'un organisme, même unicellulaire. On sait cependant qu'une mutation a plus de chances d'être létale si elle affecte une protéine centrale dans ce réseau (Jeong *et al.*, 2001). Nous pouvons donc reformuler plus précisément notre questionnement : la présence de protéines centrales peut-elle affecter indirectement la structure du génome, en raison de l'effet potentiellement dramatique de leurs mutations ? En d'autres termes, la structure du génome indirectement sélectionnée est-elle différente selon que la correspondance gènes-traits est bijective ou, au contraire, dégénérée ?

Là encore, l'expérimentation *in silico* s'avère particulièrement utile pour tester la pertinence de ce questionnement. Il est en effet beaucoup plus aisé de contrôler les propriétés générales de la correspondance gènes-traits dans un modèle tel qu'*aevol* que dans un organisme réel. Nous pouvons ici contrôler l'effet maximal des mutations géniques en fixant la pléiotropie maximale des protéines, grâce au paramètre $w_{\max}$. Celui-ci définit la taille maximale de l'ensemble de processus biologiques auxquels une protéine contribue (largeur des "triangles"). Lorsque $w_{\max}$ est faible, toutes les protéines sont spécialisées sur une gamme étroite de processus, alors que lorsque $w_{\max}$ est élevé, on autorise l'apparition de protéines qui interviennent dans de nombreux processus, interagissent avec de nombreuses autres protéines et ont un grand poids dans le phénotype de l'organisme.

Dans ce chapitre, nous étudions donc l'influence de ce paramètre sur l'évolution du génome. Après avoir présenté le plan d'expérience (section 1), nous nous assurerons qu'augmenter $w_{\max}$ permet effectivement de faire apparaître des protéines très pléiotropes et d'augmenter l'impact des mutations géniques (section 2). Nous verrons ensuite comment cela influence le nombre de gènes et la quantité de non-codant, et nous montrerons que cet ajustement de la structure génomique est, comme dans le chapitre précédent, lié à la sélection indirecte d'un niveau constant de variabilité mutationnelle (section 3). Dans une quatrième section, nous montrerons que la présence de protéines très pléiotropes influence aussi la dynamique de l'ordre des gènes. Enfin, nous discuterons ces résultats dans le contexte des génomes réels (section 5). Ce travail fait l'objet des articles 1, 4 et 9.

1 Plan d'expérience

L'objectif de cette série d'expérience est de tester si, à taux de mutation constant, le niveau de pléiotropie des protéines peut influencer la structure du génome. Nous avons choisi de fixer tous les taux de mutation à 10^{-5} par bp, ce qui permet d'obtenir des génomes avec suffisamment de gènes et suffisamment peu de chevauchements pour que

l'étude de la dynamique de l'ordre des gènes ait un sens. Nous avons testé six valeurs pour le paramètre $w_{\max}$, qui fixe la pléiotropie maximale des protéines (largeur maximale des "triangles" protéiques) : 0,01, 0,02, 0,033, 0,1, 0,2 et 0,33. Pour chaque valeur de $w_{\max}$, nous avons fait évoluer indépendamment cinq populations pendant 40 000 générations. La sélection est dans un premier temps basée sur les rangs, car comme nous l'avons vu au chapitre précédent, cela permet un meilleur contrôle et une meilleure compréhension des relations de causalité. Un second jeu de simulations menées avec une sélection "fitness-proportionate" nous montrera cependant que les principaux résultats sont indépendants de la méthode de sélection. Les autres paramètres utilisés figurent dans le tableau IV.1.

Paramètre	Symbole	Valeur
Taille de la population	N	1000 individus
Taille initiale des génomes	L_{init}	5000 bp
Séquence promotrice	-	01010110011100100101 10, $d_{\max} =$ 4 différences autorisées
Séquence terminatrice	-	De la forme $abcd***\bar{d}\bar{c}\bar{b}\bar{a}$
Signal d'initiation de la traduction	-	011011***000
Signal de terminaison de la traduction	-	001
Code génétique	-	Voir tableau II.1, p. 75
Ensemble des processus biologiques	Ω	$[0, 1]$
Pléiotropie maximale des protéines	$w_{\max}$	De 0,01 à 0,33
Distribution de possibilité de l'environnement	f_E	Trimodale, comme sur la figure II.3, et fixe au cours du temps
Méthode de sélection	-	"Linear ranking"
Intensité de la sélection	η^+	1.998
Taux de mutation ponctuelle	u_{ponct}	10^{-5} par bp
Taux de petite insertion	u_{ins}	10^{-5} par bp
Taux de petite délétion	u_{del}	10^{-5} par bp
Taux de grande délétion	u_{gdel}	10^{-5} par bp
Taux d'inversion	u_{inv}	10^{-5} par bp
Taux de duplication	u_{dup}	10^{-5} par bp
Taux de translocation	u_{transloc}	10^{-5} par bp
Loi de la longueur des petits indels	-	Loi uniforme entre 1 et 6 bp
Loi de la longueur des segments réarrangés	l_{seg}	Loi uniforme entre 1 et L

Tab. IV.1: Paramètres utilisés pour tester l'influence de la pléiotropie des protéines sur l'évolution structurelle du génome.

2 Vérification de l'effet de $w_{\max}$ sur le protéome

Même dans un modèle simple comme *aevol*, l'impact moyen des mutations géniques n'est pas contrôlable directement. Pour le modifier, nous devons donc agir à travers un autre

paramètre plus accessible et qui lui est *a priori* lié, tel que $w_{\max}$. Il faut donc avant toute chose vérifier cette corrélation, c'est-à-dire s'assurer qu'en modifiant $w_{\max}$, nous modifions effectivement l'allure du protéome et l'effet des mutations. Rappelons que fixer $w_{\max}$ n'impose pas directement une largeur donnée aux "triangles" protéiques, mais simplement une borne supérieure. En augmentant $w_{\max}$, on autorise l'apparition de protéines très pléiotropes, mais cette possibilité offerte pourrait ne pas être utilisée ou être contre-sélectionnée. Il convient par conséquent de contrôler la présence de telles protéines chez les organismes évolués sous un $w_{\max}$ élevé.

La figure IV.1 montre des exemples de protéomes obtenus après 40 000 générations. On constate que les protéines très pléiotropes sont effectivement utilisées lorsqu'on les autorise. De telles protéines interagissent fonctionnellement avec de nombreuses autres protéines (recouvrement des triangles) et apparaissent donc comme très connectées sur la représentation du réseau d'interactions fonctionnelles. Pour tester si la présence de ces protéines augmente l'effet des mutations géniques, nous avons effectué des "knock-out" systématiques sur tous les gènes du meilleur individu final de chaque run : nous avons mesuré la perte d'adaptation causée par la délétion de chaque gène. La figure IV.1 montre que certaines pertes de gènes peuvent être dramatiques si $w_{\max}$ est élevé. La figure IV.2 montre que l'impact moyen d'une perte de gène dépend clairement de $w_{\max}$. Cette analyse confirme que ce paramètre est un levier approprié pour modifier l'impact des mutations géniques. Il est donc pertinent d'étudier son effet sur l'évolution de la structure du génome.

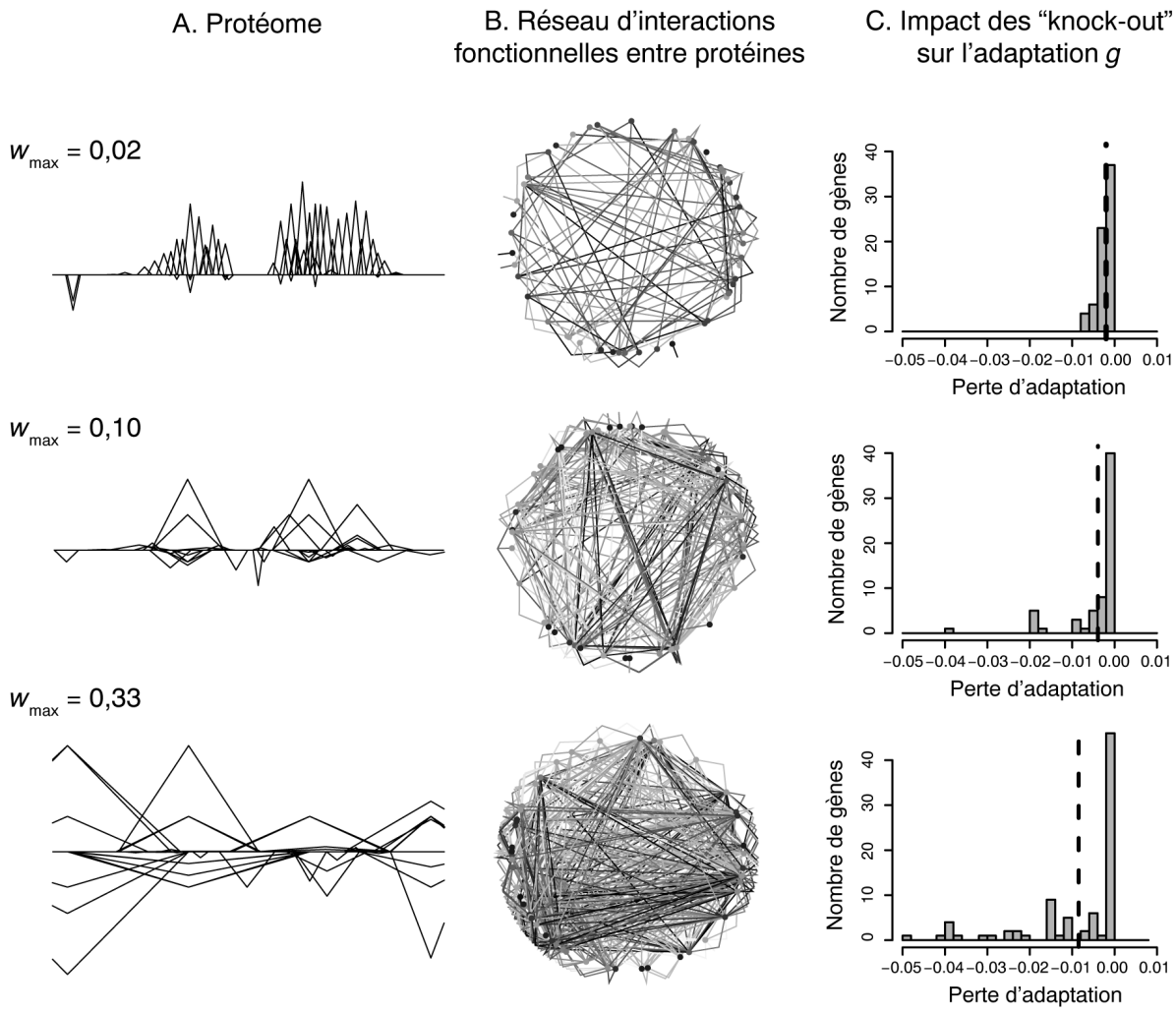


Fig. IV.1: Effet de $w_{\max}$ sur l'allure du protéome. La figure montre le protéome du meilleur individu final d'une des cinq populations, pour $w_{\max}$ =0,02, 0,1 et 0,33. **A** : Superposition des distributions de possibilité des protéines. Les triangles de hauteur négative correspondent à des protéines inhibitrices. **B** : "Réseau" d'interactions fonctionnelles entre protéines, selon les conventions de la figure II.6, p. 85. **C** : Résultats de l'expérience de knock-out systématique décrite dans le texte principal. Il s'agit de l'histogramme de la perte d'adaptation qui serait causée par la perte de chaque gène (une perte d'adaptation de -0,01 par exemple signifie que l'écart g entre le phénotype et l'optimum environnemental augmente de 0,01). La ligne verticale pointillée en indique la moyenne.

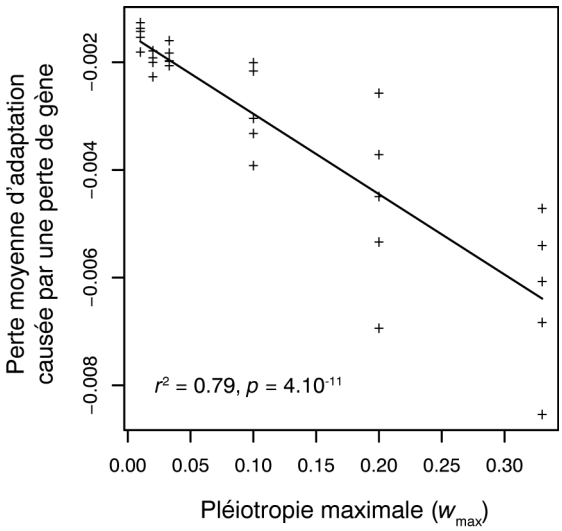


Fig. IV.2: Effet de $w_{\max}$ sur la perte moyenne d'adaptation causée par une perte de gène (résultats complets de l'expérience de knock-out systématique menée sur le meilleur individu final de chaque simulation). La p-value indiquée sur le graphique est à prendre avec précaution car l'hypothèse d'égalité des variances n'est pas vérifiée.

3 Évolution du nombre de gènes et de la quantité de non codant

Dans cette section, nous allons nous focaliser sur deux composantes de la structure du génome qui se sont déjà révélées importantes dans le chapitre précédent : le nombre de gènes et la quantité de non codant. Nous allons dans un premier temps présenter factuellement leur dépendance vis-à-vis de $w_{\max}$, puis nous montrerons que la sélection indirecte d'un niveau donné de variabilité mutationnelle est à nouveau le mécanisme sous-jacent aux relations observées. Enfin, nous commenterons l'effet de la méthode de sélection sur ces résultats.

3.1 Nombre de gènes et quantité de non codant à l'équilibre

Quelle que soit la valeur de $w_{\max}$, l'évolution temporelle de la taille du génome suit ici les phases traditionnelles d'acquisition de gènes par duplication-divergence puis de réduction de la quantité de non codant jusqu'à ce qu'une taille d'équilibre soit atteinte (figure IV.3). Le nombre de gènes se stabilise également, mais plus lentement, et après une durée variable selon $w_{\max}$. Lorsque $w_{\max}$ est élevé, le processus d'acquisition de gènes s'arrête tôt, si bien que le génome final contient d'autant moins de gènes que $w_{\max}$ est fort (figure IV.4A).

À première vue, cela semble découler logiquement du fait d'avoir autorisé de plus grands triangles : quand $w_{\max}$ est plus élevé, moins de triangles sont nécessaires pour couvrir la gamme de processus. Pourtant, comme la distribution de possibilité que nous avons choisie pour l'environnement n'est pas linéaire par morceaux (voir paragraphe II.3.3, p. 92), elle ne peut pas être réalisée exactement en sommant un petit nombre de triangles. Le phénotype peut théoriquement toujours être affiné par l'ajout de triangles activateurs ou inhibiteurs. Ces fines améliorations sont cependant de plus en plus longues à obtenir : seules quelques mutations très précises peuvent améliorer le phénotype lorsque celui-ci est déjà relativement proche de l'optimum. Si, pour une raison quelconque, un $w_{\max}$ élevé amplifie ce phénomène en rendant ces mutations encore plus rares, alors cela pourrait expliquer que le nombre de gènes diminue lorsque $w_{\max}$ augmente. Mais il semble que cette diminution reflète aussi un processus plus global de rétrécissement du génome. En effet, comme le montre la figure IV.4B, $w_{\max}$ a aussi un effet (moins fort mais statistiquement significatif) sur la quantité de non codant.

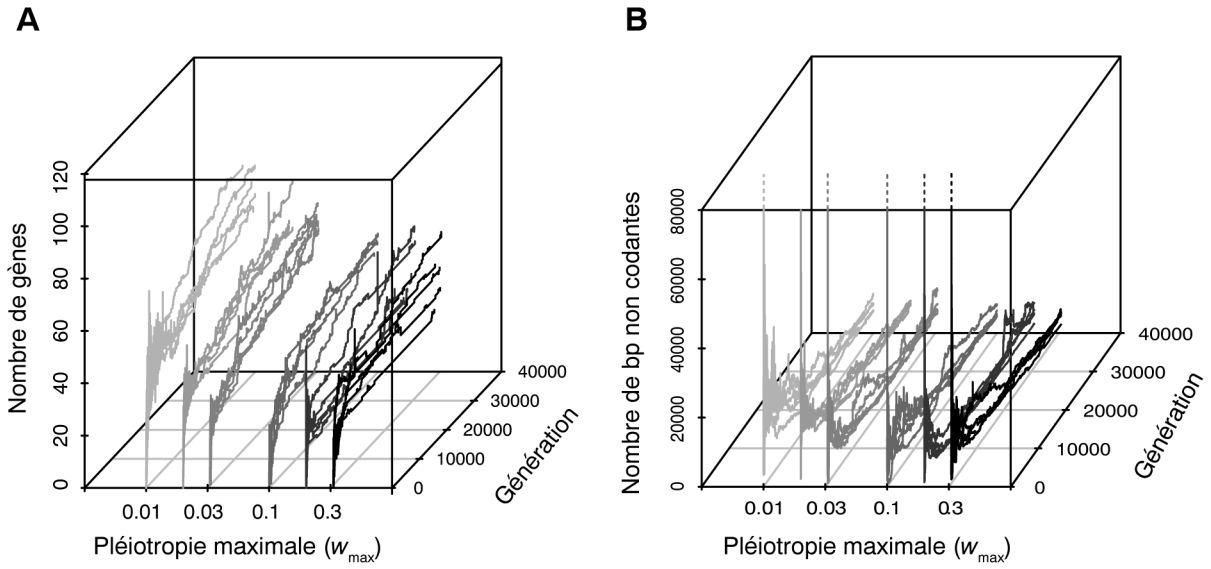


Fig. IV.3: Évolution du nombre de séquences codantes (A) et de la quantité de non codant (B) sur la lignée ancestrale du meilleur individu final, pour les différentes valeurs de $w_{\max}$ testées.

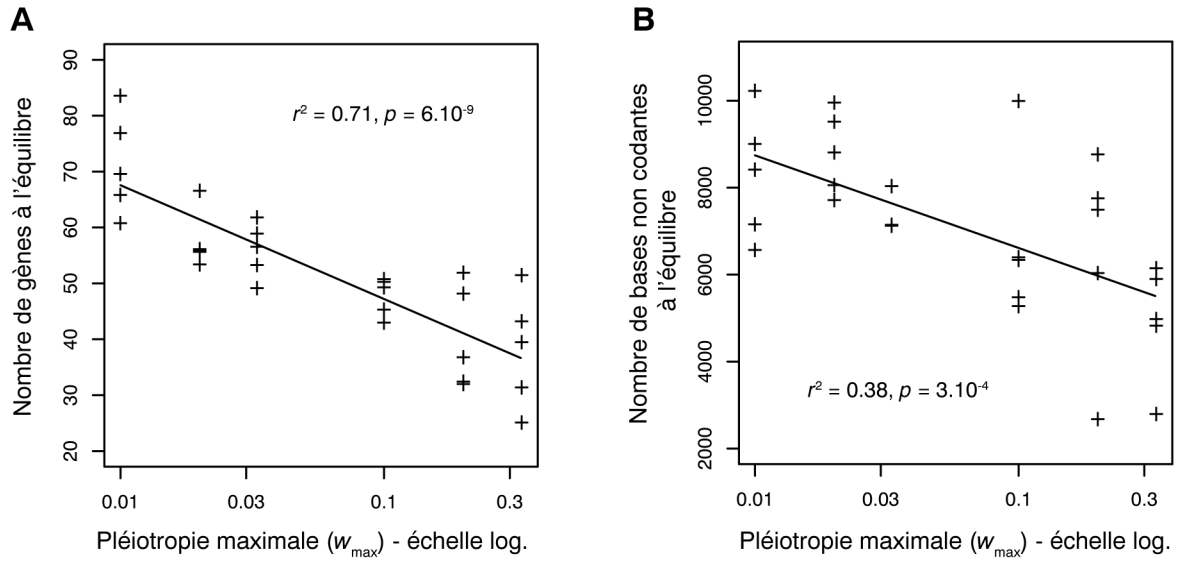


Fig. IV.4: Influence de $w_{\max}$ sur le nombre de séquences codantes (A) et sur la quantité de non codant (B) à l'équilibre. Pour chaque simulation, la valeur à l'équilibre est estimée par la moyenne des 10000 dernières générations, c'est-à-dire des 10000 derniers ancêtres du meilleur individu final. L'effet de $w_{\max}$ sur la quantité de non codant est moins fort que sur le nombre de gènes, mais il reste statistiquement significatif.

3.2 Sélection indirecte du niveau de variabilité mutationnelle

Il nous faut expliquer comment $w_{\max}$ peut agir sur un paramètre – la quantité de non codant – qui ne contribue pas à l'adaptation immédiate des individus, mais qui peut in-

fluencer leur variabilité mutationnelle. En augmentant, à travers $w_{\max}$, l'effet moyen des mutations géniques, nous avons augmenté cette variabilité mutationnelle. Les changements observés au niveau de la structure génomique pourraient-ils alors correspondre à une compensation de cette perturbation, permettant de maintenir une variabilité constante et donc un bon compromis exploitation-exploration ? Pour tester cette hypothèse, il est nécessaire d'estimer la variabilité mutationnelle globale des individus finaux. Pour cela, au chapitre précédent, nous avions pu utiliser l'indicateur simple et analytiquement prédictible qu'est la proportion F_ν de descendants neutres, mais cet indicateur partiel ne convient pas ici. En effet, il ne fait que distinguer les mutations neutres des mutations non neutres, sans prendre en compte la "gravité" des mutations non neutres. Comme c'est précisément cette gravité que nous modifions à travers $w_{\max}$, il nous faut impérativement un indicateur qui la prenne en compte, comme la différence moyenne d'adaptation entre un progéniteur et l'ensemble de ses descendants (perte moyenne d'adaptation par réplication). Nous avons estimé empiriquement cette perte moyenne d'adaptation en simulant 50000 réplications indépendantes du meilleur individu final de chaque run, avec le même taux de mutation que pendant l'évolution. Comme le montre la figure IV.5, cette perte moyenne d'adaptation ne dépend pas de $w_{\max}$, ce qui signifie que les individus finaux présentent tous le même niveau global de variabilité mutationnelle, quel que soit l'impact d'une perte de gène. La raison en est que le *nombre* de gènes mutés par réplication est plus faible lorsque $w_{\max}$ est plus élevé (figure IV.5). Rappelons que le taux de mutation par bp est le même pour toutes ces simulations : c'est donc bien au niveau de la structure du génome que l'effet se produit.

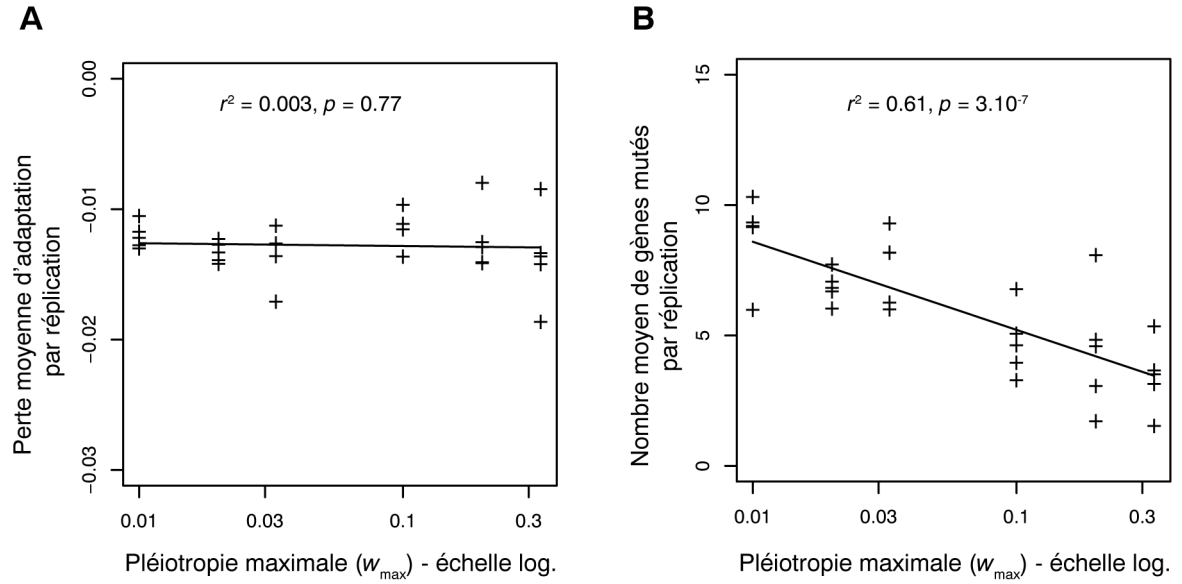


Fig. IV.5: A : Invariance de la perte moyenne d'adaptation *par réplication* vis-à-vis de $w_{\max}$, en dépit de l'influence de $w_{\max}$ sur l'impact d'une mutation génique (voir figure IV.2). **B :** Cette invariance s'explique par le fait que lorsque $w_{\max}$ augmente, la structure génomique indirectement sélectionnée est telle que moins de gènes sont mutés à chaque réplication.

Comment la structure du génome peut-elle influencer le nombre de gènes mutés à chaque réplication ? Il est clair qu'en ce qui concerne les mutations locales (mutations ponctuelles,

petites insertions et petites délétions), le nombre de gènes mutés est directement proportionnel au nombre total de gènes. Le nombre de gènes a donc une influence évidente sur le nombre de gènes qui subissent une mutation locale. Le rôle de la quantité de non codant est moins trivial, mais il existe cependant, dès lors que les grandes délétions et les duplications sont prises en compte.

Considérons deux génomes comportant un même nombre N_G de gènes (répartis régulièrement sur le chromosome), mais différant par la quantité de non codant : l'un est de longueur L et l'autre de longueur $L' > L$. Si u est le taux de grandes délétions par bp, le premier génome va subir en moyenne uL grandes délétions par réplication, contre uL' pour le second. Si la longueur des segments réarrangés suit la loi uniforme entre 1 et L, alors chaque grande délétion touche en moyenne $N_G/2$ gènes, quelle que soit la quantité d'ADN intergénique[1]. Ainsi, au total, une réplication du premier génome se solde par une perte moyenne de $uLN_G/2$ gènes, contre $uL'N_G/2$ pour le second. Moins de gènes sont donc perdus en moyenne dans le premier génome. Si, en plus de contenir moins de non codant, ce premier génome contient aussi moins de gènes, l'effet est encore renforcé. Le même raisonnement vaut pour les duplications.

En somme, tous types de mutation confondus, il est possible de modifier le nombre moyen de gènes mutés par réplication en modifiant le nombre de gènes total et/ou la quantité de non codant. Cela peut permettre de retrouver une variabilité mutationnelle globale acceptable, malgré la présence de protéines très pléiotropes, c'est-à-dire malgré l'impact potentiellement fort des mutations géniques. Dans la compétition évolutive, les individus qui présentent (par hasard) la structure génomique adéquate étant donné l'impact des mutations géniques sont alors – de fait – ceux dont la descendance se maintient sur le long terme. La sélection indirecte d'un niveau donné de variabilité mutationnelle peut donc induire un couplage inattendu entre le niveau fonctionnel (pléiotropie des protéines) et le niveau génomique (quantité de non codant).

3.3 Influence de la méthode de sélection

L'argumentation précédente repose sur l'invariance de la perte moyenne d'adaptation vis-à-vis de $w_{\max}$. On pourrait y opposer que cette invariance ne reflète pas la sélection indirecte d'un niveau donné de variabilité, mais simplement le fait que la valeur exacte de l'adaptation importe peu puisque la sélection est basée sur les rangs. Nous avons donc reconduit l'ensemble de l'expérience avec la méthode de sélection "fitness-proportionate", qui assigne directement les probabilités de reproduction en fonction des valeurs brutes d'adaptation (le paramètre k étant ici fixé à 100).

Dans ce nouveau jeu de données, l'évolution temporelle du génome suit les deux phases

[1] Comme nous l'avons noté au chapitre précédent, si la longueur moyenne des segments excisés augmente linéairement avec L (c'est le cas si l_{seg} suit la loi uniforme), la probabilité $\tilde{\nu}_{\text{gdel}}$ qu'une grande délétion soit neutre n'augmente pas avec la quantité d'ADN intergénique. Cela signifie que l'ADN intergénique ne protège pas les gènes des grands réarrangements.

habituelles de croissance puis de stabilisation. Comme précédemment, le nombre de gènes et la quantité de non codant se stabilisent après un laps de temps variable selon $w_{\max}$ (figure IV.6). On retrouve le fait que leurs valeurs à l'équilibre diminuent lorsque $w_{\max}$ augmente, bien qu'elles diffèrent quantitativement du celles du jeu de données initial. Il faut notamment souligner que l'échelle du graphe IV.6D est logarithmique : la quantité finale de non codant semble ici décroître exponentiellement plutôt que linéairement. Ces différences quantitatives entre les deux jeux de données étaient attendues, puisque le chapitre précédent a montré que l'intensité de la sélection influence aussi la compacité du génome. Nous avons choisi k, le paramètre du mode "fitness-proportionate", de sorte à obtenir approximativement la même intensité de sélection qu'en ranking une fois que la plupart des gènes sont acquis. Mais une correspondance exacte est impossible à obtenir car lorsque la sélection est basée sur les valeurs brutes d'adaptation, son intensité décroît au fur et à mesure du temps. Malgré cela, l'effet qualitatif de $w_{\max}$ sur l'allure du génome persiste, ce qui suggère que le comportement du modèle est relativement robuste vis-à-vis du mode de sélection.

Il reste à vérifier que la compacité accrue du génome lorsque $w_{\max}$ est élevé permet de compenser l'effet aggravé des mutations géniques. Nous avons testé cette hypothèse en simulant, comme précédemment, 50000 réplications indépendantes du meilleur individu final de chaque simulation. Comme le montre la figure IV.7B, la perte moyenne d'adaptation par réplication n'est pas aggravée lorsqu'une plus grande pléiotropie est autorisée. Ici, on observe même une tendance inverse[1]. Ceci est surprenant car chez ces individus, chaque perte de gène est en moyenne plus coûteuse (figure IV.7A). Cela signifie que le nombre moyen de gènes mutés par réplication – qui dépend lui-même de la structure du génome – est une composante très importante de la variabilité mutationnelle globale. La figure IV.7C montre en effet que grâce à leur structure génomique plus compacte, les individus finaux obtenus pour un $w_{\max}$ élevé subissent moins de mutations géniques à chaque réplication. Ainsi, sous un $w_{\max}$ élevé, les individus dont la structure génomique (compacte) permettait de muter moins de gènes à chaque réplication ont donc pu compenser l'impact accru des mutations géniques et sont donc ceux qui ont été indirectement sélectionnés.

En conclusion, ce jeu de données supplémentaire montre que même si le mode de sélection peut jouer sur les valeurs numériques des différents indicateurs, il ne modifie pas les grandes tendances mises en évidence précédemment. On retrouve un couplage entre les propriétés fonctionnelles des protéines et la compacité du génome, couplage dû au fait que ces deux niveaux sont impliqués dans la variabilité mutationnelle globale du phénotype.

[1] Pour être plus précis, la variabilité mutationnelle globale est à peu près constante lorsque $w_{\max}$ varie, sauf pour la plus faible valeur de $w_{\max}$ (0,01). Le niveau de variabilité indirectement sélectionnée semble plus grand dans ce cas précis, ce qui cause la "tendance" observée. Pourquoi l'exploration serait-elle davantage favorisée dans ce cas ? La comparaison des niveaux d'adaptation à $t = 40\,000$ montre que la population est moins bien adaptée lorsque $w_{\max}$ =0,01 (il est en effet difficile d'approcher l'optimum environnemental par une combinaison de triangles exclusivement étroits). Sous un mode de sélection "fitness-proportionate", cela signifie que la population subit encore une sélection directionnelle forte. Il semblerait donc logique que l'exploration de nouveaux phénotypes soit davantage favorisée dans ce cas.

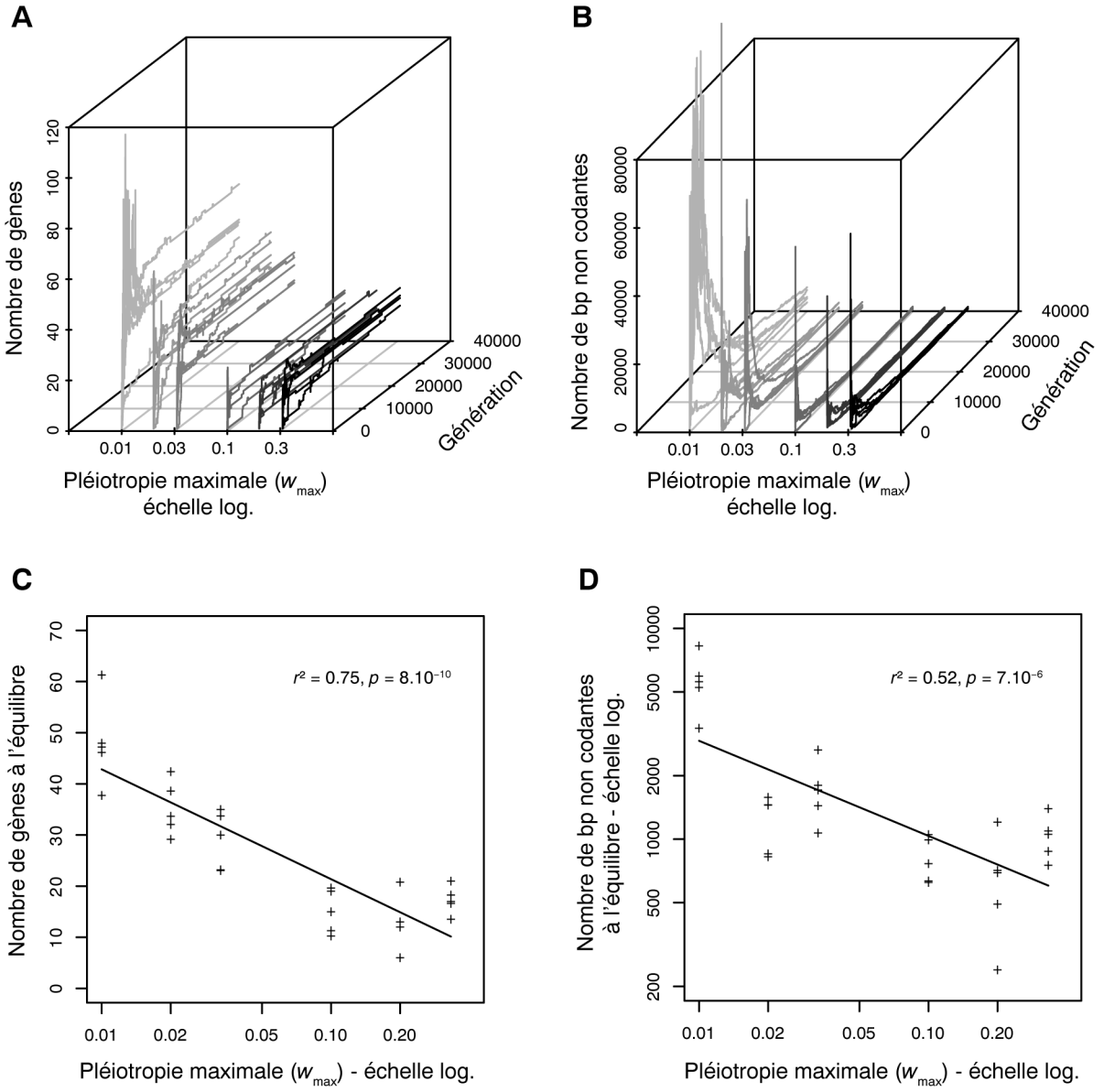


Fig. IV.6: Relation entre $w_{\max}$ et la compacité du génome sous une sélection "fitness-proportionate". **Haut :** Évolution du nombre de séquences codantes (A) et de la quantité de non codant (B) sur la lignée ancestrale du meilleur individu final, sous un régime de sélection basé sur les valeurs absolues d'adaptation. **Bas :** Influence de $w_{\max}$ sur le nombre de séquences codantes (C) et sur la quantité de non codant (D) à l'équilibre, sous un régime de sélection basé sur les valeurs absolues d'adaptation. Pour chaque simulation, la valeur à l'équilibre est estimée par la moyenne des 10000 dernières générations, c'est-à-dire des 10000 derniers ancêtres du meilleur individu final.

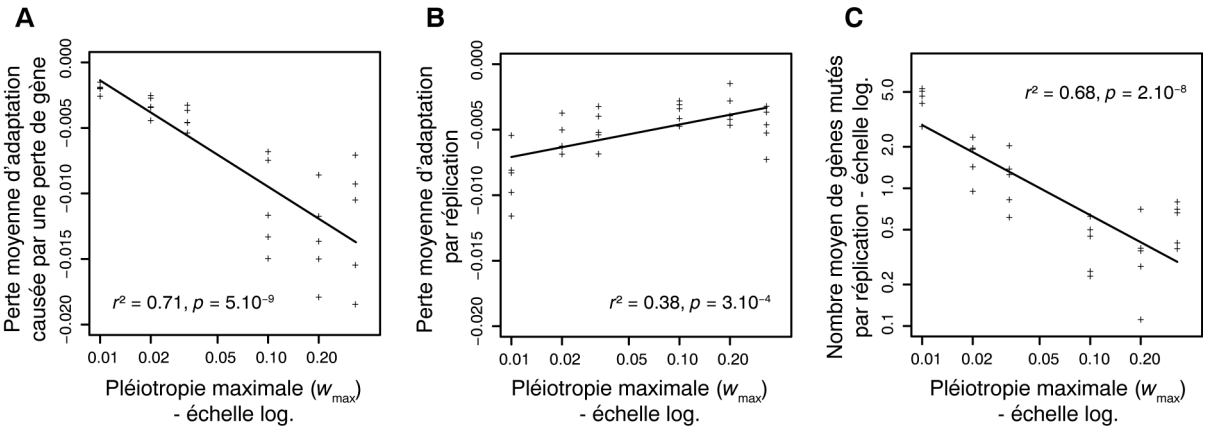


Fig. IV.7: Variabilité mutationnelle globale sous une sélection "fitness-proportionate". **A :** Effet de w_{max} sur la perte moyenne d'adaptation causée par une perte de gène. Pour chaque simulation, les gènes du meilleur individu final ont été supprimés tour à tour ; le graphe présente la moyenne sur tous les gènes des pertes d'adaptation résultantes. Les mutations géniques ont donc un impact plus fort lorsque w_{max} est augmenté. **B :** Malgré cela, la perte moyenne d'adaptation par réplication n'est pas aggravée ; elle tend même à se rapprocher de 0. **C :** Comme précédemment, cela est dû au fait que lorsque w_{max} augmente, moins de gènes sont mutés à chaque réplication.

4 Évolution de l'ordre des gènes

Nous nous sommes jusqu'à présent focalisés sur deux caractéristiques du génome qui influencent fortement la variabilité mutationnelle du phénotype, à savoir le nombre de gènes et la quantité de non codant. Cette variabilité peut néanmoins être aussi modulée plus finement par d'autres aspects de l'organisation du génome. Par exemple, la position relative des gènes détermine si leurs destins mutationnels sont liés. Ainsi, si les gènes d'un même réseau de régulation sont éparpillés sur le chromosome, alors les duplications ont de grandes chances de n'affecter qu'une partie du réseau et donc de causer des problèmes de dosage ; si, au contraire, les gènes du réseau sont rassemblés en un seul cluster, alors les duplications pourront affecter le réseau entier et donc causer moins de dommages (Wagner, 1994; Shimeld, 1999). Dans le modèle, des effets de ce type pourraient se produire, par exemple, au niveau des paires activateur-inhibiteur : l'impact phénotypique d'une duplication est moins important si elle affecte simultanément les deux gènes. La position relative des gènes sur le chromosome pourrait donc avoir son propre effet sur la variabilité mutationnelle du phénotype, et, par conséquent, subir une pression de sélection indirecte. Si de telles contraintes existent, on peut penser qu'elles sont d'autant plus fortes que les interactions fonctionnelles sont nombreuses et intenses entre les protéines. Nous avons donc cherché à caractériser la dynamique de la position relative des gènes en fonction de l'allure du protéome, en contrôlant l'allure du protéome par l'intermédiaire de $w_{\max}$, et en évaluant la dynamique de l'organisation des gènes à travers celle de leur ordre sur le chromosome. La question que nous abordons dans cette section est donc la suivante : une fois le nombre de gènes stabilisé, leur ordre est-il plus contraint lorsque $w_{\max}$ est élevé ?

La connaissance exacte des événements mutationnels fixés s'avère particulièrement utile pour répondre à cette question. Pour chaque simulation, nous avons analysé les mutations apparues sur la lignée ancestrale du meilleur individu final, en nous focalisant sur celles qui (i) se sont produites après la stabilisation du nombre de gènes et (ii) pouvaient changer l'ordre des gènes sans affecter le phénotype. Nous avons donc considéré les inversions et les translocations neutres apparues sur la lignée ancestrale entre les générations 30 000 et 40 000. Une inversion ou une translocation est considérée neutre si elle n'a pas d'effet sur l'adaptation g, ce qui est le cas ici si ses points de rupture sont situés dans des régions non fonctionnelles[1]. Nous considérons par ailleurs qu'une inversion ou une translocation est conservative si elle préserve les positions relatives des "régions fonctionnelles" (une région fonctionnelle étant ici, comme au chapitre précédent, une région transcrite contenant au moins une séquence codante).

Comme le montre la figure IV.8, le nombre de translocations ou d'inversions neutres fixées tend à décroître lorsque $w_{\max}$ augmente, ce qui reflète vraisemblablement la baisse du taux de mutation génomique due à une compacité accrue du génome. Pour caractériser la stabilité de l'ordre des gènes au-delà de cet effet, nous avons représenté sur la figure IV.9 la

[1]Si cette hypothèse est valide dans le modèle, elle est plus discutable pour un génome réel, dans la mesure où une inversion peut avoir des conséquences phénotypiques même si elle ne rompt aucun gène. En effet, dans un génome réel, le fait qu'un gène change de brin peut avoir des conséquences sur sa transcription (voir chapitre I).

proportion (notée P) de réarrangements conservant l'ordre des gènes, parmi les réarrangements neutres fixés. On observe alors que plus $w_{\max}$ est élevé, plus la proportion P_{inv} d'inversions conservatives est forte au sein des inversions neutres fixées. Ainsi, bien que l'organisation des gènes ne soit jamais complètement figée, elle est globalement plus stable lorsque le réseau d'interactions fonctionnelles entre les protéines est plus connecté : il y a moins d'inversions fixées, et celles-ci tendent à préserver plus souvent l'ordre des gènes. Cela pourrait indiquer qu'une organisation spécifique des gènes soit indirectement sélectionnée lorsque ce réseau contient des noeuds très connectés (protéines très pléiotropes).

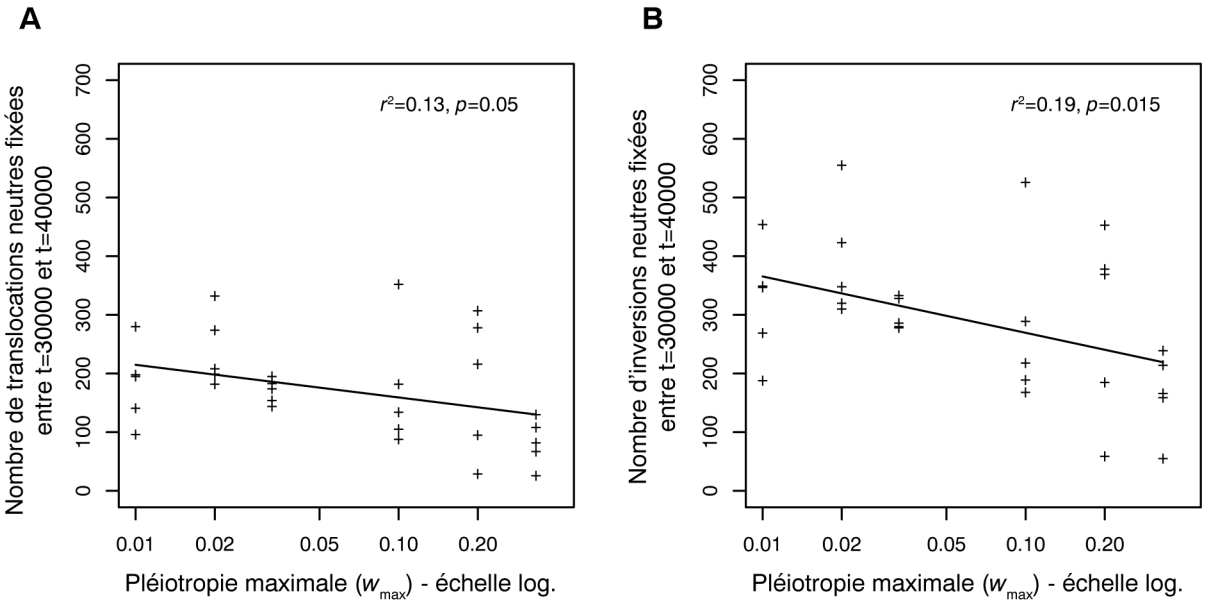


Fig. IV.8: Influence de $w_{\max}$ sur le nombre de translocations (A) et d'inversions (B) neutres qui se sont produites sur la lignée ancestrale du meilleur individu final, entre $t = 30\,000$ et $t = 40\,000$.

Deux raisons peuvent être sous-jacentes à une stabilité accrue de l'ordre des gènes : (i) les événements non conservatifs sont indirectement contre-sélectionnés, ou (ii) la structure du génome – et en particulier la variance des distances intergéniques – est telle que les événements ont spontanément plus de chances d'être conservatifs. Afin de distinguer ces deux causes possibles, nous pouvons décomposer la proportion P de réarrangements conservatifs parmi les réarrangements neutres fixés en deux rapports R_1 et R_2, de la façon suivante (voir la figure IV.10 pour une interprétation graphique) :

$$\begin{aligned} P &= \frac{p(f \text{ ET } n \text{ ET } c)}{p(f \text{ ET } n)} \\ &= \frac{p(f|(n \text{ ET } c))}{p(f|n)} \cdot \frac{p(n \text{ ET } c)}{p(n)} \\ &= R_1 \cdot R_2 \end{aligned} \tag{IV.1}$$

où $p(f)$, $p(n)$, $p(c)$ sont les probabilités qu'un réarrangement aléatoire soit respectivement

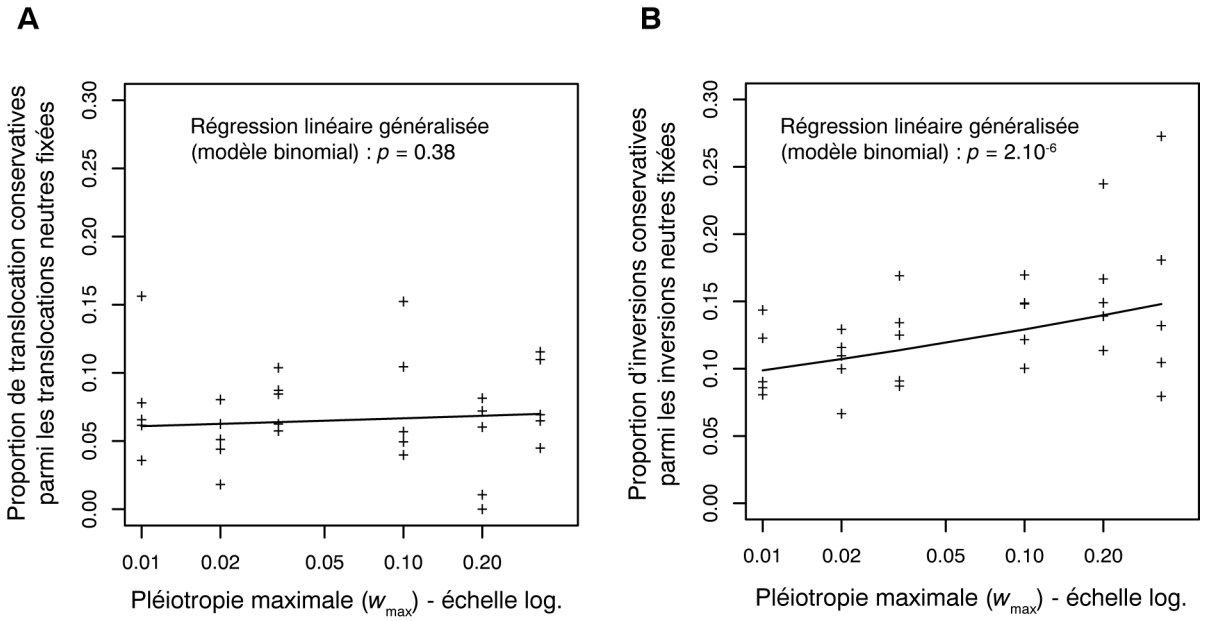


Fig. IV.9: Influence de $w_{\max}$ sur la proportion de réarrangements sans incidence sur l'ordre des gènes. **A** : Proportion P_{transloc} de translocations conservatives parmi les translocations neutres fixées. **B** : Proportion P_{inv} d'inversions conservatives parmi les inversions neutres fixées.

fixé, neutre, conservatif. Ici, nous pouvons estimer $p(n)$ et $p(n \text{ ET } c)$ par :

$$\begin{cases} p(n)_{\text{inv}} &= \left(1-\dfrac{l}{L}\right)^2 \\ p(n)_{\text{transloc}} &= \left(1-\dfrac{l}{L}\right)^3 \\ p(n \text{ ET } c)_{\text{inv}} &= \dfrac{1}{L^2}\displaystyle\sum_{i=1}^{N_G}\lambda_i\left(\lambda_{i-1}+\lambda_i+\lambda_{i+1}\right) \\ p(n \text{ ET } c)_{\text{transloc}} &= \dfrac{1}{2L^2}\left(1-\dfrac{l}{L}\right)\displaystyle\sum_{i=1}^{N_G}\lambda_i\left(\lambda_i+1\right)+\dfrac{1}{2L^2}\displaystyle\sum_{i=1}^{N_G}\lambda_i\left(\lambda_i-1\right) \end{cases} \tag{IV.2}$$

où N_G est le nombre de régions fonctionnelles, l leur longueur totale, et λ_i la distance en bp entre la fin de la région i et le début de la région $i+1$ (on a donc $\sum_{i=1}^{N_G}\lambda_i = L-l$).

Le rapport $R_1 = \frac{p(f|(n \text{ ET } c))}{p(f|n)}$, quantifie l'avantage sélectif de la "conservativité" parmi les réarrangements neutres. Notons que cet avantage ne peut être immédiat dans la mesure où l'événement est neutre. Il peut cependant apparaître à plus long terme si, par exemple, perturber l'ordre des gènes augmente l'impact des mutations lors des reproductions suivantes. Le second rapport, $R_2 = \frac{p(n \text{ ET } c)}{p(n)} = p(c|n)$, mesure la probabilité qu'un réarrangement aléatoire neutre soit conservatif. Comme le montre l'équation IV.2, ce second rapport dépend de la structure du génome, et en particulier des distances intergéniques λ_i.

Pour comprendre l'origine de la stabilité accrue du génome, nous avons estimé les deux rapports pour chaque simulation. Pour cela, nous avons d'abord calculé analytiquement

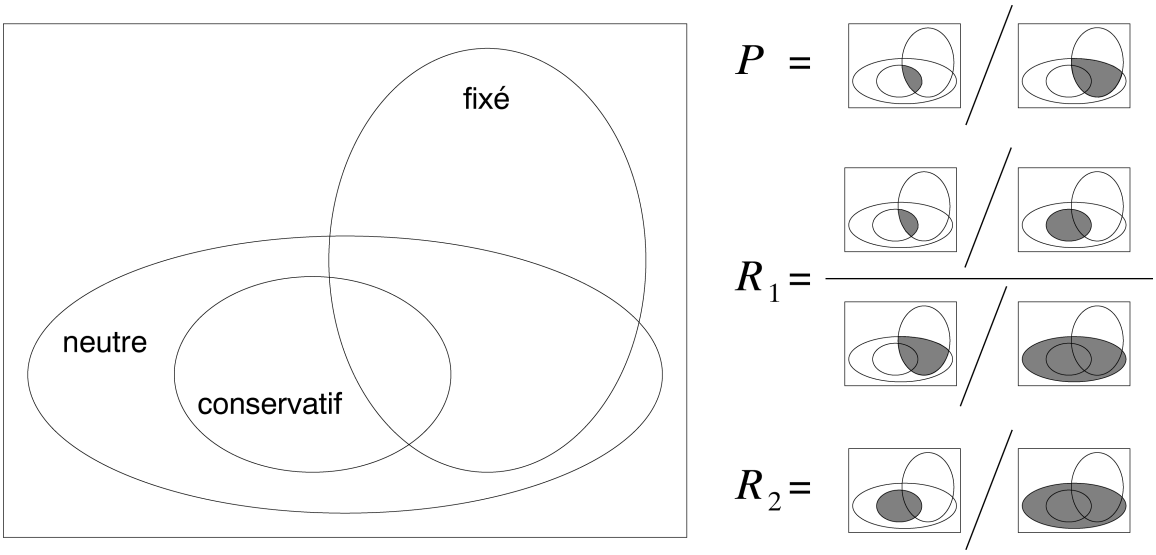


Fig. IV.10: Décomposition de P en deux rapports R_1 et R_2. Le grand rectangle désigne l'ensemble des translocations (ou des inversions) qui se sont spontanément produites dans la population entre $t = 30\ 000$ et $t = 40\ 000$. Une certaine partie d'entre elles étaient neutres, et parmi celles-ci, certaines étaient conservatives. On peut également distinguer celles qui ont été fixées, c'est-à-dire celles qui se trouvent sur la lignée ancestrale des individus finaux.

$p(n \text{ ET } c)$ et $p(n)$ pour les ancêtres successifs du meilleur individu final, à partir des caractéristiques strcuturelles de leurs génomes (équation IV.2). Nous en avons ensuite pris la moyenne sur les 10 000 générations considérées. Cela permet d'obtenir le rapport R_2. Il suffit enfin de diviser P par R_2 pour obtenir R_1.

Comme le montre la figure IV.11, le rapport R_1 est en moyenne inférieur à 1, ce qui signifie que préserver l'ordre des gènes n'augmente pas la probabilité de fixation. De plus, ce rapport n'augmente pas avec $w_{\max}$. Pour les translocations, il *diminue* même lorsque $w_{\max}$ augmente. La stabilité accrue de l'ordre des gènes ne peut donc s'expliquer par une sélection préférentielle des réarrangements conservatifs. Cette stabilité doit donc être due au fait que les réarrangements conservatifs sont spontanément plus fréquents. C'est en effet ce que montre la figure IV.12 : que ce soit pour les translocations ou pour les inversions, le rapport R_2 – c'est-à-dire la probabilité qu'un réarrangement neutre préserve l'ordre des gènes – augmente avec $w_{\max}$. La stabilité accrue de l'ordre des gènes en présence de protéines très pléiotropes résulte donc d'une structuration du génome qui réduit la fréquence des réarrangements perturbateurs.

C'est vraisemblablement au niveau de la variance des distances intergéniques que cet effet se joue : à quantité de non codant égale, une inversion a moins de chances de perturber l'ordre des gènes s'ils forment un seul cluster que s'ils sont distribués régulièrement sur le chromosome (voir l'équation IV.2). Il n'existe pas, dans notre modèle, de pression directe ni de mécanisme mutationnel (comme la duplication en tandem) susceptible de favoriser l'apparition de tels clusters. Cette structuration résulte donc, selon toute vraisemblance, d'une pression de sélection indirecte : les génomes structurés de telle sorte que les inver-

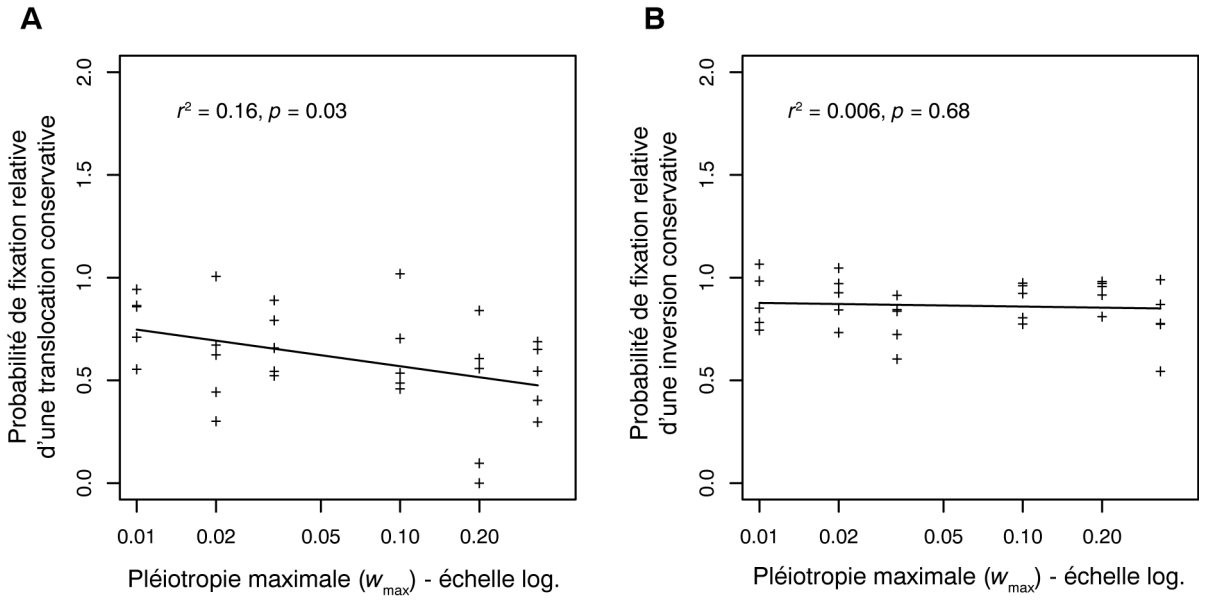


Fig. IV.11: Probabilité de fixation d'une translocation (A) ou d'une inversion (B) neutre et conservative, normalisée par la probabilité de fixation d'une translocation ou d'une inversion neutre (rapport R_1). Pour les translocations, ce rapport diminue lorsque w_{max} augmente. Pour les inversions, w_{max} n'a pas d'effet.

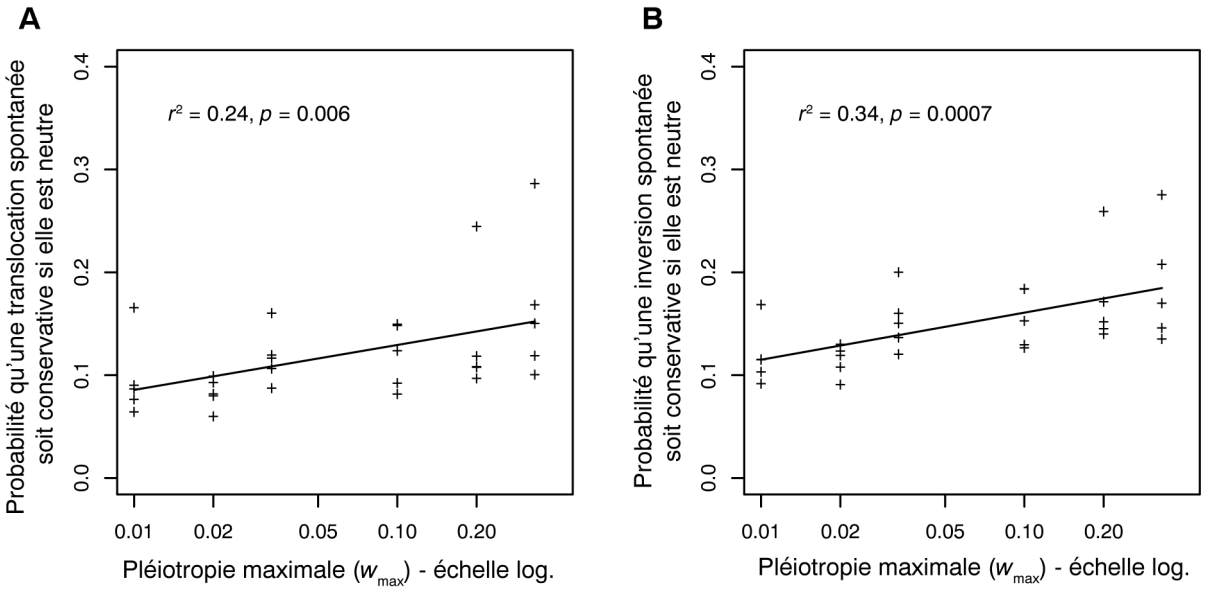


Fig. IV.12: Probabilité qu'une translocation (A) ou une inversion (B) spontanée soit conservative sachant qu'elle est neutre (rapport R_2). Dans les deux cas, l'effet de w_{max} est significatif.

sions et les translocations ont plus de chances de préserver l'ordre des gènes sont ceux qui, de fait, se maintiennent à long terme lorsque les protéines sont très pléiotropes (c'est-à-dire, ici, lorsque les mutations géniques ont potentiellement un fort impact). L'avantage sélectif indirect apporté par cette structuration n'est cependant pas évident. S'agit-il de

la préservation d'un certain ordre des gènes, qui optimiserait par exemple l'effet des duplications et des grandes délétions (regroupement des gènes en interaction fonctionnelle, éloignement des "hubs", ...)[1] ? Ou bien s'agit-il de tout autre chose, auquel cas la préservation accrue de l'ordre des gènes ne serait qu'un effet secondaire ? La première hypothèse serait particulièrement élégante : la sélection indirecte d'un niveau donné de variabilité mutationnelle ne favoriserait pas seulement une certaine taille de génome, mais aussi un certain ordre des gènes, ce qui favoriserait alors un certain pattern mutationnel au niveau des réarrangements, et donc une certaine répartition des gènes sur le chromosome. Mais il faut reconnaître que si cette hypothèse est correcte, le fait qu'en moyenne, les réarrangements conservatifs n'aient pas une plus grande probabilité de fixation est pour le moins intrigant. Une explication possible résiderait dans le conflit potentiel entre la stabilisation d'un ordre des gènes particulier et sa mise en place. Si l'organisation des gènes doit être conservée la plupart du temps, alors les génomes où les réarrangements perturbateurs sont plus rares sont indirectement sélectionnés, comme nous l'avons observé. Cependant, certains réarrangements perturbateurs peuvent être nécessaires de temps à autre pour répondre à des changements dans le réseau d'interactions fonctionnelles.

5 Discussion

Le tableau général qui se dégage de cette série d'expériences est celui d'une interaction inattendue entre les propriétés fonctionnelles des protéines (pléiotropie, poids dans le phénotype) d'une part, et la structure du génome (nombre de gènes, quantité de non codant, dynamique de l'ordre des gènes) d'autre part. Cette interaction mettant en jeu des niveaux différents, elle est moins forte que celle mise en évidence au chapitre précédent entre le taux de mutation et la structure du génome, mais elle est statistiquement significative. Les deux interactions émergent d'un même phénomène, celui de la sélection indirecte d'une certaine variabilité mutationnelle du phénotype : si cette variabilité est perturbée par une augmentation de la fréquence (chapitre précédent) ou de l'effet des mutations (chapitre courant), des modifications compensatrices de la structure du génome sont indirectement sélectionnées. La pertinence, du point de vue des génomes réels, du couplage mis en évidence ici entre l'effet des mutations géniques et la structure du génome dépend donc en partie des mêmes éléments que ceux soulignés au chapitre précédent.

Ainsi, comme nous l'avons déjà mentionné, la façon dont la sélection indirecte agit sur la structure du génome dépend de la contribution de celle-ci à la variabilité mutationnelle du phénotype. Nous avons pris en compte ici le rôle du non codant dans les réarrangements, mais il faut tout d'abord souligner que dans les génomes réels, celui-ci est souvent composé d'éléments transposables capables de s'insérer dans les gènes, formant ainsi une source supplémentaire de variabilité mutationnelle pour le phénotype. Ensuite, le rôle du non codant dans les réarrangements a été ici modélisé de façon simple, mais il peut être plus

[1] L'examen visuel des génomes finaux, en relation avec les réseaux d'interactions fonctionnelles correspondant, ne fait pas ressortir de règles évidentes pour l'ordre des gènes. Les possibilités énumérées restent, à ce stade de l'étude, des hypothèses qu'il conviendrait de tester spécifiquement.

complexe en réalité. Pour qu'une augmentation de l'effet des mutations induise, comme ici, une compaction du génome, il est nécessaire que des génomes plus compacts subissent des duplications et des délétions moins nombreuses et moins longues, si bien que moins de gènes soient mutés à chaque réplication. Nous avons établi au chapitre précédent que ces conditions pourraient être remplies dans les génomes réels, même si les corrélations en question sont certainement moins fortes que dans notre modèle : la taille moyenne des réarrangements, par exemple, n'augmente peut-être pas *linéairement* avec la taille du génome dans la mesure où elle dépend de la répartition des éléments répétés et de la portée des différents mécanismes de recombinaison. Il reste cependant probable que la taille moyenne des réarrangements dépende, d'une façon ou d'une autre, de la taille du génome. Des pressions de sélection indirecte comparables à celles que nous avons mises en évidence seraient donc en mesure de s'exercer sur la compacité du génome.

Par ailleurs, dans notre modèle, le rétablissement du niveau de variabilité après une perturbation ne peut se faire qu'à travers des ajustements de la structure du génome, car le taux de mutation est fixe dans le temps, ainsi que, dans une certaine mesure, l'effet des mutations géniques[1]. Or les organismes vivants disposent d'un plus large panel de composants "évoluables" : dans une situation comme celle que nous modélisons – l'augmentation de l'effet des mutations géniques –, l'évolution naturelle pourrait sélectionner de nouveaux mécanismes de canalisation (protéines chaperonnes, voies métaboliques alternatives, etc.) plutôt que des modifications de la structure du génome.

Dans ce contexte, et comme pour le chapitre précédent, nous ne pouvons pas prédire que la comparaison de différentes espèces réelles va révéler directement un couplage entre l'effet des mutations géniques et la structure du génome. S'il est aujourd'hui relativement aisé d'accéder aux données génomiques (au moins pour les espèces modèles), une comparaison pertinente de l'effet des mutations est beaucoup plus difficile à établir. Des expériences d'accumulation de mutations (Mutation-Accumulation, MA) réalisées chez un certain nombre d'espèces modèles semblent suggérer que les mutations géniques sont plus délétères chez les espèces qui présentent le plus grand nombre de gènes (Martin et Lenormand, 2006), ce qui contredirait les résultats de nos simulations. Cependant, cette comparaison directe d'espèces distantes peut être trompeuse, car la fitness est estimée de manière différente selon les espèces (Bataillon, 2000; Martin et Lenormand, 2006). On l'estime par exemple par le taux de croissance intrinsèque chez *Escherichia coli*, le LRS (Lifetime Reproductive Success) chez *Arabidopsis thaliana*, et par la viabilité des œufs chez *Drosophila melanogaster*. Un autre problème fondamental des approches d'accumulation de mutations est leur incapacité à détecter les petites variations de fitness (Bataillon, 2000) ainsi que, chez les micro-organismes, les variations majeures qui sont contre-sélectionnées (Kibota et Lynch, 1996; Martin et Lenormand, 2006). En dehors de ces difficultés, la comparaison d'espèces phylogénétiquement distantes met en jeu quantité de facteurs susceptibles d'avoir leur propre effet sur la taille du génome : des différences au niveau des taux de mutation, de l'activité des éléments transposables ou encore de la taille

[1] Celui-ci peut être théoriquement rétabli à un niveau plus bas si les protéines avec un grand w (proche de $w_{\max}$) sont contre-sélectionnées : cela permettrait de revenir à un protéome plus modulaire. Ce n'est cependant pas ce qui est observé en pratique dans les simulations. Les protéines très pléiotropes sont conservées, et c'est la structure du génome qui tamponne l'augmentation de la variabilité mutationnelle.

de la population peuvent masquer l'influence de l'effet moyen des mutations géniques.

Nos résultats relatifs à la dynamique de l'ordre des gènes sont plus difficiles à mettre en perspective, car il n'existe pas, à notre connaissance, d'étude mettant en relation l'effet des mutations géniques sur la fitness avec la dynamique de l'ordre des gènes. L'organisation des gènes est pourtant particulièrement étudiée chez les génomes procaryotes, en raison notamment des corrélations mises en évidence entre la proximité des gènes sur le chromosome et leurs liens fonctionnels (Dandekar *et al.*, 1998; Lathe *et al.*, 2000; Rogozin *et al.*, 2002a). La découverte récente de clusters de gènes co-exprimés dans des génomes eucaryotes (Hurst *et al.*, 2004) suggère que l'ordre des gènes n'y est pas non plus aléatoire. De nombreuses hypothèses ont été proposées pour rendre compte d'une organisation modulaire des gènes. Les hypothèses neutralistes présentent cette modularité comme native, résultant d'un processus d'acquisition de gènes par duplication en tandem. Elles sont cependant mises en défaut par le fait que la majorité des opérons bactériens sont constitués de gènes sans similarité de séquence, et par le fait qu'elles n'expliquent pas le *maintien* des groupes formés (Lawrence et Roth, 1996). Alternativement, on peut imaginer un avantage sélectif direct associé à la formation d'opérons, comme une co-régulation plus facile (Price *et al.*, 2005). Lawrence et Roth (1996) ont pour leur part proposé que le regroupement soit avantageux pour les gènes eux-mêmes plutôt que pour l'organisme (hypothèse dite de l'opéron égoïste), car cela leur permet de se propager plus efficacement par transfert horizontal : si tous les gènes nécessaires à une fonction donnée sont acquis en bloc, alors ils ont plus de chances de se maintenir dans la nouvelle cellule hôte que s'ils sont acquis individuellement. Cependant, d'après Pal et Hurst (2004) et Price *et al.* (2005), cette hypothèse ne permet pas de rendre compte des patterns concrètement observés dans les génomes bactériens. On peut enfin citer, pour les populations subissant de la recombinaison allélique, l'hypothèse de co-adaptation proposée par Fisher dès 1930. Lorsque deux gènes sont en interaction et s'il existe des combinaisons alléliques plus efficaces que d'autres, alors il peut être indirectement avantageux de rapprocher les loci en question (Sinervo et Svensson, 2002) : cela réduit la probabilité qu'un échange allélique n'affecte qu'un seul des deux gènes et conduise ainsi à deux produits géniques interagissant moins bien. L'effet de la recombinaison allélique sur le regroupement des gènes a été confirmé à l'aide d'un algorithme évolutionniste par Pepper (2000). Pourtant, aucune de ces hypothèses ne peut, en l'état, expliquer notre observation d'une stabilité accrue de l'ordre des gènes lorsque les mutations géniques ont un plus fort impact : il n'y a, dans nos simulations, ni duplications en tandem, ni régulation en réponse aux changements environnementaux, ni transfert horizontal, ni recombinaison allélique. Nos expériences suggèrent donc l'existence d'un nouveau type de pression sur l'ordre des gènes, que nous pensons lié à une optimisation de l'effet des grandes délétions et des duplications. Il reste alors à déterminer les contraintes précises que cela implique sur l'ordre des gènes : dans nos simulations, il semble que les gènes en interaction fonctionnelle ne soient pas significativement plus proches les uns des autres que des gènes pris au hasard. L'effet pourrait se manifester davantage au niveau de la position relative des gènes les plus pléiotropes. Si de telles contraintes peuvent être identifiées dans le modèle, il sera possible d'en rechercher la trace dans les génomes réels.

6 Conclusion

Les expériences *in silico* présentées dans ce chapitre montrent que, tout étant égal par ailleurs, la structure du génome peut spontanément s'ajuster en fonction de l'impact des mutations géniques. Le nombre de gènes, la quantité de non codant et la dynamique de l'ordre des gènes dépendent ainsi des propriétés de la correspondance gènes-traits (genotype-phenotype map), c'est-à-dire de la façon dont les protéines interagissent et déterminent le phénotype. Comme dans le chapitre précédent, ce couplage émerge de la sélection indirecte d'un niveau donné de variabilité mutationnelle : l'impact potentiellement fort des mutations géniques est compensé par une compacité accrue du génome, permettant de diminuer le nombre moyen de gènes mutés par réplication. Au-delà de ce mécanisme particulier, cette étude montre que des relations inattendues et complexes peuvent donc émerger entre des niveaux différents (le génome et le protéome), à partir des mécanismes évolutifs les plus fondamentaux (mutation au niveau du génotype, sélection au niveau du phénotype). Si de telles relations ont pu être isolées grâce à la simulation, elles sont difficiles à révéler par comparaison directe d'espèces phylogénétiquement distantes, car elles peuvent être masquées par d'autres facteurs. Des résultats tels que ceux présentés ici posent donc un double défi. Un défi conceptuel tout d'abord, avec la nécessité de développer une compréhension systémique des espèces en évolution. Il s'agit de prendre en compte non seulement les interactions fonctionnelles entre les produits géniques, comme nous y invite déjà la Biologie des Systèmes (Kitano, 2002), mais aussi les interactions globales entre génome et protéome à l'échelle évolutive. Un défi expérimental ensuite, avec la nécessaire mise au point de protocoles spécifiques pour isoler et analyser de telles interactions.

Conclusions et perspectives

Cette thèse est née du constat que l'évolution des génomes et l'évolution de la variabilité constituent ordinairement des problématiques disjointes, alors que la structure du génome apparaît comme une composante importante de la variabilité mutationnelle du phénotype. Le pari fut lancé de concevoir un modèle qui apporterait un éclairage nouveau sur l'évolution des génomes, parce qu'il s'appuierait sur les connaissances et le savoir-faire accumulés par les biologistes et les informaticiens dans le domaine de l'évolution de la variabilité. C'est dans ce contexte que furent développés le modèle *aevol* et la plate-forme de simulation associée, dans une démarche empruntant à l'évolution moléculaire ses notions clés, aux algorithmes évolutionnaires leurs techniques computationnelles et à la vie artificielle son point de vue épistémologique.

Après des échanges interdisciplinaires parfois difficiles mais toujours fructueux à terme, quelques dizaines de milliers de lignes de code, des semaines de simulations, quelques giga-octets de données analysées et autant d'hypothèses formulées, malmenées et parfois abandonnées, trois leçons peuvent être tirées de cette étude. La première est que, *in silico*, la sélection indirecte de la variabilité mutationnelle agit effectivement sur la structure du génome, et en particulier sur la quantité de non codant, le nombre de gènes et l'ordre des gènes. Cela suggère que les biais mutationnels et les coûts sélectifs directs ne sont pas les seules pressions à l'œuvre dans les génomes réels. La seconde est que les réarrangements chromosomiques jouent un rôle majeur dans les phénomènes mis en évidence, ce qui souligne la limite des raisonnements uniquement fondés sur les mutations locales. Enfin, la troisième est qu'à l'échelle du temps évolutif, des interactions imprévues peuvent émerger entre des niveaux différents (ici, entre l'organisation fonctionnelle du protéome et la structure du génome), sous l'effet de la sélection indirecte de la variabilité mutationnelle. Ce dernier point appelle à développer une compréhension systémique de l'évolution, prenant en compte les interactions complexes qu'elle peut générer.

Le sujet est cependant loin d'être clos. Le modèle développé permet déjà, en l'état, d'aborder de nouvelles questions. L'influence de la variabilité de l'environnement figure sans aucun doute parmi les plus pertinentes. Il est clair, cependant, que ce modèle comporte un certain nombre de raccourcis qui peuvent être levés pour plus de réalisme biologique. Étant donné le rôle majeur des réarrangements dans les phénomènes observés, une modélisation plus fine des mécanismes de réarrangements constitue un point prioritaire. Il conviendrait de prendre explicitement en compte les similarités de séquences dans le processus, malgré le temps de calcul important que cela impose. Il serait alors pertinent de

tester aussi l'effet de la recombinaison allélique, dans la mesure où les déséquilibres de liaison jouent théoriquement un rôle important dans les phénomènes de sélection indirecte. On peut également envisager de laisser le taux de mutation par base évoluer en même temps que la structure du génome, ce qui permettrait de tester si ces deux leviers peuvent être utilisés simultanément.

Mais il nous semble qu'à ce stade de l'étude, l'enjeu principal ne réside pas seulement dans le modèle lui-même. Il se trouve au moins autant dans la mise en évidence des pressions de sélection indirecte sur la structure des génomes réels. Les relations qui émergent dans le modèle – comme la dépendance de la taille du génome au taux de mutation – ne sont pas nécessairement retrouvées par les approches comparatives, car celles-ci ne peuvent isoler l'effet d'un seul facteur. Les approches d'évolution expérimentale, bien qu'extrêmement coûteuses, paraissent plus prometteuses dans le sens où elles permettent d'analyser directement et relativement finement la dynamique évolutive du génome d'une espèce donnée. L'expérience que nous proposons consisterait à suivre l'évolution de la taille du génome après avoir provoqué une augmentation constitutive du taux de mutation. Une telle expérience permettrait de bâtir entre l'évolution des génomes et celle de la variabilité le pont dont nous avons esquissé les plans.

Bibliographie

Achaz, G. (2002). *Etude de la dynamique des génomes : les répétitions intrachromosomiques*. Thèse de doctorat, Université Pierre et Marie Curie (Paris VI). 208 pages.

Achaz, G., Coissac, E., Netter, P. et Rocha, E. P. C. (2003). Associations between inverted repeats and the structural evolution of bacterial genomes. *Genetics*, 164:1279–1289.

Achaz, G., Coissac, E., Viari, A. et Netter, P. (2000). Analysis of intrachromosomal duplications in yeast *Saccharomyces cerevisiae* : A possible model for their origin. *Mol. Biol. Evol.*, 17:1268–1275.

Achaz, G., Netter, P. et Coissac, E. (2001). Study of intrachromosomal duplications among the eukaryote genomes. *Mol. Biol. Evol.*, 18:2280–2288.

Achaz, G., Rocha, E. P. C., Netter, P. et Coissac, E. (2002). Origin and fate of repeats in bacteria. *Nucleic Acids Res.*, 30:2987–2994.

Adami, C. (2006). Digital genetics : unravelling the genetic basis of evolution. *Nat. Rev. Genet.*, 7:109–118.

Ancel, L. W. et Fontana, W. (2000). Plasticity, evolvability, and modularity in RNA. *J. Exp. Zool.*, 288:242–283.

Ancel Meyers, L., Ancel, F. D. et Lachmann, M. (2005). Evolution of genetic potential. *PLoS Comput. Biol.*, 1:e32.

Andersson, J. O. et Andersson, S. G. E. (1999). Insights into the evolutionary process of genome degradation. *Curr. Opin. Genet. Dev.*, 9:664–671.

Andolfatto, P. (2005). Adaptive evolution of non-coding DNA in Drosophila. *Nature*, 437:1149–1152.

Aronson, H. R. W. et Hendrickson, W. (1994). Quantification of tertiary structural conservation despite primary sequence drift in the globin fold. *Protein Sci.*, 3:1706–1711.

Avery, O. T., MacLeod, C. M. et MacCarthy, M. (1944). Studies of the chemical nature of the substance inducing transformation of pneumococcal types - Induction of

transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.*, 79:137–158.

Bagley, J. D. (1967). *The behavior of adaptive systems which employ genetic and correlation algorithms*. Thèse de doctorat, University of Michigan, Ann Arbor, USA. University Microfilms No. 68-7556. 185 pages.

Barabasi, A.-L. et Oltvai, Z. N. (2004). Network biology : Understanding the cell's functional organization. *Nat. Rev. Genet.*, 5:101–113.

Barrell, B. G., Air, G. M. et Hutchinson, C. A. (1976). Overlapping genes in bacteriophage phiX174. *Nature*, 264:34–41.

Basten, C. J. et Moody, M. E. (1991). A branching-process model for the evolution of transposable elements incorporating selection. *J. Math. Biol.*, 29:743–761.

Bataillon, T. (2000). Estimation of spontaneous genome-wide mutation rate parameters : whither beneficial mutations ? *Heredity*, 84:497–501.

Bedau, M. A., MCCaskill, J. S., Packard, N. H., Rasmussen, S., Adami, C., Green, D. G., Ikegami, T., Kaneko, K. et Ray, T. S. (2000). Open problems in artificial life. *Artificial Life*, 6:363–376.

Bedau, M. A. et Packard, N. H. (2003). Evolution of evolvability via adaptation of mutation rates. *Biosystems*, 69:143–162.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S. et Haussler, D. (2004). Ultraconserved elements in the human genome. *Science*, 304:1321–1325.

Bengtsson, B. O. (2004). Modelling the evolution of genomes with integrated external and internal functions. *J. Theor. Biol.*, 231:271–278.

Bergthorsson, U. et Ochman, H. (1998). Distribution of chromosome length variation in natural isolates of *Escherichia coli*. *Mol. Biol. Evol.*, 15:6–16.

Betran, E. et Long, M. (2002). Expansion of genome coding regions by acquisition of new genes. *Genetica*, 115:65–80.

Blickle, T. et Thiele, L. (1994). Genetic programming and redundancy. *In* Hopf, J., éditeur : *Genetic Algorithms within the Framework of Evolutionary Computation (Workshop at KI-94, Saarbrücken)*, pages 33–38. Max-Planck-Institut für Informatik (MPI-I-94-241), Saarbrücken. 161 pages.

Blickle, T. et Thiele, L. (1996). A comparison of selection schemes used in evolutionary algorithms. *Evol. Comput.*, 4:361–394.

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Muller, S., Eils, R., Cremer, C., Speicher, M. R. et Cremer, T. (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.*, 3:e157.

Brosius, J. (2003). How significant is 98.5% "junk" in mammalian genomes ? *Bioinformatics*, 19(Suppl. 2):ii35.

Brosius, J. et Gould, S. J. (1992). On "genomenclature" : a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". *Proc. Natl. Acad. Sci. USA*, 89:10706–10710.

Burke, D. S., De Jong, K. A., Grefenstette, J. J., Ramsey, C. L. et Wu, A. S. (1998). Putting more genetics into genetic algorithms. *Evol. Comput.*, 6:387–410.

Byrne, A., de Laski, A., Courage, K. et Wallace, C. (1982). Handbook of computer models for traffic operations analysis. Rapport technique FHWA-TS-82-213, Federal Highway Administration, Washington DC.

Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D. et May, G. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol.*, 4:10.

Cavalier-Smith, T., éditeur (1985). *The Evolution of Genome Size*. John Wiley and Sons Ltd, Chichester. 534 pages.

Chao, L. et Cox, E. C. (1983). Competition between high and low mutating strains of *Escherichia coli*. *Evolution*, 37:125–134.

Charles, H., Mouchiroud, D., Lobry, J., Goncalves, I. et Rahbe, Y. (1999). Gene size reduction in the bacterial aphid endosymbiont, buchnera. *Mol. Biol. Evol.*, 16:1820–1822.

Charlesworth, B. et Charlesworth, D. (1983). The population dynamics of transposable elements. *Genet. Res.*, 42:1–27.

Chicurel, M. (2001). Genetics : Why evolution might not favor increased genetic variability. *Science*, 292:1826.

Clark, M. A., Baumann, L., Thao, L., Moran, N. A. et Baumann, P. (2001). Degenerative minimalism in the genome of a psyllid endosymbiont. *J. Bacteriol.*, 183:1853–1861.

Claverie, J. M., Ogata, H., Audic, S., Abergel, C., Suhre, K. et Fournier, P. E. (2006). Mimivirus and the emerging concept of "giant" virus. *Virus Res.*, 117:133–144.

Coghlan, A., Eichler, E. E., Oliver, S. G., Paterson, A. H. et Stein, L. (2005). Chromosome evolution in eukaryotes : a multi-kingdom perspective. *Trends Genet.*, 21:673–682.

Comeron, J. P. (2001). What controls the length of non-coding DNA ? *Curr. Opin. Genet. Dev.*, 11:652–659.

Conrad, M. et Ebeling, W. (1992). M. V. Volkenstein, evolutionary thinking and the structure of fitness landscapes. *Biosystems*, 27:125–128.

Corbara, B., Drogoul, A., Fresneau, D. et Lalande, S. (1993). Simulating the sociogenesis process in ant colonies with MANTA. *In Towards a Practice of Autonomous Systems : Proceedings of the First European Conference on Artificial Life*, pages 224–235, Paris. MIT Press, Cambridge. 533 pages.

Correns, C. (1900). Gregor Mendels Regel über das Verhalten der Nachkommenschaft der Bastarde. *Berichte der Deutschen Botanischen Gesellschaft*, 18:158–168.

Cossins, A. (1998). Cryptic clues revealed. *Nature*, 396:309–310.

Cox, E. C. (1976). Bacterial mutator gene and the control of spontaneous mutation. *Annu. Rev. Genet.*, 10:135–156.

Dandekar, T., Snel, B., Huynen, M. et Bork, P. (1998). Conservation of gene order : a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, 23:324–328.

Darwin, C. (1859). *The Origin of Species.* John Murray, Londres. 496 pages (édition Signet Classics de 2003).

Dawson, K. J. (1998). Evolutionarily stable mutation rates. *J. Theor. Biol.*, 194:143–157.

Dawson, K. J. (1999). The dynamics of infinitesimally rare alleles, applied to the evolution of mutation rates and the expression of deleterious mutations. *Theor. Pop. Biol.*, 55:1–22.

De Bolle, X., Bayliss, C. D., Field, D., Van de Ven, T., Saunders, N. J., Hood, D. W. et Moxon, E. R. (2000). The length of a tetranucleotide repeat tract in *Haemophilus influenzae* determines the phase variation rate of a gene with homology to type III DNA methyltransferases. *Mol. Microbiol.*, 35:211–222.

De Visser, J. A., Hermisson, J., Wagner, G. P., Ancel Meyers, L., Bagheri-Chaichian, H., Blanchard, J. L., Chao, L., Cheverud, J. M., Elena, S. F., Fontana, W., Gibson, G., Hansen, T. F., Krakauer, D., Lewontin, R. C., Ofria, C., Rice, S. H., Von Dassow, G., Wagner, A. et Whitlock, M. C. (2003). Perspective : Evolution and detection of genetic robustness. *Evolution*, 57:1959–1972.

De Visser, J. A., Zeyl, C. W., Gerrish, P. J., Blanchard, J. L. et Lenski, R. E. (1999). Diminishing returns from mutation supply rate in asexual populations. *Science*, 283:404–406.

De Vries, H. (1900). Das Spaltungsgesetz der Bastarde. *Berichte der Deutschen Botanischen Gesellschaft*, 18:83–90.

Deceliere, G., Charles, S. et Biemont, C. (2005). The dynamics of transposable elements in structured populations. *Genetics*, 169:467–474.

Demetrius, L. (1983). Selection and evolution in macromolecular systems. *J. Theor. Biol.*, 103:619–643.

Denamur, E., Bonacorsi, S., Giraud, A., Duriez, P., Hilali, F., Amorin, C., Bingen, E., Andremont, A., Picard, B., Taddéi, F. et Matic, I. (2002). High frequency of mutator strains among human uropathogenic *Escherichia coli* isolates. *J. Bacteriol.*, 184:605–609.

Denver, D. R., Morris, K., Lynch, M. et Thomas, W. K. (2004). High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. *Nature*, 430:679–682.

Dermitzakis, E. T., Reymond, A. et Antonarakis, S. E. (2005). Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nat. Rev. Genet.*, 6:151–157.

Dimpfl, J. et Echols, H. (1989). Duplication mutation as an SOS response in *Escherichia coli* : Enhanced duplication formation by constitutively activated RecA. *Genetics*, 123:255–260.

Doolittle, R. F. (1995). The origins and evolution of eukaryotic proteins. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 349:235–240.

Drake, J. W. (1991). A constant rate of spontaneous mutation in DNA-based microbes. *Proc. Natl. Acad. Sci. USA*, 88:7160–7164.

Drake, J. W. (1993). Rates of spontaneous mutation among RNA viruses. *Proc. Natl. Acad. Sci. USA*, 90:4171–4175.

Drake, J. W., Charlesworth, B., Charlesworth, D. et Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics*, 148:1667–1686.

Dubois, D. et Prade, H. (1980). *Fuzzy Sets and Systems, Theory and Applications.* Academic Press, New York. 393 pages.

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., de Montigny, J., Marck, C., Neuveglise, C., Talla1, E. *et al.* (2004). Genome evolution in yeasts. *Nature*, 430:35–44.

Duret, L., Dorkeld, F. et Gautier, C. (1993). Strong conservation of non-coding sequences during vertebrates evolution : potential involvement in post-transcriptional regulation of gene expression. *Nucleic Acids Res.*, 21(10):2315–2322.

Dykhuizen, D. E., Dean, A. M. et Hartl, D. L. (1987). Metabolic flux and fitness. *Genetics*, 115:25–31.

Earl, D. J. et Deem, M. W. (2004). Evolvability is a selectable trait. *Proc. Natl. Acad. Sci. USA*, 101(32):11531–11536.

Edwards, J. S. et Palsson, B. O. (2000). Metabolic flux balance analysis and the in silico analysis of escherichia coli k-12 gene deletions. *BMC Bioinformatics*, 1:1.

Eiben, A. E., Hinterding, R. et Michalewicz, Z. (1999). Paramete control in evolutionary algorithms. *IEEE Trans. Evol. Comput.*, 3:124–141.

Eichler, E. E. et Sankoff, D. (2003). Structural dynamics of eukaryotic chromosome evolution. *Science*, 301:793–797.

Eigen, M. (1971). Self-organization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58:456–523.

Eigen, M., McCaskill, J. et Schuster, P. (1988). Molecular quasi-species. *J. Phys. Chem.*, 92:6881–6891.

Elena, S. F. et Lenski, R. E. (2001). Epistasis between new mutations and genetic background and a test of genetic canalization. *Evolution*, 55(9):1746–1752.

Ellegren, H. (2004). Microsatellites : simple sequences with complex evolution. *Nat. Rev. Genet.*, 5:435–445.

Epstein, C. J. (1966). Role of the amino-acid code and of selection for conformation in the evolution of proteins. *Nature*, 210:25–28.

Esnault, C., Maestre, J. et Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.*, 24:363–367.

Eyre-Walker, A. et Keightley, P. D. (1999). High genomic deleterious mutation rates in hominids. *Nature*, 397:344–347.

Eyre-Walker, A., Keightley, P. D., Smith, N. G. C. et Gaffney, D. (2002). Quantifying the slightly deleterious mutation model of molecular evolution. *Mol. Biol. Evol.*, 19:2142–2149.

Fang, Z., Doig, C., Kenna, D. T., Smittipat, N., Palittapongarnim, P., Watt, B. et Forbes, K. J. (1999). IS6110-mediated deletions of wild-type chromosomes of *Mycobacterium tuberculosis*. *J. Bacteriol.*, 181:1014–1020.

Fares, M. A., Ruiz-Gonzalez, M. X., Moya, A., Elena, S. F. et Barrio, E. (2002). GroEL buffers against deleterious mutations. *Nature*, 417:398.

Feder, M. E. et Hofmann, G. E. (1999). Heat-shock proteins, molecular chaperones, and the stress response : Evolutionary and ecological physiology. *Annu. Rev. Physiol.*, 61:243–282.

Fell, D. (1997). *Understanding the Control of Metabolism*. Portland Press, Londres. 301 pages.

Fogel, D. B. (1995). Phenotypes, genotypes, and operators in evolutionary computation. *In Proceedings of the 1995 IEEE International Conference on Evolutionary Computation*, pages 193–198, Perth, Australie. IEEE Press, Piscataway. 851 pages.

Fontana, W. et Schuster, P. (1998). Continuity in evolution : On the nature of transitions. *Science*, 280:1451–1455.

Forbes, N. (2004). *Imitation of Life*. MIT Press, Cambridge. 176 pages.

Fox Keller, E. (2000). *The Century of the Gene*. Harvard Univeristy Press, Cambridge. Traduction française : *Le siècle du gène*, Gallimard, 2003, 173 pages.

Frank, A. C., Amiri, H. et Andersson, S. G. E. (2002). Genome deterioration : loss of repeated sequences and accumulation of junk DNA. *Genetica*, 115:1–12.

Frazer, K. A., Sheehan, J. B., Stokowski, R. P., Chen, X., Hosseini, R., Cheng, J. F., Fodor, S. P., Cox, D. R. et Patil, N. (2001). Evolutionarily conserved sequences on human chromosome 21. *Genome Res.*, 11:1651–1659.

Freeland, S. et Hurst, L. (1998). The genetic code is one in a million. *J. Mol. Evol.*, 47:238–248.

Friedberg, E. C., Wagner, R. et Radman, M. (2002). Specialized DNA polymerases, cellular survival, and the genesis of mutations. *Science*, 296:1627–1630.

Friedberg, E. C., Walker, G. C. et Siede, W. (1995). *DNA Repair and Mutagenesis*. ASM Press, Washington. 698 pages.

Fukuda, Y., Nakayama, Y. et Tomita, M. (2003). On dynamics of overlapping genes in bacterial genomes. *Gene*, 323:181–187.

Gerhart, J. et Kirschner, M. (1997). *Cells, Embryos, and Evolution*. Blackwell Scientific, Oxford. 642 pages.

Gerrish, P. J. et Lenski, R. E. (1998). The fate of competing beneficial mutations in an asexual population. *Genetica*, 102:127–144.

Gevers, D., Vandepoele, K., Simillion, C. et Van de Peer, Y. (2004). Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends Microbiol.*, 12:148–154.

Gheldof, N. et Dekker, J. (2004). Spatial organization of genomes. *Curr. Genomics*, 5:157–168.

Gil, R., Sabater-Munoz, B., Latorre, A., Silva, F. J. et Moya, A. (2002). Extreme genome reduction in *Buchnera* spp. : toward the minimal genome needed for symbiotic life. *Proc. Natl. Acad. Sci. USA*, 99(7):4454–4458.

Gilbert, N. (2004). Agent-based social simulation : Dealing with complexity. Rapport technique 12-08-04, Centre for Research on Social Simulation, University of Surrey, Guildford, Royaume-Uni. Disponible sur `http://www.complexityscience.org/modules.php?op=modload&name=Sections&file=index&req=viewarticle&artid=23&page=1` (dernière visite : le 7 novembre 2006).

Gillespie, J. H. (1981). Mutation modification in a random environment. *Evolution*, 35:468–476.

Giovannoni, H. J., Tripp, H. J., Givan, S., Podar, M., Vergin, K. L., Baptista, D., Bibbs, L., Eads, J., Richardson, T. H., Noordewier, M., Rappe, M. S., Short, J. M., Carrington, J. C. et Mathur, E. J. (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science*, 309:1242–1245.

GOLDBERG, A. L. et WITTES, R. E. (1966). Genetic code : Aspects of organization. *Science*, 153:420–424.

GOLDBERG, D., KORB, G. et DEB, K. (1989). Messy genetic algorithms : Motivation, analysis, and first results. *Complex Systems*, 3:493–530.

GOLDBERG, D. E., DEB, K., KARGUPTA, H. et HARIK, G. (1993). Rapid accurate optimization of difficult problems using fast messy genetic algorithms. *In* FORREST, S., éditeur : *Proceedings of the Fifth International Conference on Genetic Algorithms*, pages 56–64, Urbana-Champaign, USA. Morgan Kaufmann, San Mateo. 660 pages.

GRAUR, D., SHAULI, Y. et LI, W. H. (1989). Deletions in processed pseudogenes accumulate faster in rodents than in human. *J. Mol. Evol.*, 28:279–285.

GREGORY, T. R. (2001). Coincidence, coevolution, or causation ? DNA content, cell size, and the C-value enigma. *Biological Reviews of the Cambridge Philosophical Society*, 76:65–101.

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

GRIMM, V. (1999). Ten years of individual-based modelling in ecology : what have we learned and what could we learn in the future ? *Ecol. Modell.*, 115:129–148.

GU, Z., STEINMETZ, L. M., GU, X., SCHARFE, C., DAVIS, R. W. et LI, W. H. (2003). Role of duplicate genes in genetic robustness against null mutations. *Nature*, 421:63–66.

HAHN, M. W., STAJICH, J. E. et WRAY, G. A. (2003). The effects of selection against spurious transcription factor binding sites. *Mol. Biol. Evol.*, 20:901–906.

HAIG, D. et HURST, L. (1991). A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.*, 33:412–417.

HARIK, G. R. (1997). *Learning gene linkage to efficiently solve problems of bounded difficulty using genetic algorithms.* Thèse de doctorat, University of Michigan, Ann Arbor, USA. 148 pages.

HARVEY, I. (1992). Species Adaptation Genetic Algorithms : a basis for a continuing SAGA. *In* VARELA, F. J. et BOURGINE, P., éditeurs : *Toward a Practice of Autonomous Systems : Proceedings of the First European Conference on Artificial Life*, pages 346–354, Paris. MIT Press, Cambridge. 533 pages.

HERMISSON, J. et WAGNER, G. P. (2004). The population genetic theory of hidden variation and genetic robustness. *Genetics*, 168:2271–2284.

HERMISSON, J. et WAGNER, G. P. (2005). Evolution of phenotypic robustness. *In* JEN, E., éditeur : *Robust Design : A Repertoire from Biology, Ecology, and Engineering*, Santa Fe Institute Studies on the Sciences of Complexity, pages 47–70. Oxford University Press, Oxford. 320 pages.

Hogeweg, P. et Hesper, B. (1990). Individual oriented modelling in ecology. *Math. Comput. Modell.*, 13:83–90.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems.* University of Michigan Press, Ann Arbor. Seconde édition : MIT Press, 1992.

Holsinger, K. E. et Feldman, M. W. (1983). Modifiers of mutation rate : evolutionary optimum with complete selfing. *Proc. Natl. Acad. Sci. USA*, 80:6732–6734.

Houle, D. (1998). How should we explain variation in the genetic variance of traits ? *Genetica*, 103:241–253.

Hsu, T. C. (1975). A possible function of constitutive heterochromatin : the bodyguard hypothesis. *Genetics*, 79(Suppl.):137–150.

Huang, W., Petrosino, J., Hirsch, M., Shenkin, P. et Palzkill, T. (1996). Amino acid sequence determinants of beta-lactamase structure and activity. *J. Mol. Biol.*, 258:688–703.

Hughes, D. (2000). Evaluating genome dynamics : the constraints on rearrangements within bacterial genomes. *Genome Biol.*, 1:reviews0006.1–0006.8.

Hurst, L. D. (1995). The silence of the genes. *Curr. Biol.*, 5:459–461.

Hurst, L. D., Pal, C. et Lercher, M. J. (2004). The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.*, 5(4):299–310.

Huynen, M. A. (1996). Exploring phenotype space through neutral evolution. *J. Mol. Evol.*, 43(165–169).

Huynen, M. A. et Hogeweg, P. (1994). Pattern generation in molecular evolution : exploitation of the variation in RNA landscapes. *J. Mol. Evol.*, 39:71–79.

Huynen, M. A., Stadler, P. F. et Fontana, W. (1996). Smoothness within ruggedness : the role of neutrality in adaptation. *Proc. Natl. Acad. Sci. USA*, 93:397–401.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409:860–921.

Ishii, K., Matsuda, H., Iwasa, Y. et Sasaki, A. (1989). Evolutionarily stable mutation rate in a periodically changing environment. *Genetics*, 121:163–174.

Itoh, T., Martin, W. et Nei, M. (2002). Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. *Proc. Natl. Acad. Sci. USA*, 99:12944–12948.

Jeong, H., Mason, S. P., Barabasi, A. L. et Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411:41–42.

Johnson, T. (1999a). The approach to mutation-selection balance in an infinite asexual population, and the evolution of mutation rates. *Proc. R. Soc. Lond. B*, 266:2389–2397.

JOHNSON, T. (1999b). Beneficial mutations, hitchhiking and the evolution of mutation rates in sexual populations. *Genetics*, 151:1621–1631.

JOHNSON, T. et BARTON, N. H. (2002). The effect of deleterious alleles on adaptation in asexual populations. *Genetics*, 162:395–411.

JUDSON, O. P. et HAYDON, D. (1999). The genetic code : what is it good for ? an analysis of the effects of selection pressures on genetic codes. *J. Mol. Evol.*, 49:539–550.

KACSER, H. et BURNS, J. A. (1981). The molecular basis of dominance. *Genetics*, 97:639–666.

KARLIN, S. et MCGREGOR, J. (1974). Towards a theory of the evolution of modifier genes. *Theor. Pop. Biol.*, 5:59–103.

KASHI, Y., KING, D. et SOLLER, M. (1997). Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet.*, 13:74–78.

KATO-MAEDA, M., RHEE, J. T., GINGERAS, T. R., SALAMON, H., DRENKOW, J., SMITTIPAT, N. et SMALL, P. M. (2001). Comparing genomes within the species *Mycobacterium tuberculosis*. *Genome Res.*, 11:547–554.

KAWECKI, T. J. (2000). The evolution of genetic canalization under fluctuating selection. *Evolution*, 54(1):1–12.

KEIGHTLEY, P. D., KRYUKOV, G. V., SUNYAEV, S., HALLIGAN, D. L. et GAFFNEY, D. J. (2005). Evolutionary constraints in conserved nongenic sequences of mammals. *Genome Res.*, 15:1373–1378.

KIBOTA, T. T. et LYNCH, M. (1996). Estimate of the genomic mutation rate deleterious to overall fitness in *E. coli*. *Nature*, 381:694–696.

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

KIMURA, M. (1960). Optimum mutation rate and degree of dominance as determined by the principle of minimum genetic load. *J. Genet.*, 57:21–34.

KIMURA, M. (1967). On the evolutionary adjustment of spontaneous mutation rates. *Genet. Res.*, 9:23–34.

KITANO, H. (2002). Systems biology : a brief overview. *Science*, 295:1662–1664.

KLEINA, L. et MILLER, J. (1990). Genetic studies of the lac repressor. 13. extensive amino-acid replacements generated by the use of natural and synthetic non-sense suppressors. *J. Mol. Biol.*, 212:295–318.

KORNBERG, A. et BAKER, T. A. (1992). *DNA Replication*. Freeman and Co., New York. 931 pages.

KOZA, J. R. (1992). *Genetic Programming : On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge. 840 pages.

Koza, J. R. et Andre, D. (1995). Evolution of both the architecture and the sequence of work-performing steps of a computer program using genetic programming with architecture-altering operations. *In* Siegel, E. V. et Koza, J. R., éditeurs : *Working Notes for the AAAI Symposium on Genetic Programming*, pages 50–60, Cambridge, USA. AAAI, Menlo Park. 140 pages.

Krakauer, D. C. (2000). Stability and evolution of overlapping genes. *Evolution*, 54:731–739.

Krakauer, D. C. et Plotkin, J. B. (2002). Redundancy, antiredundancy, and the robustness of genomes. *Proc. Natl. Acad. Sci. USA*, 99:1405–1409.

Kuan, C. T., Liu, S. K. et Tessman, I. (1991). Excision and transposition of Tn5 as an SOS activity in *Escherichia coli*. *Genetics*, 128:45–57.

Lamb, R. A. et Horvath, C. M. (1991). Diversity of coding strategies in influenza viruses. *Trends Genet.*, 7:261–266.

Langdon, W. B. (1998). The evolution of size in variable length representations. *In 1998 IEEE International Conference on Evolutionary Computation*, pages 633–638, Anchorage, Alaska, USA. IEEE Press, Piscataway. 600 pages.

Langdon, W. B. et Poli, R. (1997). Fitness causes bloat. *In* Chawdhry, P. K. *et al.*, éditeurs : *Soft Computing in Engineering Design and Manufacturing*, pages 13–22. Springer-Verlag, Londres. 480 pages.

Lathe, W. C., Snel, B. et Bork, P. (2000). Gene context conservation of a higher order than operons. *Trends Biochem. Sci.*, 25:474–479.

Lawrence, J. G., Hendrix, R. W. et Casjens, S. (2001). Where are the pseudogenes in bacterial genomes ? *Trends Microbiol.*, 9:535–540.

Lawrence, J. G. et Roth, J. R. (1996). Selfish operons : Horizontal transfer may drive the evolution of gene clusters. *Genetics*, 143:1843–1860.

Layzer, D. (1980). Genetic variation and progressive evolution. *Am. Nat.*, 115:809–826.

LeClerc, J. E., Li, B., Payne, W. L. et Cebula, T. A. (1996). High mutation frequencies among *Escherichia coli* and *Salmonella* pathogens. *Science*, 274:1208–1211.

Leigh, E. G. (1970). Natural selection and mutability. *Am. Nat.*, 104:301–305.

Leigh, E. G. (1973). The evolution of mutation rates. *Genetics*, 73(Suppl.):1–18.

Leong, S., Chang, J., Ong, R., Dawes, G., Stemmer, W. et Punnonen, J. (2003). Optimized expressionand specific activity of il-12 by directed molecular evolution. *Proc. Natl. Acad. Sci. USA*, 100:1163–1168.

Levenick, J. R. (1991). Inserting introns improves genetic algorithm success rate : Taking a cue from biology. *In* Belew, R. et Booker, L., éditeurs : *Proceedings of the Fourth International Conference on Genetic Algorithms*, pages 123–127, San Diego, USA. Morgan Kaufman, San Francisco. 569 pages.

Loeb, D. S. R., Everitt, L., Manchester, M., Stamper, S. et Hutchinson, C. (1989). Complete mutagenesis of the HIV-1 protease. *Nature*, 340:397–400.

Ludwig, M. Z., Bergman, C., Patel, N. H. et Kreitman, M. (2000). Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature*, 403:564–567.

Luke, S. (2000). *Issues in Scaling Genetic Programming : Breeding Strategies, Tree Generation, and Code Bloat*. Thèse de doctorat, University of Maryland, College Park, USA. 178 pages.

Luke, S. (2003). Modification point depth and genome growth in genetic programming. *Evol. Comput.*, 11:67–106.

Luke, S. (2005). Evolutionary computation and the C-value paradox. *In Proceedings of Genetic and Evolutionary Computation Conference (GECCO-2005)*, pages 91–97, Washington, USA. ACM Press, New York. 2222 pages.

Lunter, G., Ponting, C. P. et Hein, J. (2006). Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput. Biol.*, 2:e5.

Lynch, M. (2006a). The origins of eukaryotic gene structure. *Mol. Biol. Evol.*, 23:450–468.

Lynch, M. (2006b). Streamlining and simplification of microbial genome architecture. *Annu. Rev. Microbiol.*, 60:327–349.

Lynch, M., Blanchard, J., Houle, D., Kibota, T., Schultz, S., Vassilieva, L. et Willis, J. (1999). Spontaneous deleterious mutation. *Evolution*, 53:645–663.

Lynch, M., Burger, D. et Gabriel, W. (1993). The mutational meltdown in asexual populations. *J. Hered.*, 84:339–344.

Lynch, M. et Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, 290:1151–1155.

Lynch, M. et Conery, J. S. (2003). The origins of genome complexity. *Science*, 302:1401–1404.

Maere, S., Bodt, S. D., Raes, J., Casneuf, T., Montagu, M. V., Kuiper, M. et de Peer, Y. V. (2005). Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. USA*, 102:5454–5459.

Maeshiro, T. et Kimura, M. (1998). The role of robustness and changeability on the origin and evolution of genetic codes. *Proc. Natl. Acad. Sci. USA*, 95:5088–5093.

Maestre, J., Tchenio, T., Dhellin, O. et Heidmann, T. (1995). mRNA retroposition in human cells : processed pseudogene formation. *EMBO J.*, 14:6333–6338.

Makalowska, I., Lin, C. F. et Makalowski, W. (2005). Overlapping genes in vertebrate genomes. *Comput. Biol. Chem.*, 29:1–12.

Maniloff, J. (1996). The minimal cell genome : "on being the right size". *Proc. Natl. Acad. Sci. USA*, 93:10004–10006.

Mao, E. F., Lane, L., Lee, J. et Miller, J. H. (1997). Proliferation of mutators in a cell population. *J. Bacteriol.*, 179:417–422.

Margulies, E. H., Blanchette, M., Haussler, D. et Green, E. D. (2003). Identification and characterization of multi-species conserved sequences. *Genome Res.*, 13:2507–2518.

Martin, G. et Lenormand, T. (2006). A general multivariate extension of Fisher's geometrical model and the distribution of mutation fitness effects across species. *Evolution*, 60:893–907.

Matic, I., Rayssiguier, C. et Radman, M. (1995). Interspecies gene exchange in bacteria : the role of SOS and mismatch repair systems in evolution of species. *Cell*, 80:507–515.

Maynard Smith, J. (1983). Models of evolution. *Proc. R. Soc. B*, 219:315–325.

Meselson, M. et Stahl, F. W. (1958). The replication of DNA in *Escherichia coli*. *Proc. Natl. Acad. Sci USA*, 44:671–682.

Meyer-Nieberg, S. et Beyer, H. G. (2006). Self-adaptation in evolutionary algorithms. *In* Lobo, F., Lima, C. et Michalewicz, Z., éditeurs : *Parameter Setting in Evolutionary Algorithms*. Springer, New York.

Mikkola, R. et Kurland, C. G. (1991). Is there a unique ribosome phenotype for naturally occuring *Escherichia coli* ? *Biochimie*, 73:1061–1066.

Miller, E. S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T. et Ruger, W. (2003). Bacteriophage T4 genome. *Microbiol. Mol. Biol. Rev.*, 67:86–156.

Miller, G. F. (1995). Artificial life as theoretical biology : How to do real science with computer simulation. Rapport technique CSRP 378, School of Cognitive and Computing Sciences, University of Sussex. Disponible sur `http://cogslib.cogs.susx.ac.uk/details.php?id=7727` (dernière visite : le 7 novembre 2006).

Miller, J. H. (1996). Spontaneous mutators in bacteria : Insights into pathways of mutagenesis and repair. *Annu. Rev. Microbiol.*, 50:625–643.

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

Miyake, T. (1960). Mutator factor in *Salmonella typhimurium*. *Genetics*, 45:11–14.

Moran, N. A. et Mira, A. (2001). The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. *Genome Biol.*, 2:R54.

Moran, N. A. et Wernegreen, J. J. (2000). Lifestyle evolution in symbiotic bacteria : insights from genomics. *Trends Ecol. Evol.*, 15:321–326.

Morgan, T. H., Sturtevant, A. H., Muller, H. J. et Bridges, C. (1915). *The Mechanism of Mendelian Heredity*. Holt, Rinehart & Winston, New York. Disponible sur `http://www.esp.org/books/morgan/mechanism/facsimile/` (dernière visite : le 7 novembre 2006).

Mouchiroud, D., D'Onofrio, G., Aissani, B., Macaya, G., Gautier, C. et Bernardi, G. (1991). The distribution of genes in the human genome. *Gene*, 100:181–187.

Moxon, E. R. et Thaler, D. S. (1997). The tinkerer's evolving tool-box. *Nature*, 387:659–662.

Moxon, R., Bayliss, C. et Hood, D. (2006). Bacterial contingency loci : the role of simple sequence DNA repeats in bacterial populations. *Annu. Rev. Genet.* In press.

Nelson, C. E., Hersh, B. M. et Carroll, S. B. (2004). The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol.*, 5(4):R25.

Nilsson, A. I., Koskiniemi, S., Eriksson, S., Kugelberg, E., Hinton, J. C. D. et Andersson, D. I. (2005). Bacterial genome size reduction by experimental evolution. *Proc. Natl. Acad. Sci. USA*, 102:12112–12116.

Nordin, P. et Banzhaf, W. (1995). Complexity compression and evolution. *In* Eshelman, L., éditeur : *Proceedings of the Sixth International Conference on Genetic Algorithms (ICGA 95)*, pages 310–317, Pittsburgh. Morgan Kaufmann, San Francisco. 624 pages.

Nordin, P., Banzhaf, W. et Francone, F. D. (1997). Introns in nature and in simulated structure evolution. *In* Lundh, D., Olsson, B. et Narayanan, A., éditeurs : *Bio-Computation and Emergent Computation : Proceedings of BCEC97*, pages 22–35, Skovde, Suède. World Scientific Publishing. 300 pages.

Normark, S., Bergström, S., Edlund, T., Grundström, T., Jaurin, B., Lindberg, F. P. et Olsson, O. (1983). Overlapping genes. *Ann. Rev. Genet.*, 17:499–525.

Nuzhdin, S. V. (1999). Sure facts, speculations, and open questions about the evolution of transposable element copy number. *Genetica*, 107:129–137.

Ochman, H. (2005). Genomes on the shrink. *Proc. Natl. Acad. Sci. USA*, 102:11959–11960.

Ochman, H. et Jones, I. B. (2000). Evolutionary dynamics of full genome content in *Escherichia coli*. *EMBO J.*, 19:6637–6643.

Ochman, H. et Moran, N. A. (2001). Genes lost and genes found : Evolution of bacterial pathogenesis and symbiosis. *Science*, 292:1096–1098.

Ofria, C. et Adami, C. (1999). Evolution of genetic organization in digital organisms. *In* Landweber, L. et Winfree, E., éditeurs : *Evolution as Computation*, pages 2296–313. Springer, New York. 300 pages.

Ofria, C., Adami, C. et Collier, T. C. (2003). Selective pressures on genomes in molecular evolution. *J. Theor. Biol.*, 222:477–483.

Ohno, S. (1970). *Evolution by Gene Duplication*. Springer-Verlag, New York. 160 pages.

Oliver, A., Canton, R., Campo, P., Baquero, F. et Blazquez, J. (2000). High frequency of hypermutable *Pseudomonas aeruginosa* in cystic fibrosis lung infection. *Science*, 288:1251–1253.

Ophir, R. et Graur, D. (1997). Patterns and rates of indel evolution in processed pseudogenes from humans and murids. *Gene*, 205:191–202.

Orr, H. A. (2000). The rate of adaptation in asexuals. *Genetics*, 155:961–968.

Pal, C. et Hurst, L. D. (2000). The evolution of gene number : are heritable and non-heritable errors equally important ? *Heredity*, 84:393–400.

Pal, C. et Hurst, L. D. (2004). Evidence against the selfish operon theory. *Trends Genet.*, 20(6):232–234.

Pavesi, A., De Iaco, B., Granero, M. I. et Porati, A. (1997). On the informational content of overlapping genes in prokaryotic and eukaryotic viruses. *J. Mol. Evol.*, 44: 625–631.

Peck, J. (1994). A ruby in a rubbish : beneficial mutations, deleterious mutations, and the evolution of sex. *Genetics*, 137:597–606.

Pepper, J. W. (2000). The evolution of modularity in genome architecture. *In* Maley, C. C. et Boudreau, E., éditeurs : *Proceedings of the Evolvability Workshop at Alife VII*, Portland, USA. Disponible sur `http://www.santafe.edu/~jpepper/papers/ALIFE7_modularity.pdf` (dernière visite : le 7 novembre 2006).

Petit, M. A., Dimpfl, J., Radman, M. et Echols, H. (1991). Control of large chromosomal duplications in *Escherichia coli* by the mismatch repair system. *Genetics*, 129:327–332.

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

Petrov, D. A. (2002). Mutational equilibrium model of genome size evolution. *Theor. Popul. Biol.*, 61:533–546.

Petrov, D. A., Sangster, T. A., Johnston, J. S., Hartl, D. L. et Shaw, K. L. (2000). Evidence for DNA loss as a determinant of genome size. *Science*, 287:1060–1062.

Plotkin, J. B., Dushoff, J. et Fraser, H. B. (2004). Detecting selection using a single genome sequence of *M. tuberculosis* and *P. falciparum*. *Nature*, 428:942–945.

Plough, H. H. (1941). Spontaneous mutability in *Drosophila*. *Cold Spring Harbor Symp. Quant. Biol.*, 9:127–137.

Price, M. N., Huang, K. H., Arkin, A. P. et Alm, E. J. (2005). Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res.*, 15:809–819.

Queitsch, C., Sangster, T. A. et Lindquist, S. (2002). Hsp90 as a capacitor of phenotypic variation. *Nature*, 417:618–624.

Radman, M., Matic, I. et Taddei, F. (1999). Evolution of evolvability. *Ann. N. Y. Acad. Sci.*, 870:146–155.

Radman, M. et Wagner, R. (1986). Mismatch repair in *Escherichia coli*. *Annu. Rev. Genet.*, 20:523–538.

Ramsey, C. L., De Jong, K. A., Grefenstette, J. J., Wu, A. S. et Burke, D. S. (1998). Genome length as an evolutionary self-adaptation. *In Proceedings of the 5th International Conference on Parallel Problem Solving from Nature*, volume 1498 de *Lect. Notes Comput. Sci.*, pages 345–356, Amsterdam, Pays-Bas. Springer-Verlag, Londres. 1041 pages.

Rayssiguier, C., Thaler, S. et Radman, M. (1989). The barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch-repair mutants. *Nature*, 342:396–401.

Rennard, J.-P. (2002). *Vie artificielle*. Vuibert, Paris. 408 pages.

Rennell, D., Bouvier, S., Hardy, L. et Poteete, A. (1991). Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.*, 222:67–87.

Reynolds, C. W. (1987). Flocks, herds, and schools : A distributed behavioral model. *In SIGGRAPH '87 Conference Proceedings*, volume 21 de *Computer Graphics*, pages 25–34, Anaheim, USA. ACM Press, New York.

Richardson, A. R., Yu, Z., Popovic, T. et Stojiljkovic, I. (2002). Mutator clones of *Neisseiria meningitidis* in epidemic serogroup A disease. *Proc. Natl. Acad. Sci. USA*, 99:6103–6107.

Rocha, E. P. C. (2003a). An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences : from duplications to genome reduction. *Genome Res.*, 13:1123–1132.

Rocha, E. P. C. (2003b). DNA repeats lead to the accelerated loss of gene order in bacteria. *Trends Genet.*, 19:600–603.

Rocha, E. P. C. et Danchin, A. (2003a). Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat. Genet.*, 34:377–378.

Rocha, E. P. C. et Danchin, A. (2003b). Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res.*, 31:6570–6577.

Rocha, E. P. C., Danchin, A. et Viari, A. (1999a). Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in bacillus subtilis and other competent prokaryotes. *Mol. Biol. Evol.*, 16:1219–1230.

Rocha, E. P. C., Danchin, A. et Viari, A. (1999b). Functional and evolutionary roles of long repeats in prokaryotes. *Res. Microbiol.*, 150:725–733.

Rogozin, I. B., Makarova, K. S., Murvai, J., Czabarka, E., Wolf, Y. I., Tatusov, R. L., Szekely, L. A. et Koonin, E. V. (2002a). Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res.*, 30:2212–2223.

Rogozin, I. B., Spiridonov, A. N., Sorokin, A. V., Wolf, Y. I., Jordan, I. K., Tatusov, R. L. et Koonin, E. V. (2002b). Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet.*, 18:228–232.

Romano, L. et Wray, G. (2003). Conservation of endo16 expression in sea urchins despite evolutionary divergence in both cis and trans-acting components of transcriptional regulation. *Development*, 130:4187–4199.

Rosenberg, S. M. (2001). Evolving responsively : adaptive mutation. *Nat. Rev. Genet.*, 2:504–515.

Rutherford, S. L. et Lindquist, S. (1998). Hsp90 as a capacitor for morphological evolution. *Nature*, 396:336–342.

Samuel, C. E. (1989). Polycistronic animal virus mRNAs. *Prog. Nucleic Acid Res. Mol. Biol.*, 37:127–153.

Sankoff, D. (2003). Rearrangements and chromosomal evolution. *Curr. Opin. Genet. Dev.*, 13:583–587.

Schaaper, R. M. (1998). Antimutator mutants in bacteriophage T4 and *Escherichia coli*. *Genetics*, 148:1579–1585.

Scharloo, W. (1991). Canalization : Genetic and developmental aspects. *Annu. Rev. Ecol. Syst.*, 22:65–93.

Schuster, P., Fontana, W., Stadler, P. F. et Hofacker, I. (1994). From sequences to shapes and back : a case study in RNA secondary structures. *Proc. R. Soc. Lond. B*, 255:279–284.

Schuster, P. et Swetina, J. (1988). Stationary mutant distributions and evolutionary optimization. *Bull. Math. Biol.*, 50:635–660.

Schutz, M. (1995). Other operators : gene duplication and deletion. *In* Baeck, T., Fogel, D. et Michalewicz, Z., éditeurs : *Handbook of Evolutionary Computation*, pages C3.4.2 :1–5. IOP Publishing Ltd et Oxford University Press, Bristol. 988 pages.

Shimeld, S. M. (1999). Gene function, gene networks and the fate of duplicated genes. *Semin. Cell. Dev. Biol.*, 10:549–553.

Silva, F. J., Latorre, A. et Moya, A. (2001). Genome size reduction through multiple events of gene disintegration in buchnera aps. *Trends Genet.*, 17:615–618.

Sinervo, B. et Svensson, E. (2002). Correlational selection and the evolution of genomic architecture. *Heredity*, 89(5):329–338.

Skipper, R. (2004). The heuristic role of Sewall Wright's 1932 adaptive landscape diagram. *Philosophy of Science*, 71:1176–1188.

Smit, A. F. A. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.*, 9:657–663.

Sniegowski, P. D., Gerrish, P. J., Johnson, T. et Shaver, A. (2000). The evolution of mutation rates : separating causes from consequences. *Bioessays*, 22:1057–1066.

Sniegowski, P. D., Gerrish, P. J. et Lenski, R. E. (1997). Evolution of high mutation rates in experimental populations of e. coli. *Nature*, 387:703–705.

Soule, T. et Foster, J. A. (1998). Removal bias : a new cause of code growth in tree based evolutionary programming. *In 1998 IEEE International Conference on Evolutionary Computation*, pages 781–186, Anchorage, Alaska, USA. IEEE Press, Piscataway. 600 pages.

Starlinger, P. (1977). DNA rearrangements in procaroytes. *Annu. Rev. Genet.*, 11:103–126.

Stearns, S. C. (1993). The evolutionary links between fixed and variable traits. *Acta Paleont. Polonica*, 38:1–17.

Stearns, S. C. (2002). Progress on canalization. *Proc. Natl Acad. Sci. USA*, 99:10229–10230.

Stone, J. et Wray, G. (2001). Rapid evolution of cis-regulatory sequences via local point mutations. *Mol. Biol. Evol.*, 18:1764–1770.

Taddei, F., Radman, M., Maynard-Smith, J., Toupance, B., Gouyon, P. H. et Godelle, B. (1997). Role of mutator alleles in adaptive evolution. *Nature*, 387:700–702.

Tamames, J. (2001). Evolution of gene order conservation in prokaryotes. *Genome Biol.*, 2:R20.

Tan, T., Bogarad, L. D. et Deem, M. W. (2004). Modulation of base-specific mutation and recombination rates enables functional adaptation within the context of the genetic code. *J. Mol. Evol.*, 59:385–399.

Teichmann, S. A. et Babu, M. (2004). Gene regulatory network growth by duplication. *Nat. Genet.*, 36:492–496.

Tenaillon, O., Toupance, B., Le Nagard, H., Taddei, F. et Godelle, B. (1999). Mutators, population size, adaptive landscape and the adaptation of asexual populations of bacteria. *Genetics*, 152:485–493.

Thornton, J. M., Orengo, C. A., Todd, A. E. et Pearl, F. M. (1999). Protein folds, functions and evolution. *J. Mol. Biol.*, 293:333–342.

Treffers, H. P., Spinelli, V. et Belser, N. O. (1954). A factor (or mutator gene) influencing mutation rates in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA*, 40:1064–1071.

Van der Woude, M. W. et Baumler, A. J. (2004). Phase and antigenic variation in bacteria. *Clin. Microbiol. Rev.*, 17:581–611.

Van Nimwegen, E., Crutchfield, J. P. et Huynen, M. (1999). Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci. USA*, 96:9716–9720.

Van Valen, L. (1973). A new evolutionary law. *Evolutionary Theory*, 1:1–30.

Veeramachaneni, V., Makalowski, W., Galdzicki, M., Sood, R. et Makalowska, I. (2004). Mammalian overlapping genes : The comparative perspective. *Genome Research*, 14:280–286.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A. *et al.* (2001). The sequence of the human genome. *Science*, 291:1304–1351.

Von Dassow, G., Meir, E., Munro, E. M. et Odell, G. M. (2000). The segment polarity network is a robust developmental module. *Nature*, 406:188–192.

Von Tschermak, E. (1900). Über künstliche Kreuzung bei *Pisum sativum*. *Zeitschrift für das landwirtschaftliche Versuchswesen in Österreich*, 3:465–555. Habilitationsschrift.

Waddington, C. H. (1957). *The Strategy of the Genes*. MacMillan, New York. 262 pages.

Wagner, A. (1994). Evolution of gene networks by gene duplications : a mathematical model and its implications on genome organization. *Proc. Natl. Acad. Sci. USA*, 91: 4387–4391.

Wagner, A. (1996). Does evolutionary plasticity evolve ? *Evolution*, 50:1008–1023.

Wagner, A. (2000). Mutational robustness in genetic networks of yeast. *Nat. Genet.*, 24:355–361.

Wagner, A. (2005a). *Robustness and Evolvability in Living Systems*. Princeton University Press, Princeton. 408 pages.

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

Wagner, G. P., Booth, G. et Bagheri-Chaichian, H. (1997). A population genetic theory of canalization. *Evolution*, 51:329–347.

Wagner, G. P. et Krall, P. (1993). What is the difference between models of error thresholds and Muller's ratchet ? *J. Math. Biol.*, 32:33–44.

Watson, J. D. et Crick, F. H. C. (1953). Molecular structure of nucleic acids : A structure for Deoxyribose Nucleic Acid. *Nature*, 171:737–738.

Weatherall, D. J. et Clegg, J. B. (1976). Molecular genetics of human haemoglobin. *Ann. Rev. Genet.*, 10:157–178.

Whitley, D. (1989). The GENITOR algorithm and selection pressure : Why rank-based allocation of reproductive trials is best. *In* Schaffer, J. D., éditeur : *Proceedings of the 3rd International Conference on Genetic Algorithms*, pages 116–121, Fairfax, USA. Morgan Kaufmann, San Mateo. 439 pages.

Wilke, C. O. (2001a). Adaptive evolution on neutral networks. *Bull. Math. Biol.*, 63:715–730.

Wilke, C. O. (2001b). Selection for fitness versus selection for robustness in RNA secondary structure folding. *Evolution*, 55:2412–2420.

Wilke, C. O. (2005). Quasispecies theory in the context of population genetics. *BMC Evol. Biol.*, 5:44.

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 the survival of the flattest. *Nature*, 412:331–333.

Woese, C. R. (1965). On the evolution of the genetic code. *Proc. Natl. Acad. Sci. USA*, 54:1546–1552.

Wooding, S. et Jorde, L. B. (2006). Duplication and divergence in humans and chimpanzees. *Bioessays*, 28:335–338.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *In* Jones, D., éditeur : *Proceedings of the Sixth International Congress on Genetics*, pages 355–366. Reproduit dans Ridley, M. (1997). *Evolution*. Oxford University Press, Oxford, pages 32-40.

Wu, A. S. et Lindsay, R. K. (1995). Empirical studies of the genetic algorithm with non-coding segments. *Evol. Comput.*, 3:121–147.

Wu, A. S. et Lindsay, R. K. (1996). A comparison of the fixed and floating building block representation in the genetic algorithm. *Evol. Comput.*, 4:169–193.

Yu, H., Wu, A. S., Lin, K.-C. et Schiavone, G. A. (2003). Adaptation of length in a nonstationary environment. *In* Cantu-Paz, E. *et al.*, éditeurs : *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2003), Part II*, volume 2724 de *Lect. Notes Comput. Sci.*, pages 1541–1553, Chicago, USA. Springer, New York. 1274 pages.

ZADEH, L. (1978). Fuzzy sets as the basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28. Réédité en 1999 dans le supplément du volume 100 de la même revue, pages 9–34.

Annexe : Calcul de F_ν

Considérons un chromosome circulaire de L paires de bases, numérotées de 1 à L. Soit N_G le nombre de "régions fonctionnelles" que ce chromosome contient, une région fonctionnelle étant définie comme une région transcrite (promoteur et terminateur inclus) contenant au moins une séquence codante[1]. Durant sa réplication, ce chromosome peut subir des mutations locales (mutations ponctuelles, petites insertions, petites délétions), ainsi que des réarrangements chromosomiques (duplications, délétions, translocations, inversions). Pour le type de mutation j, le taux par bp est noté u_j. À chaque réplication, le nombre de mutations de type j est tiré selon la loi binomiale $\mathcal{B}(L, u_j)$. Une fois le nombre de mutations ponctuelles connues, les positions du chromosome concernées sont tirées au hasard le long du chromosome (loi uniforme). Il en va de même pour les petites insertions et les petites délétions. Pour les réarrangements, nous considérons ici le cas où la loi de la longueur des segments réarrangés est uniforme entre 1 et L : les bornes d'un segment dupliqué, excisé, inversé ou déplacé sont toutes deux choisies uniformément sur le chromosome.

Connaissant la structure du chromosome, nous voulons calculer la probabilité F_ν qu'il soit répliqué sans mutation ou bien uniquement avec des mutations neutres. Cette probabilité correspond à la proportion attendue de "descendants neutres", c'est-à-dire phénotypiquement identiques au progéniteur. Nous l'estimons par la probabilité $\tilde{F}_\nu$ qu'aucune mutation n'affecte de région fonctionnelle. Nous négligeons donc (i) la probabilité qu'une mutation en dehors de toute région fonctionnelle crée un nouveau gène, (ii) la probabilité qu'une mutation dans une région fonctionnelle soit neutre, et (iii) la probabilité que plusieurs mutations dans une même région fonctionnelle se compensent exactement.

Le calcul s'effectue en deux étapes. Dans un premier temps, nous calculons, pour chaque type de mutation, la probabilité $\tilde{\nu}$ qu'une mutation réalisée aléatoirement n'affecte aucune des régions fonctionnelles. Nous calculons ensuite $\tilde{F}_\nu$ en prenant en compte les taux de mutation.

[1] Si deux régions transcrites se chevauchent, elles sont considérées comme une seule région fonctionnelle.

Probabilité qu'une mutation aléatoire n'affecte aucune région fonctionnelle

Notations :

- b_i est la position où la i-ème région fonctionnelle débute,
- e_i est la position où la i-ème région fonctionnelle s'arrête,
- $l_i = e_i - b_i + 1$ est la longueur de la i-ème région fonctionnelle,
- $l = \sum_{i=1}^{N_G} l_i$ est le nombre total de bp fonctionnelles,
- $\lambda_i = b_{i+1} - e_i - 1$ est le nombre de bp strictement comprises entre les régions fonctionnelles i et $i+1$,
- $\lambda = \sum_{i=1}^{N_G} \lambda_i = L - l$ est le nombre total de bp non fonctionnelles,
- p est la position affectée par une mutation locale,
- p_1 est la position où un segment réarrangé débute,
- p_2 est la position où un segment réarrangé s'arrête,
- p_3 est la position où un segment dupliqué ou déplacé est réinséré.

Mutation locale

Le raisonnement est le même pour les mutations ponctuelles, les petites insertions et les petites délétions. Comme la position p d'une mutation locale est choisie uniformément sur le chromosome, toutes les positions sont équiprobables. La probabilité qu'une mutation locale n'affecte aucune région fonctionnelle peut donc être simplement calculée comme le rapport du nombre de cas favorables sur le nombre de cas possibles :

$$\tilde{\nu}_{\text{ponct}} = \tilde{\nu}_{\text{del}} = \tilde{\nu}_{\text{ins}} = \frac{L-l}{L} = 1 - \frac{l}{L} \tag{IV.3}$$

Inversion

Une inversion n'affecte aucune région fonctionnelle si ses deux points de rupture sont situés dans des zones intergéniques. Les deux positions p_1 et p_2 doivent donc être à l'intérieur de régions non fonctionnelles (éventuellement différentes). Comme p_1 et p_2 sont choisies indépendamment, nous pouvons écrire :

$$\tilde{\nu}_{\text{inv}} = P\left(p_1 \text{ est dans une région non fonc.}\right) P\left(p_2 \text{ est dans une région non fonc.}\right)$$

Les positions p_1 et p_2 sont choisies uniformément sur le chromosome. Ainsi :

$$\tilde{\nu}_{\text{inv}} = \left(1 - \frac{l}{L}\right)^2 \tag{IV.4}$$

Translocation

Une translocation n'affecte aucune région fonctionnelle si p_1 et p_2 sont à l'intérieur de régions non fonctionnelles (éventuellement différentes) et si le point de ré-insertion p_3 est également situé dans une région non fonctionnelle. Le segment est d'abord excisé du chromosome, puis p_3 est choisie uniformément sur le chromosome raccourci, et enfin le segment est ré-inséré à la position choisie. Par conséquent, la probabilité que p_3 soit choisie dans une région non fonctionnelle n'est pas exactement égale à $1-\frac{l}{L}$, comme pour p_1 et p_2. Cependant, si nous supposons que la densité en gènes varie peu le long du chromosome, cela reste une bonne approximation. Sous cette hypothèse, nous avons :

$$\tilde{\nu}_{\text{transloc}} = \left(1 - \frac{l}{L}\right)^3 \tag{IV.5}$$

Grande délétion

Une grande délétion n'affecte aucune région fonctionnelle si p_1 est dans une région non fonctionnelle, et si p_2 est entre p_1 et le début de la prochaine région fonctionnelle.

$$\tilde{\nu}_{\text{gdel}} = \sum_{i=1}^{N_G} \sum_{m=e_i+1}^{b_{i+1}-1} P\left(p_2 = p_1, p_1+1, \text{ ..., ou } b_{i+1}-1 \mid p_1 = m\right) P\left(p_1 = m\right)$$

Le choix de p_2 est indépendant de la valeur de p_1. Donc :

$$\begin{aligned}\tilde{\nu}_{\text{gdel}} &= \sum_{i=1}^{N_G} \sum_{m=e_i+1}^{b_{i+1}-1} P\left(p_2 = p_1, p_1+1, \text{ ..., or } b_{i+1}-1\right) P\left(p_1 = m\right) \\ &= \sum_{i=1}^{N_G} \sum_{m=e_i+1}^{b_{i+1}-1} \left(P\left(p_1 = m\right) \sum_{k=m}^{b_{i+1}-1} P\left(p_2 = k\right) \right)\end{aligned}$$

Comme p_1 et p_2 sont choisies uniformément sur le chromosome,

$$\forall m,k \in \{1,...,L\}\ P(p_1 = m) = P(p_2 = k) = \frac{1}{L}$$

Ainsi :

$$\begin{aligned}
\tilde{\nu}_{\text{gdel}} &= \sum_{i=1}^{N_G} \sum_{m=e_i+1}^{b_{i+1}-1} \left(\frac{1}{L} \sum_{k=m}^{b_{i+1}-1} \frac{1}{L} \right) \\
&= \frac{1}{L^2} \sum_{i=1}^{N_G} \sum_{m=e_i+1}^{b_{i+1}-1} \sum_{k=m}^{b_{i+1}-1} (1) \\
&= \frac{1}{L^2} \sum_{i=1}^{N_G} \sum_{m=e_i+1}^{b_{i+1}-1} (b_{i+1} - m) \\
&= \frac{1}{L^2} \sum_{i=1}^{N_G} \left(\sum_{m=e_i+1}^{b_{i+1}-1} b_{i+1} - \sum_{m=e_i+1}^{b_{i+1}-1} m \right) \\
&= \frac{1}{L^2} \sum_{i=1}^{N_G} \left(b_{i+1}(b_{i+1} - e_i - 1) - \sum_{n=1}^{b_{i+1}-e_i-1} (n + e_i) \right)
\end{aligned}$$

Or $b_{i+1} - e_i - 1 = \lambda_i$:

$$\begin{aligned}
\tilde{\nu}_{\text{gdel}} &= \frac{1}{L^2} \sum_{i=1}^{N_G} \left(b_{i+1}\lambda_i - \sum_{n=1}^{\lambda_i} (n + e_i) \right) \\
&= \frac{1}{L^2} \sum_{i=1}^{N_G} \left(b_{i+1}\lambda_i - \frac{\lambda_i(\lambda_i+1)}{2} - \lambda_i e_i) \right) \\
&= \frac{1}{L^2} \sum_{i=1}^{N_G} \left(\lambda_i(b_{i+1} - e_i) - \frac{\lambda_i(\lambda_i+1)}{2}) \right) \\
&= \frac{1}{L^2} \sum_{i=1}^{N_G} \left(\lambda_i(\lambda_i+1) - \frac{\lambda_i(\lambda_i+1)}{2}) \right)
\end{aligned}$$

Nous avons donc finalement :

$$\tilde{\nu}_{\text{gdel}} = \frac{1}{2L^2} \sum_{i=1}^{N_G} \lambda_i(\lambda_i+1) \qquad \text{(IV.6)}$$

Duplication

Une duplication n'affecte aucune région fonctionnelle si (i) le segment copié ne comprend aucune région fonctionnelle, et (ii) le point d'insertion de la copie est dans une région non fonctionnelle. Les deux étapes sont indépendantes. La probabilité de (i) a été calculée ci-dessus et est en fait égale à $\tilde{\nu}_{\text{gdel}}$. La probabilité de (ii) est la probabilité que p_3 soit dans une région non fonctionnelle. Contrairement au cas de la translocation, p_3 est choisi sur un chromosome intact de longueur L. Nous avons donc :

$$\tilde{\nu}_{\text{duplic}} = \frac{1}{2L^2}\left(1-\frac{l}{L}\right)\sum_{i=1}^{N_G}\lambda_i(\lambda_i+1) \tag{IV.7}$$

Probabilité qu'aucune région fonctionnelle ne soit mutée pendant la réplication

Soit X_j le nombre d'événements de type j que le chromosome subit pendant sa réplication. X_j suit la loi binomiale $\mathcal{B}(L,u_j)$. La probabilité qu'aucun de ces événements n'affecte de région fonctionnelle est :

$$\begin{aligned}\tilde{F}_{\nu,j} &= \sum_{x=0}^{L} P(\text{aucun des évén. } j \text{ n'affecte de région fonc. } \mid X_j=x)P(X_j=x)\\ &= \sum_{x=0}^{L} P(\text{aucun des évén. } j \text{ n'affecte de région fonct. } \mid X_j=x)\binom{L}{x}u_j^x(1-u_j)^{L-x}\end{aligned}$$

Supposons que les événements successifs sont indépendants. C'est exact pour les mutations locales, mais c'est une approximation pour les réarrangements, puisqu'ils peuvent modifier les distances intergéniques, le nombre de gènes ou la taille du chromosome. Sous cette hypothèse simplificatrice, nous pouvons écrire :

$$\begin{aligned}\tilde{F}_{\nu,j} &= \sum_{x=0}^{L}\tilde{\nu}_j^x\binom{L}{x}u_j^x(1-u_j)^{L-x}\\ &= \sum_{x=0}^{L}\binom{L}{x}(u\tilde{\nu}_j)^x(1-u_j)^{L-x}\\ &= (u_j\tilde{\nu}_j+1-u_j)^L\\ &= (1-u_j(1-\tilde{\nu}_j))^L\end{aligned}$$

Enfin, nous devons prendre en compte tous les types de mutation que le chromosome peut subir. Si nous supposons que les événements de types différents sont indépendants, ce qui est à nouveau une simplification, alors la probabilité qu'aucune région fonctionnelle ne soit mutée durant la réplication peut s'écrire :

$$\tilde{F}_\nu = \prod_j (1 - u_j(1 - \tilde{\nu}_j))^L \tag{IV.8}$$

où les $\tilde{\nu}_j$ sont à remplacer par les expressions précédemment obtenues.

FOLIO ADMINISTRATIF

THÈSE SOUTENUE DEVANT L'INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE LYON

NOM : KNIBBE DATE de SOUTENANCE : 4 décembre 2006

Prénoms : Carole

TITRE : Structuration des génomes par sélection indirecte de la variabilité mutationnelle
Une approche de modélisation et de simulation

NATURE : Doctorat Numéro d'ordre : 2006-ISAL-0093

École doctorale : Evolution Ecosystèmes Microbiologie Modélisation (E2M2)

Spécialité : Approches mathématiques et informatiques du vivant

Cote B.I.U. - Lyon : T 50/210/19 / et bis CLASSE :

RÉSUMÉ : À long terme, le succès évolutif d'une lignée ne dépend pas seulement de la valeur adaptative de ses fondateurs. Il dépend également de la capacité des descendants à transmettre le génotype ancestral sans mutation délétère, tout en découvrant parfois des mutations favorables. Un niveau intermédiaire de variabilité mutationnelle peut donc être, de fait, indirectement sélectionné. En simulant, à l'aide d'un modèle individu-centré, l'évolution de génomes soumis à la fois à des mutations locales et à des réarrangements chromosomiques, nous montrons que la structure du génome est un levier d'ajustement du degré de variabilité : le nombre de gènes et, de façon plus surprenante, la quantité de non codant s'ajustent en fonction du taux de mutation et de l'impact moyen des mutations géniques, maintenant ainsi un niveau constant de variabilité mutationnelle. L'émergence de ces couplages surprenants suggère que les génomes ne sont pas seulement façonnés par les biais mutationnels et les coûts sélectifs directs, mais aussi, à plus long terme, par des pressions plus indirectes.

MOTS-CLÉS : bioinformatique, évolution artificielle, évolution des génomes, mutation, réarrangements, nombre de gènes, ADN non codant, modélisation individu-centrée, simulation.

Laboratoire(s) de recherche : Laboratoire de Biologie Fonctionnelle, Insectes et interactions
UMR INRA/INSA 203 - INSA de Lyon
Bâtiment Louis Pasteur
20, avenue Albert Einstein, 69621 Villeurbanne Cedex, France

Laboratoire PRISMa (Productique et Informatique des Systèmes Manufacturiers)
INSA de Lyon
Bâtiment Blaise Pascal
20, avenue Albert Einstein, 69621 Villeurbanne Cedex, France

Directeurs de thèse : Jean-Michel Fayard, Guillaume Beslon

Président du jury : Marc Schoenauer

Composition du jury :

Guillaume Beslon	Maître de conférences, INSA Lyon, directeur de thèse
Laurent Duret	Directeur de recherche, CNRS, examinateur
Jean-Michel Fayard	Professeur, INSA Lyon, directeur de thèse
Michel Morvan	Professeur, ENS Lyon, rapporteur
Eduardo Rocha	Chargé de recherche HDR, CNRS, rapporteur
Marc Schoenauer	Directeur de recherche, INRIA, président du jury
François Taddéi	Chargé de recherche HDR, INSERM, membre invité

Structuration des génomes par sélection indirecte de la variabilité mutationnelle – Une approche de modélisation et de simulation

Résumé : À long terme, le succès évolutif d'une lignée ne dépend pas seulement de la valeur adaptative de ses fondateurs. Il dépend également de la capacité des descendants à transmettre le génotype ancestral sans mutation délétère, tout en découvrant parfois des mutations favorables. Un niveau intermédiaire de variabilité mutationnelle peut donc être, de fait, indirectement sélectionné. En simulant, à l'aide d'un modèle individu-centré, l'évolution de génomes soumis à la fois à des mutations locales et à des réarrangements chromosomiques, nous montrons que la structure du génome est un levier d'ajustement du degré de variabilité : le nombre de gènes et, de façon plus surprenante, la quantité de non codant s'ajustent en fonction du taux de mutation et de l'impact moyen des mutations géniques, maintenant ainsi un niveau constant de variabilité mutationnelle. L'émergence de ces couplages surprenants suggère que les génomes ne sont pas seulement façonnés par les biais mutationnels et les coûts sélectifs directs, mais aussi, à plus long terme, par des pressions plus indirectes.

Mots-clés : bioinformatique, évolution artificielle, évolution des génomes, mutation, réarrangements, nombre de gènes, ADN non codant, modélisation individu-centrée, simulation.

Evolution of genome structure by indirect selection of the mutational variability – A computational approach

Abstract: In the long term, the evolutionary success of a lineage does not depend only on the fitness of its founders. It also depends on the ability of the descendants to pass on the ancestral genotype without deleterious mutations, while sometimes discovering beneficial mutations. An intermediate level of mutational variability can thus be indirectly selected. Here, by simulating genome evolution under both local mutations and large rearrangements, we show that genome structure is an important component of the variability level: gene number and, more surprisingly, the amount of non coding DNA are adjusted according to the mutation rate and to the deleteriousness of gene mutations. The emergence of such surprising couplings suggests that, aside from mutational biases and direct selective costs, genomes are also shaped by more indirect selective pressures.

Keywords: bioinformatics, artificial evolution, genome evolution, mutation, rearrangements, gene number, non coding DNA, individual-based modelling, simulation.