Origine evolutive de la complexite des systèmes biologiques : Une etude par evolution experimentale in silico

Vincent Liard

N° d'ordre NNT : 2020LYSEI085

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON

opérée au sein de

L'institut national des sciences appliquées de Lyon

École doctorale InfoMaths
ED 512

Informatique et applications

Soutenue publiquement le 27/10/2020, par

Vincent Liard

Origine évolutive de la complexité des systèmes biologiques
Une étude par évolution expérimentale *in silico*

Devant le jury composé de :

Schneider, Dominique	Professeur des Universités	UGA	Rapporteur
Muller, Jean-Pierre	Directeur de Recherche	CIRAD	Rapporteur
Dillmann, Christine	Professeure des Universités	Université Paris Saclay	Examinatrice
Lopez, Philippe	Professeur des Universités	UPMC	Examinateur
Beslon, Guillaume	Professeur des Universités	INSA-LYON	Directeur de thèse
Rouzaud-Cornabas, Jonathan	Maître de conférences	INSA-LYON	Co-encadrant

Département FEDORA – INSA Lyon - Ecoles Doctorales – Quinquennal 2016-2020

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
CHIMIE	**CHIMIE DE LYON** http://www.edchimie-lyon.fr Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage secretariat@edchimie-lyon.fr INSA : R. GOURDON	**M. Stéphane DANIELE** Institut de recherches sur la catalyse et l'environnement de Lyon IRCELYON-UMR 5256 Équipe CDFA 2 Avenue Albert EINSTEIN 69 626 Villeurbanne CEDEX directeur@edchimie-lyon.fr
E.E.A.	**ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE** http://edeea.ec-lyon.fr Sec. : M.C. HAVGOUDOUKIAN ecole-doctorale.eea@ec-lyon.fr	**M. Gérard SCORLETTI** École Centrale de Lyon 36 Avenue Guy DE COLLONGUE 69 134 Écully Tél : 04.72.18.60.97 Fax 04.78.43.37.17 gerard.scorletti@ec-lyon.fr
E2M2	**ÉVOLUTION, ÉCOSYSTÈME, MICROBIOLOGIE, MODÉLISATION** http://e2m2.universite-lyon.fr Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 INSA : H. CHARLES secretariat.e2m2@univ-lyon1.fr	**M. Philippe NORMAND** UMR 5557 Lab. d'Ecologie Microbienne Université Claude Bernard Lyon 1 Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69 622 Villeurbanne CEDEX philippe.normand@univ-lyon1.fr
EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTÉ** http://www.ediss-lyon.fr Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 INSA : M. LAGARDE secretariat.ediss@univ-lyon1.fr	**Mme Sylvie RICARD-BLUM** Institut de Chimie et Biochimie Moléculaires et Supramoléculaires (ICBMS) - UMR 5246 CNRS - Université Lyon 1 Bâtiment Curien - 3ème étage Nord 43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex Tel : +33(0)4 72 44 82 32 sylvie.ricard-blum@univ-lyon1.fr
INFOMATHS	**INFORMATIQUE ET MATHÉMATIQUES** http://edinfomaths.universite-lyon.fr Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 infomaths@univ-lyon1.fr	**M. Hamamache KHEDDOUCI** Bât. Nautibus 43, Boulevard du 11 novembre 1918 69 622 Villeurbanne Cedex France Tel : 04.72.44.83.69 hamamache.kheddouci@univ-lyon1.fr
Matériaux	**MATÉRIAUX DE LYON** http://ed34.universite-lyon.fr Sec. : Stéphanie CAUVIN Tél : 04.72.43.71.70 Bât. Direction ed.materiaux@insa-lyon.fr	**M. Jean-Yves BUFFIÈRE** INSA de Lyon MATEIS - Bât. Saint-Exupéry 7 Avenue Jean CAPELLE 69 621 Villeurbanne CEDEX Tél : 04.72.43.71.70 Fax : 04.72.43.85.28 jean-yves.buffiere@insa-lyon.fr
MEGA	**MÉCANIQUE, ÉNERGÉTIQUE, GÉNIE CIVIL, ACOUSTIQUE** http://edmega.universite-lyon.fr Sec. : Stéphanie CAUVIN Tél : 04.72.43.71.70 Bât. Direction mega@insa-lyon.fr	**M. Jocelyn BONJOUR** INSA de Lyon Laboratoire CETHIL Bâtiment Sadi-Carnot 9, rue de la Physique 69 621 Villeurbanne CEDEX jocelyn.bonjour@insa-lyon.fr
ScSo	**ScSo*** http://ed483.univ-lyon2.fr Sec. : Véronique GUICHARD INSA : J.Y. TOUSSAINT Tél : 04.78.69.72.76 veronique.cervantes@univ-lyon2.fr	**M. Christian MONTES** Université Lyon 2 86 Rue Pasteur 69 365 Lyon CEDEX 07 christian.montes@univ-lyon2.fr

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

Table des matières

Table des figures

Comment remercier assez mes directeurs
pour leur sagacité, leur patience et leur soutien sans faille
au cours de ce long chemin d'élaboration commune ?

Chapitre I

De la complexité des systèmes biologiques

1 Introduction

Où que se porte notre regard, tous les organismes sur lesquels il s'arrête nous frappent par leur « complexité ». Pour peu que nous y songions quelques instants, la difficulté de quantifier cette complexité nous laisse perplexe. La question même de la définir pose des problèmes qui s'avèrent d'autant plus insurmontables qu'on en énumère les divers aspects (taille des systèmes, intrication des composants, degrés de liberté, etc). Hormis des controverses de nature essentiellement religieuse, une chose cependant est établie, c'est que la complexité des systèmes biologiques, comme tout autre trait trouve sa source dans l'évolution des espèces au sens où Darwin l'entendait déjà (Darwin, 1859). *Nothing in biology makes sense except in the light of evolution* titrait Theodosius Dobzhansky en 1973 [1] (Dobzhansky, 1973). C'est ce point de vue que nous adopterons ici : partir de la théorie de l'évolution pour expliquer la complexité des systèmes vivants. En effet, au delà du constat qu'elle prend sa source dans l'évolution, la question de la « causalité évolutive » de la complexité reste controversée, et ce bien que plusieurs communautés scientifiques s'en soient emparées : paléontologie, génétique des populations, biologie computationnelle, vie artificielle ou encore physique statistique... La théorie de l'évolution, malgré la relative simplicité de sa formulation initiale, s'est en effet développée en de nombreuses circonvolutions qui ont ouvert un large éventail de possibilités pour expliquer comment l'évolution a pu conduire à tel ou tel trait spécifique et, pour ce qui nous concerne, à la complexité. En d'autres termes, il n'y a pas aujourd'hui de consensus pour expliquer la ou les « causes ultimes » (Mayr, 1961) qui rendraient compte de la complexité biologique.

L'objectif de la présente thèse est de développer des modèles qui permettent d'identifier les causes ultimes de la complexité biologique. Pour cela, nous utiliserons une plateforme multi-échelle permettant de simuler l'évolution dans des situations réclamant des degrés divers de complexité de la part des organismes. Avant cela, cependant, il apparaît nécessaire de mieux comprendre ce que recouvre la complexité d'un organisme biologique et quelles sont les différentes hypothèses qui ont pu être proposées pour expliquer son

1. Dans un article dont la postérité aura retenu le titre plus que le contenu puisqu'il s'agit d'une charge contre le créationnisme.

accumulation au cours de l'évolution. Dans la suite de ce chapitre, nous effectuerons ainsi une revue de la littérature du domaine, en abordant une stratégie convergente : partant des définitions générales de la complexité, nous convergerons ensuite vers une définition de la complexité qui soit plus spécifique aux systèmes vivants. Nous aborderons alors les liens qui unissent complexité et évolution en détaillant particulièrement les principales hypothèses causales proposées dans la littérature.

2 Qu'est-ce que la complexité ?

Plusieurs raisons contribuent aux controverses que suscite la question de la complexité biologique et de son évolution (Miconi, 2008). Deux d'entre elles sont essentielles : tout d'abord, l'absence d'une mesure (et même d'une définition) de la « complexité » qui soit universellement acceptée, ensuite le fait que les organismes biologiques sont foncièrement structurés selon des échelles différentes et peuvent donc voir leur complexité croître et décroître simultanément selon les niveaux d'organisation considérés. Avant d'entrer dans la description des différentes hypothèses proposées dans la littérature pour expliquer son origine évolutive, il est donc fondamental de clarifier ce que nous entendons ici par « complexité » et par « complexité biologique ».

2.1 Un foisonnement de définitions et de mesures

Régulièrement au cours de l'histoire des sciences, la complexité, les sciences de la complexité et la science des systèmes complexes reviennent sur le devant de la scène, parfois sous ces appellations, parfois sous des appellations différentes, comme ce fut le cas pour la « première cybernétique » (Wiener, 1948; Weaver, 1948), la « seconde cybernétique » (Simon, 1962) ou la « systémique » au tournant des années 1960-1970 (Von Bertalanffy *et al.*, 1973) – voir (Benkirane *et al.*, 2002) pour une description de ces différents courants. Pourtant, malgré cet intérêt de longue date, force est de constater que le concept de « complexité » est resté assez insaisissable et qu'aucune définition ne s'est imposée. Les principales raisons de cette difficulté sont probablement, d'une part le caractère transdisciplinaire de ce concept qui conduit de nombreuses disciplines à proposer des définitions, celles-ci peinant souvent à sortir du giron de leurs disciplines d'origine, et d'autre part la confusion fréquente entre définition et mesure. Dans la suite de cette section, nous allons successivement aborder ces deux points, sachant qu'il n'est pas ici question de proposer une vision exhaustive (cette thèse entière n'y suffirait pas) mais simplement de fixer le cadre général qui nous permettra ensuite de nous concentrer sur la complexité biologique et l'étude de son évolution.

2.1.1 Définitions qualitatives

Depuis la fin des années 1990, la notion de complexité, omniprésente dans les discours, trouve parfois son chemin jusque dans les actes. Elle a donné lieu à un foisonnement d'ouvrages plus ou moins grand public, de numéros spéciaux de revues de vulgarisation (en vingt ans, le magazine scientifique mensuel « La Recherche » a ainsi publié des dossiers sur la complexité en novembre-décembre 2002, février 2007, mai 2012 et juillet-août 2018) et la multiplication d'instituts plus ou moins officiels (dont le plus célèbre d'entre eux – et précurseur – le Santa Fe Institute) et de sites Internet... avec dans tous les cas ou presque, des définitions plus ou moins précises et plus ou moins directes de l'objet de leur attention : la complexité. Nous n'allons pas ici nous lancer dans une exégèse de ces définitions qui utilisent, pour la plupart, un vocabulaire différent pour, finalement exprimer des idées très similaires. Pour illustrer, nous allons en revanche nous concentrer ici sur l'analyse de l'une d'entre elles : la définition proposée par Alain Barrat du Centre de Physique Théorique de Marseille dans l'introduction au numéro spécial « Chaos et Systèmes Complexes » de

Juillet-Aout 2018 de *La Recherche* (Pajot, 2018) et reprise par Wikipedia [1]. Selon Alain Barrat, un système complexe est :

> « un système composé d'un grand nombre d'éléments interagissant sans coordination centrale, sans plan établi par un architecte, et menant spontanément à l'émergence de "structures complexes", c'est-à-dire de structures stables avec des motifs présentant plusieurs échelles spatiales et temporelles ».

On trouve dans cette définition les trois éléments classiques des définitions qualitatives de la complexité :

Un grand nombre d'éléments La présence d'un « grand » nombre d'éléments en interactions est une caractéristique centrale de toutes les définitions de systèmes complexes. On notera cependant que la notion de « grand nombre » est totalement subjective et qu'elle n'est jamais objectivée dans les définitions. En outre, la présence même d'un tout et d'éléments constituants est aussi une propriété subjective puisqu'elle dépend d'un choix d'observation et de description (ainsi, la cellule biologique est souvent présentée comme « l'unité élémentaire du vivant » alors qu'elle est à l'évidence elle-même composée d'un grand nombre d'éléments moléculaires).

Interactions locales Définie ici en négatif par « l'absence de coordination centrale [ou de plan établi] par un architecte », la présence exclusive d'interactions locales est utilisée par opposition à la troisième composante (le comportement global non-trivial, cf. infra). Il s'agit d'insister sur le fait que le comportement global est « émergent », c'est-à-dire produit par une dynamique qui se fonde au niveau local.

Un comportement global non-trivial Toutes les définitions qualitatives de système complexe ou de complexité utilisent d'une façon ou d'une autre une périphrase pour exprimer les caractéristiques non-triviales du comportement du système. Ici ces caractéristiques sont considérées comme la présence de « structures stables avec des motifs présentant plusieurs échelles spatiales et temporelles » mais de nombreuses définitions reposent sur des caractérisations plus implicites (le complexe étant ainsi souvent défini par opposition au « simple » [2]), voire tautologiques. Ainsi, Melanie Mitchell, du Santa Fe Institute, définit un système complexe comme « *a system in which large networks of components with no central control and simple rules of operation give rise to a complex collective behavior [...]* » (Mitchell, 2009).

Ce qui ressort fortement ici est que ces trois éléments, présents dans la très grande majorité des définitions qualitatives, sont tous, d'une façon ou d'une autre, totalement subjectifs, qu'il s'agisse du dénombrement des éléments, du caractère « local » des interactions ou du jugement relatif à la « complexité » (sic) du comportement global. Ce caractère subjectif est d'ailleurs reconnu par certains auteurs, tels Whitesides et Ismagilov (1999) qui, après avoir proposé une définition de la complexité en chimie (« *A complex system is one whose evolution is very sensitive to initial conditions or to small perturbations, one in which the number of independent interacting components is large, or one in which there are multiple pathways by which the system can evolve. Analytical descriptions of such systems typically require nonlinear differential equations.* ») ajoutent que :

1. `https://fr.wikipedia.org/wiki/Complexité`
2. Dans Wikipédia toujours, mais dans la version anglaise, on peut lire : « *"Systems exhibit complexity" means that their behaviors cannot be easily inferred from their properties* ».

« *the system is "complicated" by some subjective judgment and is not amenable to exact description, analytical or otherwise* ». En poussant la critique à l'extrême, on pourrait même objecter à ce type de définitions leur forme d'oxymores puisqu'elles impliquent qu'un même système, qui aurait la propriété de complexité à un niveau donné – micro, la perde au niveau macro puisque celui-ci, par définition, n'est plus « en grand nombre ». Cette critique, qui confine certes à l'absurde, montre cependant combien les définitions qualitatives de la complexité sont *in fine* dépendantes de l'observateur et donc du niveau d'observation. Partant de ce constat, certains auteurs ont ainsi cherché à proposer des définitions assumant pleinement le caractère subjectif de la complexité. On peut ainsi citer Bertin *et al.* (2011) qui proposent d'intégrer explicitement la notion d'observation (donc de point de vue) dans la définition : « Nous appellerons "complexe" une approche qui vise à comprendre comment la dynamique d'interaction entre des entités *micro* parvient à créer une unité à un autre niveau d'observation *macro* ». Cependant, au milieu de la profusion de définitions, celles-ci ne semblent pas s'imposer, pas plus qu'aucune autre au demeurant.

Le caractère subjectif des définitions de la complexité pose problème puisqu'il implique que tout système soit susceptible d'être considéré comme complexe [1] ou non *selon le point de vue*, c'est-à-dire à une certaine échelle, pour certaines propriétés, pour certains régimes de fonctionnement, à une certaine époque, relativement à certaines connaissances scientifiques, etc. On peut ainsi citer certains cas extrêmes, telle une pierre, « système » auquel on serait tenté de refuser le qualificatif de complexe mais qui le devient, *d'un certain point de vue*, par exemple dès lors qu'on s'intéresse, comme la *dynamique des dislocations*, au réseau de micro-fissures qui la parcourt (Desbenoit *et al.*, 2005) (ou, bien évidemment, à une échelle encore inférieure, dès lors qu'on s'intéresse explicitement aux interactions entre les atomes qui la composent).

Peut-on alors tirer quelque enseignement des définitions qualitatives de la complexité pour notre propos, à savoir l'évolution de la complexité. Pour répondre à cette question, il suffit de revenir aux trois propriétés évoquées ci-dessus et les appliquer aux systèmes biologiques. (*i.*) Pour ce qui est du grand nombre d'éléments, même si la notion de « grand nombre » est subjective quantitativement (et même qualitativement puisque le nombre incriminé dépend lui-même du niveau de description auquel on s'intéresse, cf. *infra*), tout système biologique, qu'on le considère à l'échelle moléculaire, cellulaire, physiologique ou écologique, peut être raisonnablement considéré comme constitué d'un grand nombre – et même d'un très grand nombre – d'éléments. (Le nombre de cellules dans un corps humain est de l'ordre de 10^{14}.) (*ii.*) L'absence de centralisation peut poser quelques questions pour certains systèmes biologiques (par exemple au niveau physiologique où certains organes peuvent avoir un rôle centralisateur) mais là encore, il est relativement consensuel de considérer les systèmes biologiques comme des systèmes distribués. Enfin, (*iii.*), le comportement global non-trivial, ou l'« apparence d'unité » pour reprendre la terminologie de Bertin *et al.* (2011), fait-il aussi consensus : il est indubitable que le comportement d'une cellule est non-trivial comparé aux propriétés des éléments moléculaires qui la composent et, comme nous l'avons déjà évoqué plus haut, la notion même de cellule s'est imposée comme « l'unité élémentaire du vivant » comme l'énoncent doctement les cours de Science

1. On parle ici de systèmes *réels* (au sens ils sont composés d'entités physiques). Cette affirmation ne s'applique bien évidemment pas aux systèmes formels ou simulés.

de la Vie et de la Terre (SVT).

Il ressort de cette énumération qu'on peut légitimement considérer *tous* les systèmes biologiques comme des systèmes complexes. S'agissant du questionnement central de cette thèse, ce constat nous renseigne peu : si tous les systèmes biologiques sont complexes c'est, certes, qu'ils ont acquis cette propriété au cours de leur évolution [1] mais cela implique aussi qu'une définition plus quantitative – une mesure – va nous être nécessaire pour raisonner sur la dynamique évolutive de la complexité.

2.1.2 Mesures

Plusieurs mesures de complexité ont été proposées dans différents champs disciplinaires mais les seules à s'être imposées – peut-être parce que les plus explicites – émanent des sciences de l'information [2] (Delahaye *et al.*, 1994). Dès 1963, Andrei Kolmogorov a ainsi proposé que la complexité d'un système soit mesurée par la longueur du plus petit programme capable de générer ce système (Kolmogorov, 1963). Cette définition a le mérite de la simplicité et de l'élégance mais elle pose un problème central, assez similaire à celui de la mesure d'information de Claude Shannon (1948) : dans les deux cas, en effet, la mesure est syntaxique et non sémantique. En d'autres termes, ces mesures ne considèrent pas le contenu informationnel (la sémantique) du signal analysé mais seulement la forme de ce signal (sa syntaxe). La conséquence directe est que la mesure de Kolmogorov, comme l'entropie de Shannon, est maximale pour un système (un signal) totalement aléatoire. Plusieurs auteurs ont ainsi proposé des variantes de la complexité de Kolmogorov afin d'introduire une forme de sémantique dans la mesure. On peut ainsi citer la profondeur logique de Charles Bennett (1995), basée sur la même idée que la complexité de Kolmogorov, mais qui considère non plus la longueur du plus petit programme générant le système mais le temps d'exécution de ce programme. Cependant, même ainsi corrigées, les mesures de complexité basée sur la théorie algorithmique de l'information ne recouvrent que partiellement les contours des systèmes complexes au sens des définitions qualitatives présentées ci-dessus. Même si nous avons souligné les limites de ces définitions, il est clair que celles-ci ont un sens *pour la biologie* que n'ont pas la complexité de Kolmogorov ou la profondeur logique [3]. À partir de la formulation générale des définitions qualitatives, on peut proposer des mesures plus en phase avec celles-ci tel Heylighen (2007) qui, partant d'une formulation qualitative [4], ajoute que « *To make this qualitative notion more quantitative, I add that a system becomes more complex as the number of distinctions (distinct components, states, or aspects) and the number of relations or connections increases* »,

1. Il est pour cela nécessaire de supposer que ces mêmes systèmes *n'étaient pas complexes* à l'origine. Cette supposition peut être considérée comme légitime mais elle pose néanmoins plusieurs problèmes. D'une part, la notion même d'origine des systèmes biologiques est très mal délimitée ; d'autre part, comme nous l'avons souligné plus haut, tout système réel peut être considéré comme complexe *d'un certain point de vue*. Ainsi, une molécule autoréplicatrice pourrait-elle être considérée comme « simple » du point de vue de la biologie mais elle serait probablement vue comme complexe du point de vue de la physique.

2. Nous n'aborderons pas ici la question de la mesure de la complexité algorithmique qui, malgré l'homonymie de l'appellation, recouvre un ensemble de concepts totalement différents de la complexité telle qu'elle est envisagée ici.

3. Sans compter que ces mesures, très théoriques, ne sont pas calculables en pratique (Shen, 2000).

4. « *to have a complex, you need two or more distinct components that are connected in such a way that they are difficult to separate* » (Heylighen, 2007).

mais pour souligner immédiatement que ces mesures sont elles-mêmes subjectives puisque dépendantes du point de vue (pour reprendre les termes utilisés ci-dessus).

La question de la mesure de complexité se heurte en fait a un double écueil : soit la mesure est très formalisée mais elle s'écarte alors du contenu sémantique des définitions qualitatives, soit la mesure est proche de ces dernières mais elle devient alors subjective, ce qui l'invalide en tant que mesure même ! Adami (1998) résume ce dilemme en reprochant aux mesures formelles leur absence de *contexte*. D'une façon générale, il ressort de la discussion qui précède que les définitions, qualitatives comme quantitatives, de la « complexité » ne peuvent s'extraire d'un domaine d'application spécifique et que ce n'est que relativement à un *point de vue*, donc à une discipline, à une application ou à un questionnement particulier. C'est pourquoi nous n'allons pas ici nous appesantir sur la question de la complexité « en général » mais nous focaliser sur la complexité en biologie, voire, en évolution.

2.2 La complexité en biologie

Nous l'avons déjà évoqué ci-dessus, il est généralement admis que tous les systèmes biologiques sont des systèmes complexes au sens des définitions générales. Dès lors, deux questions se posent : la question de l'origine de cette complexité et la question de la fonction de cette complexité. Ces deux questions sont évidemment intimement liées, et relèvent toutes deux de l'expression de la causalité biologique telle qu'énoncée par Mayr (1961) : la première pose la question des causes ultimes de la complexité tandis que la seconde pose la question de ses causes proximales, c'est-à-dire de sa réalisation ici et maintenant. Dans les deux cas, pour que le débat sur les causes de la complexité soit fructueux, il importe de s'accorder sur le sens de l'expression « complexité des systèmes biologiques » ainsi que sur les éventuels outils de mesure. Or, si le consensus sur la complexité des systèmes biologiques est bien établi, il résiste assez mal à la clarification du débat ou à la recherche d'une « échelle de complexité » entre les organismes, échelle qui ne serait pas sans rappeler la *Scala naturæ* et ses dimensions téléologique et anthropocentrique, voire religieuse (figure I.1).

L'une des causes principales de la difficulté qu'il peut y avoir à parler « complexité » en biologie est la nature à la fois complexe (ce qui, dans la définition, implique une forme de chaos local) et organisée de la biologie. Ce mélange hétérogène (qu'on me pardonne ici ce jeu de langage) a été très bien décrit dès 1948 par Warren Weaver (1948) dans son article *Science and Complexity*. Il y distingue en effet les problèmes de complexité désorganisée (*disorganized complexity*), qui relèvent en particulier de la physique statistique (Weaver (1948) cite Josiah Willard Gibbs comme l'un des pères de l'étude de la complexité désorganisée), des problèmes de complexité organisée (*organized complexity*) qui mêlent inextricablement les questions de complexité (toujours au sens de la physique statistique) et les questions d'organisation. Or, lorsque, dans son article, Warren Weaver (1948) donne une liste de questions classiques du champ de la complexité organisée, on ne peut qu'être frappé par le fait qu'elles relèvent toutes directement des sciences du vivant [1] : « *What makes an evening primrose open when it does ? Why does salt water fail*

1. Warren Weaver (1948) ne limite cependant pas la complexité organisée aux sciences du vivant. Il cite ainsi, plus loin dans l'article, plusieurs questions relatives aux sciences économiques et sociales.

Figure I.1 – La *Scala naturæ* selon Didacus Valades, *Rhetorica Christiana* (1579). Les êtres vivants sont ordonnés selon une échelle établie sur la base de leur degré de perfection ou de complexité qui culmine avec les anges et Dieu, témoignant ainsi de la place que revendique l'homme au sommet du monde visible, juste en dessous du divin.

to satisfy thirst ? Why can one particular genetic strain of microorganism synthesize within its minute body certain organic compounds that another strain of the same organism cannot manufacture ? Why is one chemical substance a poison when another, whose molecules have just the same atoms but assembled into a mirror-image pattern, is completely harmless ? Why does the amount of manganese in the diet affect the maternal instinct of an animal ? What is the description of aging in biochemical terms ? What meaning is to be assigned to the question : Is a virus a living organism ? What is a gene, and how does the original genetic constitution of a living organism express itself in the developed characteristics of the adult ? Do complex protein molecules "know how" to reduplicate their pattern, and is this an essential clue to the problem of reproduction of living creatures ? » (Weaver, 1948).

La nature à la fois complexe et organisée de la biologie a plusieurs conséquences qui rendent la discussion difficile. Tout d'abord, contrairement aux systèmes physiques il n'existe pas de système biologique isolé ni même isolable [1]. Ensuite, la complexité peut se penser à la fois en termes de relation entre niveaux d'organisation et en termes de

1. En pratique, il n'existe pas non plus de système physique totalement isolé. En revanche, de nombreux

structure de chacun de ces niveaux. Enfin, la complexité peut évoluer plus ou moins indépendamment sur chacun de ces niveaux, rendant une classification des organismes par ordre de complexité pour le moins hasardeuse. Sur ce dernier point, un exemple riche d'enseignements est donné par l'évolution des endosymbiontes comme fruit de l'association d'un eucaryote et d'une bactérie. De telles associations, fréquentes chez les insectes, se traduisent, pour les insectes, par l'évolution d'un nouveau type cellulaire spécialisé (le bactériocyte (Moran, 2007)) et de signatures évolutives spécifiques, incluant la perte de certaines voies de synthèse des acides aminés (Wilson et Duncan, 2015). Pour les bactéries en revanche l'association se traduit par une réduction massive de leur matériel génétique. À titre d'exemple, la bactérie endosymbiotique *Buchnera aphidicola*, symbiote obligatoire du puceron du pois a perdu au cours de cette association près de 80% du matériel génétique par rapport à son ancêtre (Moran et Mira, 2001). Enfin, globalement, cette association se traduit par l'apparition d'un nouveau niveau d'organisation, l'holobionte, caractérisé par l'association obligatoire de deux organismes. Comment dès lors qualifier une telle dynamique en termes de complexité : s'agit-il d'une augmentation de la complexité ? Apparition d'un nouveau niveau d'organisation, apparition d'un nouveau type cellulaire... Ou alors s'agit-il d'une perte de complexité ? Réduction massive du génome de la bactérie, perte de gènes essentiels de l'hôte... On le voit, le caractère organisé et multi-niveaux des systèmes biologique rend l'appréciation des variations de complexité difficile, d'autant plus que ce type d'événement n'est pas rare à toutes les échelles du vivant. Qu'on songe par exemple à l'association d'une α-protéobactérie avec une cellule eucaryote primitive qui a donné naissance à la cellule eucaryote « moderne » ou au développement de super-organismes chez de nombreux insectes sociaux. Dans tous ces exemples, l'apparition d'un nouveau niveau d'organisation s'est traduit par une perte de complexité à un autre niveau (réduction drastique du génome chez la mitochondrie, perte des capacités reproductive chez la majeure partie des membres d'une colonie, etc). Même en restreignant l'analyse aux organismes unicellulaires, rien ne nous permet de supposer que les variations de complexité se produisent de façon homogène, simultanée, aux niveaux du génome, du transcriptome, du protéome ou du phénotype. De fait, plusieurs paradoxes fameux illustrent le décalage qui peut se produire entre la quantité d'information encodée par le génome et celle que traduit le phénotype. On pense bien entendu ici au *C-Value paradox* (Thomas, 1971) et au *G-Value paradox* (Hahn *et al.*, 2002).

Afin de distinguer la complexité telle qu'elle se manifeste à travers les différents niveaux d'organisation et la complexité affichée sur tel ou tel niveau d'organisation, Daniel W. McShea a proposé de distinguer la *complexité horizontale* de la *complexité verticale* (McShea, 2017). Dans cette acception, la complexité verticale correspondrait au nombre de niveau d'organisation présents dans un système biologique, tandis que la complexité horizontale caractériserait le nombre d'éléments présent à un niveau donné. Cette distinction entre complexité verticale et complexité horizontale est bien connue dans le monde des organisations qui y ajoute d'ailleurs une notion de *complexité spatiale* (Mileti *et al.*, 1977) mais, si elle a le mérite de poser le problème, elle ne permet pas vraiment de le résoudre puisque, là encore, la définition des niveaux est éminemment subjective (Banzhaf *et al.*, 2016), tout comme la définition des éléments ou leur dénombrement (la question se

systèmes physiques sont virtuellement *isolables* au sens ou les supposer isolés n'impacte pas notre capacité à en appréhender la dynamique. Ce n'est pas le cas des systèmes biologiques qui sont sujets à des flux entrants et sortants permanents.

pose, par exemple, de considérer le nombre d'éléments ou le nombre d'éléments différents). En outre, comme nous l'avons déjà souligné à plusieurs reprise, la complexité organisée interdit de considérer une mesure unique de la complexité horizontale : les différents niveaux d'organisation d'un système biologique étant à la fois liés (par exemple par les mécanismes de décodage du génome) et indépendants (parce que ces mêmes mécanismes sont redondants et dégénérés), la complexité horizontale peut varier considérablement d'un niveau à l'autre. Dès lors, la complexité d'un système biologique ne peut plus être considérée comme une propriété atomique mais doit au contraire se comprendre comme une propriété plurielle, soumise à des choix d'observation. Nous verrons ci-dessous que ce constat aura des répercutions importantes dès lors qu'il s'agira de quantifier la complexité d'un systèmes biologique.

Malgré ces difficultés, il est possible, dans certains contextes bien délimités, de donner des définitions relativement strictes de la complexité biologique. Ainsi, Maynard-Smith et Szathmary (1997) distinguent-ils plusieurs transitions majeures dans l'histoire évolutive du vivant, ces transitions pouvant être assimilées à un accroissement de la complexité verticale au cours du temps, la notion de transition correspondant alors à l'ajout d'un niveau d'organisation. La particularité de la définition de Maynard-Smith et Szathmary (1997) est de ne pas considérer le nombre de niveaux d'organisation (critère subjectif comme nous l'avons dit) mais de considérer d'un point de vue historique le niveau d'organisation doué de la propriété de reproduction. Les auteurs utilisent donc un critère relativement objectif : le support de l'information et sa transmission de génération en génération. En effet, constatant que l'apparition d'un nouveau mécanisme de réplication s'accompagne de la domestication des mécanismes de réplication aux niveaux précédents (ceux-ci ne pouvant plus s'exprimer hors du contexte de la réplication du niveau « supérieur »), il est possible de proposer une notion de complexité ordonnée chronologiquement (sinon quantifiée), selon l'ordre historique d'apparition des nouveaux niveaux de réplication (Szathmáry, 2015). Maynard-Smith et Szathmary (1997) distinguent ainsi les huit transitions majeures suivantes :

1. Molécules auto-réplicantes → Populations de molécules
2. Réplicateurs indépendants → Chromosomes
3. ARN → ADN
4. Procaryotes → Eucaryotes
5. Reproduction clonale asexuée → Reproduction sexuée
6. Unicellulaires (protistes) → Multicellulaires (animaux, plantes, champignons)
7. Individus solitaires → Colonies
8. Sociétés primitives → Sociétés humaines, langage

Ces transitions majeures sont relativement bien caractérisées par les propriétés de réplication et correspondent à des étapes cruciales de l'accroissement de la complexité au cours de l'évolution [1] (Szathmáry, 2015) . Cependant, si elles permettent d'ordonner la complexité à grands traits, elles ne permettent pas de quantifier la complexité au sein d'une même classe d'organismes puisque ceux-ci, par définition, sont équivalents. Si nous

1. Szathmáry (2015) a plus récemment proposé une révision de cette liste mais les principes généraux restent similaires.

voulons étudier plus finement l'évolution de la complexité, il va donc falloir que nous disposions d'une mesure de la complexité biologique.

2.3 Mesurer la complexité biologique

Nous avons vu (section 2.1.2) les difficultés qu'il peut y avoir à quantifier la complexité en général. Eu égard à la difficulté que présente déjà le seul fait de proposer une définition à la complexité en biologie, il semble évident qu'en outre lui donner une mesure serait une gageure. De fait, si l'on excepte, d'une part, les mesures théoriques très générales (*p. ex.* la mesure de complexité de Kolmogorov) dont nous avons vu qu'elles ne s'appliquent que difficilement à des systèmes réels, la plupart des études se concentrent sur des comptages d'éléments à différents niveaux d'organisation d'un système biologique. On peut ainsi citer bien évidemment la taille du génome, qui conduit, on le sait, au C-Value paradox (Thomas, 1971; Petrov, 2001), ou le nombre de gènes, qui a donné le G-Value paradox (Hahn *et al.*, 2002), mais aussi le nombre d'éléments taxonomiques (Schopf *et al.*, 1975), la morphologie (McShea, 1993; Wagner, 1996; Saunders *et al.*, 1999) ou le nombre de types cellulaires (Valentine, 2000; Finlay et Esteban, 2009). Dans tous les cas, ces mesures ont permis d'identifier des tendances à l'accumulation de complexité mais, parce qu'elles sont souvent spécifiques à un clade donné, elles ne permettent pas de généraliser les observations à l'ensemble du règne vivant. De fait, ces différentes mesures, relativement triviales dans leur principe et dont nous avons vu qu'elles comportent une importante part de subjectivité, peinent à proposer un cadre conceptuel suffisamment universel pour pouvoir s'accorder sur une mesure objective de la complexité.

Un tel cadre conceptuel a été proposé par Chris Adami (2002b) pour mesurer la complexité horizontale (bien que cette distinction ne soit pas évoquée dans l'étude). Pour ce faire, l'auteur propose de suivre le raisonnement suivant. Sachant que tout système biologique est un système en évolution, il propose de considérer l'évolution comme un mécanisme conduisant à un flux d'information de l'environnement vers l'organisme. De ce fait, selon Adami (2002b) plus un organisme aura intégré d'information de son environnement vers un ou plusieurs niveaux d'organisation, plus il pourra être considéré comme complexe. Il est alors possible de définir une ou plusieurs mesures de complexité biologique à condition de pouvoir estimer la quantité d'information (ou l'entropie) à un ou à plusieurs niveaux d'organisation du système biologique.

En partant de ce principe, Chris Adami (2002a) définit une mesure de complexité pour une séquence biologique (*p. ex.* un génome), étant entendu que cette séquence est le support de l'information évolutive au sein d'une population de N séquences de taille L. En l'absence de sélection, l'entropie H d'une séquence est simplement égale à la longueur de cette séquence : $H_{\max}(X) = L$. En effet, le nombre de séquences possibles est D^L (où D est la taille de l'alphabet $\mathcal{A}$ utilisé pour construire la séquence) et l'entropie est calculée comme la somme des logarithmes (en base D) des probabilités d'occurrence des différentes séquences. En présence de sélection, l'entropie H va être inférieure à H_{max} et, si on dispose d'une population de séquences alignées, alors il sera possible de calculer l'entropie en chaque site de la séquence. En effet, en chaque site j de la séquence, l'entropie peut être estimée par :

$$H(j) = -\sum_{i \in \mathcal{A}} p_i(j) \log p_i(j) \tag{I.1}$$

où $p_i(j)$ est la probabilité (estimée à partir des alignements) de trouver le symbole i à la position j de la séquence. À partir des entropies locales, on peut approximer l'entropie globale de la séquence en faisant l'hypothèse réaliste de leur indépendance statistique :

$$H(X) \approx \sum_{j=1}^{L} H(j) \tag{I.2}$$

La complexité de la séquence est alors donnée par la différence entre l'entropie maximale de la séquence et son entropie globale, telle qu'approximée par l'équation I.2 (c'est-à-dire par la quantité d'information stockée dans la séquence) :

$$C_1(H) = L - H(X) \tag{I.3}$$

Malheureusement, la mesure de C_1 est difficile à obtenir sur une population de séquence réelle car elle suppose (*i.*) de disposer d'un échantillon suffisant de séquences dans la population et (*ii.*) d'être en mesure d'aligner toutes ces séquences de façon à pouvoir calculer les entropies locales (équation I.2). De ce fait, cette mesure de complexité a été principalement appliquée à des simulations menées avec la plateforme Avida (Adami, 2002b,a).

Dans un article récent, Moya *et al.* (2020) ont adopté une approche quantitative de la complexité des séquences en développant une série de mesures permettant de quantifier l'information contenue dans la séquence sans nécessiter de calculer les entropies locales (équation I.2 ci-dessus), donc sans nécessiter un échantillonnage de la population, et sans référer à la fonction biologique de la séquence (ce qui, selon les auteurs, permet une estimation de la complexité « pure » (McShea et Brandon, 2010) sans que cette affirmation soit étayée dans l'article – les auteurs présentent en effet ce choix comme une conjecture). Techniquement, Moya *et al.* (2020) analysent les séquences de 91 espèces de cyanobactéries et estiment leur complexité au moyen des six mesures différentes suivantes :

La Complexité de Composition de la Séquence (SCC) estime le nombre et l'entropie des domaines génomiques à partir de la composition en bases.

La Complexité de Composition binaire S(C,G) vs. W(A,T) (SCC_{SW}), comme la SCC la complexité de composition binaire estime le nombre et l'entropie des domaines génomiques mais la partition en domaines est ici déterminée à partir de la composition en paires CG vs. AT.

La Complexité de Composition binaire R(A,G) vs. Y(T,C) (SCC_{RY}). Le principe est le même que pour SCC_{SW} si ce n'est que la composition considérée est ici celle des paires AG vs. TC.

La Complexité de Composition binaire K(A,C) vs. M(T,G) (SCC_{KM}), identique à SCC_{SW} et SCC_{RY} mais qui considère cette fois les paires AC vs. TG.

La Signature Génomique (GS), qui mesure la fréquence des *k-mer* dans le génome et quantifie l'écart entre la fréquence observée et la fréquence attendue sous l'hypothèse d'une distribution aléatoire uniforme. GS correspond à l'écart maximum observé pour toutes les tailles de *k-mer*.

Le « Biobit » (BB), qui est une mesure proposée par Bonnici et Manca (2016) pour mesurer la complexité d'une séquence génomique. Le principe de la mesure Biobit est de maximiser la complexité lorsque l'entropie (donc la distance avec une séquence aléatoire) et l'« anti-entropie » (la distance avec une séquence parfaitement régulière) s'équilibrent. Ces deux valeurs (entropie et anti-entropie) sont estimées sur la base de la composition en *k-mer* de la séquence.

Ces mesures sont par ailleurs comparées à trois mesures plus classiques : la taille des génomes, le taux de GC et le nombre de gènes. On notera que, bien que les auteurs exploitent six métriques différentes, celles-ci dépendent fortement les unes des autres puisque quatre d'entre elles sont basées sur un algorithme de partitionnement de la séquence en domaines et deux sont basées sur la décomposition en *k-mer*. Les auteurs comparent ensuite ces métriques sur les 91 espèces de cyanobactéries et concluent, d'une part, que les mesures de complexité diffèrent nettement des mesures génomiques classiques, d'autre part que quatre de ces mesures (SCC, SCC_{RY}, SCC_{SW} et GS) augmentent avec la longueur des branches séparant les clades de la racine (les deux autres mesures, SCC_{KM} et BB montrant respectivement l'absence de signal et une corrélation négative) tandis que les paramètres génomiques classiques, eux ne présentent aucune tendance. On notera d'ailleurs sur ce point que les cyanobactéries comptent deux clades ayant connu des épisodes de *streamlining* (Batut *et al.*, 2014). Les auteurs concluent que leurs métriques montrent que l'accroissement de la complexité est dû à l'action de la sélection naturelle et non à l'action de la dérive génétique. Ce point est cependant sujet à caution, en particulier parce que les métriques utilisées sont probablement sensibles à la proportion codante dans les génomes ainsi qu'aux taux de mutation, paramètres qui, tous deux, ont fortement varié au sein des cyanobactéries (Batut *et al.*, 2014).

3 Origines évolutives de la complexité

3.1 Introduction

La question de l'origine évolutive de la complexité des systèmes biologiques est aussi vieille que la biologie évolutive elle-même. Ainsi, de façon clairvoyante, Charles Darwin évoque-t-il cette question dès la rédaction de l'« origine des espèce », et ce à quatre reprises, abordant ainsi les différentes facettes du problème (Darwin, 1859) :

— *« Hence, if certain insectivorous birds (whose numbers are probably regulated by hawks or beasts of prey) were to increase in Paraguay, the flies would decrease—then cattle and horses would become feral, and this would certainly greatly alter (as indeed I have observed in parts of South America) the vegetation : this again would largely affect the insects ; and this, as we just have seen in Staffordshire, the insectivorous birds, and so onwards in ever-increasing circles of complexity »* (chapitre III, page 79). La complexité est ici considérée sous un angle phénoménologique, elle est directement liée aux relations trophiques.

— *Slow though the process of selection may be, if feeble man can do much by his powers of artificial selection, I can see no limit to the amount of change, to the beauty and infinite complexity of the coadaptations between all organic beings, one with another and with their physical conditions of life, which may be effected in the long course of time by nature's power of selection* (chapitre IV, page 109). La complexité est ici vue dans une perspective évolutive, potentiellement infinie.

— *If during the long course of ages and under varying conditions of life, organic beings vary at all in the several parts of their organisation, and I think this cannot be disputed ; if there be, owing to the high geometrical powers of increase of each species, at some age, season, or year, a severe struggle for life, and this certainly cannot be disputed ; then, considering the infinite complexity of the relations of all organic beings to each other and to their conditions of existence, causing an infinite diversity in structure, constitution, and habits, to be advantageous to them, I think it would be a most extraordinary fact if no variation ever had occurred useful to each being's own welfare, in the same way as so many variations have occurred useful to man* (chapitre IV, page 127). Même si le propos n'est pas totalement explicite, Darwin pose ici clairement la complexité de l'environnement (via les relations inter-spécifiques) comme moteur de l'évolution.

— *Although the belief that an organ so perfect as the eye could have been formed by natural selection, is more than enough to stagger any one ; yet in the case of any organ, if we know of a long series of gradations in complexity, each good for its possessor, then, under changing conditions of life, there is no logical impossibility in the acquirement of any conceivable degree of perfection through natural selection. In the cases in which we know of no intermediate or transitional states, we should be very cautious in concluding that none could have existed, for the homologies of many organs and their intermediate states show that wonderful metamorphoses in function are at least possible. For instance, a swim-bladder has apparently been converted into an air-breathing lung* (chapitre VI, page 204). Charles Darwin pose ici clairement la question de l'explication évolutive de la complexité des systèmes biologiques (ici l'œil). On notera que

cette discussion entre dans le chapitre VI où Darwin discute des limites potentielles de la théorie de l'évolution. Dans ce chapitre, il se montre pleinement conscient que l'explication de la complexité apparente des systèmes biologiques est une pierre angulaire de l'acceptation de sa théorie. Aussi ajoute-t-il franchement, page 189 : *If it could be demonstrated that any complex organ existed, which could not possibly have been formed by numerous, successive, slight modifications, my theory would absolutely break down*[1].

À travers ces quatre extraits de l'« origine des espèces » on voit que la question de la complexité concerne plusieurs aspects fondamentaux de l'évolution. En effet, ils posent la question de la complexité des organismes eux-mêmes, de son origine et de ses causes, de la complexité de l'environnement – Darwin faisant ici l'hypothèse que la complexité l'environnement est cause de la complexité des organismes – et des limites à la complexité que peut accumuler l'évolution, toutes questions qui sont encore ouvertes aujourd'hui. Ainsi, la question de l'existence d'une « limite » à la complexité biologique est aujourd'hui associée à la question de l'« open-endedness » de l'évolution, une question fortement débattue dans la communauté de la vie artificielle (Banzhaf *et al.*, 2016; Taylor *et al.*, 2016; Packard *et al.*, 2019a,b). Cette question de l'open-endedness est elle-même associée à la question dite de la « flêche de la complexité » – *arrow of complexity* (Miconi, 2008), la question étant ici de savoir si on observe, ou non, une tendance générale à la complexification au cours de l'évolution. Avec quelque naïveté, cette question peut sembler triviale si l'on se croit fondé d'en juger à partir du constat que l'homme serait *le* produit de l'évolution, eu égard au haut degré de complexité qu'il a atteint. Cependant, outre que l'homme n'est pas plus le produit de l'évolution que n'importe quel autre organisme (dont bactérie et plantes) et qu'il n'a pas « plus » évolué (ou, pour le dire autrement, que les « autres » organismes n'ont pas arrêté d'évoluer), il faut noter que cette interprétation naturelle est fortement influencée par notre perception immédiate d'humains. En effet, si nous contemplons le monde qui nous entoure, l'impression de complexité est immédiate : ne sommes-nous pas entourés d'organismes situés au plus haut de « l'échelle de la complexité » ? (Ceci, quelle que soit la définition ou même l'existence de cette dernière.) Chats, chiens, araignées – ou même les puces qui occupent parfois nos maisons ! – ne sont-ils pas des organismes pluricellulaires, dotés d'un système nerveux central ? Si l'on prend comme échelle de complexité les « transitions majeures » de Maynard-Smith et Szathmary (1997), n'est-on pas entouré majoritairement d'organismes situés sur les barreaux les plus hauts de cette échelle ? Cette interprétation *naturelle*, au sens de Paul Feyerabend[2], est pourtant totalement contraire à ce que nous montre la science : en pratique nous sommes entourés – et en nombre infiniment plus grand que d'organismes dits « supérieurs » – d'organismes unicellulaires, de virus à ADN ou à ARN, de prions, etc. Nos sens nous trompent totalement quant à la composition du vivant contemporain, qu'il s'agisse du nombre d'organismes (il y a dans

1. De fait, cette argumentation, connue sous l'appellation de *irreducible complexity*, est aujourd'hui l'un des angles d'attaque favoris des créationnistes contre la théorie de l'évolution. Nous ne citerons pas de document ici pour ne pas faire de publicité à ces thèses mais, on le voit, l'enjeu de l'origine évolutive de la complexité dépasse le seul questionnement scientifique.

2. Selon Feyerabend (1979), une interprétation naturelle est une hypothèse formulée implicitement à partir de nos sensations immédiates. Faisant partie inscrite « charnellement » dans la connaissance commune à tout être humain, elle est très difficile à déceler et à remettre en question, même dans un cadre scientifique. L'exemple classique d'interprétation naturelle est l'apparente immobilité de la terre.

un seul estomac plus de bactéries que d'hommes sur terre) mais aussi, et peut-être surtout si l'on considère la question de la flèche de complexité, sur le plan du nombre d'espèces. La figure I.2 propose ainsi une représentation moderne, phylogénétique, de l'« arbre du vivant ». Une simple comparaison de cette figure avec la *Scala naturæ* (figure I.1) illustre la force de cette interprétation naturelle. Contrairement à cette dernière en effet, parmi toutes les branches de l'arbre des espèces, seule une petite fraction est composée d'organismes eucaryotes (donc ayant « franchi » la quatrième transition majeure – voir section 2.2) et, dans cette fraction pourtant déjà mineure, une grande partie des organismes sont unicellulaires et n'ont donc pas « franchi » la sixième transition majeure. Et encore, cette représentation ne tient pas compte des virus ou des phages dont plusieurs centaines de milliers de séquences sont aujourd'hui disponibles dans les bases de données génomiques. Cette vision plus objective de la répartition des espèces vivantes montre que, même s'il est indéniable que la complexité de certaines espèces a augmenté au cours de l'évolution, cette observation ne représente pas nécessairement une tendance globale. En outre, nous avons déjà évoqué les exemples d'espèces ayant connu une diminution de leur complexité suite à un changement de traits d'histoire de vie, tels les endosymbiotes ou certaines cyanobactéries marines (Giovannoni *et al.*, 2005; Batut *et al.*, 2014).

Même si les question de la flèche de la complexité et de l'open-endedness de l'évolution restent ouvertes, c'est la question de l'origine évolutive de la complexité qui reste la plus débattue. En effet, si on considère l'évolution comme un processus d'adaptation à l'environnement, il peut sembler évident que l'origine de la complexité réside dans l'adaptation à un environnement lui-même complexe. C'est d'ailleurs ce que semble suggérer Darwin (1859) lui-même dans la troisième citation ci-dessus (*then, considering the infinite complexity of the relations of all organic beings to each other and to their conditions of existence, causing an infinite diversity in structure, constitution, and habits, to be advantageous to them* [...]). Cependant, pour plusieurs auteurs, il s'agit là à nouveau d'une interprétation naturelle (Soyer et Bonhoeffer (2006) parle de « intuitive view » pour qualifier cette hypothèse) et d'autres processus, plus subtils, pourrait tout autant expliquer l'origine évolutive de la complexité. Ainsi, globalement, deux écoles se distinguent : comme nous venons de le voir, la première pose la sélection naturelle comme moteur de l'accroissement de la complexité, la seconde, en revanche, affirme que le processus de variation lui-même suffit à l'expliquer. Selon la première ce serait donc la richesse intrinsèque d'environnements exigeants qui appellerait des organismes complexes, aptes à en tirer un meilleur parti que des organismes simples. Selon la deuxième le processus de variation comporterait lui-même un biais en faveur de la complexité. Dans un cas comme dans l'autre, ces écoles sont loin d'être monolithiques. Elles représentent néanmoins des familles d'argumentations fondées sur deux principes fondamentaux divergents.

3.2 Hypothèses sélectives

Lorsque l'on pense à la profusion des formes de notre environnement, on est assez naturellement conduit à penser qu'il faut pour en tirer le meilleur profit un grand nombre d'organes, de voies de signalisations, de systèmes de réaction... De ce point de vue, la complexité d'un organisme (la multiplicité de ses organes) présenterait un bénéfice en comparaison à un organisme plus simple et lui assurerait un plus grand succès reproductif et, par voie de conséquence, évolutif. Cette hypothèse, très intuitive, fait donc de la

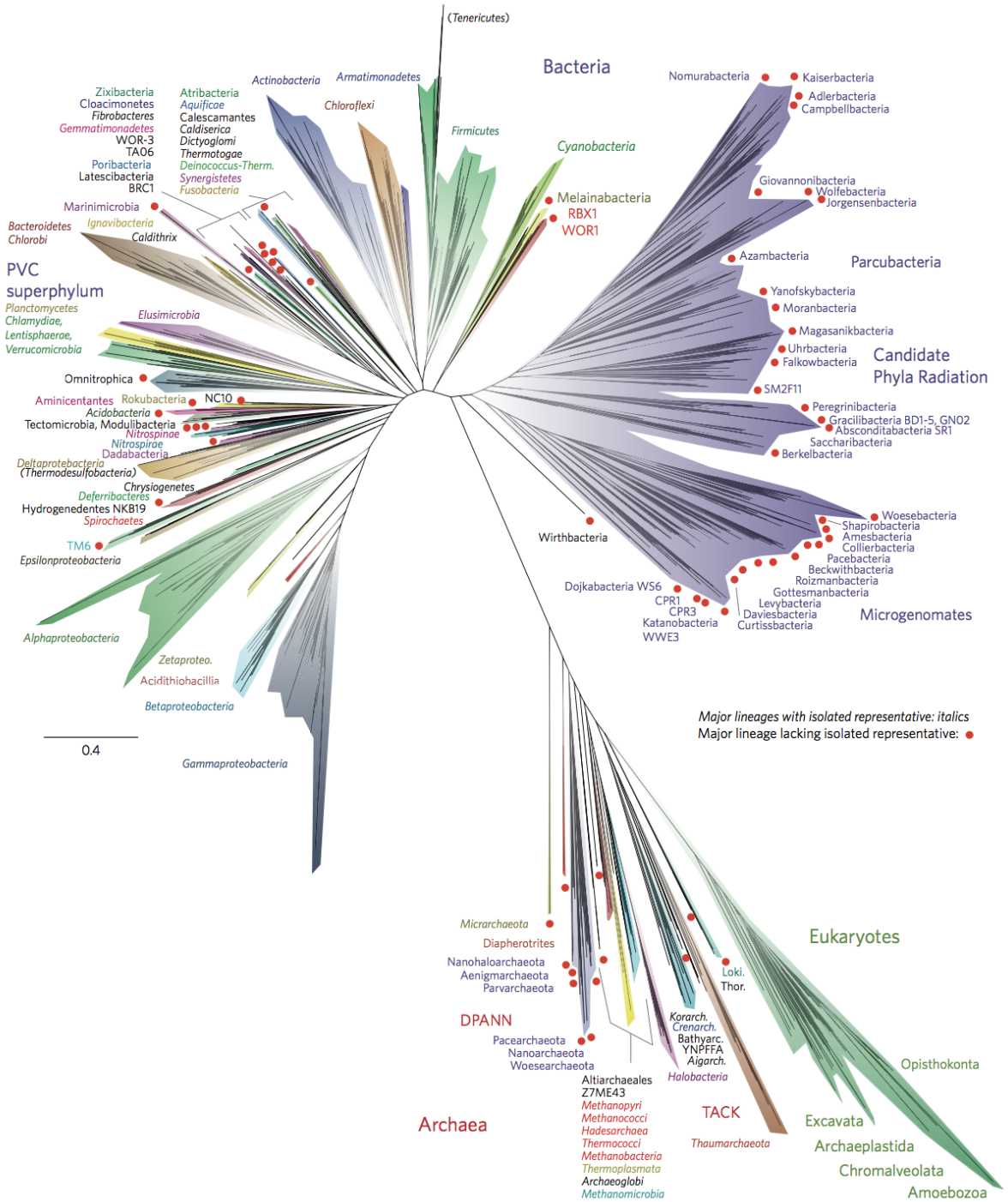


Figure I.2 – Cette représentation récente de l'arbre de la vie proposée par (Hug *et al.*, 2016) donne une idée de la distribution des complexités au sein du vivant : seule la branche « Eukaryotes » (en bas à droite) comprend des cellules à noyau – quatrième transition majeure de Maynard-Smith et Szathmary (1997) – et, au sein de cette branche, seuls certains groupes (Opisthokonta, Excavata et Chromalveolata) contiennent des organismes pluricellulaires (sixième transition majeure).

sélection le moteur de l'accroissement de la complexité. Cependant, au delà de cette formulation simple, l'hypothèse sélective n'est pas reçue sous la même acception par tous les auteurs et plusieurs mécanismes différents ont été suggérés dans la littérature pour relier sélection et complexité (Lukeš *et al.*, 2011).

3.2.1 Sélection intrinsèque de la complexité

Dans son acception la plus simple, l'hypothèse sélective énonce simplement que les organismes complexes ont intrinsèquement une plus haute fitness et ce simplement parce que la complexité est susceptible d'améliorer le fonctionnement d'un organe donné. C'est l'hypothèse classiquement adoptée pour expliquer l'évolution d'organes complexes tels que l'œil des vertébrés (Arendt *et al.*, 2009; Nilsson, 2013) ou le flagelle des bactéries [1] (Pallen et Matzke, 2006; Pallen et Gophna, 2007) : pour améliorer la fonction (ici la vision ou la motilité), l'accrétion d'un nombre toujours plus grand de composants serait nécessaire, d'où l'accroissement de la complexité. On notera ici que, comme la plupart des discours sur la complexité, c'est surtout sur la base d'exemples spécifiques qu'opère le raisonnement. Ces exemples étant contextuels, la portée de l'argument en est significativement diminuée : certes, un grand nombre d'éléments permettent de composer un œil ou un flagelle plus performants, mais ce raisonnement se base sur l'hypothèse implicite qu'un œil ou un flagelle performants représentent un avantage, ce qui suppose qu'un certain degré de complexité fonctionnelle soit déjà atteint. On notera que, selon certains auteurs, cette hypothèse se heurterait à une limite thermodynamique : si, comme l'énonce Chris Adami (2002b), la complexité est due à l'intégration, au sein de l'organisme, d'une information disponible dans l'environnement (d'une façon similaire au démon de Maxwell), alors la complexité des organismes (mesurée en termes de quantité d'information) ne peut pas dépasser celle de l'environnement au sein duquel ils évoluent (Krakauer, 2011).

3.2.2 Sélection par la complexité environnementale

La sélection intrinsèque n'est pas nécessairement le seul moteur de l'accroissement de la complexité. En effet, celle-ci peut aussi être sélectionnée – extrinsèquement – par l'environnement, pour peu que celui-ci offre lui-même une diversité de situations suffisante. Comme nous l'avons vu, cette hypothèse est déjà présente chez Darwin (1859) mais elle est implicitement ou explicitement présente chez de nombreux auteurs tels Waddington (1969), Dawkins et Krebs (1979) ou Heylighen (1999).

Plusieurs facteurs pourraient contribuer à la complexité de l'environnement, au premier titre desquels les interactions inter-spécifiques, les relations proie-prédateur (dont la « course aux armements »), les relations hôte-parasite ou la sélection sexuelle (Albantakis *et al.*, 2014; Zaman *et al.*, 2014). On notera cependant que cette hypothèse explique une complexité biologique (celle des organismes) par une autre (la complexité biologique des écosystèmes). Cependant d'autre formes de complexité environnementales existent qui ne sont pas assujetties à l'évolution, telles que la diversité des paramètres physiques des niches écologiques (températures, ensoleillement, etc), la distinction entre les milieux marins (salins ou non), terrestres ou aériens voire les variations temporelles de ces pa-

1. Il n'est d'ailleurs pas surprenant que ces deux organes soient souvent pris comme exemple par les créationnistes tenants de la « complexité irréductible » (Snyder *et al.*, 2009).

ramètres (variations journalières, marées, etc) dont il a récemment été montré qu'elles peuvent favoriser la mise en place d'écosystèmes complexes (Rocabert *et al.*, 2017).

3.2.3 Sélection indirecte

Les deux hypothèses précédentes mettaient en œuvre la sélection « directe », les organismes les plus complexes ayant simplement une valeur sélective plus élevée, que ce soit pour des raisons intrinsèques ou extrinsèques. Il est cependant désormais bien connu en biologie évolutive que le nombre de descendants n'est pas le seul facteur du succès évolutif. En effet, des facteurs dits « indirects » ou « de second ordre » peuvent aussi contribuer au succès évolutif d'un organisme ; il s'agit de la robustesse (Wilke *et al.*, 2001; Wagner, 2007, 2008) et de l'évolvabilité (Pigliucci, 2008; Woods *et al.*, 2011), c'est-à-dire respectivement de la capacité d'un organisme à se reproduire ou à muter en conservant sa fitness et de la capacité d'un organisme à explorer efficacement son voisinage évolutif et, de ce fait, à accroître sa fitness.

Le lien entre complexité, robustesse et variabilité a été établi très tôt dans l'histoire de la science des systèmes complexes (Simon, 1962) et il a été naturellement transposé à la complexité des organismes. Ainsi, une relation entre robustesse et complexité a pu être observée sur des organismes numériques (Lenski *et al.*, 1999) et il a été proposé que les systèmes biologiques complexes pourraient bénéficier d'une « robustesse distribuée » (Wagner, 2005). La question de l'évolvabilité est davantage débattue. Il a en effet été envisagé que les organismes les plus complexes seraient plus susceptibles de spéciation (Schopf *et al.*, 1975) et qu'ils présenteraient un taux d'évolution plus élevé (Wang *et al.*, 2010). Cependant, il est bien connu en génétique des populations que l'augmentation du nombre de traits sous sélection diminue le potentiel évolutif en raison du « coût de la complexité » (Orr, 2000) même si l'interprétation de ce résultat demeure débattue. De fait, il n'est pas évident que le nombre de dimensions dans un modèle mathématique tel que le Modèle Géométrique de Fisher (FGM) soit en adéquation avec la complexité d'un système biologique (Hansen, 2003; Wagner et Zhang, 2011).

3.2.4 Le modèle de la « boite à outils »

L'idée centrale du modèle de la boite à outils est qu'un système comprenant un nombre croissant de composants requiert lui-même un nombre encore plus grand de mécanismes de régulation pour contrôler et coordonner l'action de ces différents composants. Ce mécanisme est donc susceptible de provoquer un accroissement rapide de la complexité d'un système biologique, même si le nombre de composants « primaires » n'augmente, quant à lui, que graduellement.

Ce modèle a été proposé par Maslov *et al.* (2009) pour expliquer les observations de Molina et van Nimwegen (2008, 2009). En analysant les familles de gènes chez un grand nombre d'espèces bactériennes, ces auteurs ont en effet montré que le nombre de facteurs de transcription augmente quadratiquement avec la complexité du système métabolique [1].

1. Cette analyse est cependant contestée par Beslon *et al.* (2010b) qui proposent que cette augmentation quadratique peut simplement refléter l'évolution – neutre – du nombre de possibilités d'interactions entre les composants du système.

Cette hypothèse a aussi été proposée pour expliquer l'évolution de mécanismes de régulation complexes chez les eucaryotes (Mattick *et al.*, 2010).

3.2.5 Le modèle « Sélection, Pléiotropie et Compensation »

Comme le modèle de la boîte à outils, le modèle « Sélection, Pléiotropie et Compensation » (SPC), proposé par (Pavlicev et Wagner, 2012), est un mécanisme d'auto-entraînement dans la mesure où le recrutement (ou la modification) de composants dans un système nécessite l'ajout de composants supplémentaires ou la modification de composants pré-existants.

Dans le modèle SPC, le mécanisme d'auto-entraînement est dû à la pléiotropie, c'est-à-dire au fait que, dans un système biologique, un gène va affecter plusieurs caractères. De ce fait, le recrutement, au cours de l'évolution, de nouveaux composants/gènes (positivement sélectionnés pour leur contribution à une fonction donnée) va entraîner des effets pléiotropes sur d'autres fonctions. Ces effets étant généralement délétères, ils ouvrent des possibilités de compensation, éventuellement par le recrutement de nouveaux composants/gènes qui vont eux-mêmes avoir des effets pléiotropes et ainsi de suite, provoquant l'accumulation de nouveaux composants et donc l'accroissement de la complexité. On notera cependant que le modèle SPC explique l'accumulation de *mutations* compensatrices dans différents composants d'un systèmes en évolution et non nécessairement l'accumulation de *composants*. La pléiotropie constitue néanmoins un élément important de l'évolution de la complexité (Wagner et Zhang, 2011).

3.2.6 Conclusion

On le voit, bien qu'elles soient parfois considérées comme naïves ou intuitives (Soyer et Bonhoeffer, 2006), les hypothèses sélectives sont loin de former un tout homogène. En effet, la sélection peut conduire à l'accumulation de complexité en raison de causes intrinsèques, extrinsèques ou, plus subtilement par des effets d'entraînement tels que les effets « boîte à outils » ou les effets de compensation. En outre, précisément parce qu'elles sont intuitives, ces différentes hypothèses ont des répercutions importantes dans le débat opposant évolution et créationnisme. On peut d'ailleurs reconnaître derrière de nombreux défenseurs des hypothèses sélectives de fervents opposants au créationnisme, tels que Richard Dawkins (1997).

En lien direct avec les travaux présentés dans cette thèse, on peut noter que les hypothèses sélectives ont reçu le soutien de plusieurs auteurs issus du champ de la vie artificielle (Adami *et al.*, 2000; Lenski *et al.*, 2003; Yaeger *et al.*, 2008). On peut d'ailleurs noter que plusieurs de ces contributions visaient à prouver, par la simulation, qu'un mécanisme Darwinien est tout à fait à même d'atteindre des niveaux de complexité importants, même en procédant par petits incréments. Dans cette optique, ces études, plutôt que de défendre un mécanisme unique pour expliquer l'accumulation de complexité, entendent confirmer Darwin jusque dans l'affirmation que nous avons mise en exergue dans l'introduction de la présente section : « *If it could be demonstrated that any complex organ existed, which could not possibly have been formed by numerous, successive, slight modifications, my theory would absolutely break down* » (Darwin (1859), chapitre VI, page 189).

3.3 Hypothèses neutralistes

Au contraire des hypothèses sélectives, les hypothèses neutralistes posent que la complexité proviendrait de facteurs exogènes à la sélection naturelle et que la complexité augmenterait même en l'absence de sélection. Selon ces hypothèses, la complexité répondrait à un simple processus de diffusion dans un espace (génomes, protéomes, etc.) qui présenterait un biais vers la complexité. Cependant, pas plus que les hypothèses sélectives, les hypothèses neutralistes ne forment un tout homogène et l'origine ou les causes invoquées pour expliquer ce biais diffèrent selon les auteurs. On peut ainsi distinguer trois écoles de pensée : Stephen Jay Gould, fervent opposant aux hypothèses sélectives, avec son modèle de la « promenade de l'ivrogne » (*drunkard's walk*) (Gould, 1996), (McShea et Brandon, 2010) avec la *Zero-Force Evolutionary Law* (ZFEL), et Arlin Stoltzfus (1999) avec le modèle *Constructive Neutral Evolution*.

3.3.1 Drunkard Walk Model

Selon Stephen Jay Gould (1996), la question de l'évolution de la complexité est vide de sens puisque, le processus d'évolution dans la mesure où il s'apparente à un processus de diffusion dans l'espace des génotypes ou des phénotypes et où la complexité admet une borne inférieure, il va nécessairement produire une fraction d'organismes complexes – même si la majeure partie des organismes reste proche de la borne inférieure de complexité (voir figure I.3). Selon ce modèle, l'augmentation de la complexité ne serait donc pas due à quelque tendance générale mais simplement à l'existence, statistiquement certaine, d'espèces pour lesquelles la chance aura induit quelque biais dans la direction que suit la marche aléatoire, par ailleurs globalement uniforme.

Pour illuster ce mécanisme, Steven Jay Gould a proposé un modèle très parcimonieux sous le nom imagé de *Drunkard Walk Model* (la « promenade de l'ivrogne »). Dans ce modèle, la complexité varie d'une génération à la suivante en effectuant des sauts correspondants à une distribution uniforme. Si au terme du saut la grandeur est amenée à devenir négative, la complexité est maintenue à 0 jusqu'à la génération suivante. Un tel modèle conduira *nécessairement* à l'apparition d'espèce de forte complexité et à l'augmentation de la complexité moyenne. On notera au passage que cette description est totalement indépendante d'une quelconque définition de la complexité ou même d'une mesure explicite : tout trait quantitatif (par exemple la taille d'un organime) admettant une borne inférieure est ainsi susceptible d'augmenter chez certaines espèces indépendamment d'un quelconque facteur sélectif.

La figure I.3 illustre la distribution des complexités atteintes par le Drunkard's Walk Model. La courbe moyenne montre clairement comment les complexités augmentent et la distribution des complexités dessine globalement une demi-gaussienne (hormis une légère sur-représentation des faibles complexités qu'explique la frontière en 0). Ainsi, les effectifs décroissent-ils avec la complexité et cette décroissance s'accélère passé un point d'inflexion.

Il est remarquable que cette distribution s'accorde avec les observations selon lesquelles les organismes les plus complexes sont aussi les plus rares, comme l'illustre l'arbre de la vie (figure I.2). Comme nous l'avons déjà évoqué, sur celui-ci, en effet, toute la partie supérieure est constituée de bactéries tandis que, sur la branche inférieure, seul l'apex vert est constitué d'eucaryotes. Et encore ! Même dans cet apex, seule une petite fraction se trouve constituée d'organismes pluricellulaires. Or notre jugement instinctif sur la profu-

sion de la complexité se fonde essentiellement sur le fait que notre attention se focalise sur ces derniers et que nous leur attribuons un poids affectif et intellectuel disproportionné.

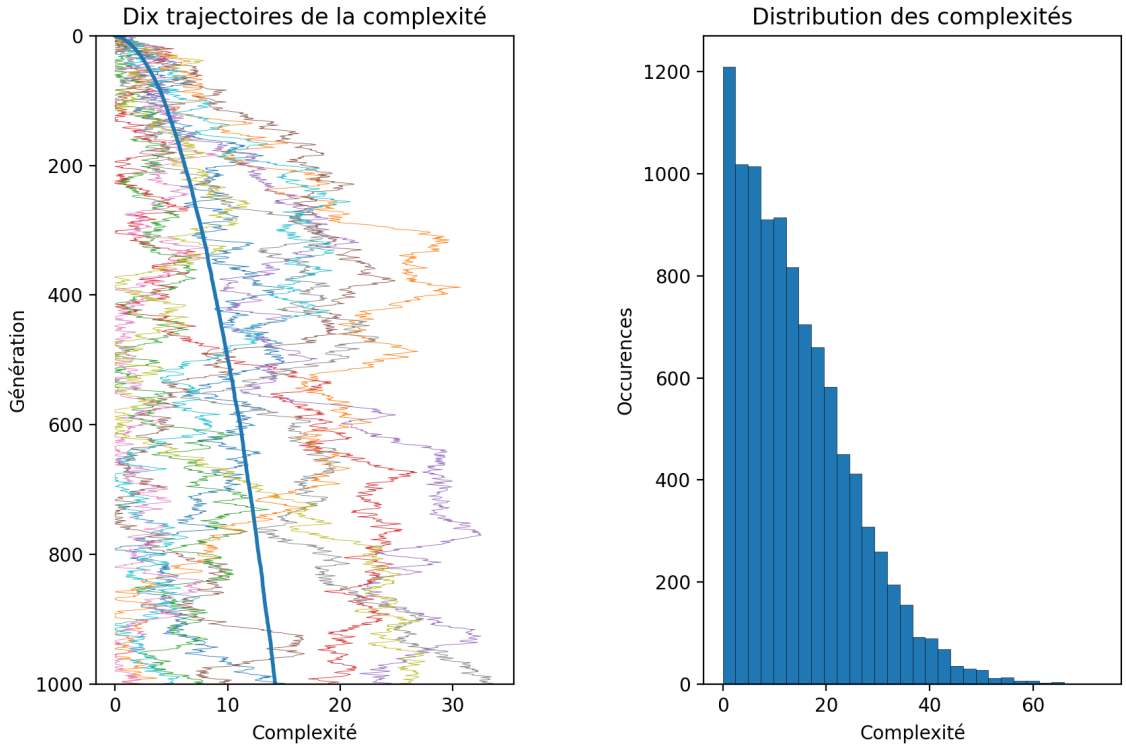


Figure I.3 – Complexités induites par le *Drunkard Walk Model* de Gould. 10 000 lignées ont été simulées à raison de 1000 générations par lignée. À gauche : exemples de trajectoires suivies par la complexité (10 premières répétitions). La courbe bleue représente l'évolution de la moyenne des complexités pour les 10 000 lignées. À droite : Distribution des complexités à la millième génération.

3.3.2 Zero-Force Evolutionary Law

Daniel McShea est un nom reconnu dans les cercles s'intéressant à l'évolution de la complexité. Paléobiologiste, il a conduit plusieurs études phénoménologiques bien reçues sur l'évolution de la complexité au cours des temps géologiques (McShea, 1993, 1996). Plus récemment, il s'est associé avec un philosophe, Robert Brandon, pour proposer la « Zero-Force Evolutionary Law » (ZFEL) dans un ouvrage au titre pour le moins ambitieux (« Biology's First Law ») édité aux presses universitaires de Chicago (McShea et Brandon, 2010). Compte-tenu de la renommée de l'auteur, cet ouvrage a reçu une audience importante malgré le fait que la ZFEL n'ait jamais été publiée dans un journal de biologie ou biologie évolutive important.

Sans nier l'effet de la sélection dans l'augmentation de la complexité, McShea et Brandon (2010) soutiennent qu'un autre facteur, plus fondamental, serait à l'œuvre. Ils postulent en effet l'existence d'une « Zero-Force Evolutionary Law » (ZFEL). Selon la ZFEL,

dans tout système évolutif (donc au sein duquel agissent un mécanisme de variation et un mécanisme d'hérédité) la diversité et la complexité augmentent en moyenne, même en l'absence de sélection naturelle, d'autres forces et de contraintes agissant sur la diversité ou la complexité (McShea et Brandon, 2010, p.3). Selon ces auteurs, la ZFEL pousserait donc la diversité et la complexité des systèmes évolutifs vers une augmentation même en l'absence de sélection (d'où la notion de « zero-force law ») et même si le processus mutationnel est sans biais ou si la complexité du système en question reste très au dessus de sa borne inférieure, faisant de la ZFEL un mécanisme universel quand – selon les auteurs – le Drunkard's walk model échoue à expliquer pourquoi la tendance à la complexité se maintient sur le très long terme.

De fait, même si la ZFEL partage avec le Drunkard's Walk Model la filiation « neutraliste », les deux modèles présentent des différences substantielles quoique discrètes. Ainsi, chez Gould, c'est la complexité elle-même qui suit une marche aléatoire alors que, chez McShea, ce serait l'organisme (par le biais d'accumulation de variation), la complexité n'en étant que le reflet. La différence entre les deux mécanismes est alors que la marche aléatoire de l'organisme se maintient sur le long terme (plus l'organisme est complexe et plus il va avoir une tendance à la variation – donc, selon McShea et Brandon (2010), plus il va devenir complexe). La ZFEL est donc bien une force au sens où elle pousse à l'augmentation de la complexité tandis que dans le Drunkard's Walk Model, l'augmentation de la complexité est passive.

La ZFEL se présente essentiellement comme une expérience de pensée soutenue par des exemples qualitatif (*p. ex.* la diversification des segments chez les vertébrés). Elle présente en ce sens une stimulation intellectuelle. On est cependant en droit d'attendre d'un véritable modèle qu'il donne des indications précises quant aux définitions de façon en particulier à pouvoir réfuter ledit modèle et, éventuellement, à le simuler. Récemment McShea *et al.* (2019) ont ainsi proposé un modèle quantitatif de la la ZFEL. Cependant cette définition assimile la complexité à la variabilité inter-constituants du système (*p. ex.* la variance pour les traits quantitatifs) ce qui, au vu des différentes définitions de la complexité semble limité. Il est important de noter ici que McShea *et al.* (2019) ne présentent pas la ZFEL comme une loi unique qui expliquerait intégralement l'évolution de la complexité. Selon les auteurs, puisque la ZFEL implique une augmentation de complexité « sans force », elle serait cependant l'hypothèse à tester avant toute autre pour expliquer une variation de complexité observée au cours de l'évolution.

3.3.3 Constructive Neutral Evolution

Proposée par Arlin Stoltzfus (1999) et fortement défendue entre autres par W. Ford Doolittle (Gray *et al.*, 2010; Lukeš *et al.*, 2011; Doolittle *et al.*, 2011; Doolittle, 2012; Brunet et Doolittle, 2018), la « Constructive Neutral Evolution » (CNE) repose sur l'idée que dans un système qui comporte plusieurs composants fonctionnellement indépendants, ceux-ci interagissent néanmoins naturellement (du simple fait de leur juxtaposition au sein d'un même système). De ce fait, l'évolution des différents composants va graduellement intégrer ces interactions au point que les différentes fonctions vont progressivement se diluer sur l'ensemble des composants. Ceux-ci deviennent alors fonctionnellement dépendants.

Comme pour le Drunkard's Walk Model ou la ZFEL, la CNE repose sur l'idée d'une diffusion mais ici la diffusion a lieu dans l'espace des interactions entre les composants

du système. Comme la taille de cet espace augmente très rapidement avec le nombre de composants, et que la fraction de cet espace qui correspond à l'absence d'interactions est infinitésimale, Lukeš *et al.* (2011) présentent la CNE comme un mécanisme de « cliquet » (*ratchet*) qui impose l'accumulation de complexité fonctionnelle au cours du temps.

Une particularité de la CNE, comparée au deux mécanismes précédents, est qu'il s'agit d'un mécanisme neutre (au sens où l'augmentation de la complexité n'est pas sélectionnée) mais nécessitant néanmoins que la sélection maintienne la fonction des composants et n'interdise pas l'établissement de nouvelles interactions (en d'autres termes, il repose sur la sélection purificatrice). En ce sens, la CNE est compatible avec les théories de Michael Lynch selon lesquelles l'accumulation de complexité pourrait s'opérer même si cette complexité va à l'encontre de la valeur sélective. En effet, toujours selon Michael Lynch, la faible taille efficace de certaines populations conduit à l'accumulation de composants faiblement délétères mais pour lesquels la sélection n'a pas de prise. Il est en effet bien connu que, dans une population de faible effectif (ou plus exactement de faible taille efficace), les mutations faiblement délétères (ou faiblement favorables) subissent un effet de dérive supérieur à l'effet de la sélection. Michael Lynch propose que ce mécanisme ait favorisé l'accumulation de structures complexes dans les génomes eucaryotes (Lynch et Walsh, 2007).

3.4 Hypothèses variationnelles

En marge des approches purement sélectionnistes ou purement neutraliste, des hypothèses d'un troisième type – souvent plus discrètes dans le débat – ont été proposées dans la littérature. Selon ces hypothèses, c'est la nature même du processus de variation qui biaiserait l'évolution vers l'accroissement de la complexité. En effet, les hypothèses neutralistes ou sélectionnistes raisonnent au niveau fonctionnel (donc indépendamment du processus variationnel sous-jacent). Or, si on se pose la question des événements mutationnels susceptibles de produire de la complexité, il semble évident que le phénomène de duplication est au cœur du processus, potentiellement accompagné d'un phénomène de divergence (Ohno, 2013). Ainsi, John-Maynard-Smith (1970) énonce, que l'évolution de la complexité admettait une « *obvious and uninteresting explanation*[1] » selon laquelle : « *The first is that processes are known (e.g. duplication) whereby the genetic material of an individual can increase. Even if the additional material is redundant or nonsensical, it does provide raw material for the evolution of increasing complexity. It is less easy to imagine processes leading to a loss of genetic material, since most losses will involve losses of functions essential for survival* ».

L'idée est donc ici que le processus d'addition de matériel génétique (envisagé implicitement comme base de tout processus de complexification) est moins délétère en moyenne que le processus de suppression de matériel génétique et que ce biais entraîne un accroissement spontané de complexité. Ainsi, selon Saunders et Ho (1976), l'accroissement de la complexité n'est pas lié à l'accroissement de fitness : selon l'auteur, accroissement de fitness et accroissement de complexité sont deux conséquences indépendantes de l'évolution. En ce sens, ce mécanisme n'est ni « sélectif », ni « neutraliste » mais « évolutif », l'évolution intègrant *à la fois* la sélection et la neutralité.

1. Affirmation sujette à caution quand si l'on considère l'intensité des débats qui traversent la biologie évolutive depuis !

Bien que de façon peu formalisée, on retrouve cette idée d'un processus intrinsèquement biaisé par la nature même des événements moléculaires chez plusieurs auteurs – voir par exemple (Heylighen, 1999). Ce n'est que plus récemment que Soyer et Bonhoeffer (2006), en modélisant l'évolution de la complexité dans les voies de signalisation cellulaires au moyen d'un modèle mathématique et de simulations, ont montré que la complexité des voies de signalisation pourrait être contingente, c'est à dire que « la complexité [d'une voie de signalisation] pourrait augmenter sans qu'aucune pression sélective ne l'impose » (Soyer et Bonhoeffer, 2006). Elle résulterait d'un déséquilibre entre les effets des mutations qui accroissent la taille du réseau métabolique et de ceux qui la réduisent car les premiers sont plus délétères que les seconds. Ces auteurs suggèrent aussi que la robustesse jouerait un rôle dans ce phénomène car les solutions les plus simples seraient également peu robustes face aux événements de délétions. Même si Soyer et Bonhoeffer (2006) ne font pas référence à la littérature antérieure, on retrouve ici l'idée que le processus évolutif est spontanément biaisé vers la complexité du fait de l'asymétrie des processus mutationnels [1].

1. On peut noter ici que cette asymétrie n'est pas systématiquement en faveur d'un accroissement de la complexité : Fischer *et al.* (2014) ont ainsi montré, au moyen d'un modèle mathématique, que l'asymétrie entre les duplications et les délétions borne l'accroissement des génomes.

4 Conclusion

Ce tour d'horizon – nécessairement partiel eu égard à la profusion des contributions dans ces domaines – autour des notions de complexité, de complexité en biologie et, finalement d'évolution de la complexité nous aura éclairé sur un certain nombre de points. Tout d'abord, force est d'admettre qu'il n'y a pas aujourd'hui de définition totale et univoque de « la complexité », pas plus qualitativement que quantitativement. Au contraire, celle-ci apparaît comme contextuelle, dépendante d'un choix d'observation ou d'analyse. En s'inspirant d'une métaphore proposée par (Heylighen, 1999), on peut illustrer ce problème par la différence entre un immeuble et un tas de briques. Cette différence est évidemment profonde, l'un étant organisé – ce qui lui confère une fonction – tandis que l'autre ne l'est d'aucune façon. Cette différence profonde s'évanouit cependant si l'observateur change d'échelle puisqu'à l'échelle atomique, l'organisation des deux systèmes est similaire.

Dans le cas des systèmes biologiques, ce premier constat – l'insaisissabilité du concept même de complexité – se double d'une seconde difficulté liée au caractère organisé propre aux systèmes biologiques qui se déploient sur plusieurs niveaux d'organisation fortement interconnectés. Dans ce cas, la complexité du système devient plurielle : observable horizontalement à chaque niveau – eux-mêmes étant fortement interdépendants et fortement indépendants du fait du caractère dégénéré du codage de l'information biologique – ou observable verticalement à travers le nombre de niveaux d'organisation. La difficulté qu'il y a à quantifier la complexité des systèmes biologiques s'en trouve encore accrue, celle-ci pouvant être exprimée relativement au nombre de composants, à l'hétérogénéité des composants ou même relativement aux nombre d'inter-dépendances entre ces composants.

Cette difficulté à définir et à quantifier la complexité biologique n'est probablement pas étrangère aux débats nourris sur son origine évolutive. À la question « Pourquoi l'évolution Darwinienne conduit-elle à un accroissement de complexité ? » de multiples réponses ont été apportées dans la littérature avec en particulier un débat autour des causes sélectives ou neutres de l'accroissement de complexité, sans oublier une « troisième voie » selon laquelle la complexité est inhérente à un processus évolutif reposant fortement, si ce n'est essentiellement, sur un mécanismes de duplication-divergence. À l'issue du recensement des différentes hypothèses, plusieurs éléments d'ensemble ressortent. Premièrement, il est clair que, malgré le débat qui les oppose, ces hypothèses ne sont pas exclusives. Au contraire, parce qu'elles n'ont pas les mêmes modalités d'action (certaines hypothèses favorisant le recrutement de nouveaux composants quand d'autre favorisent leur intégration au sein d'un tout fonctionnel intégré), il a même pu être suggéré que ces hypothèses soient complémentaires, les unes gouvernant l'émergence de la complexité quand les autres faciliteraient son accumulation ultérieure (Heylighen, 1999). En outre, certaines hypothèses tendent à expliquer la complexité du codage de l'information génétique tandis que d'autres se focalisent davantage sur les niveaux fonctionnels. Deuxièmement, la plupart de ces hypothèses reposent essentiellement sur des expériences de pensée. S'agissant d'hypothèses non-exclusives, on comprend aisément que le débat peine à être tranché, surtout que s'y immisce régulièrement la question du créationnisme qui voit dans la complexité des systèmes biologiques une opportunité pour enfoncer la théorie Darwinienne.

Face aux difficultés à discriminer parmi les différentes hypothèses, la simulation a rapidement été proposée comme un outil pour explorer l'origine évolutive de la complexité des

systèmes biologiques. La simulation présente en effet de nombreux avantages pour explorer cette question. D'une part, elle permet de réaliser des expériences idéales, dans lesquelles les pressions évolutives sont parfaitement connues et maîtrisées et dans lesquelles la complexité peut être mesurée objectivement. D'autre part, elle permet d'observer relativement aisément des temps évolutifs très longs, ce qui ici est fondamental puisque l'évolution de la complexité s'est déroulée à l'échelle des temps géologiques – macro-évolutive – et non à l'échelle micro-évolutive.

De fait, plusieurs modèles issus de la vie artificielle ont été exploités pour étudier cette question. On peut ainsi citer Avida (Lenski *et al.*, 1999; Adami, 2002a,b; Lenski *et al.*, 2003; Zaman *et al.*, 2014) et Polyworld (Yaeger, 1994; Yaeger *et al.*, 2008). Cependant, tous ces modèles se sont vus confrontés à plusieurs difficultés majeures. D'une part, dans la plupart des modèles de vie artificielle, la complexité ne peut évoluer qu'au niveau fonctionnel car la quantité d'« information génétique » (c'est-à-dire la quantité d'information génétique transmise de génération en génération) est constante. C'est par exemple le cas dans la plupart des instances d'Avida [1] : dans ce modèle le génome est constitué d'un ruban d'instructions pseudo-assembleur mais le nombre d'instructions est constant. Cette caractéristique n'interdit pas l'évolution de la complexité fonctionnelle du programme codé par cette séquence d'instructions (Lenski *et al.*, 2003) mais limite évidemment l'évolution de la complexité du génome proprement dit. En outre, le fait que l'information génétique soit prédéfinie interdit d'intégrer des opérateurs de duplication-divergence dans les mécanismes de variation alors que nous avons vu que ces opérateurs occupent probablement un rôle central dans la dynamique de la complexité (Maynard-Smith, 1970). La seconde limitation de ces modèles est qu'ils nécessitent de prédéfinir un environnement, souvent indirectement spécifié par une fonction d'évaluation (ou « fonction de fitness »). Or, dans la plupart des modèles, cette fonction d'évaluation est définie de façon à engendrer une dynamique évolutive « intéressante », c'est à dire de façon à favoriser l'émergence d'organismes complexes. En d'autres termes, les modèles de vie artificielle se placent, plus ou moins délibérément, dans une optique sélectionniste. Ce choix, ne pose pas de problème rédhibitoire lorsqu'il s'agit de démontrer que l'évolution Darwinienne peut conduire à la complexité – donc dans une optique « anti-créationniste » (Lenski *et al.*, 2003) – en revanche, il est sujet à caution dès lors qu'il s'agit de confronter les différentes hypothèses relatives à l'origine de la complexité à la « vérité des modèles ».

À partir de ces différents constats, nous avons cherché à définir un environnement de simulation permettant de distinguer les mérites des différentes hypothèses relatives à l'évolution de la complexité, c'est-à-dire un environnement qui lève les ambiguïtés propres aux expériences de pensées sans pour autant se placer d'emblée dans un cadre sélectionniste. Pour cela, nous devrons disposer d'un modèle multi-échelle (afin de pouvoir, tout au moins, étudier l'évolution de la complexité à l'échelle génomique et à l'échelle fonctionnelle) et être capable de définir des « environnements » (des fonctions de fitness) simples (*c.-à-d.* tels qu'un organisme non-complexe puisse y survivre et y prospérer) ou complexes (*c.-à-d.* requérant des fonctionnalités complexes) de façon à pouvoir aisément discriminer les hypothèses neutralistes des hypothèses sélectives. En effet, en observant l'évolution d'organismes initialement simples dans des environnements simples ou complexes, il de-

1. Notons qu'il ne s'agit pas là d'une limitation intrinsèque du modèle mais plus de ses conditions d'usage.

vrait être possible d'isoler les différentes forces poussant à l'accumulation de complexité, de les comparer et d'en estimer les effets respectifs.

Dans cette thèse, nous avons utilisé le modèle Aevol (Knibbe *et al.*, 2007a, 2008; Batut *et al.*, 2013; Rutten *et al.*, 2019) pour mettre en œuvre ce programme de recherche. Aevol est en effet une plateforme de simulation d'évolution dans laquelle les organismes sont encodés sous la forme d'un génome mais le décodage de celui-ci est directement inspiré de la structure biologique de la *genotype-phenotype map*. En outre, Aevol utilise une représentation abstraite, mathématique, des niveaux fonctionnels et de l'environnement, qui permet à la fois de quantifier la complexité fonctionnelle et de définir des environnements de complexité variable. Enfin, puisque, dans le modèle comme dans la réalité biologique, le code génétique est dégénéré et redondant, Aevol inclus de nombreux degrés de liberté permettant aux niveaux génomiques et fonctionnels d'évoluer partiellement indépendamment (bien que le premier code le second). Cette particularité nous permettra de comparer l'évolution de la complexité à différents niveaux de « nos » organismes virtuels.

Dans la suite de ce document, nous commencerons par décrire le modèle Aevol ainsi que les résultats que ce modèle a d'ores et déjà permis d'obtenir quant à l'évolution de la complexité (chapitre II). Nous verrons en particulier que, dans sa version classique, le modèle Aevol a été conçu pour forcer l'émergence évolutive d'organismes complexes via des contraintes sélective. Dans un second temps, nous définirons un plan expérimental destiné à adapter et à utiliser le modèle pour identifier et étudier les mécanismes propres à l'accumulation de complexité au cours de l'évolution (chapitre III). La première partie de ce plan expérimental sera décrite dans le chapitre IV. Nous montrerons en particulier que l'accumulation de la complexité peut être due à un mécanisme de cliquet mû par un phénomène d'épistasie de signe et nous montrerons d'une part que ce cliquet est plus puissant que les forces sélectives, directes comme indirectes, et d'autre part qu'en l'absence de sélection, le cliquet s'arrête, invalidant de fait les hypothèses neutralistes. Ce chapitre reprend la majeure partie d'une publication parue en avril 2020 dans *Artificial Life* [1]. La seconde partie du plan expérimental est présentée dans la chapitre V. Nous y montrons qu'en présence de forces sélectives, la complexité n'est pas plus importante que sous l'influence du seul cliquet identifié au chapitre IV. Nous expliquons partiellement ce résultat par l'influence d'une pression à la robustesse limitant la complexité à l'échelle des génomes et nous montrons que cette pression peut même, lorsqu'elle est suffisamment forte, contrer l'action du cliquet. Enfin, nous concluons dans le chapitre VI en discutant en particulier de la généralisation de ce mécanisme de cliquet à d'autres phénomènes historiques.

1. Liard, V., Parsons, D. P., Rouzaud-Cornabas, J., & Beslon, G. (2020). The complexity ratchet : Stronger than selection, stronger than evolvability, weaker than robustness. Artificial Life, 38-57. Cet article est lui-même une version étendue d'un article primé du best-paper award à Tokyo lors de la conférence internationale ALife en 2018 (Liard, V., Parsons, D., Rouzaud-Cornabas, J., & Beslon, G. (2018, July). The Complexity Ratchet : Stronger than selection, weaker than robustness. In *Artificial Life Conference Proceedings* (pp. 250-257). One Rogers Street, Cambridge, MA 02142-1209 USA journals-info@ mit. edu : MIT Press).

Chapitre II

Aevol : une plateforme idéale pour tester l'évolution de la complexité

1 Introduction

Aevol est un logiciel initialement développé au Laboratoire d'informatique en image et systèmes d'information (LIRIS) par Guillaume Beslon et Carole Knibbe, à partir de 2005, pour étudier l'évolution des génomes. Aevol se rattache à la famille des modèles de « génétique numérique » (ou *digital genetics*) (Adami, 2006) – voir chapitre précédent et section suivante). Il se rapproche de ce fait d'Avida (Adami, 1994; Ofria et Wilke, 2004) ou Polyworld (Yaeger, 1994). Il se distingue cependant de ces derniers par le haut degré de réalisme avec lequel il rend compte des mécanismes de transfert d'information : depuis le niveau génomique jusqu'au niveau phénotypique en passant par les niveaux transcriptomique et protéomique (la *genotype-phenotype map*). En résumé, il intègre un modèle relativement fidèle du « dogme de la biologie moléculaire » tel qu'énoncé par Francis Crick (1958). Comme nous le allons le montrer, cette caractéristique fait d'Aevol une plateforme idéale pour étudier l'évolution de la complexité des systèmes biologiques.

Les modèles de génétique numérique simulent, en un temps de calcul raisonnable, l'évolution d'une population d'organismes individuels dans un environnement contrôlé. Puisque l'usage typique de ces modèles est tout à fait comparable aux procédures de l'évolution expérimentale *in vivo*, on en est venu à parler d'évolution expérimentale *in silico*. Les populations d'organismes sont initialisées, soit à partir d'individus « naïfs » (généralement des individus générés au hasard, qui comportent un matériau génétique minimal pour « vivre »), soit à partir d'individus conçus « à la main », artificiels jusque dans l'intention qui les fait naître, pour qu'ils possèdent des propriétés spécifiques, soit encore à partir d'individus ayant déjà évolué durant un temps assez long de façon à ce qu'ils soient déjà le produit d'une évolution comme le sont tous les êtres biologiques qu'on peut rencontrer dans la nature (on parle alors de *wild-type* (WT)). Ces populations évoluent ensuite dans des conditions contrôlées, sans intervention de l'expérimentateur. C'est ensuite en observant le résultat de l'évolution, ainsi que sa dynamique évolutive, et en comparant leurs profils selon différents jeux de paramètres, que l'on peut déterminer les facteurs directs ou indirects qui entrent en jeu dans la structuration des organismes (figure II.1).

Voici donc pour les points qui rapprochent les branches *in vivo* et *in silico* de l'évo-

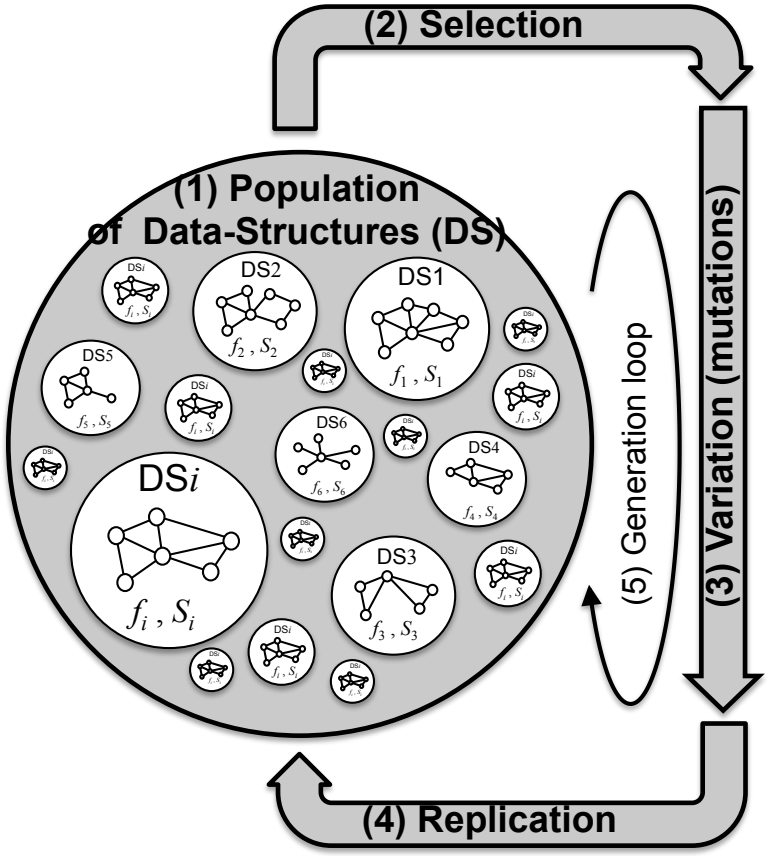


Figure II.1 – Fondement de l'évolution expérimentale *in silico*. (1) Le modèle s'appuie sur une population de N structures de données $(\mathrm{DS}_1, \mathrm{DS}_2, \cdots, \mathrm{DS}_i, \cdots, \mathrm{DS}_N)$, chacune dotée d'une fitness f_i et de propriétés systémiques quantifiables S_i. Ces structures de données sont assujetties à l'action d'un processus sélectif dont la force (2) dépend de la valeur f_i mais pas de la valeur S_i en général. Elles sont ensuite soumises à un processus mutationnel (3) qui modifie les structures de données (et donc les valeurs f_i et S_i) en même temps que les organismes se reproduisent pour constituer la nouvelle génération (4) Ces étapes de sélection, de mutation et de réplication forment ensemble une *boucle générationnelle* (5) On laisse alors la population évoluer en même temps qu'on mesure ses propriétés au cours de milliers de générations avant d'étudier la trajectoire des $f_i(t)$ et $S_i(t)$.

lution expérimentale. Il y a bien sûr des différences qui expliquent l'existence même de cette seconde branche. Comparée à l'évolution expérimentale *in vivo*, la rapidité des simulations permet de prendre en compte de très grands nombres de générations sur une durée d'expérience beaucoup plus courte (l'ordre de grandeur est de plusieurs centaines de milliers de générations par jour). Les échelles de temps simulées sont donc très différentes : il a fallu 22 ans à la *E. coli long-term evolution experiment* (LTEE) de Richard Lenski pour atteindre sa 50 000 ème génération, alors que les dernières versions d'Aevol permettent de simuler plusieurs millions de générations en 24h sur un ordinateur de bureau. Ainsi, alors qu'une expérience *in vivo* peut s'étendre sur des dizaines d'années, la

durée d'une expérience *in silico* se mesure habituellement en jours ou en semaines. Il faut mentionner aussi qu'une expérience d'évolution expérimentale *in vivo* coûte très cher en termes de matériel comme de main d'œuvre alors qu'une expérience *in silico* ne représente qu'un budget comparativement modeste, limité au fonctionnement des ordinateurs (pour autant, du moins, qu'on factorise la mise au point du simulateur comme une prestation extérieure à l'expérience). En outre, dans le cas de l'évolution expérimentale *in silico*, il est relativement aisé de multiplier le nombre de répétitions et ainsi d'obtenir une puissance statistique qu'il serait difficile d'atteindre en évolution expérimentale *in vivo*.

Mais l'avantage ne se limite pas là : là où les expériences *in vivo* permettent au mieux de conserver quelques échantillons congelés pour représenter longitudinalement l'expérience, la totalité des individus et leur généalogie est intégralement et exactement connue avec les expériences *in silico*. Pour le redire de façon imagée, alors que le calcaire permet à l'archéologue d'identifier quelquefois de beaux fossiles, notre silicium garde une collection *exhaustive* de fossiles *parfaits*, vestiges de tous les individus ayant existé (Hindré *et al.*, 2012). Vestiges qui peuvent en outre être réanimés ... sans avoir à y adjoindre de l'ADN de grenouilles (Spielberg *et al.*, 1993).

Un dernier avantage des expériences *in silico* mérite d'être traité à part : il s'agit de la possibilité de réaliser des expériences sur la base de principes qui ne peuvent être matérialisés dans un autre contexte – des « expériences impossibles » selon O'Neill (2003) [1]. C'est précisément ce que nous allons faire dans cette thèse, en tirant parti du fait que le formalisme abstrait des phénotypes Aevol permet de donner une définition concrète et opérationnelle aux notions de simplicité et de complexité.

Après un bref état de l'art de l'évolution expérimentale *in silico* (section suivante), nous décrirons de façon détaillée le modèle Aevol (section 3). Nous présenterons ensuite une brève revue des résultats significatifs qu'il a permis d'obtenir jusqu'à maintenant avant de revenir sur la question de l'intérêt du modèle Aevol pour l'étude de l'évolution de la complexité des organismes biologiques (section 4).

1. Dans un autre contexte, on peut se figurer les expériences de mécanique réputées *sans frottements* et songer au bénéfice théorique immense qu'il y a à ne considérer les frottements qu'en seconde analyse. L'approche expérimentale diffère cependant ici du fait que les phénomènes étudiés dans cette thèse ne se laissent pas circonvenir comme les expériences de mécanique classique par le fait de minimiser les frottements puis d'extrapoler.

2 État de l'art de l'évolution expérimentale *in silico*

2.1 Introduction

L'évolution expérimentale *in silico*, aussi appelée génétique numérique, est un champ de recherche qui s'est développé il y a une vingtaine d'années (Adami, 2006), même si certains modèles antérieurs pourraient aisément y être assimilés (Ray, 1991). En évolution expérimentale *in silico*, des populations d'organismes virtuels (d'un point de vue informatique, ce sont de simples structures de données) sont placées dans des environnements virtuels dans lesquels ils sont en compétition pour une ressource limitée. Selon leur succès, ils peuvent alors se reproduire ou non. Chaque organisme comporte une forme de matériau génétique que le simulateur de vie artificielle interprète pour déterminer un phénotype, lequel est soumis au processus de sélection. Enfin, ce matériau génétique est soumis à la variation, généralement au cours du processus de reproduction.

Du point de vue technique, les modèles de génétique numérique sont très proches des *algorithmes génétiques*. Ils en diffèrent cependant par leur finalité : l'état de la population finale et, en particulier, celui de son meilleur individu ne présentent en effet que peu d'intérêt pour la génétique numérique alors qu'il est la finalité même d'un algorithme génétique qui, comme algorithme d'optimisation, n'est intéressé que par un individu spécifique (généralement le meilleur individu obtenu à la dernière génération), indépendamment du processus par lequel cet individu a été obtenu. En d'autres termes, les algorithmes génétiques cherchent à résoudre un problème de conception : on soumet une population de solutions candidates au processus évolutif pour chercher une solution optimale à un problème. Du côté de l'évolution expérimentale *in silico* en revanche, l'intention est *d'étudier* le processus évolutif lui-même (la trajectoire évolutive suivie par la population plus que son point d'arrivée). L'enjeu d'un modèle de génétique numérique est donc de rassembler les éléments fondamentaux de l'évolution darwinienne pour être en mesure d'observer l'émergence des propriétés qui motivent l'étude mais, dans le même temps, de ne conserver qu'un ensemble limité de ces « éléments fondamentaux » afin de permettre l'interprétation des déterminismes qui sont à l'œuvre.

Qu'on me permette un commentaire sur l'expression « algorithme génétique » : il semble en effet regrettable qu'elle soit usitée dans une acception si restreinte. Il semblerait naturel qu'elle recouvre aussi bien les deux démarches puisque l'une comme l'autre se fondent sur des algorithmes qui mettent en œuvre des processus génétiques. Il est vrai qu'il faudrait alors occasionnellement compléter l'expression d'une épithète pour effectuer la distinction. Cela aurait malgré tout le mérite d'éviter les malentendus et de souligner la véritable différence : processus finaliste ou modèle computationnel. Point de vue latéral dans le deuxième cas, où c'est le processus lui-même qui est observé, la relation entre les générations, ou point de vue frontal dans le premier cas, où nous ne nous intéressons qu'au résultat, à la dernière génération.

Le schéma général de la génétique numérique est mis en œuvre de façons diverses selon les questions biologiques spécifiques auxquelles on veut répondre. Le processus évolutif est à l'œuvre aussitôt qu'une population est soumise à la sélection et à la variation (classiquement via un mécanisme mutationnel). Ce schéma est si général, si vaste dans son extension, que l'expérimentateur est contraint de commencer par un effort de modélisation qui implique de fixer son regard sur un aspect précis du phénomène et de s'y

limiter. En fonction de cette aspiration, il choisira un *formalisme* pour représenter le génome de ses organismes, leur phénotype ou les deux, l'idée étant ici que ce formalisme permette d'observer l'évolution des phénomènes étudiés (je devrais dire de la *scruter*). Le matériau génétique peut ainsi être représenté de façon plus ou moins réaliste comme une séquence de nucléotides, mais aussi comme une séquence d'instructions destinées à un processeur, ce peut être aussi un réservoir de gènes ou un réseau de régulation (Hindré *et al.*, 2012). L'important est que ce matériau génétique, qui (dans ce cadre) n'est qu'information, puisse ensuite être traité selon différentes formes de « chimies artificielles » (Dittrich *et al.*, 2001) pour être transformé en phénotype puis, par comparaison directe ou indirecte entre les phénotypes présents dans la population, en une valeur sélective (la *fitness*).

Une fois la représentation du matériau génétique arrêtée, il faut encore codifier sa dynamique en décidant du mécanisme de sélection et des opérateurs mutationnels.

Le mécanisme de sélection déterminera, sur la base de la fitness, la façon précise dont le patrimoine génétique se transmet d'une génération à la génération suivante. Pour le dire autrement, il s'agit de décider des critères selon lesquels la compétition entre individus s'organisera et quels seront ceux qui auront la possibilité de transmettre leur génome à la génération suivante via la réplication.

Au cours de la réplication une part du patrimoine génétique d'une génération est ainsi transmise à la génération suivante sous la forme d'un héritage génétique venu de reproducteurs sélectionnés selon un critère prédéfini. C'est à ce stade qu'intervient le deuxième processus fondamental de l'évolution Darwinienne. En effet, cet héritage ne se fait pas à l'identique : au cours de la réplication, le matériau génétique subit des altérations lors de sa transmission, ce que l'on désigne sous le nom de *mutations*. Elles sont modélisées sous la forme d'opérateurs mutationnels qui décrivent sous forme algorithmique les familles de mutations observées en biologie.

Dans les paragraphes suivants, les différentes familles de modèles sont regroupées selon la nature du formalisme génétique à partir duquel elles sont construites, c'est-à-dire selon la nature de la structure de données qui subit les variations mutationnelles : instructions pour un ordinateur virtuel, traits morphologiques, liens d'un réseau, perles représentant des gènes fonctionnels ou encore nucléotides.

2.2 Le formalisme « génome-programme »

Inspirés par la métaphore informatique, les premiers modèles de génétique numérique représentaient le « code génétique » comme un programme. Dans cette famille de modèles encore très utilisée aujourd'hui, les individus rivalisent pour gagner des fractions du temps de calcul alloué à leur génération au sein d'un ordinateur virtuel. Le génome est une séquence d'instructions d'un langage de programmation rudimentaire qui codent le programme exécuté au cours de la vie de l'individu, l'objectif de ce programme étant, à tout le moins, de s'auto-répliquer (c'est-à-dire de recopier son propre code) et, dans certains modèles, de calculer des fonctions logiques ou mathématiques. Comme le génome est représenté par une séquence fonctionnelle, ces modèles sont bien adaptés à l'étude de l'évolution de ce type de séquences – la comparaison immédiate ici serait l'évolution de séquences d'ARN dans un « monde d'ARN » (Gilbert, 1986). En revanche, comme la

structure de ces génomes est très différente de la structure d'un génome réel (du fait de l'absence de modèles de la transcription et de la traduction), il est difficile de les comparer aux génomes réels.

Avec Tierra (Ray, 1991) qui fait figure de pionnier dans cette catégorie de modèles, comme dans la génétique numérique en général, les organismes ne sont pas évalués au regard de leur succès dans l'accomplissement d'une tâche prédéfinie, ce qui fait de ce modèle un exemple d'évolution *open-ended* (Taylor *et al.*, 2016; Banzhaf *et al.*, 2016). En effet, les organismes rivalisent directement entre eux pour leur survie en faisant évoluer leur façon de se reproduire. De ce fait, comme dans le cas de l'évolution naturelle, il n'est pas nécessaire de faire intervenir une grandeur quantitative (explicite) de fitness : comme en évolution expérimentale *in vivo* la fitness n'apparaît ici que comme un outil de mesure de la valeur sélective pour l'expérimentateur. De plus, outre les optimisations que l'évolution a pu trouver (permettant une diminution du temps de réplication), les expériences menées au moyen de Tierra ont donné lieu à l'apparition spontanée de parasites et d'hyperparasites, créant ainsi une forme d'écosystème (Ray, 1991, 1992). En ce sens, Tierra peut être vu comme un modèle permettant l'accroissement de la complexité écologique.

C'est aujourd'hui à Avida (Adami et Brown, 1994; Adami, 2006) que revient le titre de leader de cette catégorie de modèles (et même de la génétique numérique dans son ensemble). Contrairement à Tierra, les organismes sont ici séparés les uns des autres, ce qui les protège d'attaques de parasites (lesdits parasites ne peuvent purement et simplement pas exister). Comme dans Tierra, les organismes ont pour tâche principale l'autoréplication, mais il peuvent aussi effectuer des opérations logiques, ce qui leur permet de gagner du temps de calcul. Au contraire de Tierra, Avida n'est donc pas un modèle d'évolution *open-ended*. De nombreux travaux se sont appuyés sur Avida pour s'intéresser en particulier à l'évolution de la robustesse (Wilke *et al.*, 2001; Wilke et Adami, 2003; Elena et Sanjuan, 2008)), mais aussi à l'évolution de la modularité des génomes (Misevic *et al.*, 2006) ou à la radiation (Chow *et al.*, 2004).

De façon intéressante pour notre étude, Avida a aussi été mis en œuvre pour montrer que des traits complexes peuvent apparaitre sous l'effet conjoint de mutations aléatoires et de la sélection naturelle (Lenski *et al.*, 2003). En partant d'organismes virtuels capables seulement de se reproduire, l'évolution fait graduellement apparaître des fonctions de complexité croissante. La mesure de fitness est ici basée sur la capacité des individus à produire 9 fonctions logiques de complexité croissante, au sens où elles nécessitent pour être codées un nombre d'instructions croissant.

Les expériences conduites exhibent deux phénomènes intéressants. D'une part, les fonctions complexes naissent à partir des fonctions simples au prix de peu de mutations alors que le nombre total de mutations (si l'on remonte jusqu'à l'individu naïf) est beaucoup plus important, ce qui illustre l'aspect graduel de l'évolution et le fait que les traits complexes n'apparaissent pas *de novo*. D'autre part, il a été constaté que les lignées ont à franchir des mutations délétères avant de parvenir à une fonction complexe. Autrement dit, l'évolution ne suit pas un processus simplement itératif.

Cette étude montre ainsi que la complexité peut apparaître au gré de l'évolution et dans quelles circonstances. Cependant, pour parvenir à cette conclusion, la fitness des individus récompense leur capacité à produire les fonctions logiques complexes qui sont attendues. C'est précisément là qu'une critique méthodologique doit être formulée : la

simulation montre l'apparition de fonctions complexes que la mesure de fitness favorise explicitement. Il en ressort finalement que la preuve qui est apportée porte essentiellement sur la possibilité, pour la sélection darwinienne, de donner naissance à des systèmes complexes. En ce sens, (Lenski *et al.*, 2003) apparaît plus comme un travail militant et anti-créationniste que comme une étude causale de l'origine de la complexité biologique.

Si Tierra et Avida utilisent tous deux un modèle de génome-programme basé sur des instruction machines simplifiées (pseudo-assembleur), Musso et Feverati (2012) ont aussi proposé un modèle utilisant des « génomes-programmes » mais en utilisant un génome représentant une machine de Turing. Ils ont ainsi montré que la quantité de code actif que la sélection est en mesure de maintenir possède une borne supérieure – le seuil d'erreur (Eigen, 1971) – et que quel que soit le taux de mutation imposé aux organismes, l'évolution pousse la proportion de séquences codantes vers le seuil d'erreur correspondant à ce taux de mutation.

2.3 Évolution de structures morphologiques et comportementales

En dehors du champ de l'évolution moléculaire – en particulier auprès du grand public – l'évolution est généralement vue comme un processus sur les caractéristiques morphologiques et comportementales des espèces, avec une attention particulière portée aux vertébrés puisque ce sont les espèces pour lesquelles la morphologie et le comportement nous sont les plus faciles à appréhender. Vue depuis une discipline telle que l'informatique, relativement étrangère à l'évolution moléculaire, il est donc assez naturel que la modélisation de l'évolution se soit souvent focalisée sur ces deux éléments que sont la morphologie et le comportement.

Lorsqu'il s'agit d'étudier le développement de traits morphologiques et comportementaux ainsi que leurs relations à l'échelle du phénotype, on peut s'abstraire des considérations moléculaires et coder plus ou moins directement les « organismes ». Les deux éléments constitutifs du modèle sont alors les structures morphologiques et leurs relations et ce sont leurs caractéristiques qui sont encodées par le génome. Le comportement de ces organismes est lui aussi encodé par le génome (via les relations entre les structures morphologiques) et il évolue donc parallèlement aux propriétés morphologiques. Les premières simulations de ce type sont dues à Karl Sims (1994a) avec son modèle « Creatures ». Ce principe a ensuite été repris dans de nombreux modèles dont« Framasticks » (Komosiński et Ulatowski, 1999). Ces deux modèles ont permis d'étudier l'évolution de la morphologie, en particulier dans le cas où deux organismes coévoluent, que ce soit par le biais de la compétition ou via des interactions proie-prédateur (Sims, 1994b; de Back *et al.*, 2006)).

Ce formalisme est très en vogue dans la communauté de la vie artificielle car il a l'avantage de conduire à des simulations très visuelles et relativement convainquantes. C'est le cas, par exemple des projets GOLEM (Pollack et Lipson, 2000), Blindbuilder (Devert *et al.*, 2006) ou des robots évolutifs de Josh Bongard (Bongard et Paul, 2001; Bongard, 2010)). Cependant, si l'objectif est d'étudier le processus évolutif et ses propriétés, ces modèles souffrent du manque de réalisme des génomes. Ceux-ci sont en effet très spécifiques et utilisent des structures de données qui n'ont que très peu de points communs avec la structure d'un génome biologique. Ainsi, dans « Creatures » (Sims, 1994b,a), le génome est un graphe dirigé récursif dont les nœuds représentent les structures morphologiques et les arêtes représentent les articulations. Ce choix, bien qu'il permette d'observer

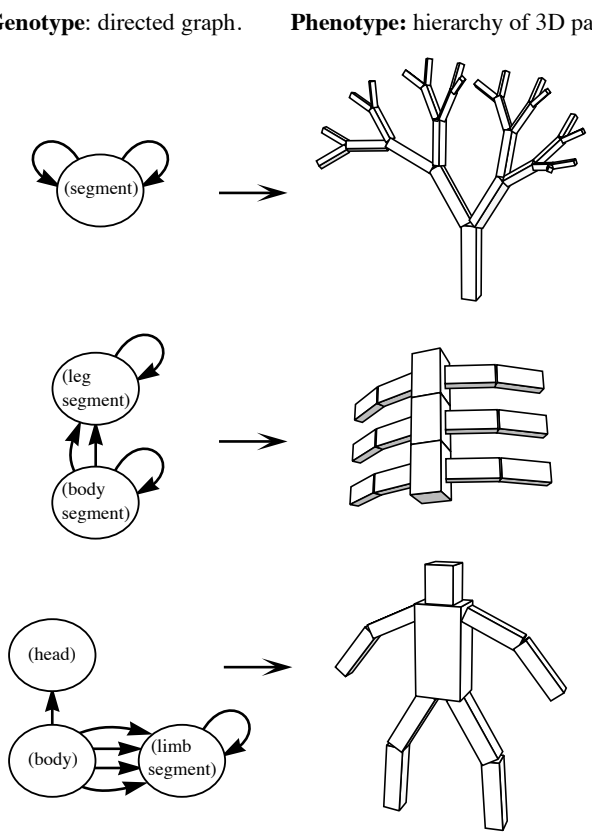


Figure II.2 – Exemples d'organismes ayant évolué dans le modèle « Creatures » (Sims, 1994a). Ces trois exemples de graphes et les organismes correspondants illustrent l'écart entre la notion de complexité si on se place du point de vue du phénotype ou du génome. L'apparente complexité du premier « organisme » traduit en fait un génome très simple. Figure d'après (Sims, 1994a).

rapidement l'apparition de structures morphologiques impressionnantes (figure II.2) entraîne une telle dissociation entre la complexité du génome et la complexité phénotypique qu'il est difficile de donner une interprétation biologique aux simulations. De fait, malgré les nombreux travaux ayant exploité ce formalisme à la suite des travaux fondateurs de Karl Sims, force est de constater que ces modèles n'ont quasiment eu aucune retombée en biologie ou en biologie évolutive.

2.4 Le formalisme « réseau »

Avec le développement de la biologie des systèmes au début des années 2000 (Ideker *et al.*, 2001; Kitano, 2002), la vision « en réseau » (réseaux de gènes, réseaux métaboliques, réseaux trophiques, etc) des systèmes biologiques s'est rapidement développée (Barabási et Oltvai, 2004) et a conduit à l'émergence de nouveaux questionnements en biologie évolutive et en évolution moléculaire. Il ne s'agissait alors plus de comprendre comment un élément du système évolue (par exemple en le comparant à un élément similaire mais présent chez d'autres organismes), mais bien de comprendre comment l'évolution structure les

interactions entre un grand nombre d'éléments dans une perspective dite d'« *Evolutionary Systems Biology* » (Soyer et O'Malley, 2013).

Lorsqu'il s'agit de s'intéresser à l'évolution d'un réseau de régulation génétique ou à l'évolution d'un réseau métabolique, chaque organisme doit comporter son propre réseau de régulation. La solution la plus simple est alors d'adopter un formalisme dans lequel le génome représente directement le réseau étudié (régulation ou métabolisme). On peut pour cela coder le génome sous la forme d'une matrice d'adjacence qui stocke l'intensité des connexions entre gènes (respectivement métabolites) ou au travers de fonctions attribuant indirectement des pondérations aux arcs du réseau. Les mutations sont alors codées comme des modifications aléatoires de certains éléments de la matrice. Ce type d'approche, très utilisé en biologie computationnelle, a permis d'étudier l'évolution de propriétés structurelles des réseaux biologiques telles que la modularité (Kashtan et Alon, 2005; Kashtan *et al.*, 2009; Espinosa-Soto et Wagner, 2010; Clune *et al.*, 2013).

Dans ces modèles, le réseau de régulation est donc codé explicitement. Les variations structurelles du réseau se produisent lors de la reproduction et nécessitent le recours à des opérateurs de mutation spécifiques qui font varier les coefficients de régulation ou, plus rarement, provoquent l'apparition ou la suppression de nouveaux liens de régulation (généralement via l'introduction d'une valeur nulle dans la matrice d'adjacence). En revanche, hormis les modèles de type NEAT (Stanley et Miikkulainen, 2002), le nombre de nœuds dans les réseaux de régulation est généralement fixé, ce qui empêche d'y recourir pour étudier l'évolution de la complexité moléculaire (même si, en théorie, elle peut s'exprimer au travers du nombre de liens du réseau).

Le formalisme réseau permet de faire facilement évoluer des réseaux d'interaction. Cependant il est important de noter que ceci se fait, là encore, au prix d'une simplification importante. En effet, les réseaux biologiques sont des abstractions de processus moléculaires beaucoup plus complexes. Une cellule « réelle » ne contient pas de réseau explicite, puisque rien, ni dans le génome, ni dans le transcriptome, ne code directement l'influence d'un facteur de transcription sur un gène. Celle-ci est issue de l'interaction entre la séquence du facteur de transcription (*Trans*) et le promoteur du gène (*Cis*) – et cette description est elle-même une simplification outrancière puisque de nombreux mécanismes différents sont à l'œuvre dans la régulation. De ce fait, le formalisme réseau simplifie considérablement les mécanismes moléculaires et leur dynamique mutationnelle en rendant compte de l'effet cumulé des différents niveaux d'organisation des réseaux biologiques. L'interprétation biologique de ces opérateurs est alors moins immédiate que celle des opérateurs réalistes utilisés par exemple dans les modèles en collier de perle (section suivante) ou, à fortiori, dans le formalisme « séquence de nucléotides » (section 2.6) qui utilise une représentation quasi-littérale des mécanismes de transcription-traduction.

2.5 Le formalisme en « collier de perles »

Partant du constat que le formalisme réseau ne permet pas une représentation suffisamment explicite des événements mutationnels modifiant un réseau moléculaire (qu'il s'agisse d'un réseau de gènes ou d'un réseau métabolique), Anton Crombach et Paulien Hogeweg, du groupe de bioinformatique et de biologie théorique de l'université d'Utrecht (NL), ont développé un formalisme alternatif permettant de représenter un réseau codé sous la forme d'une séquence d'éléments fonctionnel (Crombach et Hogeweg, 2007, 2008;

Ten Tusscher et Hogeweg, 2009; Crombach et Hogeweg, 2009). Dans ce formalisme – pour lequel la dénomination « collier de perles » a été proposée ultérieurement par Hindré *et al.* (2012) mais adoptée depuis par Paulien Hogeweg – le génome est représenté sous la forme d'une séquence ordonnée et cyclique (« un collier ») d'éléments fonctionnels (« les perles ») de différents types (figure II.3) . On trouve ainsi des perles de type « gène », « facteur de transcription », « promoteur », « élément répété », « rétrotransposons », etc. Chacun de ces types de perles est représenté sous la forme d'un n-uplet dont les valeurs indiquent les caractéristiques de l'élément fonctionnel (ainsi, pour un gène sa fonction, pour un promoteur l'identification du site de fixation correspondant, pour un facteur de transcription son site de fixation et son activité régulatrice, etc). Ce formalisme permet donc de représenter des réseaux biologiques (dont le type dépend des types de perles utilisés) via le décodage des interactions entre les perles, tout en représentant explicitement le processus de mutation en *Cis* (mutation d'une perle promoteur) ou en *Trans* (mutation d'une perle facteur de transcription). En outre, comparativement au formalisme réseau, il autorise des variations topologiques du réseau via l'ajout ou la suppression de perles : l'ordre des gènes comme leur nombre et leur degré de régulation peuvent évoluer au gré des mutations et des réarrangements.

Ce formalisme a été utilisé à l'université d'Utrecht pour étudier l'évolution de l'évolvabilité (Crombach et Hogeweg, 2007), l'évolution des réseaux de régulation dans des environnements cycliques (Crombach et Hogeweg, 2008), la spéciation (Ten Tusscher et Hogeweg, 2009) ou la rotation des ressources dans les écosystèmes (Crombach et Hogeweg, 2009). Plus récemment le formalisme en collier de perles a été utilisé par l'équipe Inria Beagle à Lyon (FR) pour étudier la spéciation dans des environnement cycliques (Rocabert *et al.*, 2017). L'équipe Beagle a par ailleurs étendu le formalisme pour développer un nouveau modèle d'algorithme évolutionnaire (Peignier *et al.*, 2015, 2020).

Comme le formalisme en collier de perles permet d'observer des variations topologiques des réseaux moléculaires (via la variation du nombre de perles sur le génome), il peut permettre d'étudier l'évolution de la complexité en fonction des contraintes environnementales. Ainsi Rocabert *et al.* (2017) ont pu observer que, lorsque deux espèces co-évoluent dans un environnement cyclique et qu'une espèce consomme la ressource primaire présente dans l'environnement alors que l'autre consomme les déchets produits à cette occasion, la première espèce présente un génome et un réseau métabolique complexe tandis que le génome et le réseau métabolique de la seconde tend à se simplifier, ce qui peut rappeler les dynamiques d'évolution réductrice observées chez les bactéries endosymbiotiques (Batut *et al.*, 2014). Cependant, la représentation du processus mutationnel reste très simplifiée ce qui est susceptible de biaiser les résultats. Ainsi, ce formalisme ne permet pas de ternir compte des séquences non codantes, alors même que ces séquences contribuent significativement aux variations de taille des génomes (Elliott et Gregory, 2015) et qu'il a été montré qu'elles biaisent significativement plusieurs mécanismes de variation (Biller *et al.*, 2016a). En outre, les mécanismes de variation des séquences fonctionnelles (et donc de la topologie des réseaux) sont considérablement simplifiés par le principe du codage par n-uplets. Cela représente un avantage indéniable en termes de calculabilité (le génome étant très simple à décoder) mais risque de biaiser les caractéristiques structurelles des réseaux, donc leur complexité.

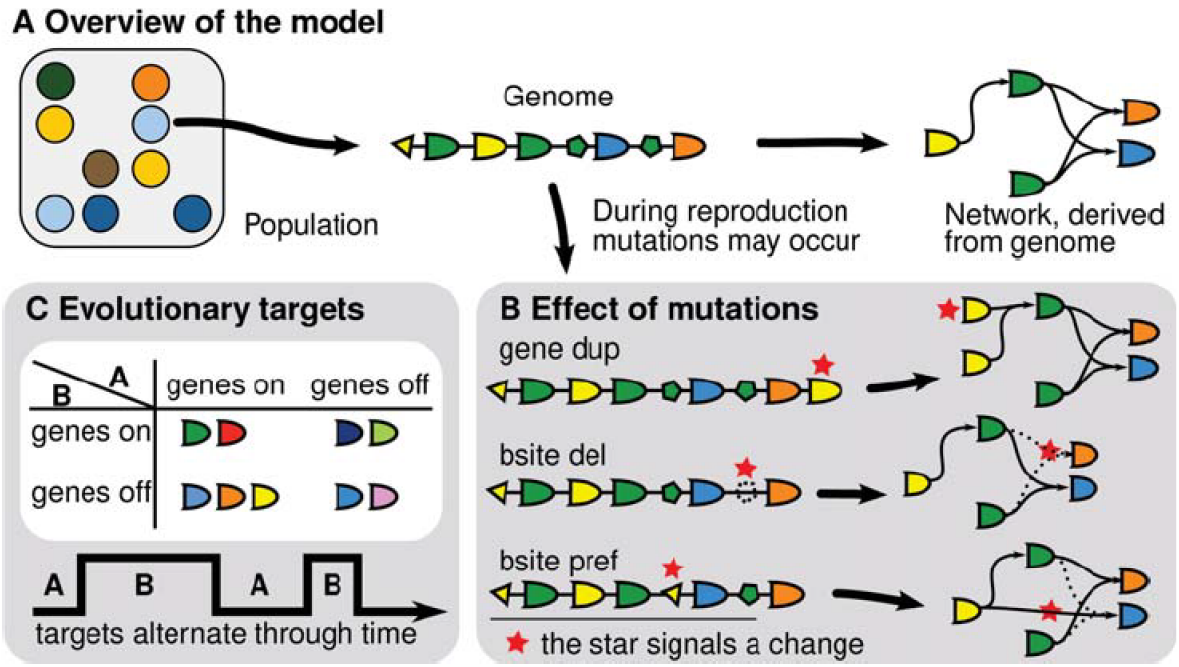


FIGURE II.3 – Le formalisme « collier de perles ». **A.** principe de codage des individus : chaque individu de la population contient un génome composé d'une séquence d'élements fonctionnels (ici des facteurs de transcription et des promoteurs représentés respectivement par une demi-ellipse et par un pentagone) et identifiés par par un entier (ici représenté par leur couleur). Lorsqu'un promoteur se situe immédiatement en amont d'un gène, il régule son activité en fonction de la présence ou non du facteur de transcription correspondant. Il en résulte un réseau qui, contrairement au formalisme « réseau » est ici codé indirectement sous la forme d'une séquence. **B.** Le codage sous la forme d'une séquence permet de modéliser les mutations de façon plus réaliste que le formalisme réseau. **C.** La sélection est basée sur l'activation des gènes. Ici les organismes peuvent rencontrer deux types d'environnement qui sont représentés indirectement par l'ensemble des gènes qui doivent être activés ou inhibés. Figure d'après (Crombach et Hogeweg, 2008).

2.6 Le formalisme « séquence de nucléotides »

Le formalisme « séquence de nucléotide » a été développé à l'Université de Lyon parallèlement – et indépendamment – au formalisme « collier de perles » mais pour répondre à la même problématique, à savoir représenter finement le processus de variation mutationnelle (qui se déroule donc sur un support séquentiel – l'ADN) indépendamment de la représentation fonctionnelle d'un organisme (qu'il s'agisse d'un réseau moléculaire ou de toute autre forme de représentation).

Le formalisme séquence de nucléotique part du principe que l'information portée par un élément nucléotidique de la séquence génétique n'est fonctionnelle que dans le contexte des nucléotides qui l'entourent. Par exemple, pour peu qu'elle ne soit pas précédée d'un promoteur puis d'un codon START et suivie, en outre, par un codon STOP puis par un terminateur, une séquence nucléotidique n'aura aucun effet sur le phénotype. Ainsi, un

« gène », dans sa définition la plus élémentaire, est une séquence traduite en protéine, ce qui signifie que les mécanismes de transcription et de traduction sont effectifs sur cette séquence particulière, impliquant que celle-ci soit entourée des séquences signal reconnues par les complexes moléculaires chargés de décoder l'information (polymérases, ribosomes, etc). De même, à l'échelle du gène lui-même, le rôle fonctionnel d'un codon particulier (donc le rôle fonctionnel d'un acide aminé particulier dans la protéine produite par ce gène), ne dépend pas (ou pas exclusivement) de sa *nature* mais aussi de son *contexte*, c'est-à-dire des codons qui l'entourent.

Sur la base de ce constat, Carole Knibbe et Guillaume Beslon ont proposé le formalisme séquence de nucléotides et le modèle Aevol (Knibbe, 2006; Knibbe *et al.*, 2007a,b). Contrairement aux formalismes précédent, dans lesquels les éléments fonctionnels sont définis intrinsèquement, ils sont ici identifiés par des séquences de signalisation de différentes natures (dans le modèle Aevol, on trouve ainsi des signaux promoteurs, Ribosome Binding Site, Start, Stop et terminateur). Dans ce formalisme, la structure génétique est donc structurellement fidèle à la structure d'un génome biologique ce qui permet de modéliser explicitement les opérateurs de variation opérant sur les génomes réels.

Dans la mesure ou le formalisme « séquence de nucléotides » est exploité par le modèle Aevol et que celui-ci sera décrit *in extenso* dans la section suivante, nous n'entrerons pas ici dans une description plus détaillée de ce formalisme et renvoyons le lecteur à la description du modèle pour de plus amples informations.

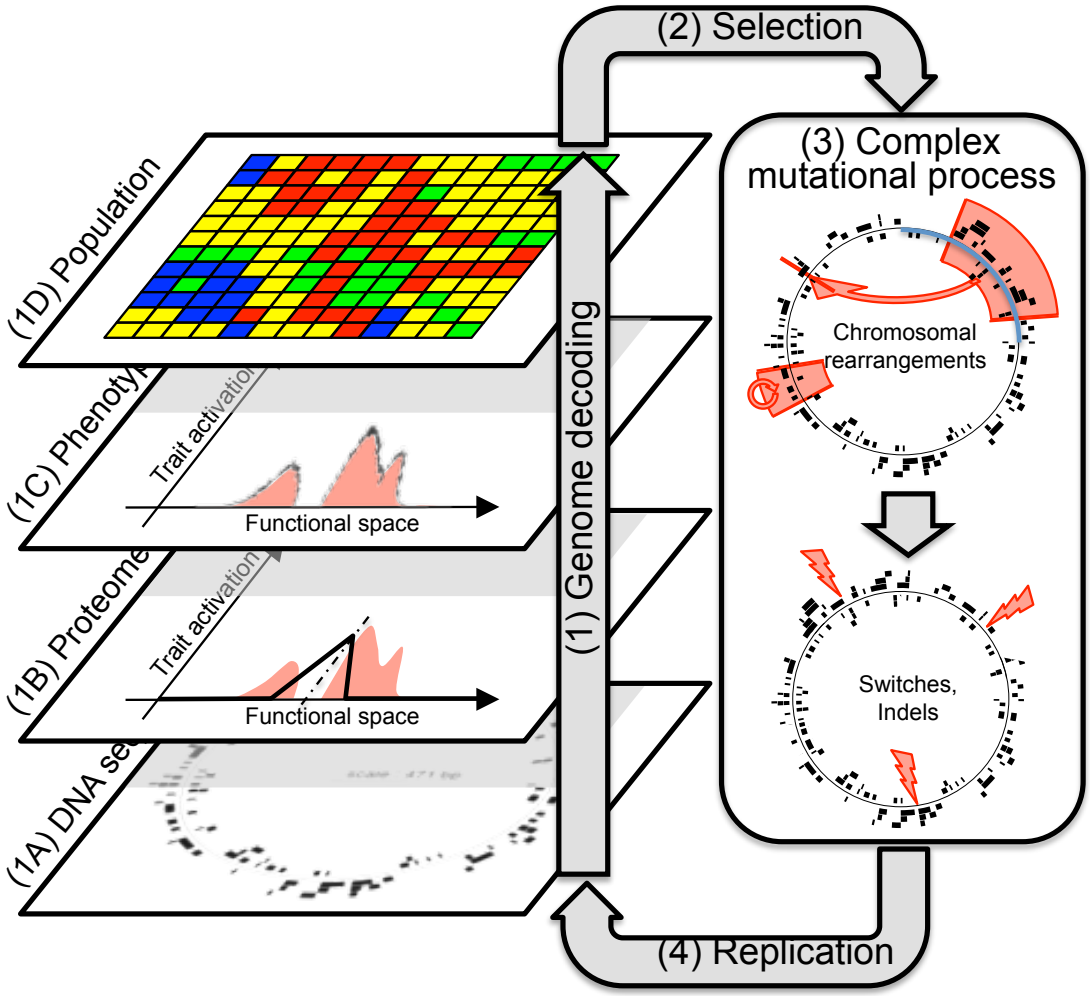


Figure II.4 – Le modèle Aevol. Les génomes (1A) comportent des gènes qui sont décodés en un ensemble de protéines. Les niveaux fonctionnels (protéines et phénotypes) recourent à une abstraction mathématique dans laquelle les protéines sont représentées par des fonctions triangles (1B) et les phénotypes sont calculés comme somme des toutes les protéines-triangles (1C). La fitness est calculée selon le principe qu'il s'agit d'une tâche d'approximation de courbe, plus le phénotype est proche d'une cible phénotypique donnée ("Phenotypic Target" (PT), en rouge sur 1B et 1C) plus organisme est réputé adapté, "fit". Aevol se fonde sur une boucle générationnelle avec un processus mutationnel qui comporte notamment les réarrangements parmi les types de mutations disponibles (3).

3 Description du modèle

3.1 Organisation générale

Lorsqu'il s'agit de décrire un modèle tel qu'Aevol, trois difficultés surgissent. La première tient à l'ambiguïté du vocabulaire si l'on désigne par le même mot la chose modélisée et le modèle associé. Ainsi le mot « nucléotide » est-il utilisé tantôt pour désigner le concept biologique, tantôt la structure de données par laquelle celui-ci est représenté dans

Aevol, à savoir un bit (seulement *un bit* car l'ADN dans Aevol est codé sur deux bases). Si l'on déplore cette ambiguïté, qu'on songe à la définition du terme ADN : *séquence de nucléotides* : elle peut aussi bien s'appliquer au concept biologique qu'au modèle, voire au formalisme de modélisation lui-même (voir section précédente). On constate alors que cette ambiguïté comporte au moins le double mérite de souligner l'intime parallèle qui relie le modèle et la chose modélisée et de permettre une communication facilitée avec « les biologistes ».

La deuxième difficulté tient à la nécessité de bien distinguer ce qui relève de la formulation explicite du modèle, et qui à ce titre est invariable dans le modèle, de ce qui n'est contraint que par l'opération des mécanismes du modèle. Comme exemple du premier, on peut citer le mécanisme de transcription de l'ADN en ARN qui est explicitement stipulé, ainsi que la séquence promotrice ou la séquence *hairpin* des terminateurs (voir section 3.2.3). À l'inverse, la taille des génomes ou la longueur des ARN ne sont pas fixées par le modèle Aevol, et n'apparaissent que comme un effet de la dynamique évolutive. De même, le nombre de séquences promotrices et terminatrices, ainsi que leurs positionnements respectifs le long du génome sont contraints par les mécanismes du modèle mais libres d'évoluer dans la limite – souvent inconnue – de ces contraintes. C'est précisement là l'intérêt de cette démarche de modélisation : en observant le résultat de l'évolution sur ces éléments, on peut observer ces contraintes donc remonter à leur origine évolutive, celle-ci étant nécessairement expliquée par tout ou partie des éléments du modèle.

De façon plus subtile, la troisième difficulté tient à distinguer les variables des paramètres. Les taux de mutation *spontanés* par exemple sont des paramètres et demeurent constants sur la durée d'une expérience (dans Aevol les taux de mutation sont exprimés, pour chaque type de mutation, en nombre de mutations par nucléotide et par génération). En revanche, le nombre de mutations *fixées*, c'est-à-dire le nombre de mutations présentes dans une lignée à l'issue de l'évolution est susceptible d'évoluer en fonction des contraintes évolutives. Là encore c'est cette distinction qui permet d'utiliser le modèle pour expliquer l'évolution des variables par rapport aux paramètres ou de définir des plans d'expériences permettant de jouer sur les variables et d'en étudier les conséquences sur le modèle.

Ayant ces trois notions présentes à l'esprit, nous pouvons poursuivre avec une vue d'ensemble du modèle Aevol. Pour cela, nous adopterons simplement le point de vue du cycle de vie d'un système biologique. Dans, Aevol chaque individu est complètement caractérisé par son ADN, lequel est ici une séquence circulaire de nucléotides binaires. L'ADN subit un processus de transcription qui produit des ARN messagers (ARNm). Chaque ARNm subit à son tour un processus de traduction qui produit des protéines (si l'ARNm est codant). L'ensemble des protéines ainsi produites donnent le phénotype de l'individu. Ce phénotype est l'élément central de l'évolution dans la mesure ou c'est l'interface au travers de laquelle un individu se trouve en contact avec son environnement. Et c'est précisément l'adéquation d'un individu avec l'environnement auquel il est confronté qui détermine son succès reproductif et sa *fitness*. De fait, c'est la bonne adaptation à l'environnement (conduisant souvent à considérer la fitness comme un « degré d'adaptation ») qui entraîne sa capacité à se reproduire et, par voie de conséquence, sa réussite évolutive.

Cette description, bien qu'elle soit fidèle, n'apprend encore que peu de choses sur Aevol. Elle donne cependant déjà une notion à gros grain des principaux aspects de la *genotype-phenotype map* d'un organisme biologique qui ont été retenus dans le modèle. Les sections suivantes permettront de donner une substance spécifique aux termes « ADN », « ARN »,

« protéine », « phénotype » et « fitness » tels qu'ils sont utilisés dans le modèle. Cette description va nous donner l'occasion de constater que le degré d'abstraction augmente brutalement avec l'étape de la traduction. Nous détaillerons ensuite le processus évolutif (variation-sélection) tel qu'il est modélisé dans Aevol.

3.2 Le codage de l'information dans Aevol

3.2.1 Introduction

Un modèle tel qu'Aevol se compose de deux éléments conceptuellement distincts, mais indissociables en pratique. Il s'agit d'abord d'un modèle mécaniste qui propose une représentation simplifiée et opérationnelle de la théorie de l'évolution. Il s'agit ensuite d'une implémentation de ce modèle sous la forme d'un simulateur. Il n'est pas inutile de faire cette distinction dans la mesure où la réflexion théorique que permet Aevol s'appuie sur les résultats obtenus grâce au simulateur mais que ces derniers sont interprétés à la lumière des hypothèses du modèle mécaniste. Ceci met en évidence le fait que le modèle et son implémentation sont censés être en parfaite adéquation, mais que cette équivalence est le fruit d'un processus humain et informatique complexe composé de développement logiciel, du recours à de nombreuses briques logicielles exogènes (*p. ex.* librairies et compilateur) et de calculs effectifs, chacun de ces niveaux présentant nécessairement des imperfections relativement au modèle mécaniste. Par ailleurs, comme nous l'avons déjà évoqué ci-dessus, dans la description du modèle, les noms utilisés sont ceux des objets biologiques modélisés. Par exemple, ce qui est désigné par génotype est en fait un *modèle de génotype*. Il ne s'agit pas de la chose mais de sa représentation, de même le terme « nucléotide » désigne en général la représentation d'un nucléotide réel dans le cadre du modèle, à savoir, du point de vue formel, un bit. Cela permet de garder à l'esprit ce qui est envisagé par le formalisme. Il n'y a pas d'ambiguïté tant qu'on garde à l'esprit que c'est ici le modèle qui est décrit, et donc le fruit de choix et d'hypothèses. Une fois passées les descriptions les plus élémentaires, les termes utilisés pourraient donc prêter à confusion et donner à penser qu'il s'agit d'une description biologique. Qu'on se garde pourtant de les confondre car le modèle ne constitue en aucun cas une description biologique exhaustive ni même une simplification à visée pédagogique.

Une deuxième mise en garde importe avant de décrire le modèle. Bien que celui-ci décrive une réalité biologique, dans le cadre d'une démarche de modélisation celle-ci ne doit en effet en aucun cas être exhaustive : le modélisateur se contentera de rechercher une coïncidence entre le modèle et un échantillon de phénomènes biologiques afin de mettre en relief certains phénomènes spécifiques et certaines questions qui suscitent son intérêt. Il en découle, d'une part, que le modèle n'est pas exportable en dehors du cadre des questions pour l'étude desquelles il a été conçu, d'autre part que certains éléments de la réalité physique ou biologique devront nécessairement être simplifiés, voire négligés (c'est à dire absents du modèle) afin de permettre, sur le modèle, un raisonnement mécaniste qu'il est impossible d'échafauder sur le système réel. On devra en particulier identifier une ou plusieurs échelles d'intérêt de façon à éviter la surcharge du modèle. Le modèle informatique relèvera donc de choix de phénomènes et d'éléments considérés comme prépondérants et d'un équilibre entre trois pôles : *la simplicité* du modèle afin que les résultats de simulations puissent être interprétés, *le réalisme* biologique afin que les

résultats aient une signification biologique qui permette le dialogue avec la communauté des biologistes, et *la calculabilité* effective, c'est à dire la possibilité pratique d'effectuer des simulations avec les contraintes contingentes que sont des moyens matériels donnés et les agendas académiques (durée d'une thèse, échéances de conférences pour ne citer que deux exemples).

Dans notre cas, Aevol a pour but d'éclairer les phénomènes en jeu dans l'évolution Darwinienne. L'accent est donc mis sur les éléments biologiques dont l'incidence est la plus probable au regard des connaissances théoriques ou des résultats expérimentaux, au détriment de facteurs *supposés* exogènes. Ainsi, dans Aevol, les populations d'organismes sont-elles représentées par une collection d'individus possédant chacun une identité propre caractérisée par son génome. Celui-ci est décodé par un ensemble de mécanismes explicites directement calqués sur le « dogme de la biologie moléculaire [1](Crick, 1968) » et les variations génétiques sont traitées de façon très réaliste, dépassant le cadre classique qui se limite aux seules mutations ponctuelles pour inclure en particulier des événements de réarrangement (variants structuraux) tels que les translocations, les duplications et les inversions de séquences arbitrairement longues. Ces choix généraux reposent sur l'idée que la dynamique évolutive est grandement dépendante de la relation entre le mode de codage de l'information sur le génome (donc de son processus de décodage – le dogme) et les opérateurs de variation opérant sur ce même génome. En d'autres termes, Aevol repose sur l'idée que la structure macroscopique du *fitness landscape* de tout organisme dépend fondamentalement de ces deux éléments que sont la structure du codage et son mode de variation. En revanche, comme énoncé ci-dessus, il sera nécessaire de simplifier d'autres éléments de la réalité biologique afin de garantir l'interprétabilité du modèle et sa calculabilité. En l'occurrence, dans Aevol, les mécanismes propres aux niveaux fonctionnels (le phénotype, sa relation à l'environnement, les interactions inter-individuelles, etc) sont le fruit d'une très forte abstraction – nous verrons que, dans notre cas, un phénotype ou un environnement seront abstraits sous la forme de fonctions mathématiques. De même, les aspects biochimiques et moléculaires seront traités de façon mécaniste via un ensemble de règles algorithmiques – ce qui aura pour conséquence directe que ces règles ne pourront pas évoluer dans la modèle.

3.2.2 Description générale

Dans Aevol, chaque individu possède un génome qui contient l'information génétique transmissible à ses descendants. Le génome est une séquence binaire double brin (les deux brins étant complémentaires). Il est décodé en deux étapes : la *transcription*, qui produit l'ARN à partir de l'ADN, et la *traduction* qui produit les protéines à partir de l'ARN. Le processus de transcription s'appuie sur des signaux consensus (les promoteurs) et sur des structures de type *hair-pin* (les terminateurs) respectivement pour l'initialisation et la terminaison de la transcription. De même, le processus de traduction met en jeu des sites d'initialisation consensus (Ribosome-Binding-Sites (RBS)) et un code génétique artificiel basé sur des codons formés de triplets (qui comportent notamment les codons Start et Stop). La séquence de codons d'un gène forme ensuite la structure primaire

1. Le « dogme de la biologie moléculaire (Crick, 1968) » énonce que l'information transite du génome au phénotype par une série de transformations comportant successivement une phase de transcription, une phase de traduction, suivies de l'interaction entre les molécules ainsi produites.

d'une protéine. Il est important de remarquer que ce processus de décodage introduit des degrés de liberté entre le génome et le protéome : un génome complexe peut coder un protéome simple (par exemple si tous les gènes ont la même séquence) et un protéome complexe peut être codé par une courte séquence (si des gènes partagent des séquences via *p. ex.* des ARNm polycistroniques – ou opérons – ou des chevauchements sur l'un ou l'autre brin). Ces degrés de liberté sont comparables à ceux qui sont observés dans les organismes réels. Comme énoncé ci-dessus, le décodage de l'information génétique, depuis la séquence d'ADN jusqu'à la séquence primaire de la protéine suit donc un ensemble de règles directement inspirées de la biologie « réelle » (et plus précisément de la génétique bactérienne).

Une fois connue la structure primaire des protéines codées sur le génome, Aevol doit calculer la contribution fonctionnelle de chacune de ces protéines ainsi que le phénotype et la fitness qui en résultent. Or, bien qu'il soit envisageable de mimer les processus biologiques au niveau des séquences, il est (jusqu'à présent) impossible de calculer la fonction d'une protéine à partir de sa structure primaire d'une façon réaliste. Pour contourner cet écueil, Aevol utilise une représentation mathématique abstraite pour décrire les niveaux fonctionnels (*c.-à-d.* la contribution fonctionnelle d'une protéine et le phénotype). Dans Aevol, toutes les fonctions sont exprimées dans un espace fonctionnel continu à une dimension (spécifiquement, sur l'intervalle $[0, 1]$) par une valeur d'activation comprise dans l'intervalle $[-1, 1]$, les bornes haute et basse correspondant respectivement aux maxima d'activation et d'inhibition de la fonction. Dans cet espace, les protéines sont décrites comme des noyaux fonctionnels en forme de triangles isocèles, l'interprétation biologique étant qu'une protéine qui active ou inhibe une fonction donnée active ou inhibe concomitamment les fonctions voisines. La largeur du support de la forme triangulaire permet donc d'implémenter une forme de pléiotropie au sein du modèle.

La description de la fonction des protéines au moyen d'un noyau fonctionnel triangulaire isocèle permet de décrire intégralement la fonction d'une protéine au moyen de trois paramètres (la moyenne m, la hauteur h et la demi-largeur w du triangle), eux-mêmes calculés au moyen de trois codes binaires de longueur variable entrelacés dans la structure primaire de la protéine (ainsi plus long est un gène plus précises sont les valeurs des paramètres m, w et h). Ce processus est l'analogue, dans Aevol, du processus de repliement des protéines puisqu'il leur confère une fonction à partir de la séquence primaire. Enfin, une fois que tous les noyaux ont été calculés à partir de l'ensemble des gènes présents sur l'ADN (voir figure II.4.1B), ils sont additionnés pour produire le phénotype (figure II.4.1C). De la même façon que le processus de transcription-traduction introduit des degrés de liberté entre le génome et le protéome, cette étape introduit des degrés de liberté entre le protéome et le phénotype. En effet, la combinaison de différentes protéines peut mener à une forme fonctionnelle simple, par exemple si les protéines partagent les mêmes valeurs pour m et w (voir section 3).

La dernière étape du processus de décodage du génome consiste à calculer la fitness de l'organisme comme exponentielle de l'écart entre la fonction phénotypique et une « cible phénotypique » représentant indirectement le contexte abiotique dans lequel ces organismes évoluent (en rouge clair sur les figures II.4.1B et 1C). Dans Aevol, la cible phénotypique est habituellement définie comme une somme de fonctions gaussiennes, ce qui nécessiterait un nombre infini de noyaux triangulaires pour s'ajuster parfaitement à la cible. Ce point fera l'objet d'un développement spécifique dans le chapitre suivant

(chapitre III, Dispositif expérimental et mesures de complexité).

Nous n'avons ici donné qu'un aperçu de l'organisation générale du processus de décodage de l'information génétique dans Aevol. Dans la suite de cette section, nous revenons sur les différentes étapes pour les décrire de façon plus détaillée et plus rigoureuse.

3.2.3 Du génome au transcriptome

Comme tout processus biologique, la « biochimie » d'Aevol repose sur deux éléments fondamentaux : le nucléotide et l'acide aminé. Au niveau du génome, c'est le nucléotide qui constitue donc l'élément de base du codage. Cependant, au contraire des nucléotides observés dans la nature (adénine, guanine, thymine, cytosine), un nucléotide Aevol est simplement un bit, au sens usuel d'un chiffre binaire : 0 ou 1. Cette simplification est rendue possible du fait que l'évolution n'est pas ici envisagée au niveau de la séquence elle-même mais au niveau de l'organisation des éléments fonctionnels (les gènes) sur la séquence.

Le génome (ou ADN) est une séquence double brin de nucléotides binaires de longueur variable. Conformément à l'inspiration bactérienne d'Aevol, cette séquence est circulaire. Du point de vue conceptuel, les deux brins sont équivalents puisque l'un se déduit directement de l'autre en calculant son complément bit à bit. Autrement dit, le génome est une séquence circulaire de paires de base, chacune d'elles étant formée par deux nucléotides complémentaires.

Il est important de souligner ici le fait que la longueur du génome est variable. C'est en effet une caractéristique essentielle d'Aevol qui le différencie en particulier de la plupart des modèles algorithmiques ou mathématiques de l'évolution. Sans ce degré de liberté, il ne serait pas possible de représenter les grandes insertions-délétions (voir section 3.3.2).

La première étape de décodage du génome consiste à identifier des séquences particulières appelées *promoteurs* à partir desquelles s'initie une transformation appelée transcription qui produit le transcriptome (les ARN messagers). Ces séquences promotrices sont définies à partir de leur ressemblance avec une séquence de référence, longue de 22 paires de base, qui constitue l'un des paramètres du modèle : toute séquence d'ADN dont la distance de Hamming à la séquence de référence (*c.-à-d.* le nombre de bits qui diffèrent entre ces deux séquences) est inférieure ou égale à 4 est considérée comme un promoteur et initie un processus de transcription de la séquence d'ADN (le sens de lecture et le sens de transcription dépendant du brin d'ADN considéré). Le processus de transcription peut cependant être plus ou moins efficace car, en fonction de la distance entre le promoteur et la séquence de référence, on calcule e, le taux de transcription de l'ARNm correspondant selon l'égalité suivante : $e = 1 - d/5$, où d est la distance de Hamming entre le promoteur et la séquence de référence [1]. Ceci conduit à une transcription maximale ($e = 1$) lorsque la séquence promotrice est exacte et nulle dès qu'elle comporte 5 erreurs ou davantage. Le taux de transcription sera ensuite utilisé pour pondérer l'efficacité des protéines traduites à partir de l'ARNm correspondant (voir section suivante).

Une fois le processus de transcription initié avec une efficacité e, il se poursuit au long

1. Formellement, la distance maximale autorisée entre la séquence promotrice et la séquence de référence – ainsi que la longueur de la séquence promotrice – sont des paramètres du modèle. Cependant, ces deux paramètres sont constants dans toutes les expériences discutées ici. C'est pourquoi nous avons pris le parti de simplifier la description du modèle en les considérant comme fixés.

de la séquence pour ne s'interrompre que lorsqu'une séquence dite « terminatrice » est rencontrée. Par analogie avec la génétique bactérienne, une séquence terminatrice est une séquence susceptible de former une structure « tige-boucle » (*hairpin*), c'est-à-dire une séquence de nucléotides telle que deux séquences complémentaires-inversées se suivent, séparées seulement par trois paires de bases arbitraires (une séquence tige-boucle est donc une séquence de la forme $abcd???\overline{d}\overline{c}\overline{b}\overline{a}$, où un nucléotide $\overline{x}$ est le nucléotide complémentaire du nucléotide x et où chaque « ? » représente un nucléotide quelconque).

3.2.4 Du transcriptome à la séquence primaire des protéines

Une fois l'étape de transcription achevée, l'information est codée sur une liste de séquences binaires simple-brin, les ARN messagers (ARNm), chacun doté d'un taux de transcription e. Chaque ARNm va alors être lui-même parcouru pour rechercher les gènes, c'est-à-dire les séquences qui coderont effectivement pour une protéine. C'est l'étape de « traduction ». Comme la transcription, la traduction repose sur la reconnaissance de signaux binaires sur la séquence (ici sur la séquence de l'ARN). L'initiation de la traduction se produit lorsque la lecture de l'ARN rencontre deux séquences signal successives séparées de trois bases : la séquence RBS (Ribosome Binding Site) correspondant à la séquence de six bases 011011 et le signal-codon Start correspondant aux bases 000. En d'autres termes, la traduction d'un gène est initiée lorsque le signal 011011???000 est détecté sur un ARNm.

À partir du signal d'initiation de la traduction, le gène est lu comme une suite de codons, chacun d'entre eux correspondant à un triplet de nucléotides consécutifs. Le code génétique associe alors chaque codon à un acide aminé particulier et le gène se termine lorsque la lecture rencontre un codon Stop sur le même cadre de lecture que le codon Start. On notera que le code génétique d'Aevol est fortement simplifié par rapport au code génétique « naturel » (figure II.5). En effet, Aevol utilisant des nucléotides binaires, le code génétique ne comporte que $2^3 = 8$ codons différents, correspondant soit à un Start, soit à un Stop, soit à un acide aminé. Conséquemment, il n'y a que 6 acides aminés différents dans la « biochimie » d'Aevol.

À l'issue du processus de traduction, l'information génétique décodée se présente sous la forme d'une liste de protéines caractérisées chacune par leur séquence primaire d'acides aminés accompagnée de leur taux de transcription (e) que nous assimilerons désormais à la concentration de la protéine correspondante.

3.2.5 De la séquence primaire à la fonction des protéines

Le passage de la séquence primaire des protéines à leur fonction est une étape clé du modèle. En effet, c'est à cette étape que s'opère la transition entre une description relativement fidèle des objets biologiques (les séquences d'ADN, d'ARNm et de d'acides aminés) et une description abstraite des fonctions biologiques, à commencer par la fonction des protéines.

Comme nous l'avons rapidement indiqué dans la section 3.2.2 ci-dessus, dans Aevol les fonctions biologiques-biochimiques sont abstraites. Les différentes fonctions biologiques sont décrites sous la forme d'une valeur numérique réelle et l'ensemble des fonctions biologiques possibles est décrit par l'intervalle $[0, 1]$. Chacune de ces fonctions peut alors

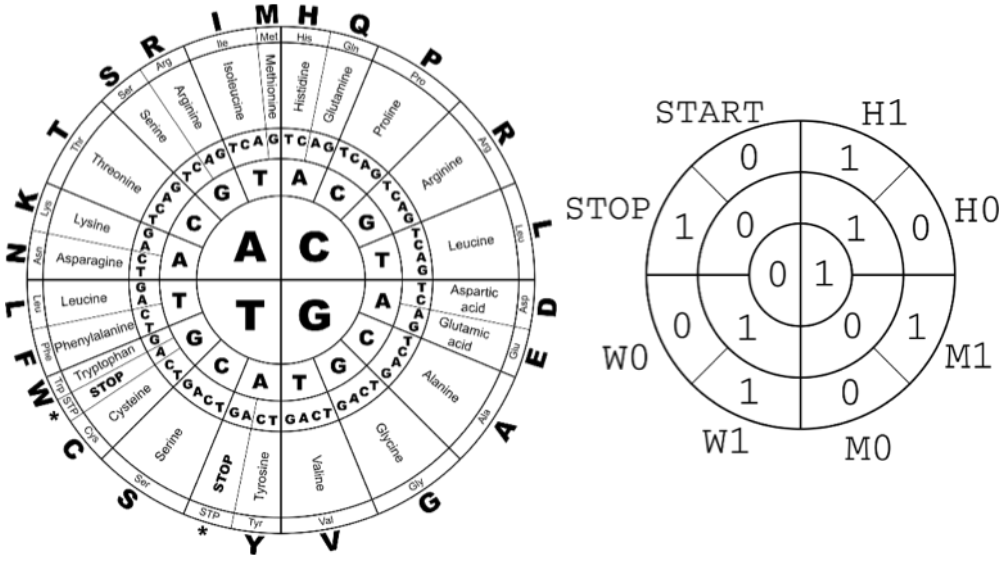


Figure II.5 – Dans le modèle Aevol comme dans la nature, les codons sont composés de trois bases. Pour Aevol cependant, les bases sont des éléments binaires, ce qui réduit fortement la combinatoire. Les deux schémas se lisent depuis le centre vers l'extérieur en choisissant les bases qui composent le codon. La roue la plus extérieure donne alors le nom du codon correspondant. On peut remarquer que dans Aevol deux codons distincts conduisent toujours à une fonction distincte alors que le code génétique « classique » est redondant. Ainsi, les codons AGA, AGG et CGT produisent tous les trois le même acide aminé (l'arginine, en l'occurrence). De même, le code génétique admet trois codons STOP (TAA, TAG et TGA) quand Aevol n'en admet qu'un (001).

être activée ou inhibée plus ou moins efficacement, l'activation parfaite d'une fonction correspondant à une valeur d'activation égale à $+1$ tandis que l'inhibition parfaite d'une fonction correspondra à une valeur d'activation égale à -1. Dans ce formalisme abstrait, une protéine correspond à une fonction élémentaire tandis qu'un phénotype (ou, comme nous le verrons, un environnement) correspondra, dans le cadre général [1] à une fonction complexe.

Dans un organisme biologique, les protéines assurent une multitude de fonctions. Dans Aevol, nous partirons du principe qu'une protéine participe à une fonction principale (soit en l'activant, soit en l'inhibant) et aux fonctions proches mais avec une contribution qui décroît linéairement selon la distance entre les deux fonctions. Exprimées dans le formalisme fonctionnel décrit ci-dessus, ces caractéristiques impliquent que les protéines correspondront à des fonctions mathématiques $f : [0, 1] \to [-1, 1]$ triangulaires et isocèles caractérisées par trois valeurs : leur moyenne m, leur demi-largeur w et leur hauteur h. La moyenne du triangle correspond à la fonction à laquelle la protéine contribue, la demi-largeur à l'étendue de ses effets pléïotropiques (en dehors de l'intervalle $[m - w, m + w]$ la contribution de la protéine est nulle) et la hauteur à l'intensité avec laquelle la protéine active (si $h > 0$) ou inhibe (si $h < 0$) la fonction.

1. Ce ne sera cependant pas le cas dans la plupart des expériences présentées ci-après, comme nous le verrons au chapitre suivant.

En vertu des principes que nous venons d'énoncer, le décodage de la fonction de la protéine revient à calculer les valeurs de m, w et h à partir de la séquence primaire de la protéine, c'est-à-dire à partir de sa séquence d'acides aminés. Cette opération s'effectue via un algorithme parfois qualifié – dans le modèle – d'algorithme de repliement : à partir de la séquence d'acide aminés (composée à partir de l'alphabet de six acides aminés $M0$, $M1$, $W0$, $W1$, $H0$ et $H1$), on commence par extraire, dans l'ordre inverse de la séquence (codage « Big-Endian »), trois sous-chaînes entrelacées correspondant respectivement aux séquences de $M0/M1$, de $W0/W1$ et de $H0/H1$. Ces trois sous-chaînes sont alors converties directement en trois séquences binaires qui donneront à leur tour trois valeurs entières (le code binaire utilisé ici étant le code de Gray, procédure classique en algorithmique évolutionnaire car elle garantit que deux valeurs consécutives soient codées par deux chaînes binaires voisines qui ne diffèrent que d'un seul bit – ou, plus exactement ici, d'un seul codon). À l'issue de cette étape, la contribution fonctionnelle de la protéine est codée sous la forme de trois entiers. Un élément fondamental du modèle est ici que les gènes ne sont pas de taille fixe (la taille d'un gène ne dépend en effet que du nombre de codons séparant le Start du Stop) en conséquence de quoi les séquences binaires correspondant aux trois entiers sont elles-mêmes de taille variable, ce qui implique que les trois entiers sont codés sur des intervalles variables en fonction de la taille de la séquence correspondante (pour une séquence binaire de taille n, l'entier correspondant sera codé sur un intervalle $[\![0, 2^n - 1]\!]$). Connaissant la valeur entière et son intervalle maximal, il est alors aisé de normaliser cette valeur pour obtenir les trois réels caractérisant le triangle fonctionnel de la protéine [1]. Le tableau II.1 illustre par un exemple les principales phases de décodage permettant de passer d'une séquence d'ADN à la fonction d'une protéine.

3.2.6 Des protéines au phénotype

Une fois connues les contributions fonctionnelles de toutes les protéines (c'est à dire, pour chacune d'elle, le triplet m, w, h) ainsi que leurs taux de transcription e, le calcul du phénotype s'effectue relativement simplement puisqu'il s'agit, en première approximation, de faire la somme des contributions de toutes les protéines. En pratique cependant le calcul du phénotype est légèrement plus compliqué. En effet, il est nécessaire de prendre en compte l'existence de deux type de contributions fonctionnelles : les activations (assurées par les protéines pour lesquelles $h > 0$) et les inhibitions (assurées par les protéines pour lesquelles $h < 0$; on notera que les protéines pour lesquelles $h = 0$ n'ont pas de contribution fonctionnelle puisque leur activité est nulle sur l'ensemble de l'espace fonctionnel). Pour ce faire, Aevol utilise des opérateurs issus de la logique floue : les opérateurs de Łukasiewicz : l'ensemble des fonctions activées (respectivement, l'ensemble des fonctions inhibées) est calculé comme la somme, bornée en 1, de tous les triangles correspondant à des protéines activatrices (respectivement inhibitrices) sur l'intervalle $[0, 1]$ (en termes de logique floue, il s'agit de l'union de tous les ensembles flous correspondants aux fonctions des protéines). Une fois ces deux ensembles calculés, l'ensemble des fonctions effectivement activées (correspondant au phénotype de l'organisme) correspond à l'ensemble des

1. La valeur w étant normalisée dans l'intervalle $[0, 1]$, la valeur h dans l'intervalle $[-1, 1]$ et la valeur w dans l'intervalle $[0, W_{max}]$, où W_{max} est un paramètre du modèle caractérisant la pléiotropie maximale des protéines.

1	ARNm	...011011???000011110101100010110101101111001...
2	Séquence de codons	START.011.110.101.100.010.110.101.101.111.STOP
3	Séquence primaire	W1 H0.M1.M0.W0.H0.M1.M1.H1
4	Sous-séquence M Sous-séquence W Sous-séquence H	M1.M1.M0.M1 W0.W1 H1.H0.H0
5	Codage binaire M Codage binaire W Codage binaire H	1101 (longueur du code : 4) 01 (longueur du code : 2) 100 (longueur du code : 3)
6	Entier M (code de Gray) Entier W (code de Gray) Entier H (code de Gray)	9 (sur l'intervalle : $[0, 2^4 - 1 = 15]$) 1 (sur l'intervalle : $[0, 2^2 - 1 = 3]$) 7 (sur l'intervalle : $[0, 2^3 - 1 = 7]$)
7	Valeur de m Valeur de w Valeur de h	9/15 $w_{max} * 1/3$ $(2 * 7/7) - 1$

TABLE II.1 – De l'ARN messager à la fonction de la protéine. La séquence de l'ARN messager (1) est lue codon par codon à partir du START et jusqu'à rencontrer un codon STOP sur le même cadre de lecture (2). Chaque codon est alors traduit en un acide aminé au moyen du code génétique, l'ensemble résultant en la séquence primaire de la protéine (3). De la séquence primaire lue du STOP vers le START (« Big-Endian ») sont extraites trois sous-séquences entrelacées, les sous-séquences M, W et H composées respectivement des suites de codons M0/M1, W0/W1 et H0/H1 (4). À partir de ces trois sous-séquences on déduit les codes binaires de M, W et H (5) puis les entiers correspondants ainsi que leurs intervalles maximaux (6). Une opération de normalisation permet alors de calculer les valeurs m, w et h correspondant à la fonction de la protéine (7).

fonctions activées ET NON inhibées. Là encore, ce sont les opérateurs de Łukasiewicz qui permettent de calculer le phénotype en soustrayant, avec une borne en 0, l'ensemble des fonctions inhibées de l'ensemble des fonctions activées (toujours sur l'intervalle $[0, 1]$).

Ce processus permet de combiner l'ensemble des fonctions activées ou inhibées par les protéines et de calculer le phénotype, qui correspondant en fin de compte à une fonction mathématique de $[0, 1] \rightarrow [0, 1]$ (et donc, suivant la représentation fonctionnelle décrite ci-dessus, à l'ensemble des fonctions activées par l'organisme suite au décodage de son génome).

3.2.7 Du phénotype à la fitness

La dernière étape du décodage du génome consiste à calculer la valeur sélective de l'organisme (sa fitness) en fonction de son phénotype. De fait, l'interaction entre un organisme et son environnement est le nœud central de tout processus adaptatif : il s'agit ici d'établir une correspondance entre les traits d'un organisme (qui sont regroupés globalement sous la notion de phénotype – « l'ensemble des traits observables » selon la définition commune [1]).

1. Le concept de phénotype frappe d'abord par son évidence si on l'aborde sous l'angle des traits objectifs, constatables voire mesurables en tant qu'observateur extérieur et indépendant. Il s'agit pourtant

Nous avons vu que, dans Aevol, le phénotype est représenté par une fonction mathématique exprimant les fonctions biologiques (abstraites) réalisées ou non par l'organisme. Le calcul de la fitness est alors opéré par comparaison entre ce phénotype et les caractéristiques abiotiques de l'environnement. Celui-ci est ici considéré comme « abiotique » car cette étape n'inclut aucune interaction entre les différents organismes composant la population. Il n'y a en particulier pas de compétition à ce niveau (celle-ci étant prise en compte par la « boucle évolutive », voir section suivante) et pas de construction de niche par exemple.

Dans Aevol, l'environnement est représenté implicitement par l'expression optimale des fonctions biologique permettant de s'y reproduire. En d'autres termes, l'environnement est représenté indirectement par un phénotype optimal, et l'adaptation, dans le modèle, correspond à un mécanisme d'ajustement de données (*curve fitting*) par rapport à une fonction mathématique cible. Cette représentation indirecte permet de mesurer le degré d'adaptation d'un organisme par une simple mesure de différence la cible et le phénotype. Techniquement, nous mesurons simplement l'intégrale de la valeur absolue de la différence entre la fonctions cible et le phénotype (ces fonctions étant toutes deux définies de $[0, 1]$ dans $[0, 1]$). Dans Aevol, cette mesure est généralement considérée comme une « erreur métabolique » ou « gap » (g). L'erreur métabolique caractérisée est donc inversement proportionnelle au degré d'adaptation : une erreur métabolique égale à zéro ($g = 0$) correspondant à un organisme parfaitement adapté à son environnement tandis qu'une erreur métabolique $g = 1$ correspond au pire organisme possible.

Le principe de la définition indirecte et abstraite de l'environnement est une très forte simplification de la réalité biologique mais elle permet une interprétation relativement facile des dynamiques évolutives puisqu'elle ne prévoit aucune interaction réciproque entre les organismes et leur environnement. Elle impose en revanche de choisir le type de fonction cible utilisé, sachant que celui-ci va déterminer la difficulté de la tâche d'identification que va devoir résoudre le mécanisme évolutif. À priori, toute fonction de $[0, 1]$ dans $[0, 1]$ peut être utilisée. Cependant, dès l'origine d'Aevol, il a été choisi de définir les cibles phénotypiques par des sommes de gaussiennes afin qu'aucune somme finie de triangles (*c.-à-d.* dans Aevol, aucun protéome et aucun génome) ne permette d'identifier exactement la fonction cible [1]. L'intention expérimentale qui dirigeait ce choix était de forcer un accroissement de la complexité qu'aucune borne n'entrave, où le terme « complexité » était entendu en un sens qualitatif (traduisant par exemple le nombre de gènes ou, indirectement, la longueur du génome). Nous verrons au chapitre suivant que ce principe, justifié dans le cadre général du modèle, devra ici être adapté afin de nous permettre d'observer des dynamiques de complexification-simplification dans des contextes environnementaux différents.

d'un concept subtil et polymorphe qu'il est difficile de définir extensivement, si bien qu'on préfère généralement se contenter d'en énumérer quelques facettes pour en étayer l'idée (taille, vitesse de réplication, éléments morphologiques, etc). Qu'on songe pourtant que l'ADN d'un organisme donné – classiquement considéré comme le génotype et non comme partie intégrante du phénotype – se rattache lui-même à son propre phénotype (en tant qu'il est le produit de l'organisme par le truchement de la méiose, qu'il est partie intégrante de l'organisme et qu'il est en contact avec l'environnement).

1. Il est à noter que, dans le cadre de l'implémentation informatique du modèle un très grand nombre de triangles pourrait techniquement permettre de représenter exactement la cible par le truchement des approximations implicites liées au type numérique utilisé. Ce point n'a cependant jamais été atteint en pratique malgré des simulations atteignant jusqu'à plusieurs centaines de triangles.

3.3 La boucle évolutive

Dans la section précédente, nous avons vu comment le codage (et le décodage via la *genotype-phenotype map*) de l'information dans Aevol permet de passer d'une séquence génétique à une valeur de sélective f. Disposant dès lors, pour chaque organisme, d'un génome et d'une fitness, nous pouvons assujettir les organismes à une boucle évolutionnaire composée d'un opérateur de sélection et d'un opérateur de variation ou, plus exactement, d'un ensemble d'opérateurs de variation comportant principalement le switch (qui inverse un bit sur les deux brins du chromosome), les petites insertions, les petites délétions (pris ensemble : les « indels ») et les réarrangements chromosomiques à grande échelle (duplications, délétions, inversions et translocations). À chacun de ces opérateurs mutationnels est associé un taux de mutation exprimé en nombre de mutations par paire de base et par génération (mut.bp^{-1}.gen^{-1}).

3.3.1 Opérateur de sélection

Une fois le phénotype connu et comparé à la cible environnementale, l'aire délimitée par ces deux courbes permet de quantifier la distance qui sépare un organisme du maximum adaptatif (le gap g ou l'erreur métabolique). À partir des gaps g_i de l'ensemble des organismes i d'une population de taille N, l'opérateur de sélection calcule le nombre de descendants d'un organisme donné et, dans le cas d'une population spatialisée, leur position spatiale.

Depuis les premières versions d'Aevol de nombreux opérateurs de sélection ont été proposés et testés (sélection sur le rang, sélection du meilleur, population spatialisée ou non, etc) mais depuis la version 5 d'Aevol, ces évolutions ont convergé vers un opérateur de sélection standardisé choisi à la fois pour des raisons techniques (liées à la parallélisation du code) et pour des raisons scientifiques (cohérence du modèle). C'est cet opérateur désormais classique que nous avons utilisé dans toutes les simulations réalisées au cours de ce travail et que nous décrirons ici, même si, en pratique, le logiciel permet de nombreuses autres possibilités.

L'opérateur de sélection classique d'Aevol est un opérateur spatialisé (les individus étant répartis sur une grille torique à mailles carrées) proportionnel à la fitness (*fitness proportionnate*). Sous ce mode de sélection, la fitness d'un organisme est calculée à partir de l'inverse de son erreur métabolique. Cependant, les tailles de population classiquement simulées dans Aevol étant de l'ordre du millier d'individus, une sélection directement proportionnelle à l'inverse du gap $(1/g)$ serait très faible et la dynamique adaptative serait essentiellement dirigée par la dérive. C'est pourquoi, dans Aevol, la fitness $f_{x,y}$ d'un organisme situé sur une case de coordonnées x, y est calculée comme l'exponentielle de son erreur métabolique : $f_{x,y} = \exp(-kg_{x,y})$ où k est un paramètre permettant de fixer l'intensité de la sélection et $g_{x,y}$ est l'erreur métabolique de l'organisme (son gap). À partir de l'ensemble des valeurs $f_{x,y}$, la sélection opère indépendamment pour chaque case de la grille pour choisir quel organisme, parmi les 9 voisins de la case concernée ainsi que celle-ci (en considérant un voisinage de Moore, voir figure II.4.1D), va coloniser cette case à la génération suivante. Pour celà, Aevol utilise un simple mécanisme de roulette biaisée dans lequel la probabilité pour un individu voisin d'une case de coordonnées x, y de coloniser cette case à la génération suivante est proportionnelle à sa fitness divisée par la somme des fitness des voisins de la case x, y. En considérant, par exemple, le voisin en

haut à gauche de la case x, y (le voisin $x-1, y-1$), sa probabilité de se reproduire en x, y est donc proportionnelle à $f_{x+1,y+1}/(f_{x,y} + f_{x+1,y} + f_{x-1,y} + f_{x,y+1} + f_{x,y-1}f_{x+1,y+1} + f_{x-1,y-1} + f_{x+1,y-1} + f_{x-1,y+1})$, où les opérateurs $x+1$ et $x-1$ (respectivement $y+1$ et $y-1$) s'entendent modulo X (respectivement modulo Y) si X (respectivement Y) est la dimension horizontale (respectivement verticale) de la grille eu égard à la topologie torique.

On notera que, en termes de théorie évolutive, ce mécanisme de sélection impose un certain nombre de contraintes. Ainsi, il simule un mécanisme dit de *soft-selection* (tous les individus sont susceptibles de se reproduire en fonction des rapports de fitness ; il n'y a donc pas de mutation létale) et impose le renouvellement total de la population à chaque génération (hypothèse de générations séparées ou modèle de Wright-Fisher). En outre, à chaque génération il limite le nombre de descendants d'un individu donné à 9 (taille du voisinage de Moore), ce qui, indirectement, ralentit la fixation des mutations les plus avantageuses.

3.3.2 Opérateurs de variation

Dans Aevol les organismes se reproduisent par clonage, c'est à dire qu'ils reçoivent l'exact patrimoine génétique de leur parent. Ils peuvent cependant subir plusieurs sortes de mutations concomitamment à leur réplication. Celles-ci relèvent de deux groupes aux effets qualitativement différents : les mutations agissant ponctuellement (ou à petite échelle) sur le génome, ce sont les switchs et les indels et les mutations de grande échelle qui changent l'organisation des séquences (au point qu'on parle parfois de « variants structurels »), ces dernières sont regroupées sous le terme de réarrangements chromosomiques. Même si plusieurs autres opérateurs de mutation ont été ajoutés à Aevol (en particulier différents opérateurs de transfert horizontal), nous ne décrirons ici que les 7 opérateurs principaux qui sont ceux que nous avons utilisés dans toutes nos expérimentations.

Substitution de base ou switch Il s'agit de la mutation la plus simple, elle consiste à basculer le bit d'une seule paire de base simultanément sur les deux brins ($0 \leftrightarrow 1$). Lorsque cette mutation affecte une région non codante, elle n'a aucun effet direct sur le phénotype (sauf si elle conduit à la création spontanée d'une séquence codante mais cet événement est hautement improbable). En revanche, lorsqu'elle affecte une séquence codante, elle peut altérer une séquence signal, par exemple un promoteur, et supprimer de ce fait un ARN et les gènes éventuels qu'il portait. Si c'est le signal d'arrêt de la traduction, qui est touché (codon Stop) le gène correspondant sera rallongé jusqu'au codon Stop suivant (si aucun codon Stop n'est présent avant le terminateur, le gène sera alors rendu non fonctionnel et donc supprimé).
À l'inverse, un switch peut créer une nouvelle séquence signal, cela ne conduit cependant que rarement à la création d'un gène nouveau. En effet les exigences pour qu'apparaisse un nouveau gène sont très contraignantes (voir sections précédentes) : il doit en effet être encadré de séquences signal de chaque sorte et dans l'ordre : promoteur, RBS, Start, Stop et terminateur. Ceci est rendu d'autant moins probable qu'aucune pression sélective ne vient accompagner les modification intermédiaires qui conduiraient à la création d'un nouveau gène puisqu'elles n'ont à elles seules aucun effet sur le phénotype.
Enfin, lorsqu'un switch se produit à l'intérieur d'un gène, il ne modifie qu'un seul

codon de ce gène, autrement dit un seul acide aminé dans la séquence primaire de la protéine correspondante. Une telle modification est de nature à modifier une ou deux des trois caractéristiques d'un triangle. L'effet d'une telle mutation dépend alors de sa position dans le gène considéré ainsi que de la taille de ce dernier. Dans tous les cas, l'effet de telles mutations, fût-il faible individuellement, peut entraîner une modification graduelle de la protéine au gré des mutations successives.

Indel Ce mot-valise regroupe les petites *in*sertions et petites *dél*étions. Les insertions consistent à insérer quelques nucléotides aléatoires dans le génome alors que les délétions suppriment quelques nucléotides. Dans les deux cas, dans Aevol le nombre de nucléotides en jeu est compris entre 1 et 6. Ces mutations peuvent avoir le même effet que les mutations ponctuelles en ce qui concerne la création ou la disparition de séquences de signalisation. Toutefois, lorsqu'elles se produisent à l'intérieur d'un gène il faut distinguer deux cas. Si le nombre de bases concernées est multiple de 3 (3 ou 6), la mutation revient à ajouter ou supprimer un ou deux codons au gène, ce qui peut n'avoir qu'un effet marginal, assez comparable à une mutation ponctuelle. Dans les autres cas, le cadre de lecture est modifié pour toutes les bases suivantes dans le gène, ce qui modifie radicalement le reste de la séquence ainsi que sa longueur puisque les bases qui constituaient le codon Stop ne sont plus alors lues sur le même cadre de lecture.

De ce fait, les indels, bien qu'ils agissent localement, peuvent avoir un effet très significatif, même s'il est généralement limité à un seul gène. En outre, les indels affectent la longueur des génomes et sont souvent évoqués pour expliquer les variations de taille des génomes observés dans la nature (Petrov, 2001). Cependant, sur ce point, il est intéressant de constater que, dans Aevol (Rutten *et al.*, 2019) ou dans les modèles construits sur la base d'Aevol (Fischer *et al.*, 2014), il a été montré a plusieurs reprise que les indels ne contribuent pas significativement aux variations de tailles des génomes, contrairement aux réarrangements chromosomiques.

Réarrangements chromosomiques Au contraire des indels, les réarrangements chromosomiques ont une action à grande échelle car ils mettent en jeu des séquences de taille arbitraire jusqu'à impliquer potentiellement la totalité du génome. Chez les bactéries réelles, plusieurs mécanismes peuvent conduire à réarranger un chromosome, notamment le mécanisme de réparation qui intervient lorsque les deux brins de l'ADN rompent ou encore lorsque la polymérase saute d'une séquence à une autre au cours de la réplication. Selon la localisation et la direction des séquences en jeu, ces mutations produisent différentes sortes de réarrangements : les duplications, les délétions, les translocations et les inversions (Higgins, 2005; Lewin, 2007). Ces quatre type d'événements sont modélisés dans Aevol, ils font l'objet des quatre points suivants.

Duplication Une duplication recopie un segment quelconque du génome soit juste à côté de son origine (on parle de duplication tandem, dans ce cas la duplication est un opérateur à deux degrés de liberté) soit en une position quelconque du génome (auquel cas la duplication a trois degrés de liberté). Dans Aevol, les duplications sont modélisées sous la forme de duplications à trois degrés de liberté.

Délétion Au cours d'une délétion, un segment aléatoire est supprimé (deux degrés de liberté).

Translocation Dans le cas des translocations, un segment de matériel génétique est déplacé (on peut penser à un couper-coller).

Inversion Les inversions, plus subtiles, portent sur un segment arbitraire qui se trouve inversé d'un brin à l'autre. Cet opérateur a donc deux degrés de liberté. Comme les deux brins d'ADN sont lus en sens opposés, dans Aevol l'inversion d'une séquence codante complète (c'est-à-dire incluant les séquences initiatrices et terminatrices) ne modifie pas l'information portée par cette séquence mais seulement les relations de synthénie sur le génome.

Tous les réarrangements impliquent de choisir des points de cassure ou d'insertion sur le chromosome. Dans la version classique d'Aevol (qui sera utilisée dans toutes les études présentées ici), ces points sont choisis aléatoirement sur le chromosome. Au cours de sa thèse, David Parsons (2011) a développé une version d'Aevol alternative dans laquelle les points de réarrangement sont choisis en fonction du degré d'homologie entre les séquences (Parsons *et al.*, 2011).

La modélisation des réarrangements chromosomiques est un des grands atouts d'Aevol. En effet, c'est l'un des seuls modèles qui rende compte de ces événements (la seule alternative étant les modèles en collier de perle, voir section 2.5). Cette propriété permet d'utiliser Aevol pour étudier les conséquences de ces événements, en particulier sur la taille du génome (Knibbe *et al.*, 2007a) ou sur l'ordre des gènes (Biller *et al.*, 2016b). Elle permet aussi d'explorer des dynamiques évolutives jusqu'ici très mal caractérisées car les réarrangements, bien qu'ils soient très fréquents dans les organismes « réels » (voir, par exemple, (Raeside *et al.*, 2014) pour une estimation de la fréquence des grands réarrangements dans la *Long-Term Evolution Experiment*) sont souvent les grands oubliés de la théorie évolutive.

Les quatre types de réarrangements modélisés dans Aevol ne sont pas équivalents quant à leurs conséquences sur le génome et sur le phénotype. Ainsi, alors que les inversions et les translocations sont dites « conservatives » (au sens où elles ne font que déplacer du matériel génétique sans le modifier – à l'exception des régions situées autour des points de coupure), les duplications et les délétions, elles, peuvent avoir un effet radical. Ainsi, une seule délétion peut supprimer un grand segment du génome (voire sa quasi-totalité). De même, une unique duplication peut dédoubler un grand nombre de gènes. Ainsi, duplications comme délétions peuvent modifier très significativement la taille du génome ainsi que la proportion de codant et donc la complexité d'un organisme.

3.4 Conclusion

Comme nous venons de le voir, Aevol est un outil de simulation relativement complexe, modélisant un grand nombre de mécanismes moléculaires et évolutifs. De ce fait, toute simulation utilisant ce modèle requiert de fixer un très grand nombre de paramètres. La table II.2 ci-dessous présente ainsi les principaux paramètres du modèle.

Paramètre	**Signification**
N	Taille de la population
X,Y	Dimensions de la grille (note : $N = X \times Y$)
$Seed$	Valeur d'initialisation du générateur aléatoire
nb_gener	Nombre de générations de l'expérience
$init\_length$	Taille des génomes initiaux
k	Pression de sélection
$E = \sum_i \alpha_i G_i$	Environnement (somme pondérée de gaussiennes $\alpha_i G_i$)
μ_{point}	Taux de mutations ponctuelles (Switch)
μ_{s_ins}	Taux de petites insertions
μ_{s_del}	Taux de petites délétions
μ_{dupl}	Taux de duplications
μ_{del}	Taux de grandes délétions
μ_{inv}	Taux d'inversions
μ_{trans}	Taux de translocations
max_indel_size	Taille maximale des indels
W_{max}	Degré maximal de pléiotropie des protéines

Table II.2 – Principaux paramètres du modèle Aevol. Nous passons ici sous silence de nombreux paramètres liés à des usages spécifiques d'Aevol ou à des prototypes (*p. ex.* régulation des gènes, alignement de séquences, environnement variable, etc). Tous les taux de mutation sont spécifiés en nombre d'événements par paire de base par génération (événements.bp^{-1}.génération^{-1}).

4 Aevol : une plateforme idéale pour tester l'évolution de la complexité ?

L'un des principaux avantages d'Aevol pour étudier l'évolution est sa capacité à simuler des organismes de complexités très différentes et, surtout, à laisser cette complexité évoluer sous l'influence d'une large gamme d'opérateurs de variations. En outre, le fait que le modèle intègre des mécanismes complexes de décodage de l'information en fait une plateforme idéale pour identifier les pressions évolutives susceptible de contraindre l'évolution de la complexité des organismes à différentes échelles (génome, transcriptome, protéome, phénotype, environnement, etc). De ce fait, Aevol apparaît comme une plateforme idéale pour étudier l'évolution de la complexité des systèmes biologiques, ce que confirment les nombreuses études réalisées ces dernières années avec ce modèle. En effet, même si la plupart de ces études n'étaient pas initialement destinées à étudier l'évolution de la complexité proprement dite, plusieurs d'entre elles ont révélé l'influence de facteurs évolutifs (*p. ex.* les taux de mutation) sur la taille des organismes (nombre de paires de bases, nombre de transcrits, nombre de gènes ou de protéines, etc). Or, nous avons vu dans le chapitre précédent que, même si une telle estimation est évidemment très grossière, ces valeurs pouvaient être rapportées à un niveau de complexité. Dans la suite de cette section, nous présenterons, par ordre chronologique, ces études en tachant d'expliquer en quoi elles éclairent, au moins partiellement, les questions d'évolution de la complexité biologique.

4.1 Evolution de la complexité des génomes

Aevol a été en très grande partie défini au cours de la thèse de Carole Knibbe (2006) et c'est aussi au cours de cette thèse que le premier résultat essentiel a été obtenu : en utilisant Aevol, Knibbe *et al.* (2007a) ont pu montrer que la taille du génome d'un organisme est inversement proportionnelle au taux de mutation sous lequel celui-ci évolue ($G \propto \mu^K$ avec $K \approx -1$).

Ce résultat obtenu au moyen de simulations conduites avec Aevol confirme ainsi la « loi de Drake » formulée par Drake (1991) (figure II.6). Fondée sur 13 observations de microorganismes (dont la bactérie *Escherichia coli*, une levure, un champignon filamenteux et des bactériophages), cette loi stipule empiriquement que $\log \mu = A \log G + B$, où μ et G sont respectivement le taux de mutation et la taille du génome et A et B des constantes. Or la valeur pour A étant proche de -1, cela revient à dire que le taux de mutation et la taille du génome sont inversement proportionnels (en log). Si l'on considère la taille du génome comme un proxy de la complexité d'un organisme – proxy au demeurant discutable, comme le montre le « C-value paradox » (Thomas, 1971) – alors la loi de Drake permet de lier phénoménologiquement la complexité des génomes à leurs taux de mutation. Le fait qu'elle se retrouve dans le contexte des simulations effectuées avec Aevol confirme l'existence d'une explication mécaniste.

L'analyse des simulations de Carole Knibbe a ainsi permis de mettre en évidence un lien causal entre le taux de mutation et la taille des génomes via la notion de robustesse (Wagner, 2007). En effet, Knibbe *et al.* (2007a) montrent que la taille des génomes converge vers un équilibre qui dépend du taux de mutation (dans Aevol il serait proba-

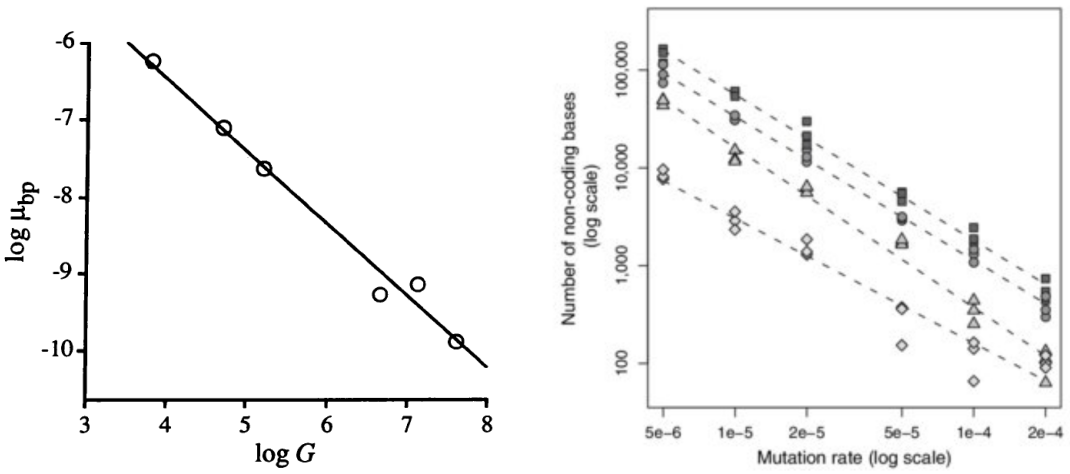


Figure II.6 – Relation entre la taille des génomes et les taux de mutation telle qu'elle a été observée dans les génomes de micro-organismes par Drake (1991) (à gauche) et dans Aevol par Knibbe *et al.* (2007a) (à droite). Les mesures de Drake (1991) ont été effectuées sur 7 microorganismes (les bactériophages M13, λ, T2 et T4, la bactérie *E. coli* et les eucaryotes unicellulaires *S. cerevisiae* et *N. crassa* – les bactériophages T2 et T4 sont regroupés sur la figure). Les simulations de Knibbe *et al.* (2007a) correspondent à 6 taux de mutation (en abscisse) et à 4 taux de sélection (symboles). Les deux courbes présentent une loi de puissance claire reliant la taille du génome et le taux de mutation (les résultats de Knibbe *et al.* (2007a) présentent la relation taux de mutation vs taille du non codant car celui-ci est le principal contributeur de la taille des génomes pour les faibles taux de mutation). figures d'après (Drake, 1991) et (Knibbe *et al.*, 2007a).

blement plus juste de parler *des* taux de mutation au vu de la diversité des opérateurs de mutation) de sorte que le nombre de descendants neutres du meilleur individu de la population se maintienne au dessus du seuil autorisant une réplication suffisamment fidèle pour assurer la transmission verticale de l'information génétique. L'idée que la robustesse contraint la quantité d'information présente sur le génome avait été proposée dès le début des années 1970 par Eigen (1971) dans sa théorie de l'*error threshold* (Biebricher et Eigen, 2005). Cependant, comme cette théorie se fonde sur un modèle théorique dans lequel seules les mutations ponctuelles sont prises en compte, elle ne parvenait pas à expliquer la façon dont la taille du génome dépend du taux de mutation car la taille des génomes est en grande partie déterminée par la quantité d'ADN dit « non codant », qui n'est pas sensible à l'*error threshold*. L'apport principal d'Aevol et des travaux de Carole Knibbe a été de montrer que la présence de réarrangements chromosomiques étend le principe de l'*error threshold* à l'ensemble de l'ADN, c'est-à-dire aux séquences codantes et non codantes (Knibbe *et al.*, 2007a).

Même si ce vocabulaire n'était pas utilisé alors, ces résultats ont une signification directe en termes de quantité d'information car la taille du génome, est l'une des mesures possibles de la quantité d'information portée par un individu. Bien entendu, les paradoxes des valeurs C et G rappellent que la complexité apparente des systèmes biologiques n'est gouvernée ni par la taille du génome ni même par le nombre de gènes (voir chapitre

précédent). Il n'en demeure pas moins que la loi mise en évidence par Carole Knibbe impose une borne supérieure à la complexité d'un organisme : à tout le moins celle-ci ne peut-elle dépasser la quantité de matériau génétique présent dans l'ADN. Ce résultat illustre par ailleurs l'intérêt du modèle Aevol pour étudier les lois régissant l'évolution de la complexité.

4.2 Limite à la complexité : le seuil d'erreur génomique

Suite aux résultats pionniers obtenus par Carole Knibbe avec Aevol (voir section précédente), Fischer *et al.* (2014) ont entrepris une étude mathématique de l'influence des réarrangements chromosomiques sur la taille des génomes. Même si ces résultats n'ont pas été obtenus en utilisant le modèle Aevol, nous les présentons ici car (*i.*) ce sont les dynamiques observées dans les simulations conduites avec Aevol qui ont motivé ce travail et (*ii.*) les résultats théoriques de Fischer *et al.* (2014) permettent de mieux comprendre les limites auxquelles la complexité génomique est assujettie telles qu'elles ont été observées dans le modèle.

Avant de décrire les travaux de Fischer *et al.* (2014), nous allons brièvement revenir sur la notion d'*error threshold* telle qu'elle a été énoncée par Eigen (1971) et reprise dans (Eigen et Schuster, 1977) et (Biebricher et Eigen, 2005).

Le principe de l'*error threshold* est relativement simple : il stipule simplement que le taux d'erreur commis lors de la réplication d'une molécule codant de l'information limite la quantité d'information transmissible d'une génération à l'autre et, par conséquent, la longueur de cette molécule. Corrélativement, pour augmenter la quantité d'information portée par une telle molécule au delà de ce « seuil d'erreur », il est donc impératif de réduire le taux de mutation (par exemple en introduisant des mécanismes de correction d'erreur). Un point fondamental de cette théorie est que le seuil d'erreur limite la longueur des séquences qui encodent de l'information génétique. Au contraire, n'étant pas contraintes au niveau de la séquence, les séquences non codantes (en particulier les séquences intergéniques) ne sont pas soumises à ce mécanisme et leur taille peut donc évoluer librement.

Afin de fournir une explication théorique aux observations de Knibbe *et al.* (2007a), Fischer *et al.* (2014) ont développé un modèle mathématique décrivant précisément l'évolution de la taille des génomes sous l'influence de différents opérateurs de mutation, dont les indels, les grandes duplications et les grandes délétions. Il est à noter que ce modèle dépasse la contrainte arbitraire généralement admise dans les modèles similaires qui interdit qu'un individu ne subisse plus d'une mutation par génération [1]. Dans le modèle proposé par Fischer *et al.* (2014), plusieurs mutations peuvent affecter un même génome à la même génération. Ce modèle montre alors qu'il existe un seuil, similaire au seuil d'erreur, au delà duquel l'accumulation de réarrangements chromosomiques à la même génération provoque un effondrement de la taille du génome sous l'effet conjoint des grandes duplications et des grandes délétions. De façon significative, cet effet est directement lié au fait que ces deux types d'événements ont un effet multiplicatif sur la taille des génomes (par opposition aux indels qui n'ont qu'un effet additif). Fischer *et al.* (2014) en concluent

1. Dans un modèle où la taille des génomes est susceptible de varier, cette contrainte est en effet peu réaliste : puisque le nombre de mutations augmente « naturellement » avec la taille du génome, il existe nécessairement un seuil au delà duquel cette limite n'est plus valide

d'une part que les indels n'ont qu'une contribution marginale à la dynamique de la taille des génomes et d'autre part que les taux de réarrangements chromosomiques bornent la taille globale des génomes de la même façon que les taux de mutation ponctuelle bornent la taille du codant : au delà d'un certain seuil, les taux de réarrangements ne permettent plus au génome de transmettre l'intégralité de son matériel génomique à la génération suivante. De façon intéressante, en comparant, pour plusieurs organismes, les tailles de génomes avec ce « seuil d'erreur génomique » estimé à partir des taux de réarrangements, Fischer *et al.* (2014) montrent que plusieurs espèces étudiées se situent juste sous le seuil, confortant ainsi empiriquement leur théorie (figure II.7).

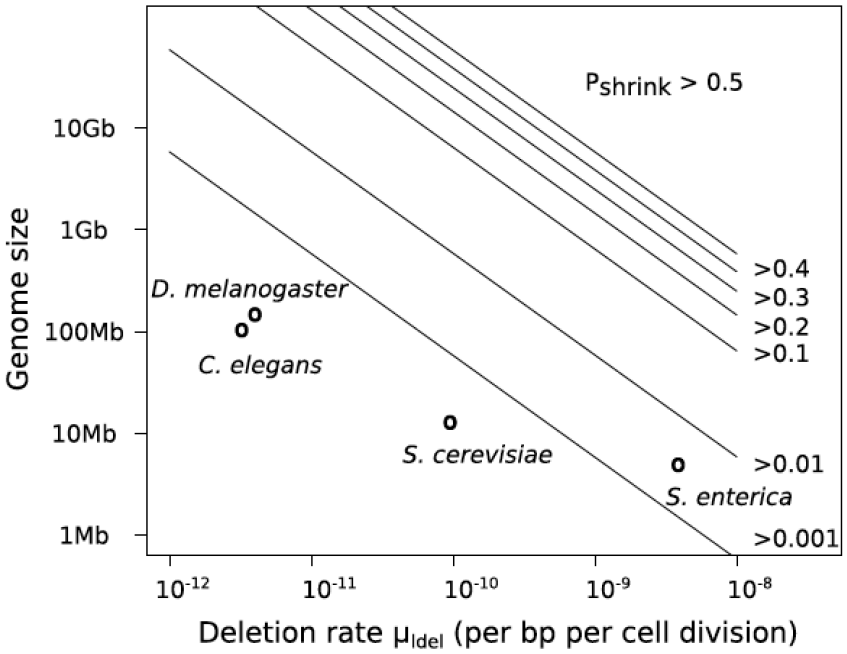


Figure II.7 – Le modèle de Fischer *et al.* (2014) prédit qu'un taux de réarrangement trop élevé comparativement à la taille du génome entraine un risque d'effondrement de celui-ci. Ce risque est ici représenté par les lignes diagonales. Si on place plusieurs espèces dont les taux de réarrangement et les tailles de génomes sont connues sur ce graphe, on constate qu'elles se situent toutes dans une zone où le risque d'effondrement est inférieur à $0,01$, voir inférieur à $0,001$. Figure d'après (Fischer *et al.*, 2014)

En termes d'évolution de la complexité, ces résultats, couplés avec ceux de Eigen (1971), montrent que plusieurs mécanismes indépendants bornent la complexité des génomes. Ils montrent en outre que la diversité de ces mécanismes est directement liée à la diversité des opérateurs de mutation (chaque famille d'opérateurs engendrant son propre mécanisme d'*error threshold*). Ils renforcent donc l'idée que l'étude de l'évolution de la complexité nécessite de prendre en compte « toute » la palette des opérateurs de variation observés dans la nature et illustrent, par là même, l'intérêt d'Aevol pour ce faire.

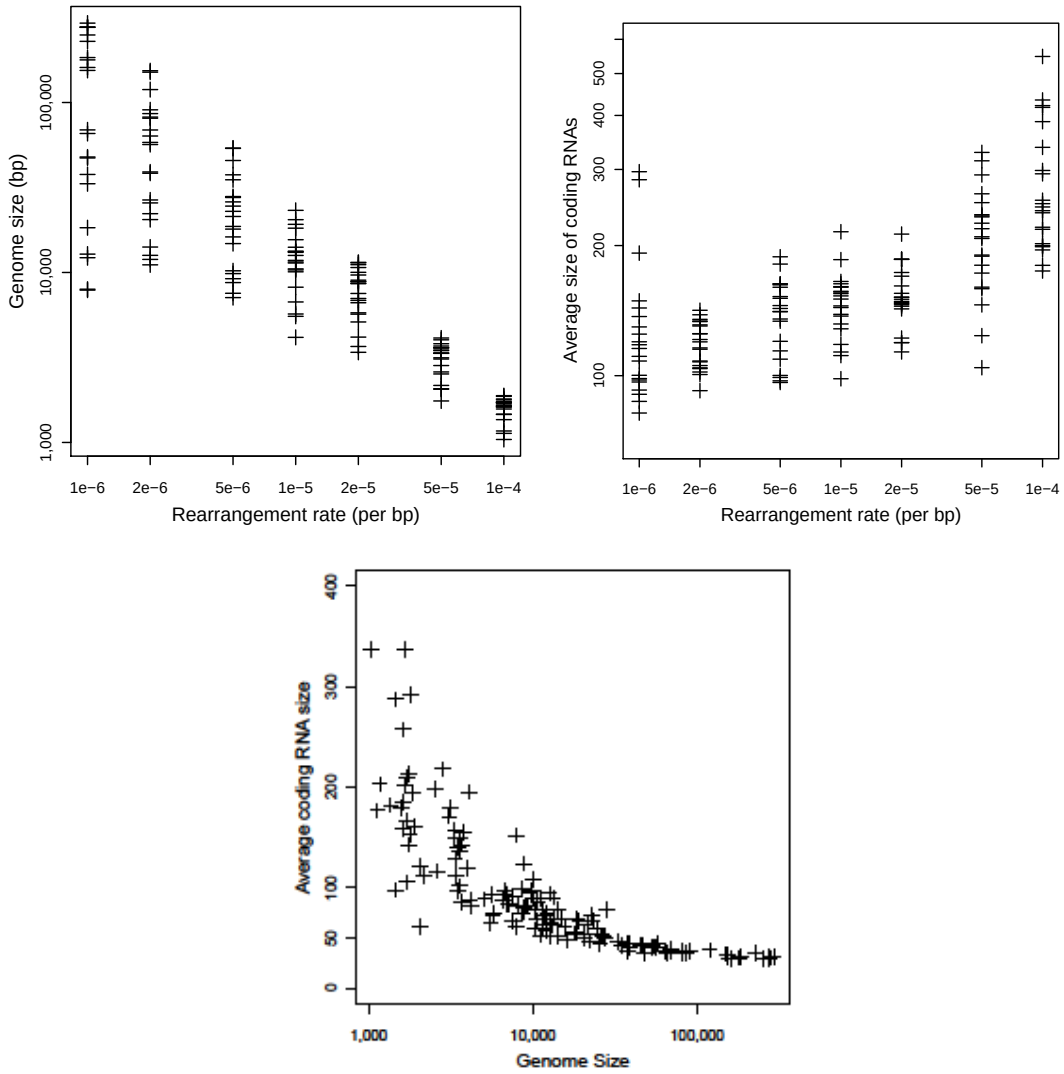


Figure II.8 – En utilisant Aevol, Parsons *et al.* (2010) ont montré que le taux de réarrangement est anti-corrélé à la taille des génomes (en haut à gauche) mais corrélé à la taille des transcrits (en haut à droite). Il en résulte (en bas) que les grands génomes portent les plus petits transcrits. Figures d'après (Parsons, 2011; Parsons *et al.*, 2010).

4.3 Complexité des transcrits

Dans sa thèse, portant sur la sélection indirecte, David Parsons (2011) a observé un effet des taux de mutation sur la structure des transcrits (Parsons *et al.*, 2010). Plus précisément, David Parsons a montré que les génomes les plus courts (ceux qui ont évolué sous les plus forts taux de mutation) portent les ARN les plus longs qui, corrélativement, portent un plus grand nombre de gènes (figure II.8). En d'autres termes, plus un génome est court, plus il est susceptible de contenir des opérons.

Ce résultat intrigant obtenu au moyen de simulations conduites avec Aevol a été par la suite confirmé indépendamment dans un article portant sur les données génétiques de trois clades α-protéobacteries, β-protéobacteries et firmicutes bactériens : Nuñez *et al.*

(2013) ont en effet montré que, chez ces bactéries, les grands génomes tendent à avoir moins d'opérons et que, de surcroît, ils sont moins bien conservés. De façon intéressante, Nuñez *et al.* (2013) évoquent une hypothèse liée à l'influence de la régulation pour expliquer leurs observations, les génomes les plus grands comprenant aussi plus de facteurs de transcription (Molina et van Nimwegen, 2008, 2009). Or, les observations de David Parsons *et al.* (2010) ont été réalisées en utilisant un modèle (Aevol, en l'occurrence) n'intégrant pas de mécanisme de régulation, ce qui conduit à proposer une hypothèse plus simple liée à robustesse : selon cette hypothèse, les taux de mutation élevés induisent une compaction du génome et l'accumulation d'opérons permet d'accentuer cette compaction en s'affranchissant de séquences de signalisation (promoteurs et terminateurs). Selon cette hypothèse, la corrélation (négative) entre la proportion d'opérons et le nombre de facteurs de transcription serait liée à un facteur causal commun : la compaction du génome. Cette conclusion est par ailleurs renforcée par les observations menées sur RAevol, une version du modèle Aevol qui comporte des mécanismes de régulation (voir section suivante).

D'un point de vue « complexité », les résultats de David Parsons *et al.* (2010) présentent notamment l'intérêt de souligner combien il peut être trompeur de juger de la complexité d'un organisme en se focalisant sur un niveau d'organisation particulier (ici, par exemple, les transcrits). Ainsi, dans leur étude, l'augmentation des taux de mutation entraîne une diminution de la complexité au niveau du génome mais concomitamment une augmentation de la complexité au niveau des transcrits.

4.4 Complexité des réseaux de régulation

Dans le modèle Aevol, utilisé dans les précédents exemples comme dans cette thèse, le phénotype d'un individu est totalement déterminé par ses gènes et les « concentrations » des protéines ne dépendent que de la qualité du promoteur associé (voir section 3.2.4). Ce choix est une profonde simplification de la réalité biologique puisqu'il néglige les mécanismes de régulation de l'expression des gènes et la dynamique du réseau de régulation génétique. Afin d'intégrer cette dynamique à Aevol, Yolanda Sanchez-Dehesa (2009) a développé au cours de sa thèse une variante du modèle, RAevol, intégrant de la régulation génétique. Dans RAevol, une protéine peut toujours avoir une activité métabolique (représentée sous la forme d'un triangle dans l'espace $[0, 1] \times [0, 1]$ mais elle peut aussi modifier l'activité d'un ou plusieurs promoteurs en fonction de l'affinité de sa séquence primaire pour les séquences précédant – *Leader* – ou suivant – *Trailer* – un promoteur (une affinité avec la séquence *Leader* augmentant l'activité du promoteur tandis qu'une affinité avec la séquence *Trailer* la diminue). En utilisant RAevol, il devient alors possible d'étudier l'évolution d'un réseau de régulation. Il est important de noter ici que RAevol diffère des modèles classiques basés sur le formalisme « réseau » (voir section 2.4) car, dans RAevol, les opérateurs de mutation n'agissent pas directement sur le réseau mais bien sur une séquence génétique qui sera décodée sous la forme d'un réseau. Il en résulte que les opérateurs de variation ont, dans RAevol, un mode d'action beaucoup plus réaliste, permettant en particulier de modifier le réseau en *Cis* ou en *Trans* et de faire varier à l'envi le nombre de nœuds et de liens dans le réseau de régulation. Ces propriétés permettent donc d'étudier le taille et la connectivité des réseaux de régulation en fonction des conditions environnementales (par exemple la complexité de l'environnement) et évolutives (par exemple les taux de mutation ou la taille de la population).

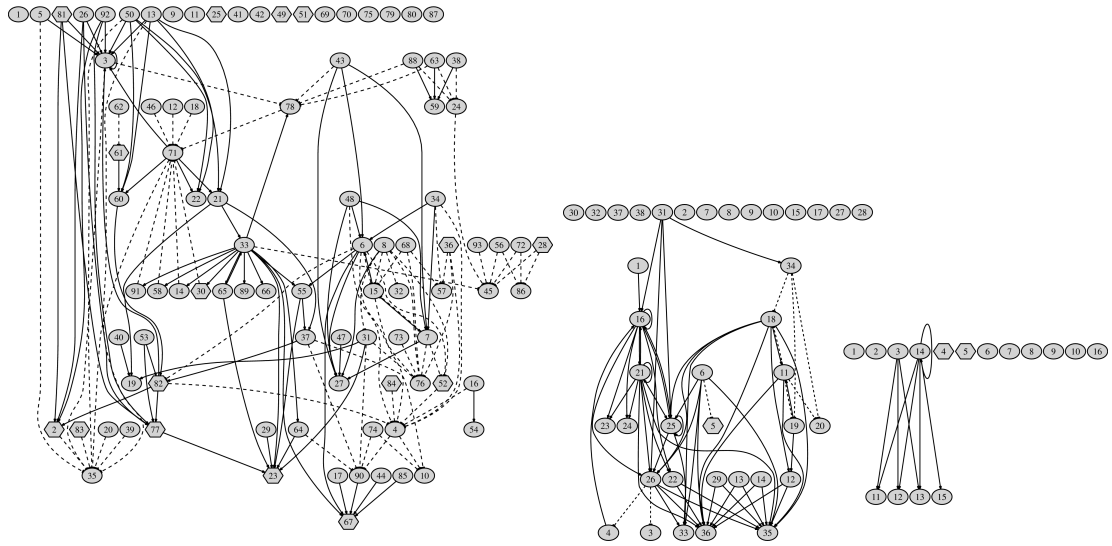


Figure II.9 – Trois exemples de réseaux de régulation obtenus avec RAevol. Ces trois réseaux ont évolué dans des conditions environnementales identiques mais avec des taux de mutation différents : 5×10^{-6} mutations.bp^{-1}.generation^{-1} (à gauche), 5×10^{-5} mutations.bp^{-1}.generation^{-1} (au centre), 2×10^{-4} mutations.bp^{-1}.generation^{-1} (à droite). Les ellipses représentent des gènes possédant une activité métabolique, les hexagones représentent des facteurs de transcription sans activité métabolique. Les traits pleins représentent des liens activateurs. les traits pointillés représentent des liens inhibiteurs. Figure d'après (Beslon *et al.*, 2010b).

Suite aux travaux de Carole Knibbe *et al.* (2007a) et de David Parsons *et al.* (2010), RAevol a été utilisé pour étudier l'influence des taux de mutation sur la complexité des réseaux de régulation (Beslon *et al.*, 2010b) : en faisant évoluer des organismes dans un environnement constant, les auteurs ont obtenu des réseaux de régulation de complexité très variable (voir figure II.9). Ils ont en particulier constaté que, dans les réseaux ainsi obtenus, le nombre de facteurs de transcription augmentait quadratiquement avec le nombre de gènes, une propriété déjà observée dans le monde bactérien (Molina et van Nimwegen, 2008, 2009) mais généralement expliquée par la complexité de l'environnement (Maslov *et al.*, 2009) tandis que, dans les résultats de (Beslon *et al.*, 2010b), cette propriété est observée dans des environnements identiques mais avec des taux de mutation différents. Il est particulièrement intéressant de noter qu'il a aussi été observé que, dans un environnement dynamique simple (à deux états), des réseaux de régulation très complexes se mettent en place de façon purement contingente, l'analyse à posteriori des réseaux montrant une dynamique très frustre malgré la densité de connexions (Beslon *et al.*, 2010a).

Les premiers travaux réalisés avec RAevol dans le cadre de la thèse de Yolanda Sanchez-Dehesa (2009) étaient restés limités car ils n'avaient pas permis d'explorer l'influence de la complexité de l'environnement sur la complexité des réseaux. Cette étude a été abordée par Vadée-Le-Brun *et al.* (2016) qui ont exploré l'effet conjoint des taux de mutation, de la pression de sélection et de la complexité de l'environnement sur l'évolution des réseaux de régulation. Dans cette étude, la complexité de l'environnement est considérée comme le nombre d'états possibles (entre 2 et 16 états dans cette étude) et l'objectif est de quan-

tifier la complexité du réseau (nombre de nœuds, densité de connexion) en fonction de celle de l'environnement. Dans ce cadre, les résultats de Knibbe (2006) sont à nouveau vérifiés : c'est le taux de mutation qui détermine le nombre de nœuds du réseau (et donc la complexité à ce niveau). En outre, de façon étonnante, il est aussi montré que le nombre de gènes est indépendant du nombre d'états de l'environnement mais que la densité de connexion (*c.-à-d.* le nombre de liens divisé par le nombre de nœuds au carré) dans le réseau de régulation augmente avec la complexité de l'environnement. Comme pour les résultats de Parsons *et al.* (2010), ce travail illustre bien la façon dont s'affrontent les degrés de complexité aux différents niveaux (nombre d'états de l'environnement, nombre de gènes, densité du réseau de régulation). Malheureusement RAevol est un modèle complexe a manipuler en particulier en raison des temps de calcul en jeu. De ce fait, les résultats de Vadée-Le-Brun *et al.* (2016) n'ont été obtenus que sur un petit nombre de simulations ce qui limite la portée statistique des résultats. Des études de plus grande ampleur sont actuellement en cours pour confirmer ces résultats préliminaires.

4.5 Le codant et le non codant n'évoluent pas de la même façon selon le type de mutations

Tous les résultats présentés jusqu'à maintenant ont mis l'accent sur l'influence des taux de mutation sur la complexité moléculaire à tous les niveaux. À l'échelle génomique, en particulier, ils montrent que la taille des séquences codantes comme celle des séquences non codantes sont directement influencées par les taux de réarrangements chromosomiques. Suite aux travaux de Rutten *et al.* (2019) cependant, le tableau apparaît beaucoup plus complexe. En effet, dans cette étude, les auteurs montrent que les tailles de ces deux compartiments peuvent évoluer indépendamment et de manière spécifique selon le type de mutation. Rutten *et al.* (2019) montrent que la perte de fitness initiale causée par l'apparition d'un phénotype mutateur [1] est rapidement compensée par l'entremise de deux mécanismes distincts : d'une part par la compaction du codant (ce qui restreint sa surface d'exposition aux mutations), d'autre part par l'accroissement de la quantité de non codant (qui biaise l'effet des mutations au profit de mutations très délétères, initiant ainsi une stratégie d'anti-robustnesse). Une fois de plus, cette dynamique illustre combien la complexité des différents compartiments moléculaires est susceptible d'évoluer selon des dynamiques différentes.

Un résultat similaire a été observé dans le cas de la sélection pour la robustesse. Comme nous l'avons déjà vu (section 4.1), les forts taux de mutation conduisent au raccourcissement des génomes du fait de la sélection indirecte pour la robustesse (Knibbe *et al.*, 2007a). On aurait alors tendance à penser que la taille de population devrait produire un effet simulaire de compaction des génomes, dans la mesure où les grandes populations accentuent l'efficacité de la sélection, donc l'efficacité de la sélection pour la robustesse. Ce raisonnement a été testé par Carde *et al.* (2019). Dans un travail préliminaire, les auteurs ont comparé les effets, d'une part de l'accroissement du taux de mutation, d'autre part de l'accroissement de la taille de la population. Conformément à l'intuition première, ils ont

1. Un organisme « mutateur » est un organisme dont le taux de mutation ponctuelle (switch) est très fortement accru par rapport au « type sauvage ». On parle généralement d'un accroissement de l'ordre d'un facteur 100.

bien observé une compaction du génome. Cependant il apparaît que cette compaction ne s'effectue pas de manière identique dans les deux cas : alors que l'augmentation des taux de mutation entraîne une compaction du codant et du non codant, l'augmentation de la taille de la population n'a d'effet que sur le non codant et la taille codante est conservée. Alors que tous les résultats présentés ci-dessus illustraient l'influence de la sélection pour la robustesse (donc des taux de mutation) sur la complexité des organismes, le résultat préliminaire de Carde *et al.* (2019) semble montrer qu'il existe plusieurs modes de sélection pour la robustesse et que l'augmentation des taux de mutation et l'augmentation de la taille de la population n'agissent pas de la même façon. En conclusion les auteurs émettent l'hypothèse que, à l'instar de la sélection directe, la sélection indirecte pourrait agir de façon positive (en sélectionnant les mutants favorables) ou négative (en éliminant les mutants défavorables) et que ces deux modes de sélection auraient des conséquences différentes sur les deux compartiments génomiques (ADN codant et non-codant).

Si la publication récente de ces résultats n'a pas permis de les prendre en compte dans la présente thèse (puisque les campagnes expérimentales avaient déjà été menées), il n'en demeure pas moins qu'ils suggèrent d'intéressantes perspectives sur l'évolution de la complexité pour lesquelles les cadres d'analyse proposés ici pourraient apporter un éclairage supplémentaire. Il faudrait en particulier les rapprocher des résultats présentés au chapitre V où nous montrons que des organismes peuvent évoluer d'une structure complexe à une structure simple sous l'effet d'un fort accroissement de leur taux de mutation.

5 Conclusion

Dans ce chapitre, nous avons tout d'abord rapidement présenté le principe de l'évolution expérimentale *in silico* ainsi que les principaux formalismes utilisés pour développer les modèles numériques correspondants. Nous avons ensuite décrit en détail le modèle Aevol, basé sur le formalisme « séquence de nucléotides » qui nous semble être le plus adapté pour étudier l'évolution de la complexité biologique. Enfin, nous avons présenté les principaux résultats obtenus avec Aevol, en insistant tout particulièrement sur ceux qui présentent une relation directe avec l'évolution de la complexité. Comme on a pu le voir, plusieurs travaux ont mis en relief la question de la complexité et en particulier le lien quasi ubiquitaire qui semble lier celle-ci aux taux de mutation. Cependant, on pourrait dire que cette ubiquité se double d'un aspect fuyant : la complexité peut s'exprimer à de nombreuses échelles et, semble-t-il, se transférer d'une échelle à une autre lorsque les contraintes changent.

Cependant, aucun des travaux précédents n'a spécifiquement concentré son attention sur la notion de complexité et, par voie de conséquence, aucun n'a établi un protocole précis et spécifique pour observer si la complexité, au sens où nous l'avons définie dans le chapitre précédent, s'accumule ou non. Ceci peut s'expliquer notamment par l'impossibilité (du moins une difficulté que nous ne savons pas lever) de caractériser, à aucune échelle, ce que serait la simplicité. Et ceci pour une simple raison : le besoin qu'ont ces expériences d'observer la dynamique évolutive sur des échelles de temps arbitrairement longues exige un dispositif expérimental qui permette sans cesse l'apparition de nouvelles innovations. De là, dans Aevol, ce choix d'une cible phénotypique incommensurable avec les protéines. (Comme on l'a déjà souligné, abstraction faite des approximations numériques, aucune somme finie de triangles ne saurait atteindre une somme de gaussiennes.) Autrement dit, le cadre expérimental commun à toutes les études précédentes impose une inflation de la complexité de telle façon qu'il devient impossible de distinguer si celle-ci résulte ou non d'un effet de la sélection et réduit les observations à ce sujet à des conjectures. Il n'en demeure pas moins que ces dernières attisent la curiosité et laissent pressentir qu'il existe là un angle d'attaque qui permettrait des interprétations fructueuses. C'est celui-ci que se propose d'adopter la présente thèse en renversant la logique d'accumulation de complexité au profit d'une logique de « simplicité » à partir de laquelle il sera possible d'explorer, de mesurer et de comparer les conditions qui, dans les simulations, conduisent ou non à l'accroissement de la complexité biologique.

Chapitre III

Dispositif expérimental et mesures de complexité

1 Introduction

Dans les deux chapitres précédents, nous avons tout d'abord présenté la controverse opposant les tenants d'une origine sélective de la complexité biologique à ceux qui défendent une origine neutre (chapitre I). Nous avons ensuite présenté une approche par simulation (l'évolution expérimentale *in silico*) et une plateforme de simulation (Aevol) que nous nous proposons d'utiliser pour trancher cette controverse (chapitre II). L'étape suivante consiste à définir un plan expérimental adossé à ce modèle qui, en engendrant des comportements symptomatiques, permette de répondre de façon incontestable (si tant est qu'un modèle puisse être incontestable) en faveur de l'une ou l'autre de ces hypothèses.

Deux difficultés centrales entravent l'expérimentateur qui voudrait explorer l'évolution de la complexité biologique. La première tient à la difficulté de quantifier la complexité, ou même seulement de comparer la complexité de systèmes différents. (Comment comparer la complexité de l'homme et du riz quand on sait que le second possède un génome dix fois plus long que le premier ?) Nous allons revenir plus précisément sur ces deux aspects. La seconde difficulté vient du fait que dans des environnements où la complexité foisonne, la fitness tirera profit d'améliorations infinitésimales au prix d'une augmentation de la complexité. En d'autres termes : la complexité sera sélectionnée. Comment, en effet, trancher entre les hypothèses neutralistes, « à la Gould » et les hypothèses sélective, « à la Dawkins » ? La question est loin d'être triviale alors que pourtant l'une et l'autre donnent à la complexité des explications très différentes. Pour mémoire, si l'on en croit Gould, la complexité augmente du simple fait que l'on considère une grandeur positive soumise à une marche aléatoire. Pour Dawkins, la complexité est le prix à payer pour tirer avantage d'un environnement complexe et gagner la compétition évolutive. Mais dans un contexte où la complexité abonde, les effets de l'une comme de l'autre sont indiscernables en pratique, aussi ne peut-on pas déterminer laquelle de ces hypothèses est la plus pertinente. Il faudrait pour cela remonter à la source historique de cette complexité pour voir comment tout cela a commencé ; il faudrait en quelque sorte, recommencer l'évolution à partir d'une situation *simple* pour donner sa chance à la simplicité. Car, si Gould prévoit que, dans un contexte simple, on verra émerger autant d'organismes simples que d'organismes complexes, Dawkins, quant à lui, s'attendrait plutôt à ce que

	Environnement simple	**Environnement complexe**
Hypothèses neutralistes	organismes simples et complexes	organismes complexes
Hypothèses sélectives	organismes simples	organismes complexes

Table III.1 – Afin de distinguer les hypothèses neutralistes des hypothèses sélectionnistes, une solution consiste à laisser évoluer des organismes (ici *in silico*) dans un environnement simple. En effet, dans ce cas, les deux hypothèses sont censées donner des résultats différents alors que dans un environnement complexe on ne peut pas les distinguer.

les organismes simples se satisfassent d'un environnement simple. C'est en suivant cette logique que nous nous proposons d'établir la distinction entre ces deux hypothèses. En effet, comme le montre très simplement la table III.1, c'est uniquement en faisant évoluer des organismes dans un environnement simple qu'il est possible de distinguer leurs causes sélectives et neutralistes de l'accumulation de la complexité biologique.

Cette approche présente cependant encore une difficulté centrale qui tient à la nécessité de définir ce qu'est un environnement *simple*. Parmi les plus marquants des résultats obtenus avec Aevol, le précédent chapitre nous a permis de considérer ceux qui ont un lien avec une interprétation de l'évolution en termes de complexité. Comme nous l'avons déjà souligné, ces résultats, malgré tout leur intérêt, ont été obtenus dans un environnement construit pour *sélectionner* la complexité (au sens ici du nombre de gènes). En outre, aucun d'eux ne s'étant focalisé spécifiquement sur la notion de complexité, aucune mesure rigoureuse n'a été définie jusqu'alors, comme si la complexité ne devait rester qu'une métaphore. La réticence à embrasser la notion de complexité peut s'expliquer par la difficulté que présente le fait de lui donner une définition objective pour un système biologique (et peut-être aussi par quelque scepticisme inhérent au vertige qu'inspire la multiplicité de la notion). Nous allons ici tâcher de montrer comment on peut tirer parti de l'approche d'évolution expérimentale *in silico* en général et d'Aevol en particulier pour nous donner un cadre d'étude qui permette de rendre compte de façon objective et quantitative de la complexité. Signalons d'emblée qu'il ne s'agit pas d'affirmer que ce serait *la bonne*, moins encore *l'unique*, mesure de complexité qui serait adéquate aux systèmes biologiques. Comme nous l'avons vu au chapitre I, la complexité ne peut pas être réduite à une unique grandeur scalaire car elle s'exprime à de multiples niveaux.

Avant même d'envisager une approche conceptuelle de la complexité des systèmes biologiques, on est envahi par la multiplicité des formes que prend la vie. Si l'on prend le point de vue du naturaliste, on peut s'étonner des différences entre les règnes autant que de la façon dont leurs interactions s'équilibrent, des symbioses, des dynamiques, des comportements que cela induit. Mais même à l'échelle supposée plus élémentaire de la microbiologie, on peut voir déjà la difficulté que pose la notion de complexité : un organisme procaryote peut coloniser de multiples environnements, les génomes procaryotes varient d'au moins un ordre de grandeur... Même pour des organismes unicellulaires tels que les bactéries, proposer une définition quantitative de la complexité reste une gageure, ce qui rend difficile l'établissement d'une méthodologie pour identifier son origine évolutive.

Si l'on considère des organismes *in silico* évoluant dans Aevol, nous avons montré au

chapitre précédent combien la complexité pouvait être multiforme et s'exprimer différemment aux différents niveaux de la *genotype-phenotype map*. La situation est cependant déjà meilleure : dans le cas des modèles *in silico* on a en effet une connaissance exhaustive des interactions entre ces différents niveaux auxquels cette complexité peut se développer. Même si la complexité n'est pas spécifiée à priori (elle résulte ici aussi des hasards de l'évolution), au moins connaît-on jusqu'aux plus infimes détails de sa mécanique. Demeure, à première vue, la difficulté centrale qui est que, même dans Aevol, il est quasiment impossible de définir ce que serait un organisme *simple* dans un environnement défini pour forcer l'accumulation théoriquement infinie de gènes (rappelons que ce n'est qu'au prix d'un somme *infinie* de triangles qu'on peut atteindre une cible phénotypique composée de gaussiennes). En revanche, si l'on accepte que la notion de simplicité (ou de complexité) doit être entendue relativement à un environnement donné (voir table III.1), alors il devient possible de définir concomitamment un *environnement* simple et un *organisme* simple. C'est ce que nous nous proposons de faire en définissant un protocole expérimental basé sur une forme d'environnement alternative à la classique somme de gaussiennes à laquelle Aevol recourait dans les précédents travaux.

2 Design expérimental

2.1 Définition de la cible phénotypique

Afin de distinguer les hypothèses sélectives des hypothèses neutralistes, nous avons vu (table III.1) qu'il est possible de faire évoluer des organismes simples dans un environnement simple. Si une telle expérience semble impossible à réaliser *in vivo* [1], elle est en revanche réalisable *in silico*, en particulier dans un modèle tel qu'Aevol. Il faut toutefois définir ce que serait un environnement simple dans le modèle.

Dans Aevol, l'individu le plus rudimentaire [2] possible comporte un seul gène, il produit alors une unique protéine-triangle et c'est cette dernière qui sera comparée à la cible phénotypique. Ainsi, pour peu que cette cible phénotypique soit triangulaire elle aussi, l'aspect rudimentaire d'un organisme à un seul gène pourrait lui valoir la qualification de *simple* car non seulement il userait de peu de moyens mais en outre il remplirait de façon totalement satisfaisante sa fonction métabolique (pour peu que les paramètres de la protéine-triangle soient adaptés à l'environnement).

Nous nous proposons donc de modifier le modèle Aevol de telle façon qu'un organisme rudimentaire puisse être parfaitement adapté. Pour cela il suffit de modifier le modèle au seul niveau de la forme de sa cible phénotypique : alors que la cible phénotypique traditionnelle est une somme de gaussiennes [3], nous avons défini une variante du modèle dans laquelle la cible phénotypique sera une fonction triangle isocèle. De cette façon, non seulement la comparaison entre un organisme et la cible devient-elle possible mais elle est alors même relativement triviale : la notion de simplicité peut en effet se restreindre, comme suggéré ci-dessus, aux organismes rudimentaires ne comportant qu'un seul gène. Nous verrons cependant (dans la section 3) que cette définition ne sera pas utilisée par la suite car elle confond la notion de gène et la notion de fonction.

Comme nous l'avons vu, dans Aevol, la notion d'environnement est modélisée par une

1. Essentiellement pour des raisons de temps mais on peut aussi remarquer qu'aucun organisme contemporain n'est simple. On peut objecter à cette remarque que certains viroïdes sont extrêmement frustres mais ceux-ci ne sont pas capables de se répliquer de façon autonome et nécessite donc un environnement complexe (une cellule hôte) pour proliférer.

2. Précisons quelque peu la distinction proposée entre le rudimentaire et le simple. La qualification de rudimentaire s'apparente ici à un simple jugement de valeur. En l'occurrence, ce qui est appelé organisme rudimentaire correspond à l'organisme le plus facile à produire. Par contraste, le sens qui est donné ici à la notion de simplicité a un aspect téléologique : c'est seulement relativement à une finalité qu'on peut être simple ou complexe. Dans l'absolu, on pourrait penser de façon plus élémentaire à procéder par comparaison : tel organisme serait plus simple ou au contraire plus complexe que tel autre. On n'obtiendrait cependant alors qu'une notion de complexité relative qui ne suffirait pas pour définir la simplicité ou la complexité, sauf à se donner une référence – qui pourrait être ici rapportée à cette notion d'organisme « rudimentaire ».

3. La cible phénotypique traditionnellement utilisée dans Aevol est la somme de trois gaussiennes définies comme suit : $G_1 \ : h = 1,2; m = 0,52; \sigma = 0,12$, $G_2 \ : h = -1,4; \sigma = 0,5; w = 0,07$ et $G_3 \ : h = 0,3; m = 0,8; \sigma = 0,03$. Les fonctions gaussiennes $x \mapsto h \exp(-(x-m)^2/2\sigma^2)$ étant ici caractérisées par leurs paramètres h (extremum de la fonction), m (abscisse de l'extremum), et σ (écart type de la fonction). Le fait qu'il s'agisse d'une fonction qui ne peut être atteinte par une somme de fonctions triangles (donc par une « somme » de protéines dans Aevol) interdit qu'aucun organisme ne puisse atteindre la fitness optimale. Ceci permet d'observer une dynamique évolutive sans limite de durée car il reste toujours un potentiel évolutif. En outre, ce paramétrage permet d'avoir deux lobes déconnectés qui peuvent être interprétés comme deux traits phénotypiques indépendants (voir figure III.1)

fonction mathématique qui associe un degré de réalisation d'une fonction biologique (un nombre réel dans l'intervalle $[0, 1]$) à l'ensemble des traits de l'espace fonctionnel (envisagé lui-même comme un nombre l'intervalle $[0, 1]$). Pour définir une cible phénotypique particulière, il suffit donc de définir une fonction mathématique de $[0, 1]$ dans [0,1]. Dans le cadre de notre exploration expérimentale de la complexité, nous allons définir une fonction simple (un triangle isocèle donc) mais aussi une fonction complexe afin de pouvoir comparer la complexité des organismes obtenus dans l'environnement simple avec une expérience « nulle » et estimer ainsi la part relative des hypothèses sélectives et non-sélectives dans l'accumulation de complexité biologique. En effet, comme nous l'avons déjà mentionné au chapitre I, la virulence du débat qui oppose Gould à Dawkins masque le fait que leurs hypothèses ne sont en pratique pas exclusives ! Cette comparaison sera effectuée via la définition d'une cible complexe composée d'une unique gaussienne.

Une fois ces choix généraux effectués, il reste à déterminer les paramètres de ces deux fonctions. Pour cela, il est nécessaire de mobiliser les connaissances détaillées du modèle afin de garantir la « simplicité » de la première mais aussi de garantir que la complexité d'individus ayant évolué en parallèle dans ces deux environnements soient effectivement comparable. Pour cela, nous avons défini ces deux fonctions comme suit :

Cible phénotypique simple La cible phénotypique utilisée pour notre première expérience est une fonction affine par morceaux, nulle sur l'intervalle $[0; 0, 4]$ et sur l'intervalle $[0, 6; 1]$ et qui forme sur $[0, 4; 0, 6]$ un triangle isocèle centré en $0, 5$ où elle atteint la valeur $0, 5$ (voir figure III.1 au centre). Ces paramètres ont été choisis de façon à garantir que cette cible soit identifiable par un seul gène (dans toutes les expériences conduites ici, on utilisera $w_{max} = 0, 1$ (w_{max} étant le paramètre qui limite la pléiotropie maximale des protéines dans le modèles – voir chapitre II). Cependant, cette cible reste difficile à atteindre car il est pour cela nécessaire que le gène soit long pour atteindre une précision suffisante. En effet, la valeur $m = 0, 5$ est difficile à atteindre et demande un code binaire virtuellement infini (voir la description du modèle, chapitre II). La cible est donc simple mais elle demande tout de même que le processus évolutif fasse son œuvre.

Cible phénotypique complexe La cible complexe est constituée d'une gaussienne centrée en $0, 5$, et normalisée en $0, 5$ (figure III.1 à droite). En outre, afin de permettre une comparaison optimale entre les cibles simple et complexe, l'écart-type de la gaussienne a été choisi de sorte que les aires des deux fonctions cibles soient identiques. En effet, l'interprétation géométrique des fonctions biologiques utilisées dans Aevol rend la complexité d'un organisme dépendante de l'aire de la fonction cible. La cible phénotypique retenue est donc la gaussienne $1/2 \cdot \exp(-100\pi(x - 1/2)^2)$ soit $\sigma \approx 0, 039$.

Ces deux fonctions, ainsi que la cible phénotypique classiquement utilisées dans Aevol, sont présentées figure III.1.

2.2 Exploration paramétrique

En dehors de la définition des deux cibles phénotypiques spécifiques présentées ci-dessus, le protocole expérimental est similaire à celui classiquement utilisé dans les expérience d'évolution expérimentale *in silico* utilisant Aevol (le lecteur pourra consulter la

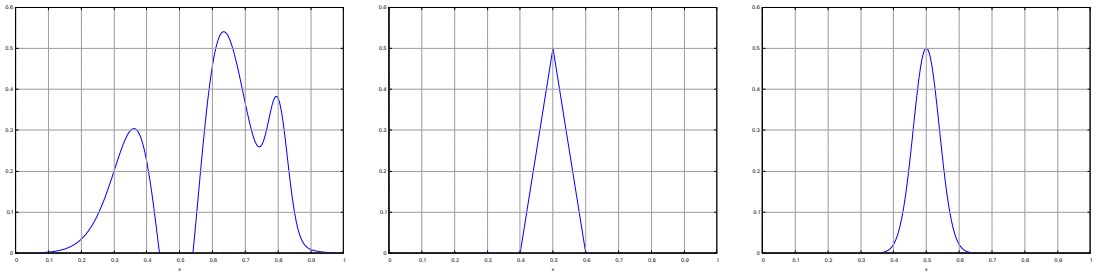


Figure III.1 – Les trois cibles phénotypiques d'Aevol mentionnées dans cette thèse. À gauche, la cible phénotypique classiquement utilisée dans Aevol composée d'une somme de trois gaussiennes. Cette cible a en particulier été utilisée dans toutes les expériences présentées au chapitre II. Au centre, la cible « simple » (triangulaire) définie dans le protocole expérimental (voir texte principal pour la justification). À droite, la cible « complexe », composée d'une unique gaussienne, définie pour conduire des expériences témoins dans un contexte où la complexité n'est pas bornée par la structure de l'environnement.

table II.2 pour la signification précise des différents paramètres.)

Nombre de générations : Toutes les simulations ont été conduites sur 270 000 générations et analysées sur 250 000 générations (voir section suivante).

Taille de population : Toutes les expériences ont été conduites avec une population de taille constante $N = 1\,024$ composée d'individus disposés sur une grille torique 32×32. Les expériences antérieures menées avec Aevol ont en effet montré que cette taille de population, bien que très faible au regard des populations biologiques réelles, limite les effets de la dérive.

Taux de sélection : Toutes les expériences ont été conduites avec un taux de sélection moyen $k = 1\,000$.

Taux de mutation : Les résultats précédents obtenus avec Aevol ont tous montré la forte influence des taux de mutation sur la complexité des organismes. Nous avons donc testé trois taux de mutation : $\mu = 10^{-4}$, $\mu = 10^{-5}$ et $\mu = 10^{-6}$ mut.bp^{-1}.gen^{-1}. Un quatrième taux de mutation avait initialement été inclus dans le protocole expérimental ($\mu = 10^{-7}$ mut.bp^{-1}.gen^{-1}) mais l'évolution s'est avérée trop lente pour que les données soient exploitables après « seulement » 270 000 générations.

Procédure d'initialisation : Pour notre étude, la procédure d'initialisation revêt une importance cruciale. En effet, elle est susceptible d'introduire un effet fondateur qui pourrait fausser nos résultats. Nous avons donc choisi d'initialiser les simulations par des populations clonales d'individus générés par *bootstrap*. Nous recherchons au hasard un génome de 5 000 bp contenant un gène fonctionnel favorable. De la sorte, nous garantissons que toutes les simulations sont initialisées avec une population viable composée d'individus les plus simples possibles.

Nombre de répétitions : Afin d'obtenir une puissance statistique raisonnable, nous avons conduit 100 répétitions pour chaque jeu de paramètre, soit un total de 300 simulations (100 par taux de mutation) pour chaque type d'environnement.

2.3 Analyse post-évolutive des lignées

Une simulation avec Aevol nous donne à voir une population d'individus avec généralement un fort polymorphisme. Dès lors, il est nécessaire, pour analyser la dynamique évolutive, de choisir ceux des individus qu'il faut considérer pour en évaluer la complexité. Traditionnellement deux approches sont utilisées dans la littérature : soit l'analyse, à chaque génération, du meilleur individu de la population, soit l'analyse des données moyennes dans la population. Dans notre cas, aucune de ces deux approches n'est acceptable. Les études antérieures avec Aevol ont en effet montré que (*i.*) le meilleur individu est souvent éliminé en raison d'une robustesse trop faible, rejoignant là les résultats présentés par Wilke *et al.* (2001), ce qui tendrait à biaiser la complexité vers le haut, et que (*ii.*) les valeurs moyennes sont souvent biaisées car les mécanismes de réarrangement ont tendance à supprimer des gènes, donc à biaiser la complexité vers le bas. Citons ici une troisième option, utilisée, par exemple dans le cadre de la plateforme Avida, qui consiste à analyser le plus grand groupe phylogénétique. Cette approche n'est cependant pas exploitable dans Aevol du fait du fort polymorphisme lié à la présence de nombreux éléments non-fonctionnels dans le génome.

Afin d'analyser, à chaque génération, un individu clairement représentatif de la dynamique évolutive, nous avons choisi d'extraire de la simulation la lignée des ancêtres de la population finale. Pour cela, au cours des simulations, Aevol enregistre toutes les relations de parenté dans un arbre phylogénétique (cette approche est simplifiée par l'absence de transfert horizontal dans nos simulations). À l'issue des 270 000 générations de chaque simulation, cet arbre est remonté à partir du meilleur individu final ce qui nous permet de reconstruire sa lignée et de mesurer les statistiques de fitness, de taille du génome, de taille du codant, de nombre de gènes, de structure du réseau de protéines... tout au long de cette lignée depuis la génération 0 jusqu'à la génération 270 000. Enfin, nous supprimons de ce jeu de données tous les enregistrements correspondant aux générations 250 000 et suivantes afin d'avoir l'assurance de n'étudier que ceux des organismes qui ont été fixés dans les lignées de la population finale (20 000 génération étant très largement supérieur au temps de coalescence pour une population de 1 024 individus – la notion même de lignée fixée n'ayant pas de sens lorsqu'on s'approche de la génération finale au delà du temps de coalescence de l'arbre phylogénétique). Cette procédure nous garantit que toutes les statistiques sur des ancêtres communs de la population finale sont enregistrées ainsi que leur représentativité relativement à la dynamique évolutive. Dans la suite du texte, l'ancêtre commun à la génération 250 000 sera appelé le « pseudo-coalescent » (en toute rigueur il ne s'agit pas du coalescent qui, lui, est inconnu puisque nous ne connaissons pas le temps de coalescence exact dans les simulations).

3 Mesures de complexité

Comme nous l'avons vu au chapitre 2.2, il n'y a pas de consensus quant aux mesures de complexité, ni parmi les biologistes, ni même dans la communauté de la vie artificielle. De plus, comme Aevol est un modèle multi-échelle, il faut de prime abord définir des mesures spécifiques à chaque niveau d'organisation car, si la complexité verticale – au sens défini par McShea (2017), c'est à dire le nombre de niveaux sur lesquels la complexité peut s'étendre – est fixée dans le cadre de notre modèle (génome, protéome, phénotype, environnement), il n'en demeure pas moins que la complexité dite horizontale – toujours selon McShea (2017) – s'exprime sur chacun de ces niveaux, potentiellement avec des disparités. Les mesures de complexité utilisées dans notre étude devront donc porter sur les séquences génétiques mais aussi sur les niveaux fonctionnels (en particulier le protéome). Il est à noter que, en vertu de notre protocole expérimental, la complexité n'est pas mesurée à l'échelle du phénotype puisque c'est à ce niveau précisément que la sélection exerce son contrôle direct et impose au phénotype de demeurer « simple » ou « complexe » suivant l'environnement choisi.

Même s'il n'y a pas de consensus établi sur les mesures de complexité, Chris Adami (2002b) a proposé une mesure de complexité intéressante, qui a été adoptée ou discutée par de nombreux auteurs, en particulier dans le domaine de la vie artificielle (Hintze et Adami, 2008; Edlund *et al.*, 2011; Auerbach et Bongard, 2014). En outre, cette mesure a été proposée dans un contexte proche du nôtre, à savoir la simulation de l'évolution. La proposition de Chris Adami (2002b) est que la complexité d'un système biologique soit quantifiée par la quantité d'information contenue dans ce système. L'idée sous-jacente de cette mesure est que, en termes d'information, le processus évolutionnaire correspond à un transfert d'information depuis l'environnement vers les individus. C'est sur cette idée que nous nous sommes appuyés pour quantifier la complexité dans nos simulations : nous avons adapté les principes de Adami *et al.* (2000) à Aevol de façon à obtenir des mesures quantitatives aux niveaux du génome et du protéome en estimant la quantité d'information stockée dans chacune de ces structures (voir sections 3.1 et 3.2). Cette approche nous a permis de disposer d'une mesure quantitative de complexité. Cependant, afin de distinguer les individus « simples » des individus « complexes » sans avoir à définir un seuil de complexité (qui serait nécessairement arbitraire) , nous avons complété ces mesures quantitatives par une classification qualitative des organismes *simples* ou *complexes* fondée sur la connaissance à priori de la structure du modèle et non sur la quantité d'information (voir section 3.3).

Notons qu'il est crucial, dans Aevol, d'analyser la complexité simultanément à différents niveaux de la *genotype-phenotype map* car, comme nous l'avons vu dans le chapitre II, celle-ci présente de nombreux degrés de liberté qui permettent de faire varier la quantité d'information transmise d'un niveau à l'autre (ou, pour le voir autrement, de compresser ou décompresser l'information codée aux différents niveaux). Dans cette étude, nous nous sommes concentrés sur deux niveaux : la séquence d'ADN (complexité « génomique ») et le protéome (complexité « fonctionnelle »). Nous avons considéré que le transcriptome était, dans Aevol, trop proche de la séquence d'ADN pour présenter un intérêt (rappelons qu'il n'y a pas de mécanisme de régulation génétique – donc pas de réseau de gènes – dans la version d'Aevol utilisée ici). Enfin, dans tout autre dispositif expérimental, un troisième niveau aurait pu présenter un intérêt, à savoir celui du phénotype. Cependant,

le contexte particulier de notre étude rend l'analyse de la complexité du phénotype inutile puisque celui-ci est astreint par la sélection à demeurer aussi simple que possible (dans le cas de la cible triangulaire « simple ») ou, au contraire, d'être le plus complexe possible (dans le cas de la cible gaussienne). Pour quantifier le phénotype, nous nous contenterons donc d'analyser sa fitness, c'est-à-dire sa distance à la fonction cible.

3.1 Mesure quantitative au niveau des séquences

Si on suit les préconisations de Chris Adami (2002b) pour estimer la complexité au niveau de la séquence, la mesure peut sembler triviale puisque, à première vue, il suffirait de mesurer la longueur du génome (équivalent donc à la valeur « C » en biologie). Cependant, comme d'ailleurs en biologie, cette mesure ne reflète en rien la quantité d'information présente sur le génome en raison, à la fois, de la dégénérescence du codage (le fait que plusieurs séquences différentes peuvent coder une même information) et de sa redondance (le fait qu'une même séquence puisse coder plusieurs informations).

Dégénérescence du codage : Dans Aevol, la dégénérescence du codage de l'information est relativement triviale. En effet, comme, dans le modèle, le code génétique n'est pas redondant (voir chapitre II, section 3.2.4 et figure II.5), il n'est pas nécessaire de tenir compte de la dégénérescence du code génétique [1]. La prise en compte de la dégénérescence doit donc uniquement être effectuée sur les séquences d'ADN, les séquences intergéniques ne contribuent pas au codage de l'information, et au niveau des séquences d'ARN, les séquences *Leader* (entre le promoteur de l'ARN et la séquence initiatrice du gène) et *Trailer* (entre le codon Stop d'un gène et le terminateur de l'ARN) ne contribue pas non plus au codage de l'information [2]. Enfin, une dernière situation doit être prise en compte, à savoir les ARN non codants. Il s'agit de séquences transcrites qui ne comportent pas de séquence initiatrice de traduction (ou pas de séquence Stop de traduction). Dans ce cas, cet ARN ne contribue pas au codage de l'information et ne doit donc pas être pris en compte dans la mesure de complexité.

Redondance du codage : Dans Aevol, il existe de nombreux mécanismes de redondance du code via le « partage de séquence ». Celui-ci peut se produire à l'échelle de la transcription, dans le cas de séquences polycistroniques (opérons) où dans le cas particulier où plusieurs promoteurs transcrivent une même séquence. Une deuxième situation de redondance se produit lorsque plusieurs gènes sont chevauchants, soit sur le même brin (lorsqu'ils sont lus sur des cadres de lecture différents), soit sur les brins complémentaires (ils sont alors lus dans des sens différents et sur des bases complémentaires). On notera que, comme pour la dégénérescence, certaines situations plus complexes peuvent se produire, par exemple si un gène chevauche un promoteur.

1. En toute rigueur, il reste quelques possibilités de dégénérescence sur la séquence de codons. Ainsi, si la séquence primaire d'une protéine comporte un codon Start, celui-ci sera ignoré (contrairement à la biologie où le codon Start est traduit en méthionine). Ces situations sont cependant très rares et ne modifient que très marginalement la mesure ; c'est pourquoi nous avons choisi de les négliger ici.

2. La situation décrite ici correspond au cas d'un ARN monocistronique – ne comportant qu'un unique gène. La situation peut cependant être beaucoup plus complexe, par exemple dans le cas d'un ARN polycistronique ou dans le cas d'un chevauchement de gènes, voir de promoteurs.

Selon Adami (2002b), la complexité d'une séquence génétique est assimilable à la quantité d'information codée par cette séquence. Compte tenu du codage binaire de l'ADN (un bit – donc ici une paire de base – correspondant à une information élémentaire), de la dégénérescence du code et de sa redondance, nous considérerons, pour un génome Aevol, que la quantité d'information correspond au nombre de bases dites « essentielles », c'est-à-dire celles qui changeraient le phénotype de l'organisme si elles mutaient. Cette mesure est directement utilisée pour quantifier la quantité d'information stockée dans le génome, que nous appelons complexité génomique, notée $\mathcal{C}_G$ dans la suite de ce texte. Cette définition tient bien compte de la dégénérescence (une base n'est comptabilisée que si elle contribue à l'information, soit en participant à la séquence d'un ou plusieurs gènes, soit en participant à une ou plusieurs séquences de signalisation, soit – plus rarement – les deux) et de la redondance (une base contribuant à coder plusieurs information n'étant comptabilisée qu'une fois [1]).

3.2 Mesure quantitative au niveau fonctionnel

La question de la mesure adéquate au niveau du protéome, que nous appelons *complexité fonctionnelle*, se montre un peu plus retorse. Il serait tentant de considérer que la complexité fonctionnelle est donnée par le nombre de protéines, ou du moins par le nombre de protéines *non dégénérées* [2] puisque les protéines dégénérées n'ont aucune contribution phénotypique et ne doivent donc pas être prises en compte dans la mesure. Cependant, une telle mesure sur-estimerait la complexité car, dans Aevol comme dans un organisme réel, il arrive fréquemment que deux gènes contribuent à une même fonction, suite à une duplication, par exemple. Dans ce cas, le modèle comptabiliserait deux protéines différentes alors alors qu'il ne s'agit que d'une amplification, *c.-à-d.* une augmentation de la concentration de la protéine dans la « cellule » sans changement de fonction. En d'autres termes, la duplication d'un gène ne modifie que le niveau d'expression de ce gène sans modifier la quantité d'information codée dans le protéome.

Comme il a été indiqué lors de la description du modèle (chapitre II, section 3.2.5), dans Aevol le protéome est un objet très abstrait, défini mathématiquement comme une collection de fonctions triangles. En nous inspirant de la complexité de Kolmogorov (Kolmogorov, 1963), nous proposons de rendre compte de la complexité de cette collection par le nombre de paramètres nécessaires et suffisants pour l'encoder. La mesure de complexité fonctionnelle, $\mathcal{C}_P$, est alors la somme du nombre de valeurs différentes pour m, w et h utilisées pour coder l'ensemble du protéome (chacune considérée avec une tolérance $\varepsilon = 0,001$ pour tenir compte des erreurs d'arrondi liées à la traduction de la séquence primaire en valeurs numériques). On notera que cette définition de la complexité fonctionnelle n'est pas sans rappeler la notion de « complexité pure », élaborée par McShea et Brandon (2010), qui se fonde sur le nombre de types de parties différentes que comporte un organisme.

La mesure ainsi définie prend donc le point de vue de la quantité d'information pour rendre compte de l'effort fourni par un individu pour tirer le meilleur parti de son envi-

1. Ce qui contribue à une forme de compression de l'information à l'échelle génomique comme observé, par exemple par Parsons *et al.* (2010) ou Rutten *et al.* (2019).

2. Une protéine dégénérée code un triangle dont l'aire est nulle, ce qui se produit dans deux situations : soit si $h = 0$ soit si $w = 0$.

ronnement, c'est à dire pour mimer au mieux sa cible phénotypique.

3.3 Classification qualitative

Des deux mesures que nous venons de proposer, ni l'une, ni l'autre ne permettent de proposer une distinction intrinsèque des organismes « simples » ou « complexes». Pourtant, une telle classification qualitative est nécessaire si nous voulons étudier les différences entre les dynamiques évolutives propres à la simplicité ou à la complexité.

L'idée première d'une classification des organismes serait de classer les phénotypes. Cependant, dans Aevol, la cible environnementale produit une contrainte au niveau du phénotype, contrainte forçant à la simplicité ou à la complexité dans la cas de notre protocole expérimental (voir ci-dessus). Cela nous interdit d'utiliser la fonction phénotypique pour classer les organismes car cela reviendrait à classer comme simples (respectivement complexes) les organismes adaptés et comme complexes (respectivement simples) les organismes mal adaptés ; la relation fitness-complexité étant alors imposée par la mesure. Nous avons de ce fait choisi de classer les organismes selon leur structure fonctionnelle, mettant ainsi l'accent sur le protéome. Une solution triviale aurait été de définir un seuil sur la mesure quantitative $\mathcal{C}_P$ mais ce seuil aurait été arbitraire. Afin donc d'éviter d'avoir à définir, de façon tautologique, la simplicité comme une faible complexité (*c.-à-d.* une complexité inférieure à un seuil), nous avons choisi de revenir à la définition du modèle et à la définition du protocole expérimental pour définir les deux « classes de complexité ».

Pour délimiter la classe des organismes que nous appellerons simples, commençons par considérer un organisme Aevol qui aurait fait évoluer un gène lui permettant de produire une protéine en correspondance avec la cible phénotypique mais dont le degré d'activation ne serait pas idéal. Graphiquement, on aurait une protéine-triangle centrée sur la cible phénotypique, de même largeur mais dont la hauteur serait différente. Trois chemins évolutifs permettent à cet organisme de passer de cette situation mal adaptée à un phénotype conforme à la cible. Premièrement, une ou plusieurs *mutations du promoteur* entraînant un changement du taux de transcription (valeur e, chapitre II, section 3.2.4) et *in fine* celui de la concentration de la protéine. Deuxièmement, l'activité intrinsèque de la protéine codée par le gène lui-même peut également évoluer via un changement dans la séquence des codons H0 et H1 entraînant une modification de la hauteur du triangle (chapitre II, 3.2.5). Ces deux premiers chemins ont la particularité de modifier l'unique gène de notre organisme. En revanche, le troisième chemin recourt à une stratégie très différente. Il consiste à adjoindre à l'organisme un nouveau gène et la nouvelle protéine associée répondant à la même fonction. Cela peut se produire notamment par duplication du gène précédent selon un mécanisme *d'amplification* couramment observé en biologie lorsque des organismes s'adaptent à un nouvel environnement [1]. Or, dans la mesure où ces trois chemins sont équivalents du point de vue de la fonction de l'organisme, nous devons trouver un critère de classification tel que ces trois organismes, génétiquement différents et résultant d'histoires évolutives distinctes, soient tous les trois considérés comme simples. Cette remarque conduit à la définition suivante :

1. Ce troisième chemin évolutif permet de revenir de façon plus précise sur la distinction faite précédemment entre ce que nous avons appelé organisme *rudimentaire* et ce que nous voulons maintenant appeler organisme *simple*. Il nous montre qu'un organisme simple peut comporter un nombre arbitraire de gènes, donc une complexité génomique et une complexité fonctionnelle arbitraire.

seront considérés comme « *simples* » les organismes Aevol dont toutes les protéines non dégénérées ont même moyenne m et même demi-largeur w.

En effet, dans cette configuration, leurs fonctions s'additionnent linéairement pour donner un phénotype triangulaire avec les mêmes caractéristiques de moyenne et de demi-largeur [1] En d'autre termes, dans cette configuration *simple*, toutes les protéines de l'organisme répondent à la même fonction et ne peuvent différer tout au plus que par leur degré d'activation h. Dans le cas contraire, il seront considérés comme « *complexes* ». Nous verrons au chapitre suivant que le bien fondé de ces catégories est confirmé à posteriori par la cohérence observée entre ces deux classes et les mesures de complexité $\mathcal{C}_G$ et $\mathcal{C}_P$. En particulier, on constatera à quel point la *stratégie évolutive* (si l'on se permet ce raccourci de langage) des deux catégories est différente. On notera que tout organisme ne comportant qu'un seul gène sera systématiquement classé comme « simple » mais aussi qu'un organisme simple n'est pas nécessairement adapté à la fonction cible (ce qui est d'ailleurs le cas pour les organismes générés aléatoirement à l'initialisation).

3.4 Interactions entre les trois mesures de complexité

Ces trois mesures de complexité vont nous permettre de suivre et de comparer l'évolution de la complexité entre les différentes conditions expérimentales et, pour une même condition, entre les répétitions. La classification qualitative en particulier va permettre de suivre à gros grain le comportement des différentes répétitions. En outre, pour un même organisme, d'éventuelles différences de comportement entre ces différentes mesures seront susceptibles de nous renseigner sur l'origine des pressions évolutives auxquelles elles sont soumises. En effet, s'il est évident que les mesures de complexité ne sont pas indépendantes les unes des autres, il est important de garder à l'esprit qu'elles ne sont pas non plus équivalentes, car il existe des degrés de liberté dans les mécanismes de décodage de l'information génétique. Qu'on pardonne ici une anticipation sur les résultats du chapitre IV, qui semble assez éclairante pour qu'on fasse l'économie d'une description qui semblerait abstraite : la figure III.2, compare les complexités fonctionnelle et génomique entre elles pour les 300 simulations d'évolution présentées au chapitre IV (sous trois taux de mutation différents, voir section 2.2). Elle représente la complexité fonctionnelle du pseudo-coalescent (voir la définition en section 2.3) de chaque répétition en fonction de sa complexité génomique. C'est d'abord la corrélation qui est remarquable et qui confirme, sans qu'il y ait lieu de s'en étonner, la façon dont ces grandeurs dépendent l'une de l'autre. Les franges colorées que constituent les nuages de points aux différents taux de mutation suscitent aussi la curiosité : on y retrouve, de fait, le résultat de Knibbe *et al.* (2007a) (voir chapitre II, section 4.1) dont il découle que plus le taux de mutation est fort moindre est la complexité génomique. Mais ce que montre aussi cette figure c'est qu'au delà de ces déterminations, on note la dispersion des nuages de points : pour une complexité génomique donnée on observe une large plage de complexités fonctionnelles. Ceci souligne combien complexité fonctionnelle et complexité génomique gardent l'une par rapport à l'autre une liberté que nous allons tâcher d'éclaircir dans les prochains chapitres.

1. Comme pour la mesure quantitative de complexité fonctionnelle, ces égalités s'entendent aux approximations de calcul près. Nous considérerons donc ici l'égalité avec une tolérance $\varepsilon = 0,001$.

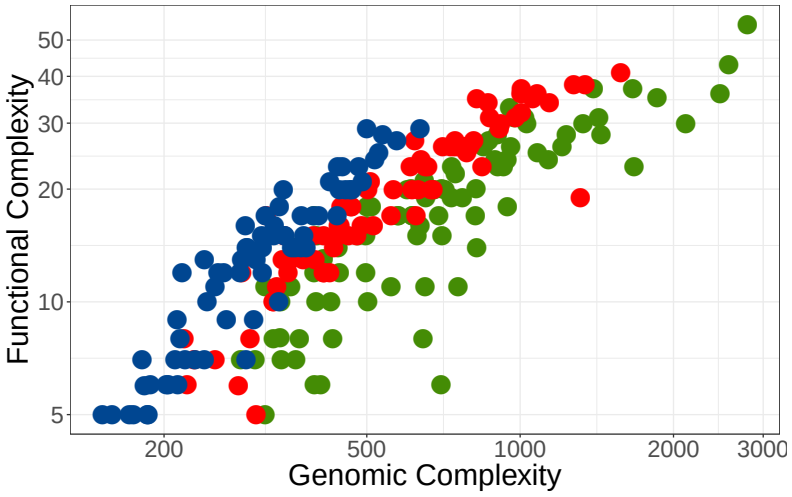


Figure III.2 – Complexité fonctionnelle en fonction de la complexité génomique. Chaque point représente le pseudo-coalescent (voir section 2.3) d'une répétition. Les couleurs indiquent les différents taux de mutation : Bleu : 10^{-4} mut.bp^{-1}.gen^{-1} ; Rouge : 10^{-5} mut.bp^{-1}.gen^{-1} ; Vert : 10^{-6} mut.bp^{-1}.gen^{-1}

4 Conclusion

Dans ce chapitre, nous avons proposé une méthodologie expérimentale se fondant sur Aevol pour éclairer la controverse entre explications sélectives et neutralistes de la complexité biologique. Nous avons aussi proposé trois mesures de complexité qui permettent de comparer entre elles les différentes expériences proposées. Forts de ces contributions, nous avons développé une version spécifique d'Aevol, qui inclut une nouvelle cible phénotypique et différents post-traitements destinés à l'analyse post-évolution des simulations (reconstruction des lignées, mesure des différentes complexités, etc). Nous n'entrerons pas dans les détails de la modification du code qui relève de l'ingénierie informatique plus que du travail scientifique (le code de la version modifiée d'Aevol est disponible en ligne : `http://www.aevol.fr/publications/ressources/Liard2018_ALife_src.tgz`). Dans les prochains chapitres, nous présenterons en détail l'analyse des résultats des simulations. Dans le chapitre IV, nous étudierons spécifiquement les simulations conduites dans un environnement simple (cible triangulaire) de façon à identifier les causes évolutives de la complexité. Dans le chapitre V, nous étudierons les simulations conduites en environnement complexe (cible gaussienne) pour, en les comparant à celles conduites en environnement simple, identifier d'éventuelles limites à l'accumulation de complexité au cours de l'évolution.

Chapitre IV

L'origine de la complexité : le cliquet épistatique

1 Introduction

Devant l'impossibilité de départager les explications sélectives et non-sélectives de l'accumulation de complexité au cours de l'histoire évolutive (chapitre I), nous avons proposé d'utiliser une approche d'évolution expérimentale *in silico* (ou *digital genetics*) avec la plateforme de simulation Aevol (chapitre II) pour conduire une série « d'expériences impossibles » (O'Neill, 2003) permettant de distinguer ces deux catégories d'hypothèses. Nous avons ensuite défini un protocole expérimental qui consiste à faire évoluer des populations dans des conditions environnementales telles que les plus rudimentaires de leurs individus (donc les plus simples [1]) soient parfaitement en mesure de survivre et de se reproduire (chapitre III, section 2). Nous avons ensuite défini un cadre d'analyse de ces simulations appuyé sur trois mesures de complexité (chapitre III, section 3). Dans le présent chapitre, nous allons analyser les résultats des simulations d'évolution dans un environnement simple. Dans un premier temps (section 2) nous verrons que, malgré la simplicité de la fonction cible nous observons une tendance « naturelle » à la complexification dans la grande majorité (75% environ) des répétitions et ce pour les trois mesures de complexité et pour les trois taux de mutation.

Dans les sections suivantes, nous analyserons plus finement les caractéristiques des populations et organismes afin d'identifier l'origine évolutive de cette tendance à la complexification. Nous testerons en particulier les hypothèses sélectives (section 3.1), sélectives indirectes (sélection de la robustesse ou de l'évolvabilité, section 3.2) et non-sélectives (en particulier la « Zero-Force Evolutionary Law » (McShea et Brandon, 2010), section 4). Devant l'absence d'explication probante, nous analyserons la dynamique évolutive des organismes complexes (section 5), ce qui nous permettra d'identifier un mécanisme de « cliquet de la complexité » (*complexity ratchet*) expliquant les dynamiques observées dans nos simulations. Nous conclurons alors en discutant cette hypothèse.

1. Voir chapitre précédent pour un discussion sur la différence entre un individu « rudimentaire » et un individu « simple ».

2 Même en environnement « simple » la complexité est la norme

Dans cette expérience, nous avons fait évoluer des populations dans un environnement simple sous trois taux de mutation à raison de 100 répétitions pour chaque taux. Parmi ces 300 simulations, 229 répondent à la définition des organismes complexes (voir chapitre III, section 3) à la génération 250 000. La table IV.1 montre la répartition des organismes simples et complexes pour les trois taux de mutation analysés et la figure IV.1 présente deux organismes types (un simple, figure IV.1.B, et un complexe, figure IV.1.C et D) ayant évolué sous un taux de mutation intermédiaire ($\mu = 10^{-5}$ mut.bp^{-1}.gen^{-1}).

	Nombre de *simples*	Nombre de *complexes*	Total
$\mu = 10^{-4}$	32 [24 – 43]	68 [58 – 76]	100
$\mu = 10^{-5}$	25 [18 – 34]	75 [66 – 82]	100
$\mu = 10^{-6}$	14 [9 – 22]	86 [78 – 91]	100

Table IV.1 – Nombre de lignées simples et complexes à la génération 250 000 pour les trois taux de mutation testés, exprimés en mut.bp^{-1}.gen^{-1}. Les nombres entre crochets indiquent les intervalles de confiance à 95% ($CI_{95\%}$) estimés par la méthode de Wilson à partir du nombre d'exemples dans chacune des deux classes.

L'observation d'une très large proportion d'organismes complexes malgré la simplicité de l'environnement pousse intuitivement en faveur d'hypothèses non-sélectives à l'accumulation de complexité. Cependant, avant de rechercher plus rigoureusement les causes de cette accumulation, nous avons d'abord vérifié que les organismes qui répondent à notre définition du *complexe* (respectivement du *simple*) correspondent bien à ceux qui accumulent de l'information (respectivement qui n'en accumulent pas) au niveau du génome (complexité génomique $\mathcal{C}_G$) et du protéome (complexité fonctionnelle $\mathcal{C}_P$).

Les figures IV.2 et IV.3 montrent la quantité d'information stockée dans les génomes (complexité génomique, $\mathcal{C}_G$, figure IV.2) et dans les protéomes (complexité fonctionnelle, $\mathcal{C}_P$, figure IV.3) pour pour les organismes simples et complexes et à tous les taux de mutation. Il est à noter que les complexités $\mathcal{C}_G$ et $\mathcal{C}_P$ ne peuvent pas être comparées quantitativement dans la mesure où elles se fondent sur des mesures essentiellement distinctes : respectivement, la quantité d'information portée par une séquence binaire et la quantité d'information dans un ensemble discret de valeurs réelles.

Les figures IV.2 et IV.3 montrent clairement que les organismes simples ont tendance à accumuler moins d'information dans leurs génomes et dans leurs protéomes. On notera cependant que la différence entre les organismes simples et complexes est moins prononcées pour les génomes (figure IV.2) que pour les protéomes (figure IV.3).

Cette observation n'est pas étonnante étant donné que notre classification qualitative est fondée sur la structure du protéome et qu'Aevol offre des degrés de liberté entre l'information codée sur le génome et sur le protéome (voir la description du modèle, chapitre II). Ces deux mesures montrent aussi que le taux de mutation a un effet important : plus la pression mutationnelle est forte, plus les complexités $\mathcal{C}_G$ et $\mathcal{C}_P$ sont faibles. Là encore, il n'y a pas lieu de s'en étonner puisque l'effet du taux de mutation sur la structure des génomes a déjà été décrit dans la littérature (Knibbe *et al.*, 2007a; Fischer *et al.*, 2014).

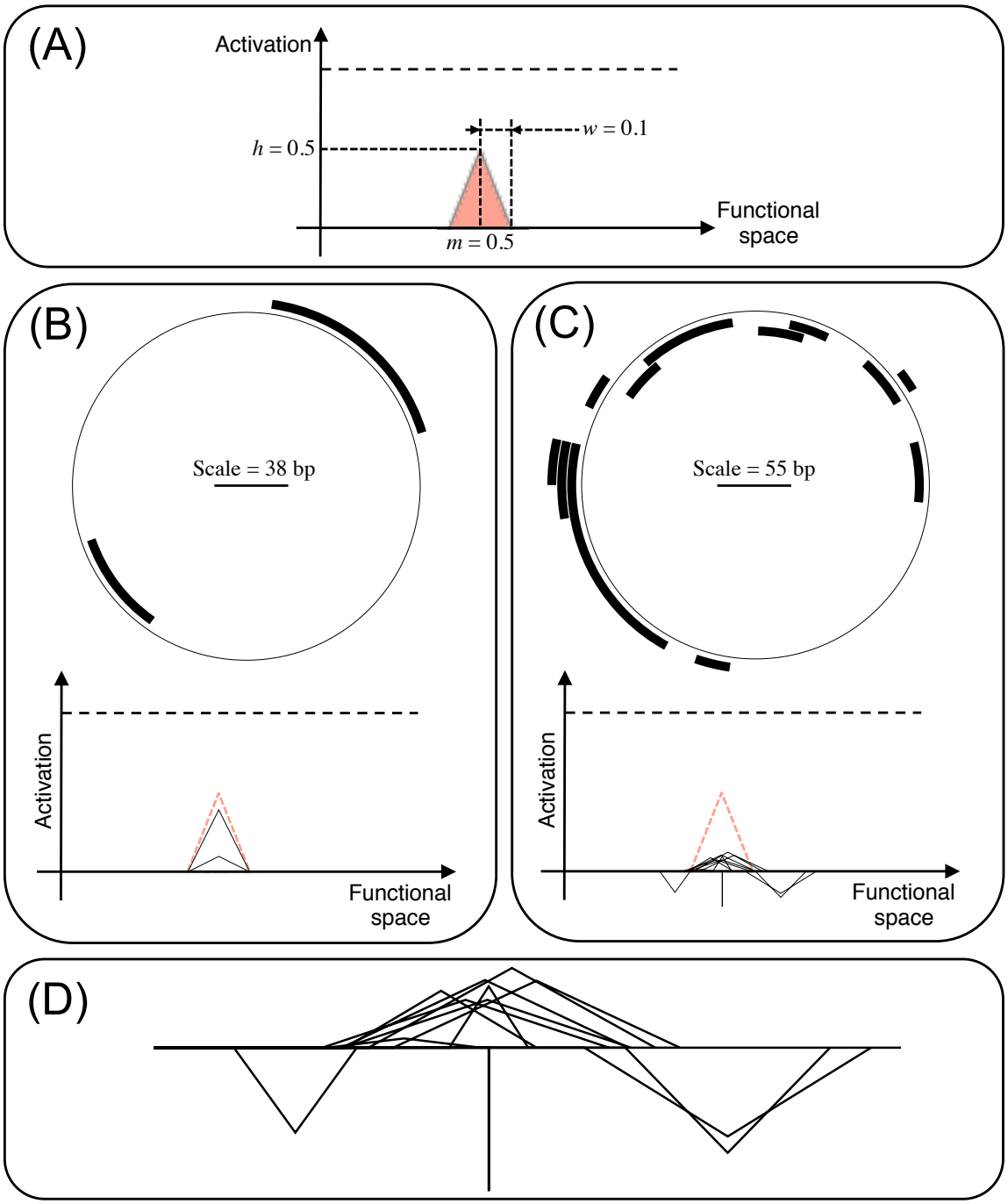


Figure IV.1 – (A) Cible phénotypique utilisée pour les expériences. (B, C) Génomes (en haut, les arcs noirs représentent les segments codants – les gènes – sur les deux brins du chromosome circulaire) et le protéome (en bas, la ligne rouge pointillée indique la fonction cible et les triangles noirs indiquent la fonction des protéines) d'un individu simple (B) et d'un individu complexe (C) (tous deux ayant évolué exactement dans les mêmes conditions ; $\mu = 10^{-5}$ mut.bp^{-1}.gen^{-1}). (D) Gros plan sur le protéome de l'individu complexe montré en (C).

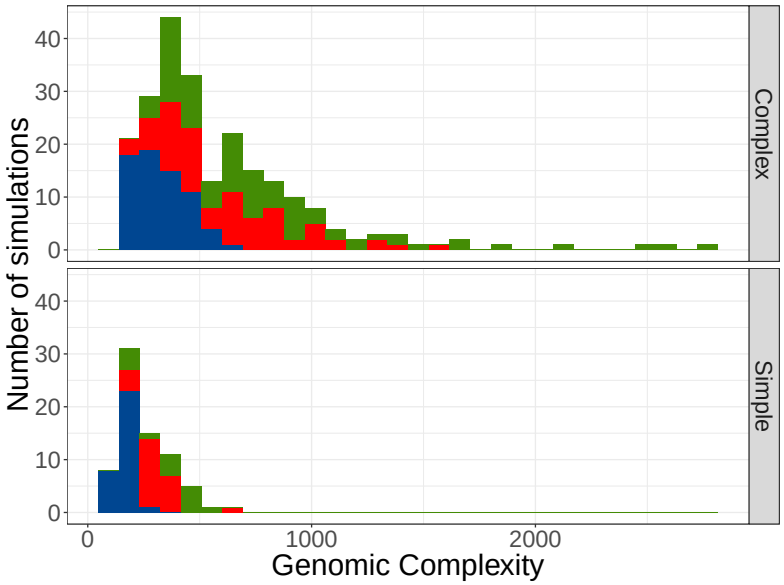


FIGURE IV.2 – Distribution de la complexité génomique ($\mathcal{C}_G$) pour les organismes complexes (en haut) et les *simples* (en bas) pour les pseudo-coalescents à la génération 250 000. Les couleurs se rapportent aux taux de mutation : bleu : 10^{-4} mut.bp^{-1}.gen^{-1} ; rouge : 10^{-5} mut.bp^{-1}.gen^{-1} ; vert : 10^{-6} mut.bp^{-1}.gen^{-1}.

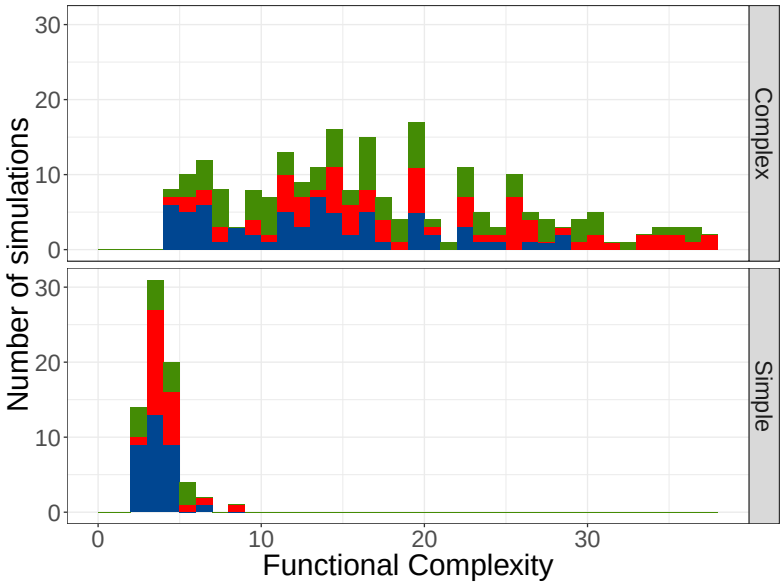


FIGURE IV.3 – Distribution de la complexité fonctionnelle ($\mathcal{C}_P$) pour les organismes complexes (en haut) et les *simples* (en bas) pour les pseudo-coalescents à la génération 250 000. Les couleurs se rapportent aux taux de mutation : bleu : 10^{-4} mut.bp^{-1}.gen^{-1} ; rouge : 10^{-5} mut.bp^{-1}.gen^{-1} ; vert : 10^{-6} mut.bp^{-1}.gen^{-1}.

Cependant, nous notons que, contrairement à la tendance remarquée sur la quantité d'information, cet effet est plus prononcé sur le génome (figure IV.2) que sur le protéome

(figure IV.3), probablement du fait que l'effet de taux de mutation est dû à un mécanisme de robustesse (Knibbe *et al.*, 2007a) et qu'à ce titre, il affecte directement le génome mais seulement indirectement le protéome. On notera d'ailleurs que cet effet était déjà perceptible sur la fraction d'organismes simples et complexes. En effet, nous observons (table IV.1 ci-dessus) que la proportion de simples baisse avec le taux de mutation, la variation étant significative, quoique relativement faible, du moins si l'on compare les taux de mutation les plus éloignés.

3 Hypothèses sélectives

3.1 Sélection directe

Ayant observé que, dans un même environnement simple, des organismes apparentés peuvent évoluer vers une structure simple ou complexe la question essentielle qui se pose est alors de déterminer si la complexité est contrôlée par l'effet de la sélection. La figure IV.4 montre la fitness du dernier ancêtre commun à la génération 250 000 (le « pseudo-coalescent ») pour chaque répétition en fonction de la complexité génomique $\mathcal{C}_G$. La figure IV.5 présente quant à elle la fitness en fonction de la complexité fonctionnelle $\mathcal{C}_P$.

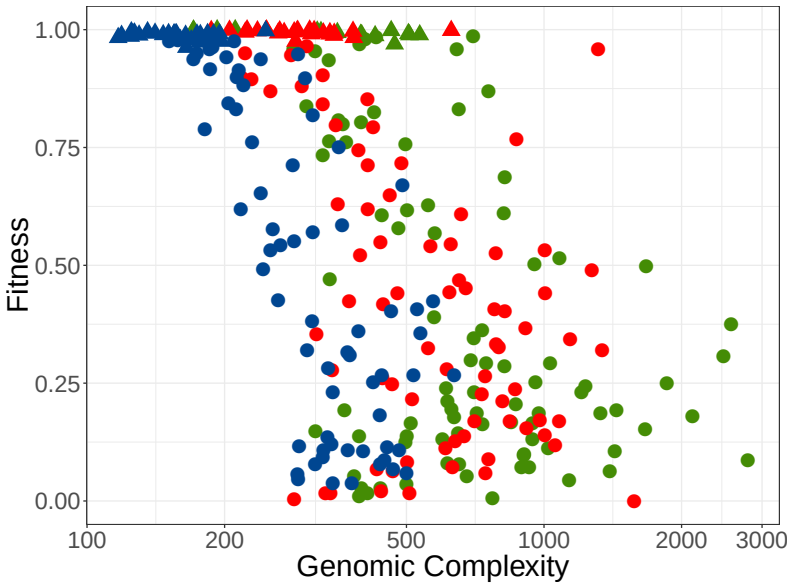


Figure IV.4 – Fitness du pseudo-coalescent (ancêtre commun à la génération 250 000) en fonction de la complexité génomique $\mathcal{C}_G$ (en échelle logarithmique). Les triangles et les cercles indiquent respectivement les pseudo-coalescents classés comme simples ou complexes. Les couleurs se rapportent aux taux de mutation : bleu : 10^{-4} mut.bp^{-1}.gen^{-1} ; rouge : 10^{-5} mut.bp^{-1}.gen^{-1} ; vert : 10^{-6} mut.bp^{-1}.gen^{-1}.

Au vu des figures IV.4 et IV.5 il apparaît clairement que les organismes simples ont une fitness plus élevée que les organismes complexes quelle que soit la mesure de complexité. Ceci est en outre confirmé par la distribution des fitness sur les deux classes qualitatives de complexité : la figure IV.6 montre que les organismes simples atteignent tous une fitness proche de 1, la meilleure fitness possible dans Aevol (fitness moyenne des organismes simples : $0,97 \pm 0,02$). Au contraire, les organismes complexes sont très disparates mais peinent à atteindre une fitness qui dépasse les $0,5$ (fitness moyenne des organismes complexes : $0,38 \pm 0,04$).

Ces résultats montrent que dans nos simulations ce n'est pas la sélection qui pousse les populations vers la complexité génomique ou fonctionnelle. Au contraire, la complexité fonctionnelle évolue ici *malgré* la sélection qui favorise clairement les « simples ».

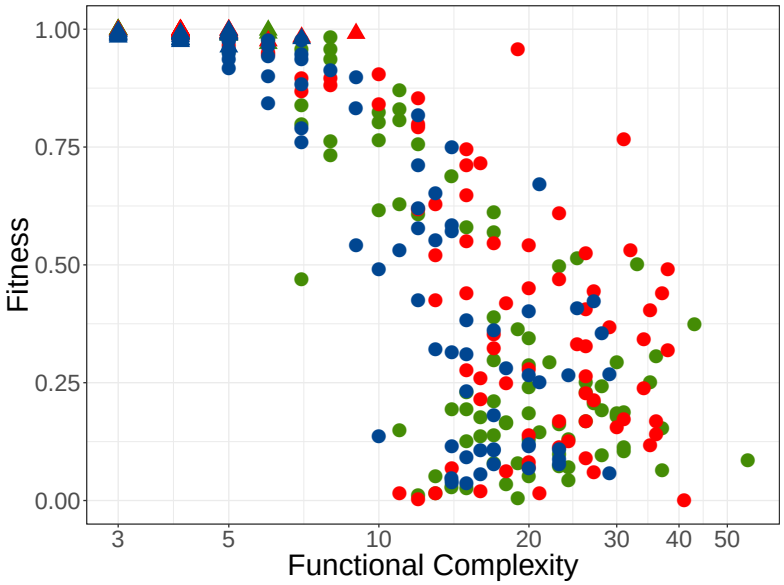


Figure IV.5 – Fitness du pseudo-coalescent (ancêtre commun à la génération 250 000) en fonction de la complexité fonctionnelle $\mathcal{C}_P$ (en échelle logarithmique). Les triangles et les cercles indiquent respectivement les pseudo-coalescents classés comme simples ou complexes. Les couleurs se rapportent aux taux de mutation : bleu : 10^{-4} mut.bp^{-1}.gen^{-1} ; rouge : 10^{-5} mut.bp^{-1}.gen^{-1} ; vert : 10^{-6} mut.bp^{-1}.gen^{-1}.

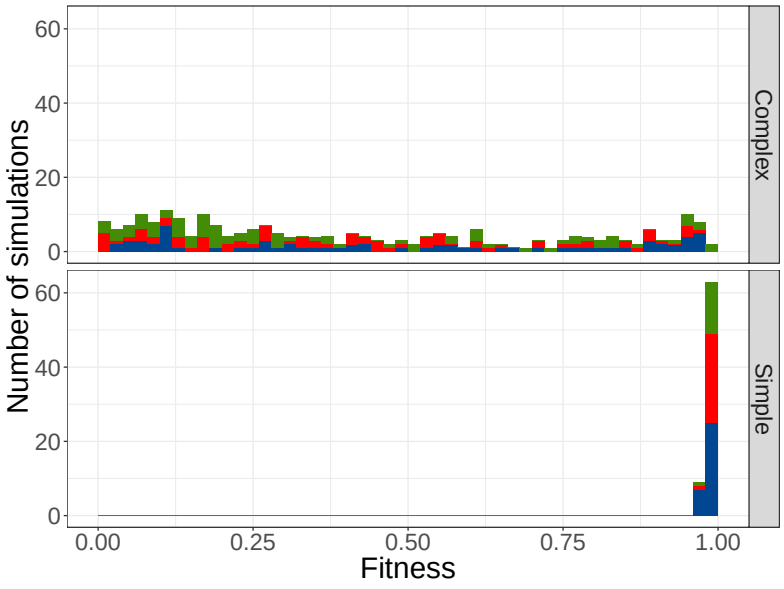


Figure IV.6 – Distribution des valeurs de fitness du pseudo-coalescent pour les organismes classés comme complexes (en haut) et simples (en bas). Les couleurs se rapportent aux taux de mutation : bleu : 10^{-4} mut.bp^{-1}.gen^{-1} ; rouge : 10^{-5} mut.bp^{-1}.gen^{-1} ; vert : 10^{-6} mut.bp^{-1}.gen^{-1}.

3.2 Sélection indirecte

Les résultats précédents tendent à invalider l'hypothèse sélectionniste pour expliquer l'accumulation de complexité, à tout le moins dans nos simulation. Il serait cependant prématuré de formuler cette conclusion. En effet, il a été montré que, dans certaines situations, la sélection indirecte (c'est à dire la sélection pour la robustesse ou pour l'évolvabilité) peut être assez forte pour dépasser la sélection directe en faveur de la fitness (Wilke *et al.*, 2001). De façon à invalider totalement l'hypothèse sélective, nous avons donc estimé la robustesse et l'évolvabilité des pseudo-coalescents simples et complexes à la génération 250 000 afin de vérifier si la sélection indirecte pouvait expliquer la tendance à la complexification.

3.2.1 Analyse de la robustesse

La robustesse est une caractéristique des organismes biologiques qui a fait l'objet, ces dernières années, d'une grande attention (Van Nimwegen *et al.*, 1999; Wilke *et al.*, 2001; De Visser *et al.*, 2003; Wilke et Adami, 2003; Lenski *et al.*, 2006; Wagner, 2007; Elena *et al.*, 2007; Félix et Wagner, 2008; Wagner, 2008). Malgré celà, les effets de la robustesse sur la dynamique évolutive sont souvent mal compris, entre autre parce que la robustesse elle-même est – selon nous – souvent mal définie. En effet, la robustesse est souvent définie comme la capacité à conserver la fitness ancestrale en présence de mutations. Ainsi Wagner (2008) écrit : « I call a biological system mutationally robust if its function or structure persists after mutations in its parts » (Wagner, 2008). Or, cette forme de robustesse – la robustesse *mutationnelle* – n'est finalement que peu significative pour caractériser la sélection indirecte. La robustesse est aussi parfois définie comme l'absence de variation. Ainsi, dans le même article, Wagner (2008) écrit aussi que : « high robustness implies low production of heritable phenotypic variation » (Wagner, 2008). Or, s'agissant d'évolution, ce qui importe n'est pas tant la robustesse vis à vis des mutations ou l'absence de variation phénotypique mais « simplement » la capacité d'un organisme à conserver sa fitness ancestrale au cours du temps et au cours des réplications. Or, cette propriété n'est pas directement équivalente à l'absence de variation phénotypique. En outre, elle dépend certes de la robustesse mutationnelle de l'organisme mais aussi de la probabilité de mutation, celle-ci pouvant elle-même dépendre des caractéristiques physiologiques ou moléculaires des individus (en particulier la taille de son génome) ou de l'environnement (avec *p. ex.*la présence d'éléments mutagènes ou de virus).

Par la suite, nous définirons donc la robustesse comme *la capacité d'un organisme à se reproduire de façon neutre* et, en cas d'ambiguïté, nous appellerons cette propriété la « robustesse réplicative » (par opposition à la robustesse mutationnelle). Nous mesurerons la robustesse réplicative d'un organisme de fitness f comme la fraction de descendants neutres F_ν (Knibbe *et al.*, 2007a), c'est à dire « la fraction de descendants dont la fitness est égale à f ». On notera que cette fraction inclut (*i.*) la fraction de descendants ayant subi une ou plusieurs mutations neutres (la robustesse mutationnelle) et (*ii.*) la fraction de descendants neutres n'ayant subi aucune mutation.

La robustesse peut être difficile à estimer *in vivo*. *In silico*, en revanche, il est généralement possible (même si ce n'est pas toujours simple) de développer des outils d'analyse permettant de quantifier ce type de propriété pour un organisme ou pour une lignée. Ainsi,

dans Aevol, nous avons défini un outil numérique de mesure de la robustesse en utilisant une approche de Monte-Carlo (Metropolis et Ulam, 1949) : partant d'un organisme donné, nous pouvons utiliser Aevol pour produire 10 000 000 (dix millions) de descendants de cet individu et mesurer la fitness de chacun de ces descendants. En comparant la fitness de l'individu initial avec la fitness de chacun de ses descendants, nous pouvons estimer F_ν, la fraction de descendants neutres, et donc la robustesse sélective. On notera que la même procédure peut être utilisée pour évaluer la robustesse mutationnelle, à condition de distinguer les descendants ayant subi une ou plusieurs mutations.

Dans le cas présent, nous avons utilisé cette méthode pour estimer la robustesse réplicative du pseudo-coalescent. Nous pouvons ainsi estimer si les individus complexes ont une meilleure robustesse que les *simples*, ce qui pourrait expliquer l'accumulation de complexité malgré le coût sélectif de celle-ci (voir section précédente).

La figure IV.7 présente les mesures de robustesse réplicative pour les organismes simples (en bleu) et complexes (en rouge) pour les trois taux de mutation (étant donnée la définition de la robustesse réplicative, agréger les données issues de taux de mutation différents n'aurait aucun sens). Elle montre que les organismes simples sont significativement plus robustes que les organismes complexes quel que soit le taux de mutation.

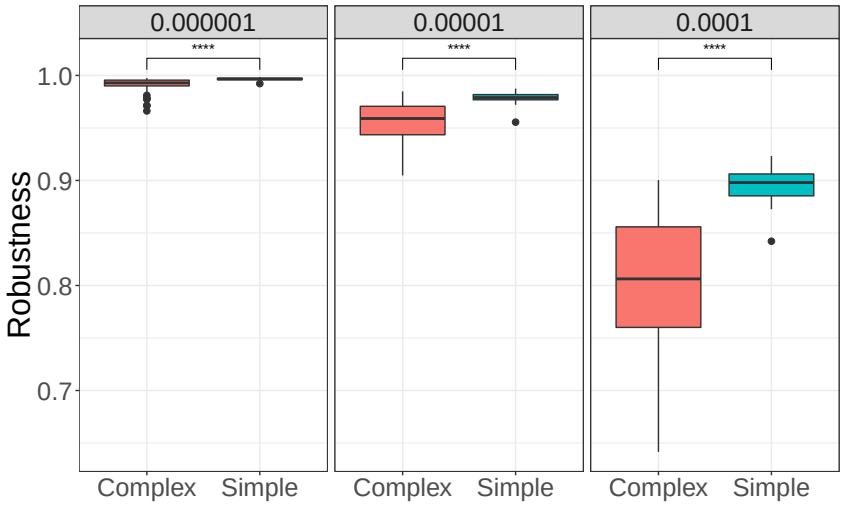


Figure IV.7 – Estimation de la robustesse des pseudo-coalescents simples (en bleu) et complexes (en rouge) pour chaque taux de mutation. Tous les écarts sont statistiquement significatifs (test de wilcoxon), p-value $< 4,0 \times 10^{-5}$ pour 10^{-4} mut.bp^{-1}.gen^{-1} , p-value $< 5,8 \times 10^{-9}$ pour 10^{-5} mut.bp^{-1}.gen^{-1} et p-value $< 5,7 \times 10^{-13}$ pour 10^{-6} mut.bp^{-1}.gen^{-1}.

La différence de robustesse réplicative entre les organismes simples et complexes trouve une explication relativement immédiate. En effet, le génome des individus complexes est plus long et contient davantage d'information (voir les figures IV.2, IV.3 et IV.8). Par conséquent, ces génomes présentent une cible mutationnelle plus grande et, pour un taux de mutation fixé, un plus grand nombre d'événements mutationnels. Il en résulte que leur robustesse réplicative est plus faible (Eigen et Schuster, 1977; Knibbe *et al.*, 2007a; Fischer *et al.*, 2014). Nous pouvons donc conclure que l'accumulation de complexité observée dans la majorité de nos simulations n'est pas du à un avantage sélectif indirect lié à la

robustesse.

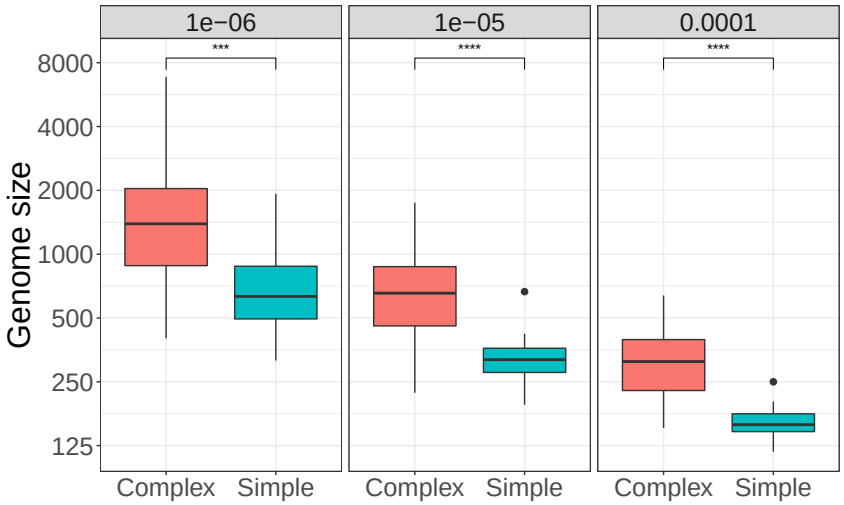


FIGURE IV.8 – Distribution de taille de génomes des pseudo-coalescents simples (en bleu) et complexes (en rouge) pour chaque taux de mutation. Quel que soit le taux de mutation, les génomes des organismes complexes sont significativement plus longs que les génomes des organismes simples. Tous les écarts sont statistiquement significatifs (test de wilcoxon), p-value $< 1,0 \times 10^{-4}$ pour 10^{-4} mut.bp^{-1}.gen^{-1} , p-value $< 7,0 \times 10^{-9}$ pour 10^{-5} mut.bp^{-1}.gen^{-1} et p-value $< 7,5 \times 10^{-13}$ pour 10^{-6} mut.bp^{-1}.gen^{-1}).

3.2.2 Analyse de l'évolvabilité

L'évolvabilité est un concept souvent associé à la robustesse, ces deux concepts étant même souvent considérés comme l'opposé l'un de l'autre puisque le premier est considéré comme favorable à l'évolution quant le second lui serait défavorable. Cette vision est cependant simpliste, car la robustesse peut, dans certains cas, augmenter l'évolvabilité en autorisant des variations génétiques masquées au niveau phénotypique et donc en permettant d'augmenter la connectivité du *fitness landscape* (Wagner, 2008; Ten Tusscher et Hogeweg, 2011). Globalement l'évolvabilité est considérée comme la faculté d'un système à évoluer. Mais au delà de cette vision générale et difficile à quantifier, de nombreuses définitions ont été proposées dans la littérature (Pigliucci, 2008), allant de la capacité à générer de la variation à la capacité à s'adapter rapidement à des changements environnementaux. Il a par ailleurs été montré que cette dernière propriété pouvait être sélectionnée dans des environnements fluctuants (Crombach et Hogeweg, 2007).

Ici, nous utiliserons une définition inspirée de Woods *et al.* (2011). Pour ces auteurs, l'évolvabilité (ou « potentiel évolutif ») correspond à l'espérance de gain de fitness d'un organisme à la génération suivante (ou aux générations suivantes). Cette définition ouvre pour nous la perspective de quantifier le potentiel évolutif en utilisant la procédure d'échantillonnage de Monte-Carlo que nous avons déjà utilisée pour la robustesse (section précédente) : l'évolvabilité *d'un individu* serait alors quantifiée comme le gain de fitness

moyen à la génération suivante, c'est-à-dire, à partir des 10 millions descendants générés, comme :

$$\begin{aligned} \text{Evolvabilité} &= \mathbb{E}\left(f - f_{\text{ancêtre}} \text{ si } \geq 0\right) \\ &= {}^{1}\!/_{K} \sum_{j=1}^{K} \begin{cases} f_j - f_{\text{ancêtre}} & \text{si } \geq 0 \\ 0 & \text{sinon} \end{cases} \end{aligned} \tag{IV.1}$$

où $K = 10^7$ est le nombre de descendants générés, $f_{\text{ancêtre}}$ est la fitness de l'organisme dont l'évolvabilité est estimée, $j \in [1 : K]$ est l'indice du descendant généré et f_j est la fitness de ce descendant La contrainte « ≥ 0 » exprime quant à elle le fait que ce descendant n'est pris en compte que s'il a une fitness supérieure à celle de son parent, donc qu'il a subi une ou plusieurs mutations favorables). Cependant, il paraît évident que le potentiel évolutif d'un organisme dépend de sa distance à l'optimum. Or, dans notre cas, nous avons vu que la fitness (donc la distance à l'optimum) des organismes simples est largement supérieure (respectivement inférieure) à celle des organismes complexes (figure IV.6, section 3.1). Dès lors, la comparaison de l'évolvabilité de ces deux types d'organisme risque de ne refléter que la distance à l'optimum.

Pour éviter cet écueil, nous avons comparé le potentiel évolutif des *simples* et des *complexes* dans un environnement modifié, en partant du principe que, dans ce nouvel environnement, la distance à l'optimum sera équivalente pour des *simples* et des *complexes*. Plus précisément, nous avons défini une nouvelle cible phénotypique (voir chapitre II, section 3.2.7 et chapitre III, section 2) constituée d'un triangle isocèle de même hauteur (h) et de même demi-largeur (w) que la cible initiale mais dont la moyenne a été légèrement décalée, de $m_{\text{initial}} = 0,5$ à $m_{\text{nouveau}} = 0,495$. Ce nouvel environnement présente donc le même degré de simplicité que l'environnement initial mais les fitness des pseudo-coalescents simples (qui ont évolué pendant 250 000 générations dans le premier environnement) y sont indiscernables des fitness des pseudo-coalescents complexes. Nous avons ensuite renouvelé l'échantillonnage et évalué l'évolvabilité en utilisant l'équation (IV.1).

La figure IV.9 présente les mesures d'évolvabilité des pseudo-coalescents simples (en bleu) et complexes (en rouge) pour les trois taux de mutation.

Les mesures d'évolvabilité sont difficiles à interpréter dans la mesure où la différence d'évolvabilité entre les *simples* et les *complexes* varie d'un taux de mutation à l'autre : l'écart est non-significatif pour $\mu = 10^{-4}$ mut.bp^{-1}.gen^{-1}, faiblement significatif en faveur des simples pour $\mu = 10^{-5}$ mut.bp^{-1}.gen^{-1} et faiblement significatif en faveur des complexes [1] pour $\mu = 10^{-6}$ mut.bp^{-1}.gen^{-1} ! Il semble donc raisonnable de conclure ici que ni les *simples* ni les *complexes* ne sont globalement avantagés sur le plan de l'évolvabilité.

Ces résultats méritent une attention particulière dans la mesure où il est souvent admis que les structures complexes sont plus évolvables que les *simples* (Simon, 1962). Pourtant,

1. On notera que la procédure d'échantillonnage de Monte-Carlo utilisée pour estimer l'évolvabilité est de moins en moins efficace à mesure que le taux de mutation décroît. En effet, la proportion de descendants ayant subi une ou plusieurs mutation parmi les 10 millions de descendants est directement proportionnelle au taux de mutation (pour un génome donné). En conséquence, pour un faible taux de mutation, le résultat de l'échantillonnage est susceptible d'être très variable, en particulier pour l'évolvabilité qui repose généralement sur une proportion très faible de descendants positifs (voir figure IV.10).

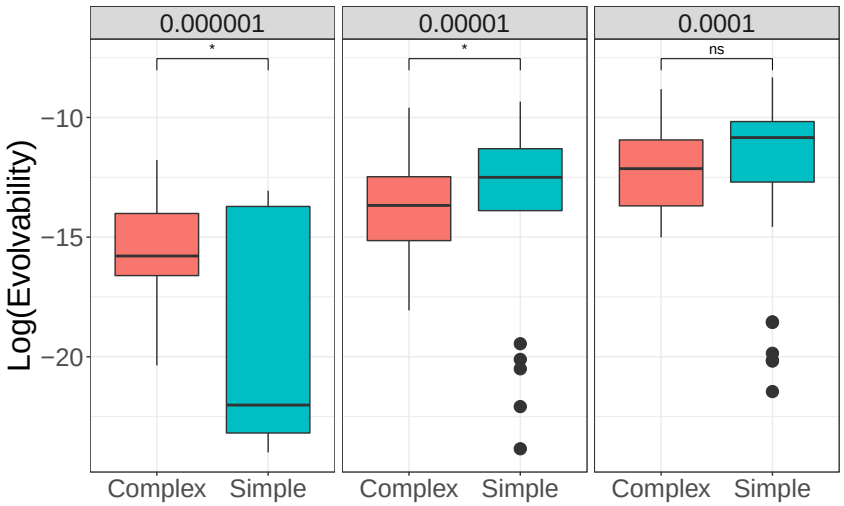


FIGURE IV.9 – Estimation de l'évolvabilité des pseudo-coalescents simples (en bleu) et complexes (en rouge) pour chaque taux de mutation (en échelle log). Les différences sont soit non-significatives, pour $\mu = 10^{-4}$ mut.bp^{-1}.gen^{-1}), soit très faiblement significatives, pour $\mu = 10^{-5}$m̃bg et $\mu = 10^{-6}$ mut.bp^{-1}.gen^{-1} (test de wilcoxon). On note par ailleurs que pour ces deux taux de mutation, les différences d'évolvabilité s'inversent.

cette supposition courante s'appuie souvent sur l'idée que les structures complexes sont modulaires, donc plus labiles et plus faciles à modifier (Clune *et al.*, 2013). Or, cette propriété n'a aucune raison d'évoluer ici. En effet, l'environnement est ici constant alors qu'il a été montré que la modularité (Kashtan et Alon, 2005) et même l'évolvabilité (Crombach et Hogeweg, 2007) sont sélectionnées en environnement variable. De façon à mieux comprendre ce résultat, nous avons distingué les deux composantes de l'évolvabilité (voir équation V.1) : la fraction des descendants positifs [1] et la gain de fitness moyen des descendants positifs. La figure IV.10 présente ces deux mesures pour les pseudo-coalescents simples et complexes. Elle montre que les pseudo-coalescents simples ont une fraction de descendants favorables significativement plus faible que les pseudo-coalescents complexes (figure IV.10, en haut) mais que le gain de fitness des simples est plus important que celui des complexes (sauf pour les taux de mutation les plus faibles). En d'autres termes, les *complexes* ont de plus nombreuses opportunités d'augmenter leur fitness – ce qui est cohérent avec leur plus grande cible mutationnelle et leur plus faible robustesse – mais ces opportunités ne permettent, individuellement, qu'un moindre gain de fitness.

3.2.3 Conclusion

En conclusion, nos mesures de robustesse et d'évolvabilité démontrent que, dans nos simulations, ce n'est pas la sélection indirecte qui préside à l'évolution de la complexité fonctionnelle au choix entre la simplicité et la complexité fonctionnelles. Au contraire, au

1. On défini ici un descendant positif comme un descendant dont la fitness est supérieur à celle de son parent. On notera que cette définition est légèrement différente de la notion de « mutant positif » (ou mutant favorable) au sens où un descendant positif peut avoir subit une ou plusieurs mutations.

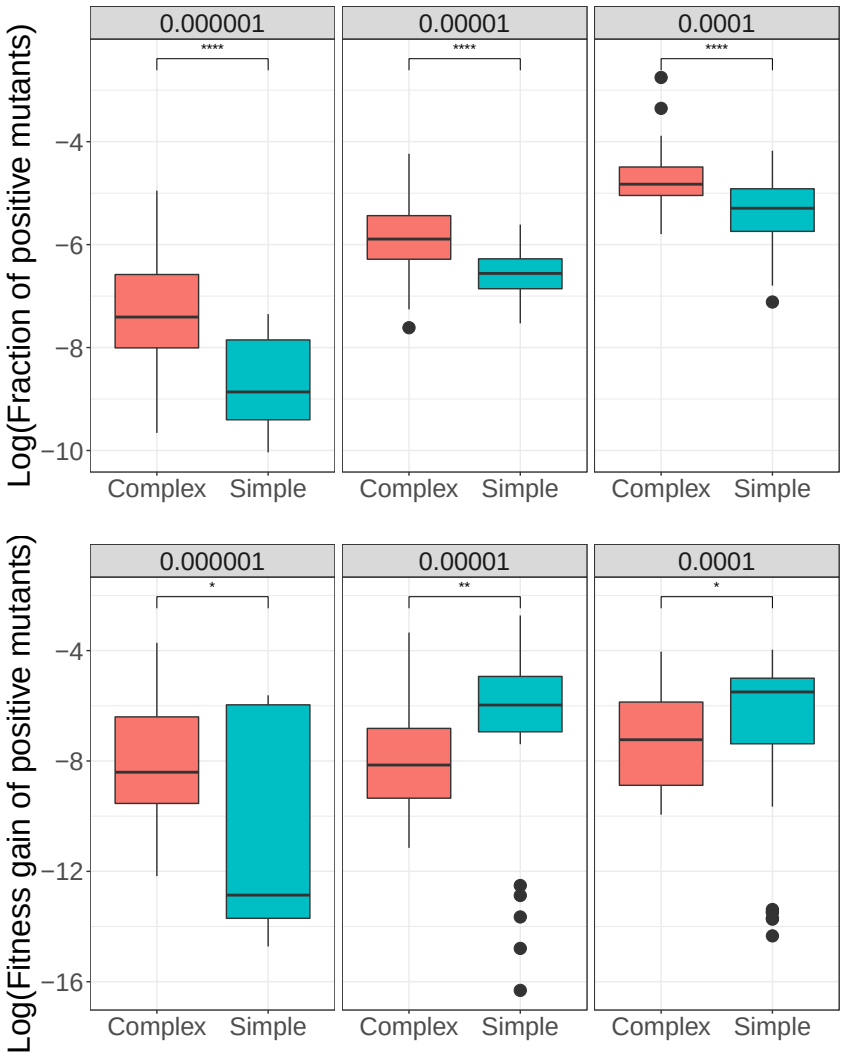


Figure IV.10 – Les deux composantes de l'évolvabilité : fraction de descendants positifs (en haut) et gain de fitness moyen des descendants positifs (en bas) pour les pseudo-coalescents simples (en bleu) et complexes (en rouge). Les organismes simples ont une plus faible fraction de descendants positifs mais – sauf pour les taux de mutation les plus faibles – un plus fort gain de fitness moyen.

vu des différences de robustesse entre les *simples* et les *complexes*, il serait plus juste de dire que la complexité apparaît ici *malgré* l'avantage sélectif indirect de la simplicité.

Au délà de cette conclusion générale, les résultats présentés ici peuvent être interprétés en termes de paysages adaptatif ou paysage de fitness (*fitness landscape*). Ce concept, initialement proposé par Wright (1932) est depuis devenu central dans l'interprétation et l'analyse des dynamiques évolutives. Le paysage adaptatif associe en effet un phenotype à une valeur sélective et un déplacement (d'un phénotype à un autre) à une variation de valeur sélective. La structure du paysage adaptatif permet donc d'interpréter plus ou moins empiriquement les dynamiques évolutives. Or, dans une interprétation en termes de paysage adaptatif, notre méthode d'estimation de la robustesse (et dans une moindre me-

sure, de l'évolvabilité) peut être interprétée comme un outil d'estimation de la structure locale du paysage adaptatif (via l'échantillonnage d'un grand nombre de voisins autour du pseudo-coalescent). En outre, pour une même expérience, tous les individus partagent le même *fitness landscape* puisque leur *genotype-phenotype map* et leur cible phénotypique sont les mêmes (il existe par ailleurs un chemin adaptatif permettant de passer de n'importe quel individu de nos simulations à n'importe quel autre, ce qui implique que leurs paysages adaptatifs sont connectés [1]).

En termes de paysages adaptatifs, nos résultats montrent que, quoique ce paysage soit le même dans toutes les expériences, les *simples* et les *complexes* se situent sur des régions très différentes de ce paysage. Les simples se trouvent sur des régions hautes et « pentues » mais qui n'ont qu'une faible connectivité alors que les *complexes* se trouvent sur des régions basses, globalement planes et fortement connectées [2]. Cette différence entre les *simples* et les *complexes* montre qu'au-delà de leurs différences en termes de complexité génomique et fonctionnelle, ces deux types d'organismes ont suivi des dynamiques évolutives différentes qui les ont conduits vers des régions très différentes (et très éloignées) du paysage adaptatif global.

1. La question pourrait cependant être débattue si on considère des taux de mutation différents. On entrerait cependant là dans une discussion qui dépasserait le cadre de cette thèse.

2. Ce résultat pourrait sembler en contradiction avec la plus grande robustesse des simples puisque la robustesse mutationnelle est plus grande dans les régions plates du paysage (Wilke *et al.*, 2001), mais il faut garder à l'esprit que nous avons ici spécifiquement mesuré la robustesse *réplicative*, propriété qui dépend non seulement du paysage de fitness mais aussi de la taille et de la structure du génome (Knibbe *et al.*, 2007a).

4 Hypothèses neutralistes

Dans la section précédente nous avons montré que l'évolution d'organismes majoritairement complexes dans nos simulations n'était pas due à un mécanisme sélectif, qu'il s'agisse de sélection directe (sélection pour la fitness, section 3.1) ou indirecte (sélection pour la robustesse ou pour l'évolvabilité, section 3.2). Dès lors, il semble logique de considérer que l'augmentation de la complexité provient de forces non sélectives et donc de privilégier l'hypothèse neutraliste, même s'il peut paraître surprenant qu'un processus neutre puisse aller à l'encontre d'un processus sélectif, du moins si celui-ci est suffisamment puissant ce qui est à priori le cas au vu des différences de fitness constatées ici (section 3.1). Les résultats observés pourraient en particulier découler d'un processus de dérive tel que le « Drunkard Walk » de Gould (1996) ou la « Zero-Force Evolutionary Law » (ZFEL) de McShea et Brandon (2010).

4.1 Dispositif

Afin de tester l'hypothèse neutraliste, nous avons évalué les effets de la sélection en laissant évoluer (pour autant que ce mot convienne encore dans un tel cadre) des organismes en l'absence de sélection, de façon à ne plus observer que l'effet de la dérive. Partant de deux populations clonales (l'une composée d'individus complexes et l'autre d'individus simples), nous les avons laissé évoluer sans pression sélective et nous avons mesuré la diversité et la complexité aux niveaux génomique et fonctionnel. En effet, les hypothèses neutralistes reposent sur l'idée que la complexité est seulement due à l'augmentation de diversité dans la population sous l'effet du processus mutationnel (McShea et Brandon, 2010; Bapteste, 2017).

Afin de choisir deux individus représentatifs (un *simple*, un *complexe*) pour ensemencer les deux populations initiales, nous avons extrait chacun des pseudo-coalescents (ancêtres communs à la générations 250 000) des lignées ayant évolué sous un taux de mutation de 10^{-6} mut.bp^{-1}.gen^{-1}. Ces 100 individus ont ensuite été classés en « simples » et « complexes » puis, pour chacun de ces groupes, nous avons calculé la médiane de complexité fonctionnelle ($\mathcal{C}_P$). Enfin, nous avons isolé, dans chacune de ces deux population, l'individu dont la complexité fonctionnelle est la plus proche de la médiane ce qui nous a permis d'isoler un représentant des *simples* et un représentant des *complexes*. Ces deux individus ont été utilisés pour initialiser 20 populations clonales de 1 024 individus (10 à partir de l'individu simple et 10 à partir de l'individu complexe) et ces 20 populations ont évolué dans Aevol pendant 10 000 générations sous un taux de mutation de 10^{-6} mut.bp^{-1}.gen^{-1} mais sans sélection (tous les individus de la population ayant donc la même probabilité de réplication, quelle que soit leur distance au phénotype cible). Pour chaque population et à chaque génération, nous avons mesuré (*i.*) la proportion d'individus complexes dans la population (cette proportion étant initialement nulle pour les 10 simulations initialisées à partir d'un individu simple), (*ii.*) la moyenne de complexité génomique ($\mathcal{C}_G$) et de complexité fonctionnelle ($\mathcal{C}_P$) et (*iii.*) la variance de la complexité génomique et de la complexité fonctionnelle au sein de chacune des vingt populations (la variance étant ici utilisée comme mesure de diversité).

La figure IV.11 présente l'évolution de la proportion d'individus complexes pendant les 10 000 générations pour les 20 populations suivies au cours de cette expérience (en haut,

les 10 populations initialisées avec un individu complexe ; en bas, les les 10 populations initialisées avec un individu simple). Elle montre que, lorsqu'une population d'individus complexes se reproduit en présence du phénomène de mutation mais en l'absence de sélection, la proportion d'individus complexes décroît rapidement pour atteindre zéro en quelques milliers de générations. En revanche, et de façon intéressante, lorsque la population est initialement composée d'individus simples (courbes du bas), la proportion d'individus complexes augmente rapidement pour atteindre environ 20% après quelques centaines de générations. Ce résultat pourrait sembler conforme avec les prédictions de la ZFEL ; cependant, après cette croissance initiale, la complexité diminue rapidement pour attendre zéro après quelques milliers de générations. Globalement, sur les vingt populations simulées, aucun individu complexe ne subsiste au bout des 10 000 générations de l'expérience.

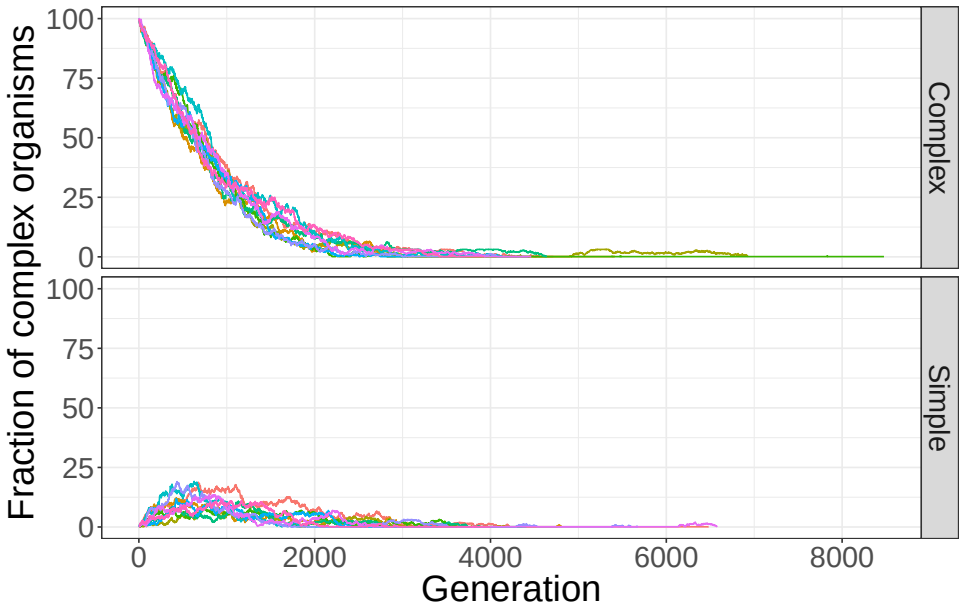


Figure IV.11 – Évolution de la proportion d'individus complexes en l'absence de sélection pour des populations initialisées à partir d'un individu complexe (en haut) et simple (en bas). Les couleurs indiquent les différentes répétitions.

4.2 Analyse de la diversité et de la complexité

Pour comprendre cette dynamique, nous avons observé les variations de diversité et de complexité génomique et fonctionnelle au sein des vingt populations. La figure IV.12 présente l'évolution des variances (à gauche) et des moyennes (à droite) de complexité génomique (en haut) et fonctionnelle (en bas) pour les 10 populations initialisées à partir d'un individu complexe. La figure IV.13 présente, quant à elle, les mêmes données mais pour les 10 populations initialisées à partir d'un individu simple.

De façon similaire à la figure IV.11, l'évolution des variances de complexité au cours du temps semble soutenir l'hypothèse neutraliste et tout particulièrement la ZFEL. Pour McShea et Brandon (2010), l'accroissement de la complexité serait due à l'augmentation de

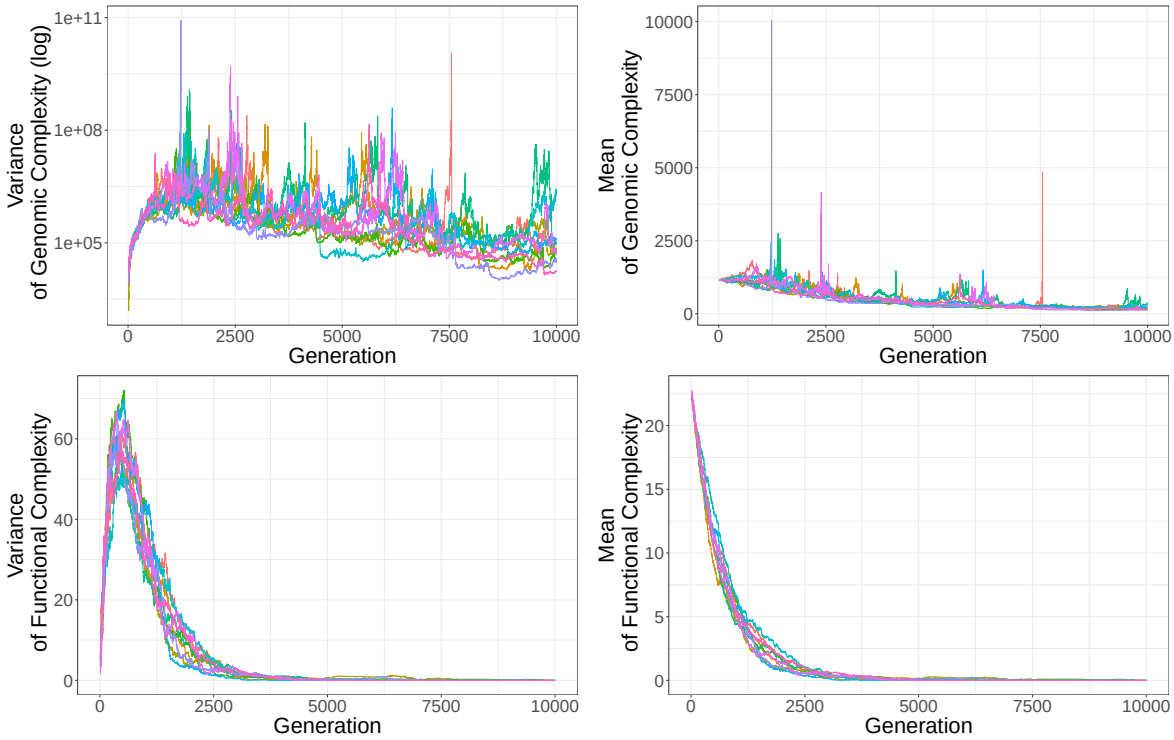


Figure IV.12 – Évolution de la variance (à gauche) et de la moyenne (à droite) de complexité génomique ($\mathcal{C}_G$, en haut) et de complexité fonctionnelle ($\mathcal{C}_P$, en bas) au sein des dix populations initialisées à partir d'un individu complexe. Les couleurs indiquent les différentes répétitions. Pour les deux mesures, la variances augmentent rapidement au début de l'expérience mais pour décroître ensuite de façon régulière. Les complexité moyennes de la population, quant à elles, décroissent tout le long de l'expérience.

la variabilité sous l'influence des différentes mutations. Or, c'est bien ce que montrent les figures IV.12 et IV.13 : quelles que soient les populations considérées (initialement simples ou complexes) et quelle que soit la mesure de complexité ($\mathcal{C}_G$ ou $\mathcal{C}_P$), la variance augmente rapidement au cours des premières centaines de générations de l'expérience, confirmant l'augmentation spontanée de la diversité. Cependant, contrairement aux hypothèses de la ZFEL, cette augmentation de la variance ne s'accompagne pas d'une augmentation de la complexité moyenne, même pour les simulations initialisées à partir d'une population clonale d'individus simples pour lesquelles la complexité initiale est relativement faible. Au contraire, pour toutes les simulations et pour les deux mesures de complexité, la complexité décroît régulièrement (même si on observe de fortes fluctuations transitoires pour la complexité génomique – en haut à droite sur les deux figures), la complexité fonctionnelle moyenne étant même nulle pour la grande majorité des simulations à la fin de l'expérience [1]. En outre, après quelques centaines de générations, la variance elle-même décroît, relativement lentement pour la complexité génomique ($\mathcal{C}_G$, en haut à gauche sur les figures), mais très rapidement pour la complexité fonctionnelle ($\mathcal{C}_P$, en bas à gauche

1. Les deux mesures de complexité étant positives (ou nulles), on notera qu'une complexité moyenne nulle implique que tous les individus de la population ont une complexité égale à zéro.

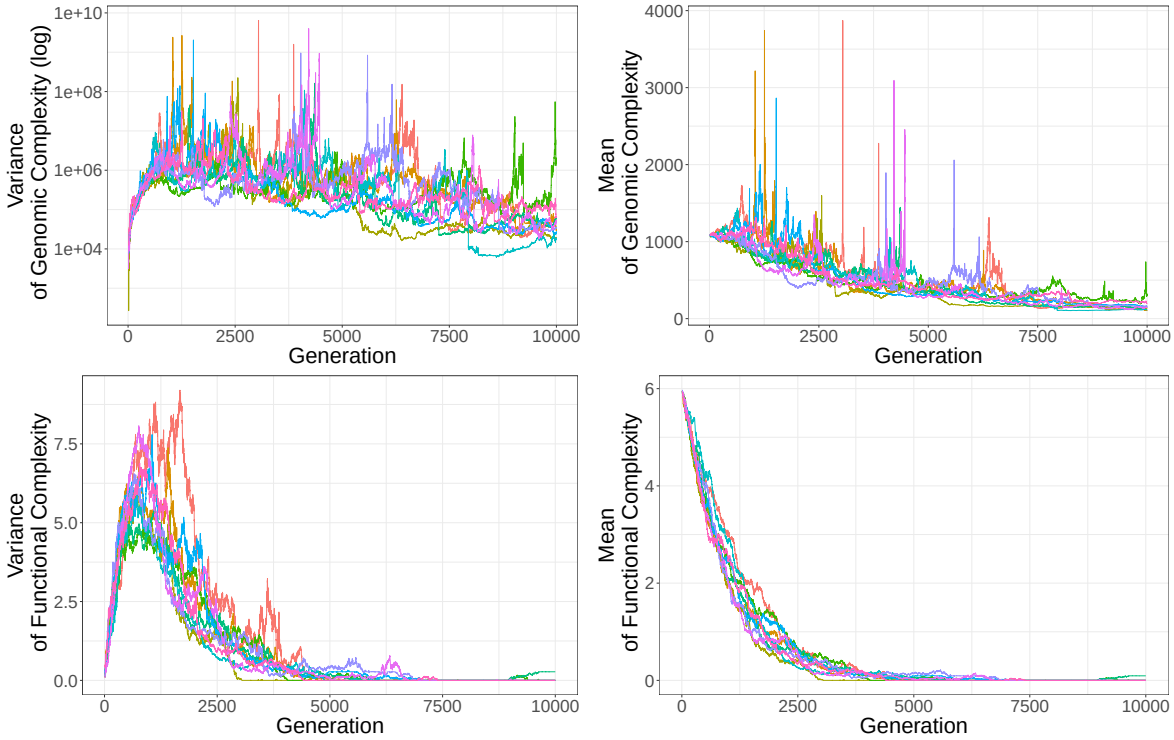


FIGURE IV.13 – Évolution de la variance (à gauche) et de la moyenne (à droite) de complexité génomique ($\mathcal{C}_G$, en haut) et de complexité fonctionnelle ($\mathcal{C}_P$, en bas) au sein des dix populations initialisées à partir d'un individu simple. Les couleurs indiquent les différentes répétitions. Pour les deux mesures, la variances augmentent rapidement au début de l'expérience mais pour décroître ensuite de façon régulière. Les complexité moyennes de la population, quant à elles, décroissent tout le long de l'expérience.

sur les figures) qui retourne rapidement à zéro.

4.3 Conclusion

Comme il a été montré, en l'absence de sélection, la variance et la moyenne des complexités diminuent dans toutes les populations testées. Ces deux constats invalident les hypothèses neutralistes en montrant que la sélection est nécessaire à l'augmentation de la complexité (nos résultats montrent même qu'elle est aussi nécessaire ne fut-ce qu'à son maintien). Devant la simplicité des explications neutralistes (souvent considérées comme « allant de soi » sur la base d'expériences de pensée ou de simples métaphores [1]), il sera utile de remonter à la causalité de nos observations.

S'agissant de populations initialement clonales, l'augmentation initiale de la variance est simple à expliquer : comme les différents individus subissent des mutations différentes (et que le processus mutationnel est plus rapide que la dérive), la diversité augmente dans la population car les individus empruntent des chemins mutationnels différents. En

1. McShea et Brandon (2010) évoquent par exemple la différentiation « naturelle » des segments d'un organisme vertébré pour expliquer l'origine de complexité de ce clade ; Éric Bapteste (2017) convoque quant à lui l'image de fils électriques qui finissent toujours par s'entrelacer et former un tout complexe.

revanche, ce que montre la diminution constante de la complexité moyenne, c'est que, contrairement aux prédictions des hypothèses neutralistes, les différents chemins empruntés par les individus sont tous dirigés vers la perte de de complexité. Les fluctuations observées sur les moyennes (en particulier les moyennes de complexité génomique) ne sont en effet que transitoires. L'explication de la diminution graduelle de la complexité est relativement simple : il s'agit de remarquer que le maintien ou l'augmentation de complexité ne résultent pas de la seule augmentation de la diversité. En effet, la structure du codage moléculaire de l'information (dans Aevol autant que dans la nature), rend nécessaire de limiter la diversité. De fait, pour qu'un élément moléculaire (*p. ex.* un gène) contribue à la complexité, qu'elle soit génomique ou fonctionnelle, il doit être identifié sur le génome ou sur l'ARNm au moyen de plusieurs séquences signal (promoteur, RBS, codons Start et Stop et terminateur – voir chapitre II). Or, en l'absence de sélection, l'accroissement de diversité affecte indifféremment les gènes (et donc les protéines) mais aussi les signaux initiateurs et terminateurs de la transcription ou de la traduction. Il en résulte que la complexité décroît, non pas en vertu de l'homogénéisation des éléments mais du simple fait de la disparition progressive des éléments sous l'effet de la dégradation des séquences signal qui, l'une après l'autre, perdent leur fonction.

Nos résultats soulignent ainsi toute la difficulté des expériences de pensée en évolution : parce qu'elles ne mettent généralement en jeu qu'un seul niveau d'organisation (le niveau phénotypique, pour ce qui concerne les exemples déjà évoqués, (McShea et Brandon, 2010) et (Bapteste, 2017)), considéré à l'exclusion des autres, elles échouent à comprendre comment l'effet d'un processus à un niveau donné peut être compensé voire contrebalancé par les effets de ce même processus à un autre niveau. En l'occurrence, l'effet de la diversité au niveau phénotypique est ici anéanti par l'effet de cette même diversité au niveau moléculaire. En d'autre termes, dans les exemples extraits de (McShea et Brandon, 2010) et (Bapteste, 2017) et cités plus haut, l'augmentation de la diversité des segments vertébraux, ou des fils électriques (selon l'auteur), ne provoque pas d'augmentation de la complexité car elle s'accompagne inexorablement de la disparition progressive de ces segments eux-mêmes, ou des fils électriques ; la sélection est donc une composante *nécessaire* à l'évolution de la complexité.

5 Vers une troisième hypothèse : le cliquet de la complexité

À première vue, les résultats présentés dans les sections 3 et 4 semblent contradictoires : d'une part nous venons de voir que la sélection est nécessaire au maintien et à l'accroissement de la complexité (section 4), d'autre part l'accumulation de complexité se fait *malgré* l'avantage sélectif des organismes simples (section 3). Cependant, ces résultats ont été obtenus en analysant uniquement les pseudo-coalescents. En d'autres termes, nous n'avons analysé jusqu'à maintenant qu'un *instant* dans l'évolution : la génération 250 000. Afin de comprendre comment nos simulations ont pu conduire à l'accumulation d'organismes complexes, c'est la *dynamique* évolutive qui a conduit à ce résultat qu'il nous faut mieux comprendre. Pour cela, nous allons analyser les trajectoires évolutives des simples et des complexes au cours de l'expérience [1].

5.1 La simplicité et la complexité sont des identités stables

Comme nous l'avons vu précédemment (section 2), à la génération 250 000, sur 300 pseudo-coalescents, 229 (soit 76%) sont complexes. Si nous observons cette même proportion à la génération 10 000 nous observons que, sur la lignée, 236 individus sur 300 sont complexes (79%). En outre, la table IV.2 ci-dessous montre que les *complexes* à la générations 250 000 sont, pour une écrasante majorité, les descendants des complexes à la génération 10 000.

	$\mu = 10^{-4}$	$\mu = 10^{-5}$	$\mu = 10^{-6}$
$P_{S\to S}$	100% [100% – 87,9%] (28/28)	100% [100% – 85,7%] (23/23)	92,3% [98,6% – 66,7%] (12/13)
$P_{C\to C}$	94,4% [97,8% – 86,6%] (68/72)	97,4% [99,3% – 91%] (75/77)	97,7% [99,4% – 92%] (85/87)

Table IV.2 – Fraction de lignées ayant conservé leur identité *simple* ou *complexe* entre les générations 10 000 et 250 000 (respectivement $P_{S\to S}$ et $P_{C\to C}$) pour les trois taux de mutation testés, exprimés en mut.bp^{-1}.gen^{-1}. Les valeurs entre crochets indiquent les intervalles de confiance à 95% ($CI_{95\%}$) calculés par la méthode de Wilson ; les valeurs entre parenthèses indiquent le nombre de lignées ayant conservé leur identité simple (respectivement complexe) à la génération 250 000 (la somme n'est donc pas égale à 300) et le nombre de lignées simples (respectivement complexes) à la génération 10 000.

Ces valeurs sont à rapprocher du fait que les simulations sont initialisées, à la génération zéro, par des populations clonales d'organismes *simples* (et même, plus précisément, d'organismes *rudimentaires*). Ces résultats indiquent donc que la transition de *simple* vers

1. Nous appelons ici « trajectoire évolutive » la suite de variations observée entre les générations 0 et 250 000 (génération du pseudo-coalescent) sur la lignée du meilleur individu à la génération 270 000 (voir chapitre III, section 2.3). Tous les événements présentés dans les sections suivantes correspondent donc à des événements *fixés* par l'évolution.

complexe se produit très tôt au cours de l'évolution (la table IV.2 montre que la transition se produit avant la génération 10 000 pour tous les taux de mutation mais, en pratique, cette transition a lieu beaucoup plus tôt pour les forts taux de mutation). Ils montrent aussi et surtout que cette transition est quasiment irréversible. En d'autres termes, l'appartenance à l'une des deux classes (*simple* ou *complexe*) fait partie de l'identité des individus et elle se conserve tout au long de leur évolution.

5.2 Dynamique des lignées *simples* et *complexes*

La stabilité de l'identité *simple* ou *complexe* des lignées au cours de l'évolution montre que notre classification qualitative capture effectivement un élément important de la dynamique évolutive. En outre, elle nous autorise à comparer les dynamiques de ces deux classes d'individus sur le plan de la complexité (génomique et fonctionnelle) et sur le plan de la fitness.

Les figures IV.14 et IV.15 présentent l'évolution des mesures de complexité au cours du temps pour les 300 lignées simulées. Les mesures sont moyennées par classe pour six classes en fonction des taux de mutation (représentés par type de courbe) et des identités *simple* ou *complexe* des lignées (représentées par couleurs : rouge pour les *complexes*, bleu pour les *simples*).

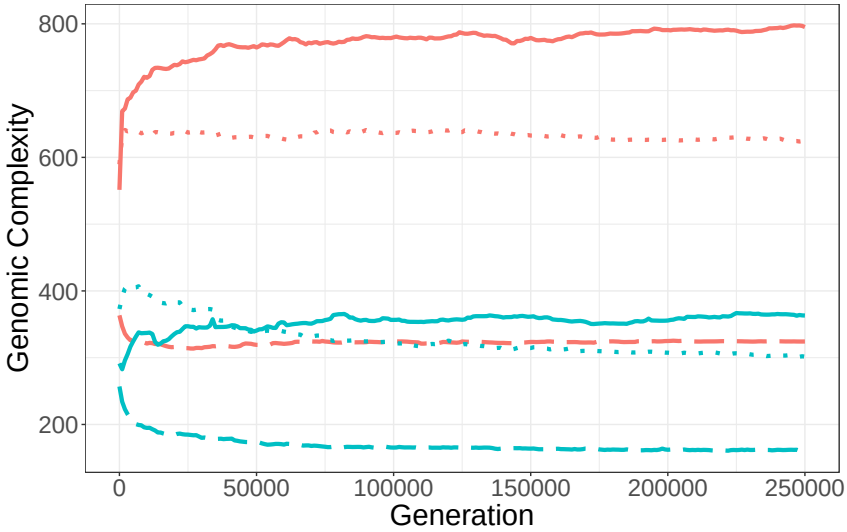


Figure IV.14 – Évolution de la complexité génomique ($\mathcal{C}_G$) (moyenne), entre les générations 0 et 250 000 pour les organismes complexes (en rouge) et simples (en bleu). Lignes pleines : faible taux de mutation (10^{-6} mut.bp^{-1}.gen^{-1}) ; points : taux de mutation intermédiaire (10^{-5} mut.bp^{-1}.gen^{-1}) ; tirets : fort taux de mutation (10^{-4} mut.bp^{-1}.gen^{-1}).

Ces deux figures montrent clairement que la dynamique d'évolution de la complexité est totalement différente pour les deux classes d'individus, en particulier pour la complexité fonctionnelle $\mathcal{C}_P$: alors que, pour les *simples*, $\mathcal{C}_P$ décroît très rapidement pour se stabiliser aux alentours de $\mathcal{C}_P = 4$, pour les *complexes*, $\mathcal{C}_P$ augmente continûment tout au long de l'expérience (entre les générations 10 000 et 250 000, $\Delta\mathcal{C}_P = -0,32 \pm 0,29$ pour les individus simples, tous taux de mutation confondus, tandis que $\Delta\mathcal{C}_P = +3,58 \pm 0,28$

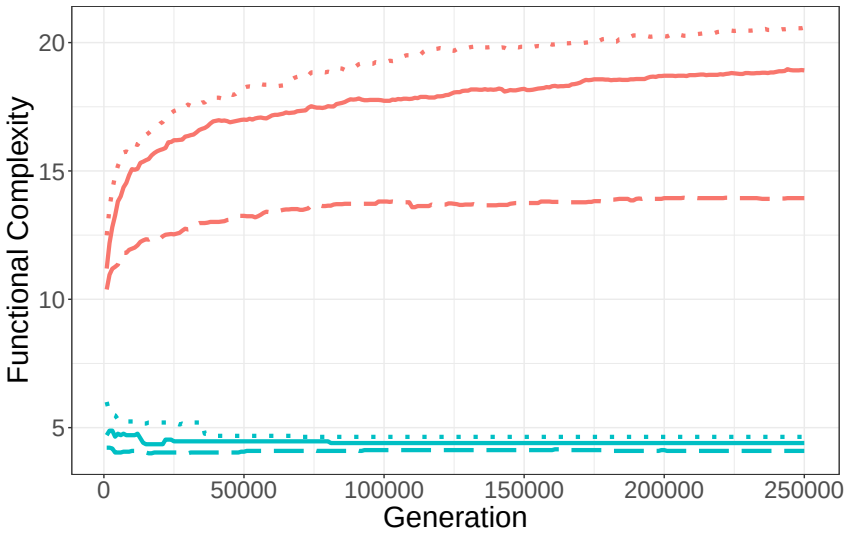


FIGURE IV.15 – Évolution de la complexité fonctionnelle ($\mathcal{C}_P$) (moyenne), entre les générations 0 et 250 000 pour les organismes complexes (en rouge) et simples (en bleu). Lignes pleines : faible taux de mutation (10^{-6} mut.bp^{-1}.gen^{-1}) ; points : taux de mutation intermédiaire (10^{-5} mut.bp^{-1}.gen^{-1}) ; tirets : fort taux de mutation (10^{-4} mut.bp^{-1}.gen^{-1}).

pour les individus complexes. Dans le cas de la complexité génomique (figure IV.14), la différence entre les deux classes reste marquée (entre les générations 10 000 et 250 000, $\Delta\mathcal{C}_G = -43,8 \pm 1,6$ pour les individus simples, tous taux de mutation confondus, tandis que $\Delta\mathcal{C}_G = +25,3 \pm 2,09$ pour les *complexes*) mais elle est dominée par la différence de taux de mutation : les *complexes* ont globalement – ce qui est cohérent – une mesure de $\mathcal{C}_P$ plus importante que les *simples*, mais, contrairement à ce qui est observé pour $\mathcal{C}_G$, les populations évoluant sous un taux de mutation faible (10^{-6} mut.bp^{-1}.gen^{-1}) voient leur complexité génomique augmenter et ce pour les individus complexes (entre les générations 10 000 et 250 000, $\Delta\mathcal{C}_G = +3,1 \pm 1,86$ pour le fort taux de mutation (10^{-4} mut.bp^{-1}.gen^{-1}), $\mathcal{C}_G = -12,7 \pm 3,13$ pour le taux de mutation intermédiaire (10^{-5} mut.bp^{-1}.gen^{-1}) et $\mathcal{C}_G = +76,1 \pm 2,80$ pour le taux de mutation faible (10^{-6} mut.bp^{-1}.gen^{-1})) comme pour les individus simples (entre les générations 10 000 et 250 000, $\Delta\mathcal{C}_G = -34,8 \pm 1,94$ pour le fort taux de mutation (10^{-4} mut.bp^{-1}.gen^{-1}), $\Delta\mathcal{C}_G = -94,0 \pm 3,91$ pour le taux de mutation intermédiaire (10^{-5} mut.bp^{-1}.gen^{-1}) et $\Delta\mathcal{C}_G = +25,4 \pm 4,27$ pour le taux de mutation faible (10^{-6} mut.bp^{-1}.gen^{-1})).

La figure IV.16 montre l'évolution de la valeur de fitness pour les lignées simples et complexes et pour les trois taux de mutation. Elle montre que, pour tous les taux de mutation, la fitness des lignées simples augmente très rapidement dès le début de l'expérience, pour se stabiliser à une valeur proche de 1, fitness maximale possible dans Aevol (gain de fitness moyen des organismes simples entre les générations 10 000 et 250 000 : $\Delta\text{Fitness}_{\text{simples}} = +0,06 \pm 0,02$). En comparaison, les organismes complexes n'améliorent leur fitness que très lentement, et surtout continûment, au cours de l'expérience (gain moyen de fitness des complexes entre les générations 10 000 et 250 000 : $\Delta\text{Fitness}_{\text{Complexes}} = +0,16 \pm 0,14$).

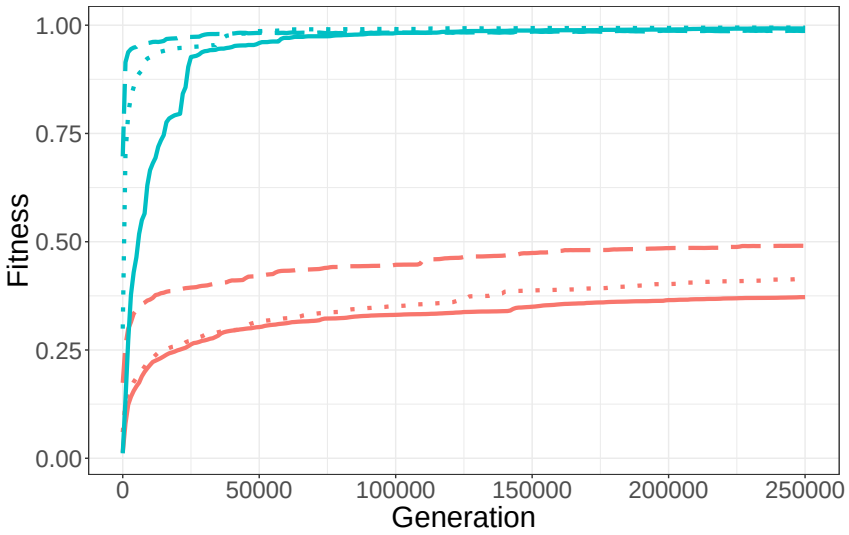


Figure IV.16 – Évolution de la Fitness (moyenne) entre les générations 0 et 250 000 pour les lignées complexes (en rouge) et simples (en bleu). Lignes pleines : faible taux de mutation (10^{-6} mut.bp^{-1}.gen^{-1}) ; points : taux de mutation intermédiaire (10^{-5} mut.bp^{-1}.gen^{-1}) ; tirets : fort taux de mutation (10^{-4} mut.bp^{-1}.gen^{-1}).

5.3 Discussion

L'évolution de la complexité génomique ($\mathcal{C}_G$), de la complexité fonctionnelle ($\mathcal{C}_P$) et de la fitness des lignées simples s'explique relativement aisément par la combinaison de deux facteurs. Durant les premières générations des simulations, l'évolution est principalement dirigée par la sélection directe et la fitness des organismes simples augmente très rapidement pour atteindre un niveau quasi-optimal (figure IV.16). Après cette première phase, la sélection devient de moins en moins efficace puisque les organismes sont proches de l'optimum. L'évolution des lignées simples est alors dirigée par la sélection indirecte et plus précisément par la sélection pour la robustesse réplicative. La conséquence première de cette sélection est une compaction du génome en général et des séquences codantes en particulier (donc une diminution de $\mathcal{C}_G$ – figure IV.14). En outre, cette dynamique est d'autant plus forte que le taux de mutation est plus élevé (Knibbe *et al.*, 2007a; Carde *et al.*, 2019). Cependant, toujours dans les lignées simples, ce mécanisme n'a pas d'effet sur $\mathcal{C}_P$ (figure IV.15) car, d'une part, $\mathcal{C}_P$ est déjà très faible et, d'autre part, les lignées simples ne subissent aucune pression évolutive pour accroître leur complexité puisqu'elles sont déjà à l'optimum.

L'évolution de $\mathcal{C}_G$ et $\mathcal{C}_P$ dans les lignées complexes s'explique aussi par une combinaison des effets de la sélection directe et indirecte mais les mécanismes par lesquels ces deux formes de sélection opèrent sont différents. En effet, comme les organismes complexes sont loin de l'optimum, la sélection est active pendant toute la durée de l'expérience, ce qui provoque un accroissement continu (quoique de plus en plus lent) de $\mathcal{C}_P$. Dans ces lignées cependant, comme dans les lignées simples, la sélection indirecte pour la robustesse exerce une contrainte sur le génome, contrainte d'autant plus forte que le taux de mutation est élevé. Cette contrainte limite la quantité d'information que les organismes peuvent

accumuler sur leur génome, soit en limitant la taille des séquences codantes (Eigen et Schuster, 1977), soit en limitant la taille totale du génome (Knibbe *et al.*, 2007a; Parsons *et al.*, 2010; Fischer *et al.*, 2014). De fait, cette limite imposée aux séquences codantes est clairement visible sur la dynamique de $\mathcal{C}_P$ (figure IV.14) pour les deux plus forts taux de mutation. Elle est aussi visible sur la taille totale des génomes (figure IV.8, page 100). Cependant, comme, dans Aevol, $\mathcal{C}_G$ et $\mathcal{C}_P$ ne sont que faiblement inter-dépendants (voir chapitre II et chapitre III, section 3.4), les contraintes affectant $\mathcal{C}_G$ n'interdisent pas l'accumulation de complexité fonctionnelle et nous observons effectivement que, pour les organismes complexes, $\mathcal{C}_P$ augmente au cours du temps (figure IV.15) malgré la borne imposée sur $\mathcal{C}_G$ par la sélection indirecte. Nous observons d'ailleurs, comme illustré dans les figures IV.17 et IV.18 que les lignées évoluant sous un faible taux de mutation (10^{-6} mut.bp^{-1}.gen^{-1}) ont la plus forte valeur moyenne de $\mathcal{C}_G$ mais que, pour $\mathcal{C}_P$, ce sont les lignées évoluant sous un taux de mutation intermédiaire qui ont la plus forte complexité, illustrant bien ici le découplage partiel entre ces deux niveaux d'organisation.

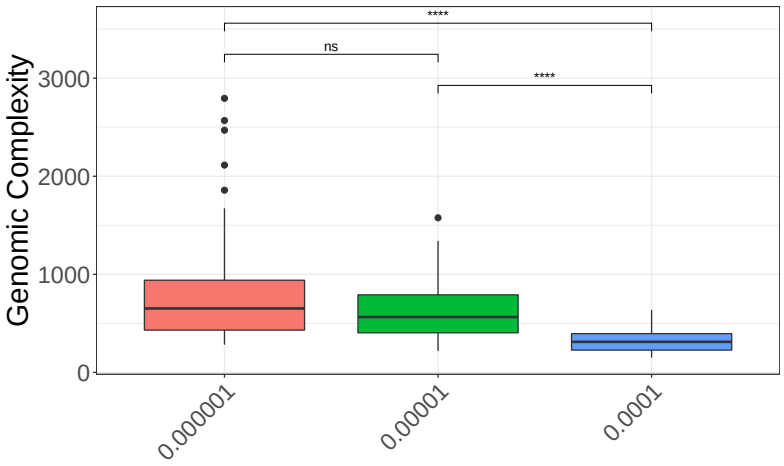


Figure IV.17 – Estimation de la complexité génomique $\mathcal{C}_G$ des pseudo-coalescents complexes pour chaque taux de mutation. Les écarts sont non significatifs entre les taux de mutation 10^{-5} mut.bp^{-1}.gen^{-1} et 10^{-6} mut.bp^{-1}.gen^{-1} et significatif pour les autres (test de wilcoxon). Le code couleur indique le taux de mutation Les couleurs se rapportent aux taux de mutation : bleu : 10^{-4} mut.bp^{-1}.gen^{-1} ; rouge : 10^{-5} mut.bp^{-1}.gen^{-1} ; vert : 10^{-6} mut.bp^{-1}.gen^{-1}..

Si la dynamique de $\mathcal{C}_G$ est ainsi explicable, il reste difficile d'expliquer pourquoi nous observons une augmentation de $\mathcal{C}_G$ au cours de l'évolution. On pourrait être tenté d'invoquer un biais mutationnel direct ou indirect (Soyer et Bonhoeffer, 2006), cependant, non seulement les taux de mutation utilisés lors des expériences sont-ils parfaitement équilibrés, mais, de plus a-t-il été montré que l'effet couplé des duplications et des délétions provoque un biais indirect vers la réduction du génome (Fischer *et al.*, 2014). Dès lors, si sélection directe, sélection indirecte et biais mutationnels poussent tous trois l'évolution vers la simplicité, quelle peut être est la force qui les contrebalance et préside à l'évolution de la complexité ?

Pour répondre à cette question, il suffit de mettre en regard les dynamiques de fitness

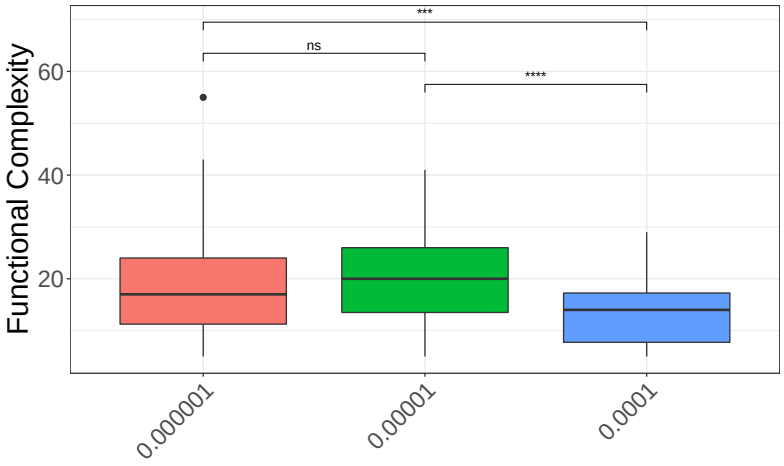


Figure IV.18 – Estimation de la complexité fonctionnelle $\mathcal{C}_P$ des pseudo-coalescents complexes pour chaque taux de mutation. Les écarts sont non significatifs entre les taux de mutation 10^{-5} mut.bp^{-1}.gen^{-1} et 10^{-6} mut.bp^{-1}.gen^{-1} et significatif pour les autres (test de wilcoxon). Le code couleur indique le taux de mutation (Bleu : 10^{-4} mut.bp^{-1}.gen^{-1} ; Rouge : 10^{-5} mut.bp^{-1}.gen^{-1} ; Vert : 10^{-6} mut.bp^{-1}.gen^{-1}).

des simples et des complexes (figure IV.16). En effet, d'une part les *simples* évoluent très rapidement (en quelques centaines de générations) vers une fitness quasi-optimale, d'autre part les *complexes*, même si ils sont, de loin, moins adaptés que les *simples* (comme le montrent les mesures de fitness des pseudo-coalescents – section 3.1) voient leur fitness moyenne augmenter régulièrement tout au long de l'expérience : même si la complexité apparaît *malgré* la sélection (du fait de l'avantage sélectif des simples), c'est bien la sélection qui semble être le moteur de l'augmentation de complexité au cours du temps.

Ce double effet de la sélection implique que, pour les organismes rudimentaires de la génération 0, deux chemins mutationnels peuvent être empruntés. L'un les conduit rapidement à une excellente adaptation et à un organisme simple, l'autre, beaucoup plus lent, les mène à une organisation complexe. En outre, les résultats présentés section 5.1 montrent que, lorsque l'évolution est engagée sur le chemin de la complexité, elle ne peut plus revenir en arrière, ceci malgré l'avantage sélectif que cela conférerait.

Collectivement, ces observations ne peuvent s'expliquer que par un mécanisme d'épistasie de signe [1]. L'épistasie est très largement documentée, aussi bien sur le plan théorique que sur le plan expérimental (Phillips, 2008). En outre, il a été montré que, dans les populations naturelles, la forme des interactions épistatiques est corrélée avec la complexité génomique (Sanjuán et Elena, 2006). Ici, des mutations bénéfiques dans le cas d'organismes rudimentaires et conduisant à des lignées simples, s'avèrent délétères dans le contexte génétique d'organismes complexes. Ce mécanisme interdit l'acquisition de gènes qui, sans cela, seraient très favorables. Comme, en outre, la suppression d'un gène serait elle aussi

1. L'épistasie correspond à une interaction entre mutations. En présence d'épistasie, l'effet d'une mutation dépend de l'occurrence préalable d'une autre mutations (Phillips, 2008). Dans le cas d'une épistasie de signe, une mutation favorable peut devenir défavorable si elle se produit après un autre événement moléculaire (Weinreich *et al.*, 2005).

défavorable, le seul chemin évolutif qui s'ouvre aux complexes est une fuite en avant vers l'augmentation de la complexité – permettant certes un gain de fitness mais beaucoup plus limité que celui qu'aurait autorisé l'évolution vers la simplicité. Cette dynamique correspond à un mécanisme de cliquet (*ratchet*) : en raison de l'épistasie de signe, un organisme engagé sur la voie de la complexité se voit astreint à accumuler toujours davantage de gènes. Ce faisant il creuse le fossé qui le sépare de la simplicité, le rendant rapidement si profond qu'il devient totalement improbable qu'une ou plusieurs mutations permettent de revenir en arrière. L'existence d'un tel « cliquet de la complexité » (*complexity ratchet*) a été évoquée par plusieurs auteurs (Cairns-Smith, 1995; Bapteste, 2017) ; cependant, à notre connaissance, notre étude est la première à mettre en évidence le rôle central de l'épistasie de signe dans le fonctionnement du cliquet.

6 Conclusion

En faisant évoluer, dans un environnement simple, des organismes « numériques » dont la complexité génomique et fonctionnelle peut varier au cours de l'évolution, nous avons obtenu des résultats très éclairants quant à l'évolution de la complexité des systèmes biologiques.

Tout d'abord, l'accroissement continu de la complexité, dans un environnement où rien ne l'exige, constitue un argument fort en faveur d'un « cliquet de la complexité », c'est-à-dire d'un mécanisme irréversible qui ajoute de nouveaux composants (gènes-protéines) à un systèmes en évolution sans pouvoir parallèlement supprimer les composants préalablement existants, même quand cette suppression pourrait être plus favorable (Cairns-Smith, 1995). De fait, une des observations les plus surprenantes dans nos résultats est que la dynamique de complexification persiste malgré les avantages sélectifs directs et indirects des organismes simples, ce qui exclut les hypothèses sélectives pour expliquer l'accumulation de complexité.

Ces résultats pourraient être interprétés en faveur d'un mécanisme non-sélectif tel que la « Zero Force Evolutionary Law » (McShea et Brandon, 2010). Cependant, en l'absence de sélection, la complexité décroît rapidement pour toutes les conditions testées et pour toutes les mesures de complexité, ce qui disqualifie les hypothèses non-sélectives pour expliquer l'accumulation de complexité.

L'invalidation des hypothèses sélectives et non-sélectives nous a conduit à étudier plus finement la dynamique évolutive d'accumulation de complexité. Nous avons alors pu montrer que celle-ci était due à un avantage sélectif des organismes les plus complexes sur les moins complexes... Mais pas sur les organismes simples ! Ce constat paradoxal nous montre que la dynamique de complexification est due à l'existence de relations d'épistasie de signe entre les mutations orientant vers une structure fonctionnelle simple et les mutations orientant vers une structure fonctionnelle complexe.

L'interprétation géométrique des structures fonctionnelles et du phénotype dans Aevol permet d'illustrer le mécanisme de cliquet que nous avons mis en évidence. Dans cette expérience, la cible phénotypique est identifiable par une ou plusieurs protéines-triangle P_{optimale} de moyenne $m = 0,5$ et de demi-largeur $w = 0,1$. Cependant, dès que le protéome contient une protéine $P_{\overline{\text{optimale}}}$ telle que $m \neq 0,5$ ou $w \neq 0,1$, il n'est plus possible d'identifier simplement la fonction car l'évolution n'a plus accès à la cible phénotypique initiale mais seulement à la cible phénotypique résiduelle, « ce qui reste à identifier » c'est-à-dire la cible triangulaire *moins* la protéine-triangle $P_{\overline{\text{optimale}}}$. En d'autres termes, dès qu'une protéine non-optimale est présente dans le protéome d'un organisme, la cible résiduelle n'est plus triangulaire... et le cliquet commence à cliquer : toute nouvelle protéine est ainsi susceptible de morceler un peu plus la cible, éloignant un peu plus l'organisme de la simplicité et créant de nombreuse opportunités pour recruter de nouvelles protéines non-optimales, accroissant progressivement par ce fait-même la valeur sélective de l'organisme, ce qui creuse un peu plus la vallée de fitness qui le sépare d'une solution simple !

La table IV.1 montre cependant que le cliquet ne s'enclenche pas systématiquement puisque près d'un quart des simulations conduisent à des organismes simples. De plus, nos résultats montrent que le « choix » entre simplicité et complexité s'opère très tôt au cours de l'évolution (souvent sur les premières centaines de générations), ce qui confirme que

le cliquet s'engage dès que les organismes recrutent leurs tout premiers gènes. Ce constat pose la question des événements qui gouvernent cette orientation et de leur contingence. Là encore, l'interprétation géométrique permet de bien comprendre le mécanisme : partant d'un organisme rudimentaire comprenant un seul gène (à priori non-optimal – voir chapitre III, section 2), l'évolution peut emprunter deux types de chemins : (1) optimiser ce premier gène par mutation et le rapprocher de P_{optimale}, (2) recruter de nouveaux gènes, principalement par duplication-divergence. En fonction de cette alternative – correspondant à ce que Crick (1968) a appelé un « *frozen accident* », l'évolution va s'engager vers une identité simple (situation 1) ou complexe (situation 2). Cette hypothèse suggère que l'épistasie de signe provoquant l'accumulation de complexité pourrait être due à des événements de réarrangement chromosomique et tout particulièrement aux événements de duplication de gènes. Nous avons testé cette hypothèse en faisant évoluer 100 populations en l'absence de réarrangements chromosomiques mais en fixant le taux de mutation des switch et des indels de sorte que la variabilité soit équivalente à celle des expériences initiales (sous un taux de mutation $\mu = 10^{-5}$ mut.bp^{-1}.gen^{-1}). Sur ces 100 répétitions, 98 ont conduit à des structures simples (à rapprocher des 25 structures simples observées dans l'expérience initiale, voir table IV.1 page 92). Ce résultat confirme le rôle primordial des réarrangements chromosomiques dans l'évolution de la complexité. Il explique aussi pourquoi les résultats présentés ici n'ont pas été observés jusqu'à présent. En effet, à notre connaissance, Aevol est le seul simulateur capable d'intégrer l'effet des réarrangements chromosomiques dans des simulations d'évolution. Nos résultats appellent donc à une meilleure prise en compte de ces événements, trop souvent délaissés dans les modèles, dans les théories, voire jusque dans les observations.

Dans ce chapitre, nous avons montré, au moyen d'une expérience atypique, spécifiquement conçue, que les hypothèses sélectives et non-sélectives ne parviennent pas à rendre compte de l'émergence d'organismes numériques complexes et que celle-ci est due à un mécanisme de cliquet qui résulte de l'épistasie de signe. Cependant, les figures IV.14, IV.15 et IV.16 traduisent un net ralentissement de la tendance, voire, dans le cas de la complexité génomique, une saturation (figure IV.14). Ces observations posent la question des facteurs qui limiteraient les effets du cliquet de la complexité : le ralentissement (pour $\mathcal{C}_P$ et la fitness) et la saturation (pour $\mathcal{C}_G$) seraient-ils les mêmes dans une situation moins spécifique ? En effet, comme nous l'avons souligné au chapitre I, les différentes hypothèses proposées pour expliquer l'origine de la complexité ne sont pas exclusives. En particulier, l'intuition nous pousse à penser que, dans un contexte de sélection directe favorable à la complexité, celle-ci sera amenée à croître, à fortiori dans le cas d'une sélection directionnelle. Il serait donc intéressant de comparer la dynamique d'évolution de la complexité observée ici avec une situation de sélection directionnelle de la complexité pour estimer les « puissances respectives » du cliquet et de la sélection. C'est cette comparaison qui fera l'objet du prochain chapitre.

Chapitre V

La complexité sélectionnée

1 Introduction

Chez les organismes complexes, le chapitre précédent nous a montré que la complexité quantitative augmente indéfiniment au cours de notre expérience d'évolution *in silico*, du moins en ce qui concerne la complexité fonctionnelle $\mathcal{C}_P$ (chapitre IV, figure IV.15). Car, sur le plan de la complexité génomique $\mathcal{C}_G$ (chapitre IV, figure IV.14), en revanche, l'augmentation semble bornée. Ces deux constats s'accompagnent de gains de fitness réguliers (chapitre IV, figure IV.16) mais en forte décélération au cours des expériences. Dans la mesure où, dans ces expériences, la cible phénotypique est simple, on peut supposer que ces phénomènes d'accumulation ou de ralentissement sont dus au fait que cette cible n'appelle que des complexités limitées. En effet, ces résultats, s'ils montrent l'existence d'un « cliquet de la complexité » ne permettent cependant pas de quantifier la force de ce cliquet par rapport à un mécanisme de sélection directe de la complexité (rappelons que, comme nous l'avons souligné au chapitre I, les différents mécanismes ne s'excluent pas mutuellement). En outre, il est tout à fait envisageable que l'efficacité du cliquet puisse varier au cours du temps évolutif. Pour étayer cette question nous avons mis en place une expérience qui, au contraire du chapitre précédent, encourage l'augmentation de la complexité par l'effet de la sélection. Pour cela, nous utiliserons une cible complexe (voir chapitre III et section suivante), tous les autres paramètres de l'expérience étant quant à eux conservés. Dans cette expérience, l'effet du cliquet de la complexité se superposera donc à l'effet direct de la sélection. Dès lors, en comparant les complexités qu'on y obtiendra, nous espérons pouvoir quantifier les effets respectifs de ces deux forces (section 3). Dans le cadre de cet environnement « exigeant », on constatera à nouveau que la complexité et la fitness suivent des trajectoires similaires. Ayant ainsi établi que le ralentissement qui marque l'acquisition de complexité ne peut être imputé à l'environnement, nous tâcherons de rendre compte des mécanismes qui expliquent ce phénomène (section 3.1). Cette discussion nous amènera à tester l'effet de la sélection pour la robustesse sur le niveau de complexité (section 3.2), ce qui nous permettra de conclure quant aux bornes qui contraignent l'accumulation de complexité.

2 Dispositif expérimental

Pour mémoire (voir chapitre III), la cible phénotypique triangulaire qui constituait un *environnement simple* est ici remplacée par une cible phénotypique gaussienne qui constitue un *environnement complexe*. En quelque sorte, on revient à un type d'expérience classique dans Aevol, où la cible phénotypique est impossible à atteindre avec un nombre fini de protéines. La spécificité est ici que l'on souhaite s'astreindre à un cadre expérimental qui permette la comparaison de l'évolution des complexités.

Bien qu'elle soit – par principe – de forme différente de la cible triangulaire utilisée précédemment, les paramètres de la cible phénotypique complexe ont été choisis de façon à favoriser des individus similaires (mais complexes). L'objectif est ici de limiter les facteurs confondants lors de la comparaison entre les deux expériences. C'est pourquoi nous avons utilisé une fonction gaussienne unique (contrairement à la cible phénotypique classiquement utilisée dans Aevol – voir chapitre III, figure III.1). En outre, dans Aevol le nombre de triangles nécessaire pour fournir une bonne approximation de la cible phénotypique est partiellement déterminé par deux facteurs : la forme de la cible, bien entendu, mais aussi l'aire sous la courbe. En effet, puisque les protéines-triangles recouvrent une aire moyenne donnée (égale à $w_{max}/2$), plus la surface délimitée par la cible phénotypique est importante, plus le nombre moyen de triangles est important, ce qui est susceptible d'influer sur la mesure de la complexité de l'organisme (même si, en pratique, le fait que l'aire des triangles soit fixée par le processus évolutif l'éloigne notablement de $w_{max}/2$). L'aire de la cible est donc un facteur confondant potentiel. Le second facteur, la forme de la cible, traduit, quant à elle, l'effort que l'évolution doit fournir pour se rapprocher de la cible phénotypique au moyen de protéines-triangles (donc de fonctions linéaires par morceaux) : alors qu'une unique protéine peut, dans l'absolu, non seulement approcher mais atteindre exactement une cible triangulaire, aucun ensemble fini de protéines ne suffira à atteindre une gaussienne. C'est le second de ces deux facteurs que nous souhaitons étudier ici de façon à comprendre comment et pourquoi la complexité s'accumule dans un environnement complexe relativement à un environnement simple. C'est pourquoi la nouvelle cible gaussienne a été définie de façon à recouvrir la même aire que la cible triangulaire. Elle est ainsi constituée d'une gaussienne centrée en $0,5$, normalisée en $0,5$ et d'écart-type $\sigma \approx 0,03989$. Spécifiquement, il s'agit donc de la gaussienne $1/2 \cdot \exp(-100\pi(x - 1/2)^2)$ (figure V.1).

Pour permettre la comparaison des simulations conduites en environnement complexe par rapport à l'environnement simple, nous utiliserons les mêmes mesures de complexité $\mathcal{C}_G$ et $\mathcal{C}_P$ déjà définies. Cependant il est important de souligner que la notion d'individus « simples » et « complexes », telle qu'elle a été définie dans le cadre d'un environnement simple n'a plus de sens ici dans la mesure où la cible n'est plus atteignable par un nombre fini de paramètres w, aussi grand soit-il. Il serait possible de définir une classification qualitative exploitant le seul critère de « simplicité » de la cible gaussienne (la symétrie axiale) en définissant comme simples des individus dont tous les gènes partageraient une unique valeur m. Cependant une telle mesure ne permettrait pas davantage la comparaison avec les résultats obtenus au chapitre précédent. En conséquence, au vu de la complexité imposée au phénotype, nous considérerons ici que les individus sont par nature complexes et nous comparerons les complexités des individus obtenus dans les simulations en environnement complexe avec complexités des les individus complexes obtenus dans

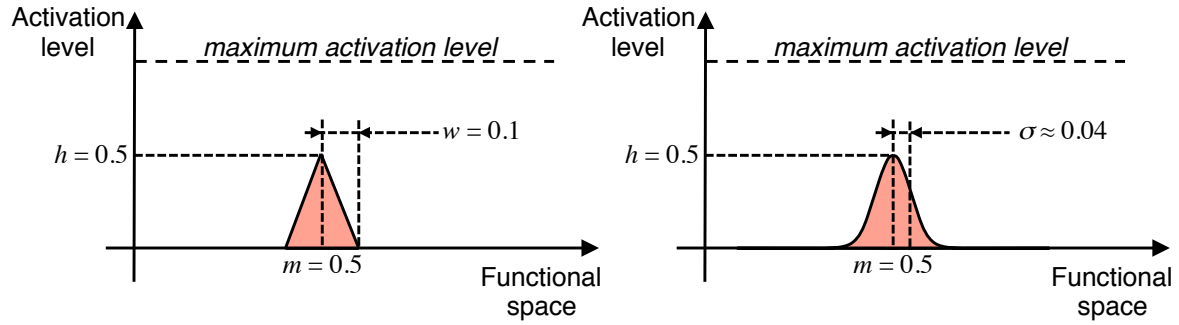


Figure V.1 – Les deux cibles phénotypiques (*phenotypic targets*) utilisées dans les expériences. À gauche : cible simple. À droite : cible complexe. La cible simple est un triangle isocèle centré en $m = 0,5$, de demi-largeur $w = 0,1$ qui a pour hauteur $h = 0,5$. Cette forme peut être parfaitement identifiée par une protéine-triangle unique (voir chapitres précédents). La cible complexe est une courbe gaussienne centrée sur $m = 0,5$, d'écart-type $\sigma \approx 0,03989$, qui culmine à $h = 0,5$. Étant donné la structure triangulaire des protéines Aevol, cette cible ne peut être parfaitement reproduite par un nombre fini de protéines.

les simulations en environnement simple au chapitre IV.

3 Résultats

3.1 Comparaison entre environnements simple et complexe

Les figures V.2 et V.3 présentent les distributions de complexités génomiques $\mathcal{C}_G$ et fonctionnelles $\mathcal{C}_P$ obtenues pour les simulations en environnements simple et complexe pour les différents taux de mutation. Sur le plan de la complexité génomique (figure V.2), les distributions sont tout à fait similaires. Sur le plan de la complexité fonctionnelle (figure V.3), les distributions sont similaires mais on observe néanmoins une légère sous-représentation des organismes les plus simples lorsque la sélection pousse à l'accumulation de complexité (environnement complexe).

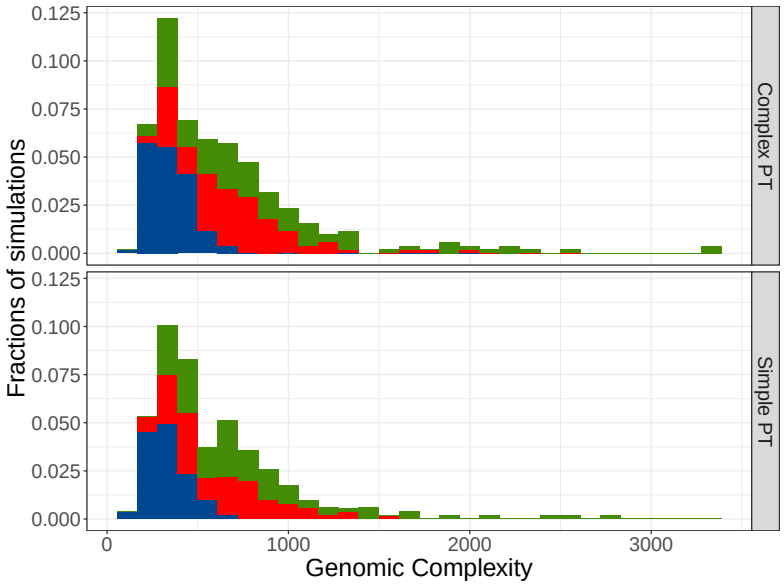


Figure V.2 – Distribution des complexités génomiques $\mathcal{C}_G$ en environnement complexe (*Complex PT*, en haut) et simple (*Simple PT*, en bas). Dans le cas de l'environnement simple, seuls sont pris en compte ici les organismes complexes. Le code couleur indique le taux de mutation (Bleu : 10^{-4} mut.bp^{-1}.gen^{-1} ; Rouge : 10^{-5} mut.bp^{-1}.gen^{-1} ; Vert : 10^{-6} mut.bp^{-1}.gen^{-1}).

Ces observation qualitatives sont confirmées par les comparaisons de moyenne entre les deux expériences (figures V.4 et V.5). En effet, nous n'observons aucune différence significative de complexité entre les organismes ayant évolué dans les deux types d'environnement alors que, dans le même temps, la fitness atteinte dans l'environnement complexe est beaucoup plus faible que la fitness atteinte dans l'environnement simple, ce qui témoigne de la difficulté qu'ont les individus à atteindre la cible complexe (figure V.6).

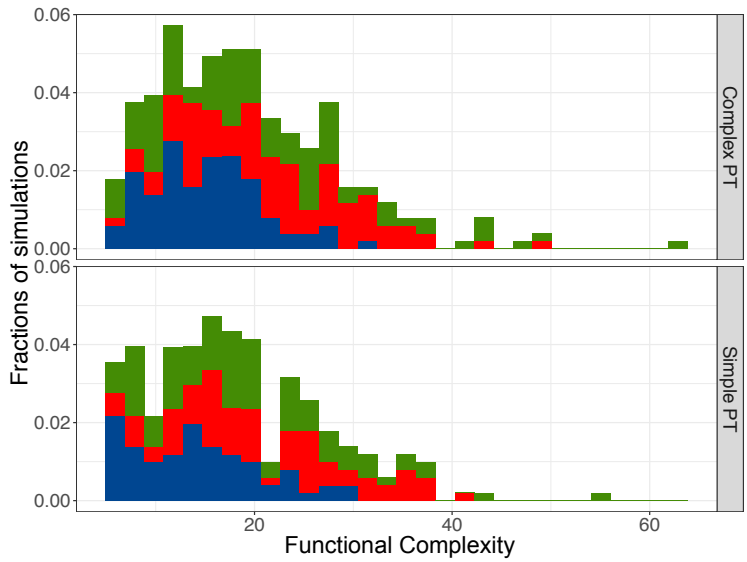


FIGURE V.3 – Distribution des complexités fonctionnelles $\mathcal{C}_P$ en environnement complexe (*Complex PT*, en haut) et simple (*Simple PT*, en bas). Dans le cas de l'environnement simple, seuls sont pris en compte ici les organismes complexes. Le code couleur indique le taux de mutation (Bleu : 10^{-4} mut.bp^{-1}.gen^{-1} ; Rouge : 10^{-5} mut.bp^{-1}.gen^{-1} ; Vert : 10^{-6} mut.bp^{-1}.gen^{-1}).

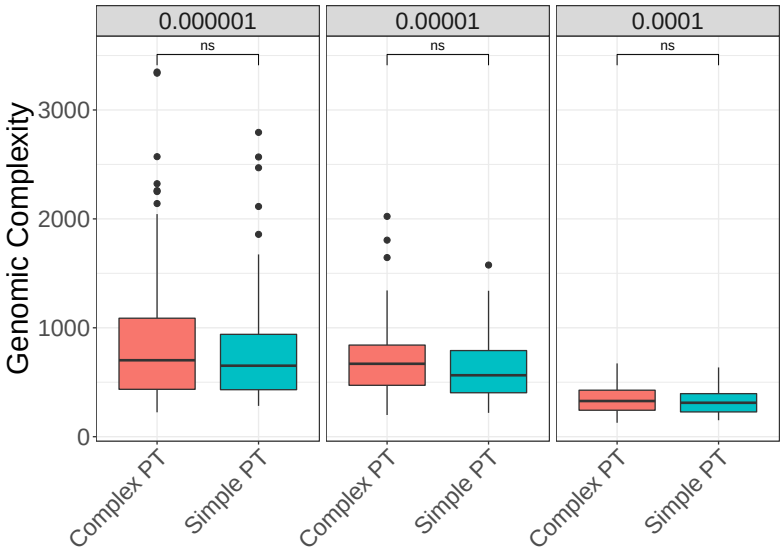
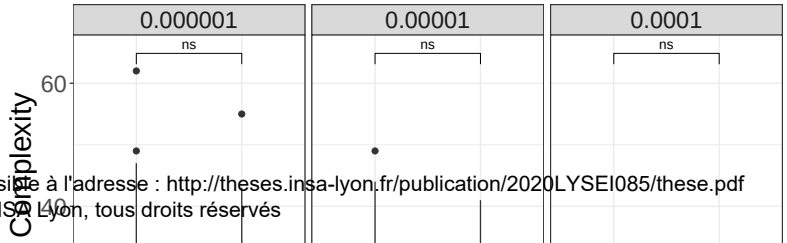


FIGURE V.4 – Comparaison des complexités génomiques $\mathcal{C}_G$ en environnements simple (*Simple PT*, en bleu) et complexe (*Complex PT*, en rouge) à la génération 250 000 pour les trois taux de mutation. L'environnement n'a pas d'effet significatif sur la complexité génomique (Wilkoxon rank test).



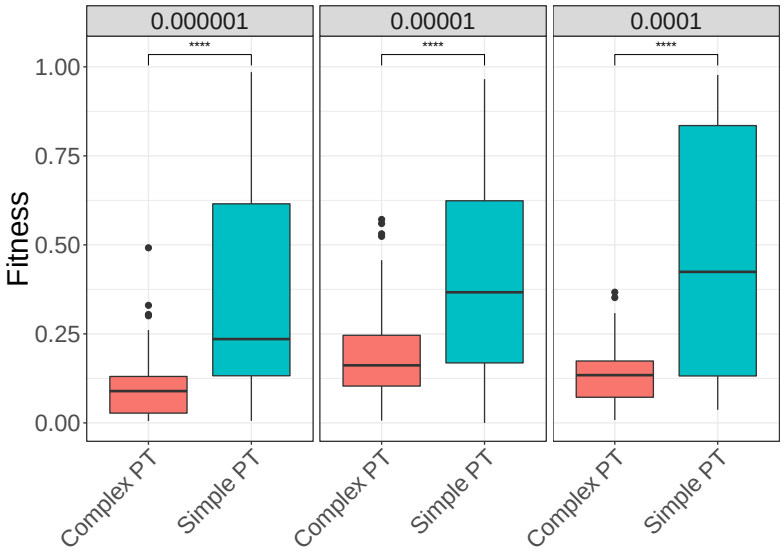


Figure V.6 – Comparaison des fitness en environnements simple (*Simple PT*, en bleu) et complexe (*Complex PT*, en rouge) à la génération 250 000 pour les trois taux de mutation. En environnement simple la fitness des individus finaux est significativement plus élevée qu'en environnement complexe (Wilkoxon rank test).

Les résultats présentés jusqu'ici peuvent paraître déroutants. En effet, en dehors de la différence de fitness, il semble que la différence qualitative entre les deux environnements n'entraîne aucune différence quantitative sur le plan de la complexité des organismes. Cependant ces observations cachent une réelle différence qualitative. En effet, alors que, en environnement simple, nous avions montré que la fitness des individus était corrélée négativement avec leur complexité, en environnement complexe nous observons, inversement, une corrélation positive (figures V.7 et V.8). Cette observation montre clairement que l'environnement complexe requiert une plus grande complexité pour s'approcher efficacement de la cible gaussienne. En environnement complexe, la complexité des organismes est donc bien dirigée par la sélection – en plus du cliquet – même si celle-ci ne conduit pas les organismes à augmenter leur complexité au-delà du niveau atteint par le cliquet en environnement simple.

Comme au chapitre précédent, il semble que l'analyse des individus à la génération 250 000 ne suffise pas à comprendre la dynamique évolutive. La figure V.9 présente l'évolution moyenne de la complexité génomique et de la complexité fonctionnelle au cours du temps, dans les environnements simples et complexes, en séparant les trois taux de mutation. Là encore, étonnamment, les résultats ne montrent pas de différence de dynamique entre les environnements (on notera que les différences de moyenne observées, par exemple, sur la complexité génomique ne sont pas significatives, comme l'avaient montré les figures V.4 et V.5). En revanche, ce qui frappe immédiatement c'est la différence de dynamique observée entre la complexité génomique ($\mathcal{C}_G$, à gauche) et la complexité fonctionnelle ($\mathcal{C}_P$, à droite). En effet, alors que le complexité génomique atteint très rapidement son maximum pour, ensuite, stagner durant toute l'évolution, la complexité fonctionnelle croit continûment durant toute la simulation et ce pour tous les taux de mutation. En

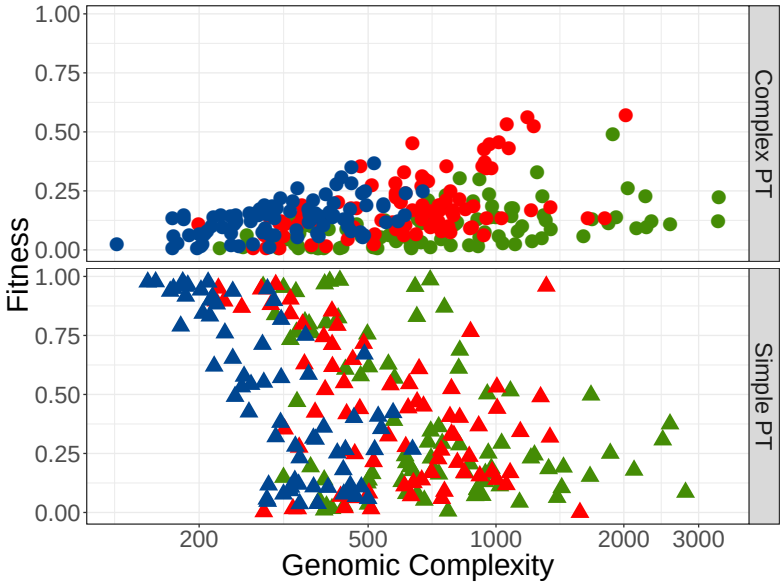


Figure V.7 – Fitness en fonction de la complexité génomique $\mathcal{C}_G$ en environnements complexe (*Complex PT*, en haut) et simple (*Simple PT*, en bas) à la génération 250 000.

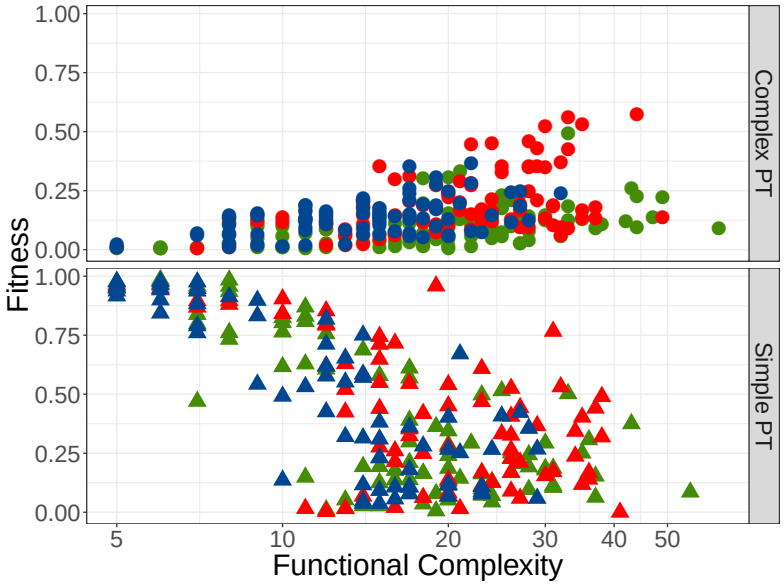


Figure V.8 – Fitness en fonction de la complexité fonctionnelle $\mathcal{C}_P$ en environnements complexe (*Complex PT*, en haut) et simple (*Simple PT*, en bas) à la génération 250 000.

outre, alors que pour la complexité génomique on constate une relation d'ordre claire entre les taux de mutation et la complexité, celle relation d'ordre n'est pas visible pour la complexité fonctionnelle puisque la plus grande complexité fonctionnelle est obtenue avec le taux de mutation intermédiaire (on notera cependant que la différence entre les complexités obtenues pour les taux de mutation de 10^{-5} et 10^{-6} n'est pas significative).

L'observation précédente conduit à formuler une nouvelle hypothèse sur la dyna-

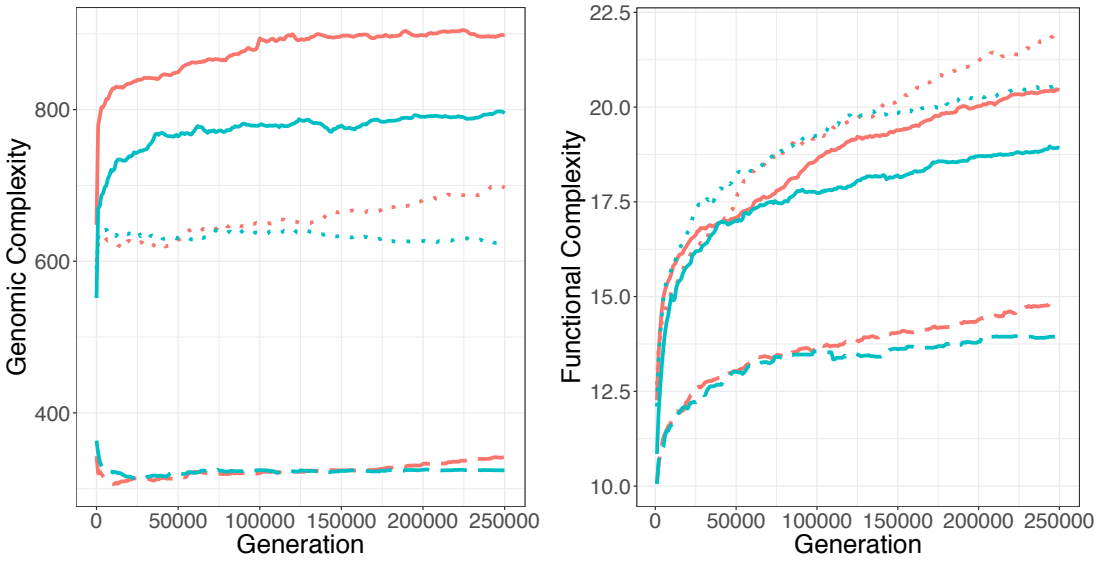


Figure V.9 – Évolution des mesures moyenne de complexité génomique ($\mathcal{C}_G$, à gauche) et fonctionnelle ($\mathcal{C}_P$, à droite) pour les deux types d'environnement (en bleu : environnement simple ; en rouge : environnement complexe) pour les trois types de mutation (tirets : $\mu = 10^{-4}$; pointillés : $\mu = 10^{-5}$; trait plein : $\mu = 10^{-6}$).

mique d'accumulation de la complexité. Cette hypothèse repose sur deux mécanismes. (1) Comme nous l'avons souligné au chapitre I, les organismes biologiques sont multi-échelles et cette propriété est aussi présente dans le modèle Aevol (chapitre II). Or, comme nous l'avons montré au chapitre précédent dans le cas de l'environnement simple, et comme nous pouvons l'énoncer de façon évidente pour l'environnement complexe, c'est la sélection qui entraîne l'augmentation de complexité, soit par un mécanisme de cliquet, soit par sélection directe. Or la sélection s'exerce sur les échelles « supérieures » de l'organisme ; en particulier sur le phénotype, lui-même étant le produit quasi-direct du protéome. Il en résulte que la pression à la complexification de l'organisme s'exerce « de haut en bas » : issue de échelles supérieure, elle se propage vers les échelles inférieures (c'est à dire, en particulier, vers le génome) mais en perdant progressivement son intensité puisque les différentes échelles présentent chacunes des degrés de liberté les unes par rapport aux autres (voir chapitre II). Il s'agit là de la première composante de l'hypothèse. (2) La seconde composante de notre hypothèse est elle aussi basée sur la structure multi-échelle des organismes mais elle repose sur les mécanismes de variation et non plus sur le mécanisme de sélection. En effet, qu'il s'agisse d'observations empiriques (Drake, 1991; Gago *et al.*, 2009) ou de résultats de modélisation (Eigen, 1971; Knibbe *et al.*, 2007a; Fischer *et al.*, 2014; Rutten *et al.*, 2019), plusieurs auteurs ont montré que la taille des génomes est directement liée aux taux de mutation subis par les individus. Comme le montre la figure V.10, cette dépendance est aussi présente dans nos simulations puisque la taille des génomes est déterminée par les taux de mutation et non par la complexité de l'environnement. Or, Knibbe *et al.* (2007a) puis Fischer *et al.* (2014) ont montré que cette dépendance était due à un mécanisme de sélection de la robustesse limitant la taille du génome en fonction des taux de mutation. Au vu de la figure V.9, nous proposons l'hypothèse que la limite

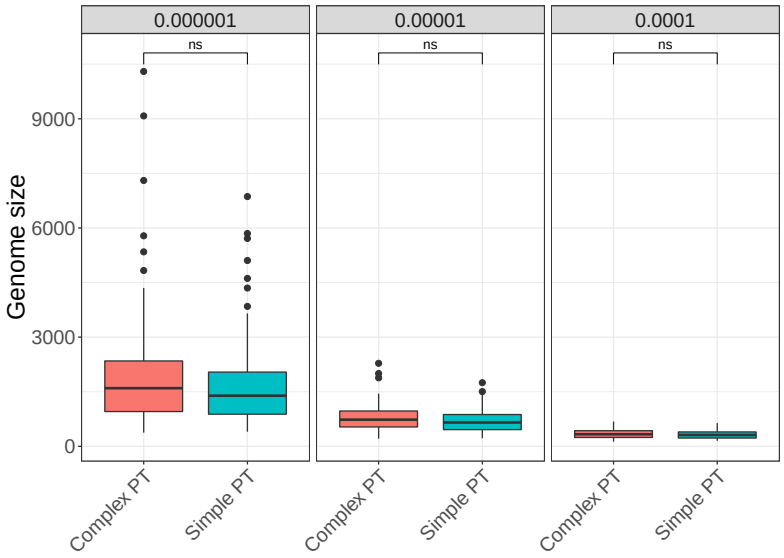


Figure V.10 – Comparaison des tailles moyennes des génomes en environnements simple (*Simple PT*, en bleu) et complexe (*Complex PT*, en rouge) à la génération 250 000 pour les trois taux de mutation. Le type d'environnement n'a pas d'influence significative sur la taille des génomes (Wilkoxon rank test) mais ceux-ci sont significativement impactés par les taux de mutation.

de complexité serait imposée « par le bas » (donc ici à l'échelle génomique) par un tel mécanisme de robustesse : en limitant la complexité génomique, ce mécanisme contraindrait la complexité des organismes. Comme pour la contrainte sélective à l'augmentation de complexité, la contrainte de robustesse se propagerait ensuite « vers le haut » mais en s'affaiblissant en raison des degrés de liberté que présente la *genotype-phenotype map*.

Cette nouvelle hypothèse repose donc sur deux mécanismes opérant à différentes échelles de l'organisme : un mécanisme de sélection de la fitness, opérant principalement sur le niveau fonctionnel, donc sur $\mathcal{C}_P$, et un mécanisme de sélection de la robustesse, opérant principalement sur le niveau génomique, donc sur $\mathcal{C}_G$. Ces deux mécanismes poussant respectivement à la croissance ou la décroissance de la complexité, la complexité effective des organismes proviendrait de l'équilibre entre ces deux forces sélectives. Cette hypothèse nous permet d'énoncer la prédiction suivante : si, pour un organisme de complexité donnée, nous augmentons la contrainte de robustesse (via l'augmentation des taux de mutation), alors nous devrions observer une diminution de complexité, voire, en environnement simple, la transition de phénotypes complexes vers des phénotypes simples. Cette prédiction sera testée dans la section suivante.

3.2 Effet des contraintes de la robustesse sur la complexité

Il est établi que, sous l'effet d'une puissante pression mutationnelle, des lignées robustes peuvent être sélectionnées au détriment de lignées de plus grande fitness (Wilke *et al.*, 2001) et que la compacité d'un génome est un facteur primordial de la robustesse mutationnelle (Knibbe *et al.*, 2007a). En outre, nous avions montré (chapitre IV, figure

IV.7) que, dans nos expériences, les organismes simples ont une plus grande robustesse que les organismes complexes. Enfin, nous avons précédemment formulé l'hypothèse que la robustesse peut permettre de réduire la complexité des organismes en imposant une forte limite à la taille des génomes donc à la complexité génomique $\mathcal{C}_G$.

Pour vérifier cette hypothèse, nous avons soumis les 300 populations ayant évolué en environnement simple à un fort accroissement des taux de mutation pendant 100 000 générations. Plus précisément, nous avons laissé chaque population poursuivre son évolution depuis la génération 250 000 mais sous des taux de mutation μ_{new} 10, 100 et 1 000 fois plus grands que les taux initiaux (sans dépasser le taux maximal $\mu_{new} = 10^{-3}$).

La table V.1 montre, pour les différentes augmentations du taux de mutation, la proportion des organismes qui, alors qu'ils étaient complexes à la génération 250 000, ont basculé vers la simplicité (C→S) à la génération 350 000.

	$\mu = 10^{-4}$	$\mu = 10^{-5}$	$\mu = 10^{-6}$
$\mu_{new} = 10^{-3}$	45,9% [58,3% – 34%] (28/61)	64,4% [74,4% – 52,9%] (47/73)	81,2% [88,1% – 71,6%] (69/85)
$\mu_{new} = 10^{-4}$	/	2,7% [9,3% – 0,7%] (2/74)	10,6% [18,9% – 5,7%] (9/85)
$\mu_{new} = 10^{-5}$	/	/	1,2% [6,4% – 0,2%] (1/85)

Table V.1 – Fraction de lignées ayant effectué une transition de complexe vers simple pour tous les couples de taux de mutation *initial* × *final* (respectivement colonnes et lignes). Les valeurs entre crochet indiquent les intervalles de confiance à 95% ($CI_{95\%}$) estimés par la méthode de Wilson à partir du nombre d'exemples dans chacune des deux classes. Les valeurs entre parenthèses indiquent le nombre de lignées ayant effectué une transition de complexe vers simple à la génération 250 000 par rapport au nombre de lignées complexes à la génération 250 000.

Sur les 600 expériences, 463 ont commencé avec des organismes complexes, parmi lesquelles 156 ($33,7 \pm 4,3\%$) ont basculé de complexe à simple au cours de leur évolution (table V.1). De façon étonnante, alors que ces organismes qui effectuent une transition complexe vers simple ont subi une très forte contrainte de robustesse, leur fitness a fortement augmenté (avec une variation moyenne de $+0,69 \pm 0,055$ au cours des 100 000 générations de l'expérience). Par contraste, les 307 organismes restants (restés complexes tout au cours de leur évolution) ont connu une variation moyenne de $+0,17 \pm 0,34$. En outre, même si ces derniers ont conservé leur identité complexe, il est à noter qu'ils ont subi une forte baisse de complexité en réaction à l'augmentation de la pression pour la robustesse (avec des variations moyennes de $\mathcal{C}_G$ et $\mathcal{C}_P$ de $-126,1 \pm 25,6$ et $-4,28 \pm 0,93$ respectivement).

Rapporté à la proportion des organismes qui, complexes à la génération 10,000, étaient devenu simples à la génération 250,000 (2 individus sur 100 pour un taux de mutation $\mu = 10^{-5}$, voir chapitre IV, section 5.1 et table IV.2), la proportion d'individus passant de

complexe à simple du fait de l'augmentation des taux de mutation est notable, à fortiori si on observe le taux de mutation (certes extrême) $\mu_{new} = 10^{-3}$. On notera cependant que la pression à la robustesse (et donc le taux de mutation) doit être particulièrement forte pour observer un effet massif (table V.1). Cela est probablement du à la sélection pour la robustesse qui, opérant déjà tout au long de l'évolution (donc, ici, depuis 250,000 générations) a produit des organismes dont la robustesse est suffisamment élevée (voir chapitre IV, section 3.2.1 et figure IV.7) pour qu'ils puissent résister sans remaniement majeur de leur génome à un accroissement raisonnable de la pression mutationnelle.

4 Discussion

Dans le chapitre précédent, nous avions montré l'existence d'un « cliquet de la complexité » poussant à l'accroissement de la complexité au cours de l'évolution, même dans des conditions « simples ». En reproduisant la même expérience dans un environnement sélectionnant la complexité, nous espérions pouvoir estimer les effets respectifs du cliquet et de la sélection directe. Nous avons cependant obtenu un résultat contre-intuitif puisque, contre toute attente, les individus qui ont évolué en environnement complexe ne se distinguent pas des individus qui ont évolué en environnement simple (pour autant qu'on exclue ceux qui ont emprunté la « voie de la simplicité »). Ce résultat montre qu'un troisième phénomène entre en ligne de compte dans dans nos simulations, qui freine l'accroissement de la complexité et ce de manière similaire en environnement simple ou complexe.

En observant les différences d'évolution de la complexité génomique $\mathcal{C}_G$ et de la complexité fonctionnelle $\mathcal{C}_P$, nous avons pu proposer un premier mécanisme limitant l'accroissement de la complexité, à savoir la pression pour la robustesse qui opère sur la taille des génomes. Un test expérimental (section 3.2) basé sur l'accroissement de la pression mutationnelle nous a montré que le facteur robustesse avait effectivement un effet (comme le montrent aussi, par ailleurs, les variations de complexité observées sous différents taux de mutation). Cette expérience a cependant montré que, à l'exception peut-être de ceux ayant évolué sous les taux de mutation les plus élevés ($\mu = 10^{-4}$), les organismes de nos simulations sont loin d'atteindre le seuil critique au-delà duquel la robustesse contraindrait fortement la complexité. De fait, il faut multiplier les taux de mutation par un facteur 100 pour observer un effet notable, sauf, là encore, pour les taux de mutation les plus élevés (voir section précédente, table V.1). Il semble donc que les contraintes de robustesses ne soient pas les seules à ralentir l'accroissement de la complexité dans nos expériences.

Au bout du compte, si la contingence peut expliquer l'initiation du cliquet de la complexité, si l'épistasie de signe peut en être le mécanisme, la question de son comportement à long terme demeure. Dans nos simulations, la complexité génomique $\mathcal{C}_G$ semble plafonner et l'accroissement de la complexité fonctionnelle $\mathcal{C}_P$ semble ralentir, ceci malgré l'existence d'une large marge d'amélioration possible chez la plupart des organismes complexes (figures IV.6 et IV.16). Nous pouvons proposer plusieurs hypothèses permettant d'expliquer ce phénomène de plafonnement-ralentissement :

1. La complexité du protéome doit être codée dans le génome mais il existe une borne supérieure à la quantité d'information que le génome, et à d'autant plus forte raison le protéome, peuvent porter, étant donnés un taux de mutation (Eigen et Schuster, 1977) et un taux de réarrangement (Knibbe *et al.*, 2007a; Fischer *et al.*, 2014). C'est le mécanisme de robustesse dont nous avons vu qu'il influe sur la complexité mais qu'il ne permet pas d'expliquer totalement le ralentissement.
2. À mesure que la complexité augmente, le bénéfice que peuvent apporter de nouveaux gènes se réduit au point qu'il peut devenir trop faible pour que la sélection permette leur fixation (on peut parler ici de « seuil de dérive »). En effet, il a été suggéré que la complexité génomique pourrait être essentiellement déterminée par des effets qui relèvent de la génétique des populations (Lynch et Conery, 2003). Ce mécanisme dépendant en premier lieu de la taille des populations, il pourrait être ac-

tif ici puisque nous simulons des populations de taille très limitée (1 024 individus). Notons toutefois que, dans nos simulations, les organismes complexes disposent encore d'une grande marge d'amélioration (voir figure IV.6) ce qui ce limite l'effet de ce mécanisme.

3. Le délai d'attente entre une innovation (correspondant ici à une augmentation de la complexité fonctionnelle $\mathcal{C}_P$) et la suivante augmente à mesure que les organismes deviennent plus complexes. En effet, le « coût de la complexité » est un mécanisme bien connu en évolution (Orr, 2000) ; il stipule que plus le nombre de traits sous sélection est important, plus l'évolution ralentit. Or, dans nos simulations, l'ajout d'un nouveau triangle (ce qui correspond à une innovation fonctionnelle) pourrait être comparé à l'accroissement du nombre de traits sous sélection : alors que, dans nos simulations, les organismes simples remplissent la cible de façon globale – c'est à dire en tant qu'un seul trait –, les organismes complexes découpent (virtuellement) la cible en éléments plus ou moins indépendants les uns des autres. De ce fait, ils sont susceptibles d'endurer un coût de la complexité accru : en même temps que la complexité augmente, l'évolution ralentit d'une façon telle qu'il faudrait attendre potentiellement un temps infini avant de pouvoir observer des organismes complexes dont la fitness soit équivalente à celle des organismes simples.

Ces mécanismes interviennent probablement tous les trois à quelque moment, pour un taux de mutation ou pour un autre. Ainsi, même si le coût à la complexité ralentit l'évolution, les organismes peuvent néanmoins lentement accroître leur complexité (il a ainsi été montré que, dans Aevol, les organismes continuent à acquérir de nouveaux gènes, même après une évolution prolongée pendant 10 millions de générations) jusqu'à atteindre l'un ou l'autre des seuils de robustesse ou de dérive. Une solution simple mais coûteuse permettrait donc de mieux caractériser les limites de la complexité : il *suffirait* de laisser évoluer les organismes pendant plusieurs dizaines, voire centaines, de millions de générations pour observer si le ralentissement de l'évolution est continu ou si nos organismes finissent par atteindre une borne stricte de complexité.

Chapitre VI

Conclusion et Perspectives

1 Bilan

Le questionnement initial de cette thèse était l'origine évolutive de la complexité – supposée – des systèmes biologiques et la méthodologie proposée à priori était la modélisation computationnelle, plus précisément l'évolution expérimentale *in silico*, au moyen de la plateforme de modélisation Aevol. Trois années et cinq chapitres plus loin, que pouvons-nous conclure ?

Nous avons commencé ce document par un tour d'horizon de la complexité des systèmes biologiques et des différentes hypothèses proposées dans la littérature pour expliquer son origine évolutive (chapitre I). Cet état de l'art nous aura conduit à souligner plusieurs points. D'abord, le manque d'uniformité dans les définitions de « la complexité », même si quelques critères (le grand nombre d'éléments, les interactions locales, l'« émergence ») sont partagés par la plupart des définitions. À l'inverse, il semble clairement établi que, quelle que soit la définition de la complexité, les systèmes biologiques sont indubitablement complexes, au sens où ils remplissent tous les critères énoncés ci-dessus. Enfin, nous avons pu constater que l'origine évolutive de cette complexité faisait profondément débat dans la communauté, une ligne de fracture forte séparant en particulier les hypothèses sélectives des hypothèses neutralistes. Nous avons conclu ce premier chapitre en proposant de distinguer ces deux familles d'hypothèses en les mettant en compétition par l'entremise de la simulation. Cependant, à l'inverse de la plupart des travaux ayant exploité la simulation pour étudier l'évolution de la complexité – par exemple, (Lenski *et al.*, 1999; Adami, 2002a,b; Lenski *et al.*, 2003; Yaeger, 1994; Yaeger *et al.*, 2008) – qui se placent dans une optique délibérément sélectionniste en soumettant des individus initialement naïfs à des environnements complexes, nous souhaitons nous positionner dans un cadre plus neutre en soumettant des individus naïfs à des environnements simples afin d'observer si, oui ou non, la complexité augmente dans un tel contexte et pour cela nous nous proposons d'utiliser la plateforme de simulation Aevol (`http://www.aevol.fr`) développée à Lyon par l'équipe Inria Beagle.

Ce choix méthodologique conduit à présenter l'approche d'évolution expérimentale *in silico* et la plateforme Aevol dans le second chapitre de ce manuscrit. Ce deuxième chapitre nous aura permis de montrer en quoi Aevol est un modèle particulièrement adapté pour étudier l'évolution de la complexité biologique. Sans revenir ici sur la structure du modèle, on notera en particulier qu'Aevol est un modèle multi-échelles (génome-transcriptome-

protéome-phénotype-fitness) et que les différentes échelles du modèle sont reliées par une *genotype-phenotype map* (on pourrait même ici parler d'une *genotype-fitness map*) présentant un grand nombre de degrés de liberté ce qui permet une évolution de la complexité différenciée aux différentes échelles (ou du moins ne l'empêche pas), de façon similaire à ce qui peut être observé en biologie, par exemple dans le cadre du « C-value paradox » ou du « G-value paradox » (Thomas, 1971; Hahn *et al.*, 2002). Nous concluons ce deuxième chapitre par une revue des résultats déjà obtenus avec la plateforme en focalisant sur les résultats reliés à notre problématique (Knibbe *et al.*, 2007a; Fischer *et al.*, 2014; Parsons *et al.*, 2010; Beslon *et al.*, 2010b; Rutten *et al.*, 2019; Carde *et al.*, 2019).

Si la question initiale et la plateforme de simulation utilisée faisaient partie des à priori de ce travail, ce n'est pas le cas du dispositif expérimental. Celui-ci est présenté au chapitre III repose sur une idée relativement élémentaire : partant d'organismes dits « rudimentaires » (dans Aevol, il s'agira d'organismes comportant un seul gène), nous pouvons utiliser Aevol pour soumettre ces organismes à deux processus évolutifs, l'un dans un environnement où la complexité sera sélectionnée (au sens où l'accroissement de la complexité d'un organisme sera nécessaire à l'accroissement de sa fitness) et un second où la complexité ne sera pas sélectionnée (au sens où un organisme « simple » pourra avoir une fitness au moins aussi élevée qu'un organisme complexe). Outre ce dispositif expérimental, nous proposons plusieurs mesures de complexité. En particulier Aevol nous permet de quantifier la complexité des organismes à différentes échelles. Nous proposons ainsi deux mesures de complexité, la « complexité génomique » ($\mathcal{C}_G$) et la « complexité fonctionnelle » ($\mathcal{C}_P$), ainsi qu'un critère qualitatif permettant de distinguer des individus « simples » et « complexes ».

Après ces trois premiers chapitres visant à définir la question, les outils et le dispositif expérimental, les chapitres IV et V présentent les résultats. Pour des raisons en grande partie historiques, nous avons choisis de présenter les résultats en séparant les deux situations expérimentales (environnement simple vs. environnement complexe, correspondant aux chapitres IV et V respectivement). En effet, ces deux situations ont été calculées successivement et présentées dans des publications différentes : les résultats en environnement simples ont été présentés dans (Liard *et al.*, 2018) et (Liard *et al.*, 2020) tandis que les résultats en environnement complexe sont décrits dans (Beslon *et al.*, 2020).

L'objectif initial des expériences décrites dans le chapitre IV était de discriminer les hypothèses sélectionnistes des hypothèses neutralistes en observant l'évolution d'organismes élémentaires dans un environnement lui-même simple. Ayant observé que la grande majorité des simulations conduisent à l'émergence d'organismes complexes, au niveau génomique comme au niveau fonctionnel, nous montrons que, contre toute attente, ni les hypothèses selectionnistes ni les hypothèses neutralistes ne permettent d'expliquer l'accumulation de complexité : d'une part les organismes simples ont une fitness bien plus élevée que les organismes complexes (invalidant les hypothèses sélectives), d'autre part, le relachement de la sélection conduit à une baisse de complexité (invalidant les hypothèses neutralistes). Une analyse plus fine de la dynamique nous montre qu'une troisième voie, distincte des deux hypothèses « classiques » de la littérature, explique l'accumulation de complexité dans nos expériences : le « cliquet de la complexité ». Nous montrons en outre que ce cliquet est dû aux interactions épistatiques entre deux types de mutations, les mutations ponctuelles et les réarrangements chromosomiques (en particulier les duplications). Alors que les premières permettent d'optimiser un gène unique et donc d'obtenir un or-

ganisme simple dans un environnement simple, les secondes provoquent un accroissement de complexité via la duplication, puis à la divergence, des gènes. Cependant l'existence de ces deux types de mutations ne permet pas d'expliquer à elle seule le fonctionnement du cliquet de la complexité. Le mécanisme de cliquet est ici du à l'existence d'une épistasie de signe entre ces deux catégories de mutations : une mutation ponctuelle favorable avant une duplication va devenir délétère, voire fortement délétère après la duplication, interdisant le retour à la simplicité, d'où le mécanisme de cliquet.

Les expériences décrites dans le chapitre V étaient censées répondre à celles décrites dans le chapitre IV : en permettant la comparaison entre l'accumulation de complexité en environnement simple et en environnement complexe nous espérions pouvoir quantifier les effets du cliquet de la complexité comparativement à une sélection directe de la complexité fonctionnelle. Cependant, contrairement aux attendus, nous n'avons pas observé une complexité plus élevée en environnement complexe. Ce résultat, étonnant, est aussi décevant de prime abord puisqu'il ne nous permet pas de quantifier les effets relatifs de ces deux modes d'accumulation de la complexité. Il nous aura cependant mis sur la voie d'un résultat inattendu puisque nous observons que, si le type d'environnement n'a pas d'influence visible sur la complexité, le taux de mutation, lui, en a une. En effet, plus le taux de mutation est faible, plus la complexité est élevée. Le comportement lié aux taux de mutation semble aussi lié à l'échelle à laquelle la complexité est mesurée : à l'échelle génomique ($\mathcal{C}_G$) la complexité apparaît comme directement reliée aux taux de mutation mais à l'échelle fonctionnelle ($\mathcal{C}_P$) la contrainte semble relaxée, sauf pour les taux de mutation les plus élevés. Ce résultat nous montre que les contraintes de robustesse mutationnelles sont susceptibles de limiter l'accumulation de complexité, dans un premier temps au niveau génomique et, par propagation via la *genotype-phenotype map* au niveau fonctionnel. Nous avons vérifié cette hypothèse en élevant le taux de mutation des organismes ce qui a conduit d'une part à la simplification de ces derniers et, d'autre part, à une augmentation de fitness, augmentation surprenante dans la mesure où les contraintes de robustesses sont souvent présentées comme antagonistes des contraintes de fitness.

Nos résultats forment un tout intrinsèquement cohérent qui diffère de toutes les hypothèses que nous avons identifiées au chapitre I (section 3). En ce sens, même si le terme « cliquet de la complexité » (ou « complexity ratchet ») est présent plus ou moins explicitement dans plusieurs références (Cairns-Smith, 1995; Lukeš *et al.*, 2011; Doolittle, 2012; Bapteste, 2017), nous n'avons pas trouvé dans la littérature de mécanisme équivalent à un cliquet mû par l'épistasie de signe. Ce résultat semble cependant cohérent avec des observations récentes de Nghe *et al.* (2018) qui montrent que l'épistasie de signe est fréquente dans les voies de signalisation, des systèmes cellulaires connus pour développer une complexité supérieure à la complexité minimale nécessaire pour remplir leur fonction (Soyer et Bonhoeffer, 2006).

Cette capacité à identifier des mécanismes inattendus est une des forces de l'évolution expérimentale *in silico* : en observant un système en évolution dans des conditions expérimentales parfaitement maîtrisées (et souvent inaccessibles *in vivo*), il permet de révéler des mécanismes qui avaient échappé à l'intuition. En ce sens, les programmes d'évolution expérimentale *in silico*, bien que conçus par l'homme, ont le pouvoir étonnant de surprendre leur concepteur. Ce pouvoir de surprise ne doit cependant pas faire oublier leur caractère artificiel et leur statut de modèle. Or, lors des démarches expérimentales fondées sur des modèles d'évolution *in silico*, c'est toujours une question difficile de séparer les tendances liées à l'évolution des artefacts inhérents au modèle. Ici, nous nous sommes appuyés sur Aevol, un modèle qui a déjà fait preuve de sa cohérence, mais cela n'élimine pas ses limites. Au moins trois d'entre elles sont susceptibles d'interférer avec nos résultats. En premier lieu, de la même façon que pour la plupart des modèles ALife, nous travaillons avec des populations de taille très petite en regard des tailles de populations observées dans la nature. Or, comme l'orientation vers la simplicité ou vers la complexité dépend d'un accident initial (« frozen accident »), la taille de population peut influer sur les probabilités d'observer cet accident au sein de la population. Cependant, puisque la sélection ne peut pas inverser le cliquet, nous formulons l'hypothèse que nos conclusions générales ne seront pas affectées, du moins qualitativement, par la taille de population. Des tests préliminaires ont été effectués pour vérifier cette hypothèse mais une étude plus poussée resterait à entreprendre. Deuxièmement, les propriétés de notre chimie artificielle sont susceptibles de différer de la biochimie réelle. En particulier, les effets de dosage sont plus importants dans Aevol que dans la nature. Cette propriété a cependant plus de chances de limiter la complexité que l'inverse puisque les duplications de gènes sont plus délétères dans le modèle. Il en résulte que ce point non plus ne devrait pas affecter nos conclusions. Enfin et surtout, Aevol a beau être un modèle multi-échelle il lui manque tout de même certains niveaux qui sont susceptibles de jouer un rôle crucial dans l'évolution de la complexité. Ainsi la version d'Aevol utilisée ici n'intègre ni écosystème complexe ni réseau de gènes. De ce fait, il est intrinsèquement impossible d'observer des phénomènes de construction de niche qui pourraient exercer un rôle significatif dans l'évolution de la complexité (Rocabert *et al.*, 2017). Pour ce qui est des réseaux de gènes, il est notable que nos résultats recoupent largement ceux que nous avons obtenus avec RAevol (la version d'Aevol dotée de réseaux de régulation) lorsque nous avons fait évoluer les réseaux de gènes dans un environnement soit constant, soit variable (Beslon *et al.*, 2010b; Knibbe *et al.*, 2011; Vadée-Le-Brun *et al.*, 2016). En effet, ces expériences montrent que la complexité des réseaux de gènes est principalement dirigée par les taux de mutation, au point que des réseaux de gènes complexes se sont développés même en environnement constant. Une perspective intéressante serait donc de reproduire les expériences décrites ici en utilisant RAevol plutôt que Aevol.

Enfin, en guise de conclusion, nous aimerions souligner que les résultats que nous avons rassemblés sur la base d'un modèle nul ne disent rien quand à l'existence ou à l'inexistence d'un phénomène de sélection intrinsèquement favorable à la complexité dans la Nature. Ils montrent, en revanche, qu'il n'est pas nécessaire d'en passer par une telle sélection de la complexité pour que celle-ci s'accumule. Ainsi, des structures biologiques complexes peuvent abonder jusque dans les conditions où la complexité est purement fortuite. Réciproquement, la fonction globale des structures complexes qu'on peut observer pourrait tout à fait être simple. Nous pensons qu'un tel résultat est d'une grande importance pour

la biologie évolutive, pour la biologie des systèmes et pour les sciences du vivant dans leur ensemble.

2 Perspectives

Notre travail ouvre de nombreuses perspectives. Nous allons ici décrire les trois directions qui semblent les plus prometteuses. La première (section suivante) est une évidence : il reste de nombreuses questions en suspens que nous pourrions étudier directement avec Aevol. La seconde consisterait à prendre du recul pour chercher à formaliser le mécanisme de cliquet afin de mieux identifier les conditions dans lesquelles il est susceptible d'opérer (section 2.2). Enfin, la troisième direction consisterait à élargir le point de vue disciplinaire pour englober le champ des sciences humaines et sociales. En effet, ce champs disciplinaire dispose depuis plus de trente ans, du concept particulièrement fructueux de « dépendance au chemin » ou « *path dependence* » (David, 2007). Ce concept questionne nos résultats quand à sa possible similitude avec notre cliquet de la complexité. Nous conclurons donc cette thèse en interrogeant cette similitude (section 2.3).

2.1 Toujours un peu plus loin

La perspective première serait de prolonger l'analyse des expériences présentées dans les chapitres IV et V afin de clarifier ou d'approfondir différents éléments. Comme nous l'avons déjà évoqué en conclusion du chapitre V, le premier d'entre eux serait la cause du ralentissement de la « complexification » fonctionnelle au cours du temps. Pour celà il serait nécessaire d'analyser les trajectoires individuelles (quand nous nous sommes ici concentrés sur les trajectoires moyennes) afin de relier l'évolution de la complexité avec différents paramètres de individus (taille des génomes, nombre de gènes...). Une telle démarche pourrait permettre de relier le ralentissement à certaines caractéristiques (en particulier la complexité génomique, la complexité fonctionnelle ou la taille des génomes) et donc de mieux en identifier la ou les causes. Les résultats du chapitre V nous ont en effet à la fois montré que les contraintes de robustesse pouvaient être cause de ralentissement mais aussi que, dans le cas des taux de mutation les plus faibles, ces contraintes ne suffisent probablement pas à expliquer le ralentissement constaté.

Le second point à approfondir serait une identification plus précise des interactions épistatiques activant le cliquet de la complexité. En effet, dans le chapitre IV, ce sont des indices indirects (les variations concomitantes de complexité et de fitness) qui nous ont permis d'identifier l'épistasie de signe comme moteur de la complexification. Il serait particulièrement intéressant d'analyser la dynamique évolutive de nos populations à l'échelle des mutations elles-mêmes. Cela permettrait d'analyser l'effet spécifique que chaque type mutation peut avoir sur la complexité, ainsi que sur la fitness, l'évolvabilité et la robustesse, et de quantifier les différences entre les mutations ponctuelles et les réarrangements. Ceci permettrait de distinguer bien plus clairement les interactions épistatiques qui se produisent dans ce modèle. Une deuxième approche, très similaire aux méthodes employées en évolution expérimentale *in vivo* serait de reconstruire/déconstruire les différentes mutations observées dans les lignées de façon à en inverser l'ordre d'arrivée. Cette approche permettrait de quantifier précisément les liens épistatiques. En effet, dans le cas d'une épistasie positive ou négative faible (donc n'inversant pas le signe des effets sur la fitness), l'ordre de survenue de deux mutations influe sur leurs avantages sélectifs respectifs mais finalement peu sur l'ensemble des mutations fixées à terme (les deux mutations se fixeront, quel que soit leur ordre d'arrivée). En revanche, une épistasie de signe entraîne une

forte contingence puisqu'une seule des deux mutations (celle qui, par hasard, arrivera en premier) va se fixer. Cet effet devrait donc être particulièrement visible via des reconstructions ou des déconstructions de mutations. On notera cependant qu'une telle démarche se heurterait ici aux caractéristiques particulières des mutations incriminées (mutations ponctuelles et réarrangements chromosomiques). En effet, dans le cas des réarrangements (ou variants structurels), il arrive fréquemment qu'il ne soit pas possible d'inverser l'ordre d'arrivée des mutations, tout simplement parce que c'est le réarrangement lui-même qui a l'opportunité de créer la séquence qui va muter par la suite [1].

Le troisième élément sur lequel un approfondissement pourrait apporter un éclairage supplémentaire sur l'évolution de la complexité serait l'identification plus précise des causes initiales du cliquet, c'est à dire du ou des événements qui conduisent certaines lignées vers la complexité ou d'autres vers la simplicité (dans le cas de l'environnement « simple » étudié au chapitre IV). Les caractéristiques propres de ce « frozen accident » (Crick, 1968) permettraient de mieux comprendre l'incidence des différentes mutations sur l'identité simple ou complexe des individus, dans la simulation et, à terme, dans des populations réelles. Une question se pose ici avec une acuité particulière : étant donné l'avantage sélectif des individus simples sur les individus complexes, la probabilité de ce *frozen accident* dépend-elle de la taille de la population incriminée ? Le cas échéant, les résultats présentés dans cette thèse seraient-ils conservés pour les tailles de population beaucoup plus importantes ? Des études préliminaires menées avec Aevol tendent à montrer que la fraction d'individus complexes ne dépend pas de la taille de la population, ces résultats préliminaires demanderaient néanmoins à être confirmés.

Enfin, toujours dans une perspective d'approfondissement des résultats présentés dans cette thèse, nous avons déjà évoqué la possibilité de reproduire tout ou partie des expériences mais en utilisant le modèle RAevol plutôt que Aevol. Pour mémoire, en plus des différents niveaux d'organisation présents dans Aevol, RAevol inclut un réseau de régulation génétique dont la complexité pourrait être étudiée au long de l'évolution (Beslon *et al.*, 2010b,a; Knibbe *et al.*, 2011; Vadée-Le-Brun *et al.*, 2016). Étant donné l'attention dont les réseaux biologiques (réseaux de gènes, réseaux métaboliques) ont été le sujet dans une perspective de « biologie des systèmes », une meilleure compréhension des règles régissant leur complexité serait particulièrement intéressante. En effet, la logique globale dans laquelle se place généralement la biologie des systèmes est que la complexité des réseaux biologiques reflète la complexité de l'environnement des organismes. Si, comme tendraient à le montrer nos résultats, cette complexité est en grande partie indépendante de la complexité environnementale, c'est tout le paradigme d'analyse des réseaux (et en particulier des réseaux de gènes) qui serait alors questionnée.

2.2 Prendre de la hauteur

Quoique nous ayons travaillé dans cette thèse sur un modèle, celui-ci était basé sur la simulation et, en tant que tel, il ne permettait qu'un raisonnement empirique, à la manière des sciences expérimentales. Ce caractère empirique rend en outre difficile de

1. Un exemple d'une telle situation serait un réarrangement de type « duplication » qui insère une nouvelle séquence, laquelle pourrait subir par la suite une ou plusieurs mutations ponctuelles. Ces dernières ne pourraient pas être reconstruites *avant* le réarrangement puisque la séquence correspondante n'existe alors tout bonnement pas.

généraliser des résultats à d'autres systèmes car celle-ci demande que les résultats soient formalisés de façon à identifier précisément leurs conditions d'occurrence. Dans ce cadre, une perspective particulièrement enthousiasmante serait de formaliser le mécanisme de cliquet identifié ici pour mieux en cerner les causes et les conséquences.

Quelques auteurs se sont attachés à formaliser la question de l'évolution de la complexité dans les systèmes biologiques, en adoptant en particulier le point de vue de la physique statistique ou des sciences de l'information (Adami, 2002a; Krakauer, 2011). Or, en adoptant ce point de vue, ils conjecturent que la complexité d'un organisme ne peut dépasser la complexité de l'environnement au sein duquel il évolue. Cette conjecture, qui semble tout à fait raisonnable au vu du second principe de la thermodynamique, est manifestement violée dans nos expériences en environnement simple, ce qui pose évidemment question. Là encore, une démarche de formalisation de nos résultats permettrait probablement d'identifier pourquoi cette conjecture ne s'applique pas dans le cadre d'Aevol et de vérifier, surtout, si elle s'applique dans le cas d'organismes biologiques réels ou non. Dans la suite de cette section, nous allons ébaucher quelques éléments de réponse à cette question en comparant en particulier Aevol avec les descriptions classiques de l'évolution (sur lesquelles se fondent en particulier Adami (2002a) et Krakauer (2011)).

Un premier élément de réponse à la question de la complexité de l'organisme relativement à la complexité de l'environnement serait de mieux définir ce qu'est l'environnement d'un organisme et, surtout, de prendre conscience que l'environnement de l'organisme (au sens écologique statique) ne correspond pas à son « environnement évolutif ». En effet, si l'environnement écologique correspond à l'ensemble des interactions potentielles entre un organisme et son milieu extérieur, l'« environnement évolutif » correspond lui à l'ensemble des interactions accessibles, par mutation, en lien avec leurs effets sur la fitness de l'organisme. En d'autres termes, l'évolution ne perçoit l'environnement dans lequel vivent les individus non pas directement et globalement mais seulement localement au travers du prisme de *l'inadaptation* des organismes, de la différence entre, d'un côté, le phénotype des organismes et, de l'autre, les ressources que propose l'environnement. Dans le cadre d'Aevol, cette différence entre « environnement écologique » et « environnement évolutif » est directement interprétable en termes de cible environnementale. Alors que l'environnement écologique correspondrait à la fonction objectif définie comme paramètre du modèle et figée, du moins dans les simulations présentées ici, l'environnement évolutif correspond à une fonction différente que nous appellerons « cible phénotypique résiduelle » et qui correspond à la différence entre le phénotype d'un organisme et la fonction objectif [1]. Or, de par sa définition même, cette cible phénotypique résiduelle n'a pas une complexité propre mais bien une complexité qui dépend à la fois de la cible « absolue » et de l'organisme lui-même. On comprend dès lors que la complexité de l'environnement évolutif n'est pas figée mais qu'elle peut augmenter ou diminuer au gré de l'évolution de l'organisme lui-même. En particulier, un organisme complexe est susceptible de provoquer une forte augmentation de la complexité de la cible résiduelle (de l'environnement évolutif), même dans le cas d'une cible phénotypique (d'un environnement écologique) simple. Ce principe est au coeur des mécanismes d'épistasie et apparaît ici comme un des déterminants théoriques du cliquet de la complexité. Il serait donc particulièrement utile de formaliser ce principe

1. On retrouve ici une idée déjà présente chez Saunders et Ho (1976) : « *The increase in complexity is shown to be a consequence of the process by which a self organizing system optimizes its organization with respect to locally defined fitness potential.* »

de cible résiduelle dans des situations plus génériques que le seul modèle Aevol.

Si ce principe de cible résiduelle apparaît comme un des moteurs du cliquet de la complexité, il n'en est cependant pas le seul, ni même le principal. En effet, nous avons vu au chapitre IV que la présence de réarrangements chromosomiques était indispensable à l'évolution de la complexité. Or, même en l'absence de réarrangements, le principe de la cible résiduelle reste valide, ce qui montre que ce principe est nécessaire mais qu'il n'est pas suffisant – il faut au moins lui adjoindre des types de mutation spécifiques. Il convient donc de formaliser ce que l'introduction de réarrangements chromosomiques dans le modèle change à la dynamique évolutive. Un premier élément de réponse est que les réarrangements (et en particulier les duplications) introduisent une dimension historique dans la dynamique évolutive des génomes puisque, comme nous l'avons déjà fait remarquer ci-dessus, certaines mutations sont rendues possibles du fait de la duplication préalable d'un segment chromosomique. En outre, du fait des duplications, la complexité des organismes n'est plus seulement liée à leur interaction immédiate avec leur environnement, mais aussi à l'historique de ces interactions. En effet, les duplications entraînent une accumulation d'information dans le génome. De ce fait, la complexité de l'organisme ne reflète plus la complexité de son environnement « immédiat » mais la somme des complexité des environnements perçus par l'organisme au cours de son histoire évolutive, somme dont on comprend immédiatement qu'elle dépasse largement la complexité de l'environnement « ici et maintenant ». Enfin, un troisième élément de réponse est que les réarrangements modifient considérablement la structure du *fitness landscape* d'un organisme. En effet, alors que le *fitness landscape* est généralement perçu comme basé sur des déplacement locaux (via les mutations), les réarrangement introduisent des « sauts » au sein du *fitness landscape* en connectant des points distants. Or, ces sauts sont susceptibles de changer considérablement la nature du *fitness landscape* (Kauffman et Levin, 1987) et donc la dynamique du processus évolutif. Une question ouverte serait de savoir si le *fitness landscape* est porteur d'information et si celle-ci peut être intégrée par l'organisme au sens d'Adami (2002b). Cette question théorique mériterait probablement à elle seule une réflexion approfondie.

Tous les résultats présentés ici ont été obtenus sur le modèle Aevol. Or, même s'il reste loin de la complexité des systèmes biologiques réels, ce modèle est déjà beaucoup trop compliqué pour se prêter à une formalisation. Afin de formaliser les résultats présentés dans cette thèse, une approche possible consisterait donc à proposer des modèles simplifiés (par rapport à Aevol), plus propices à la formalisation, mais préservant une dynamique similaire en termes de cliquet de la complexité. Un exemple de simplification pourrait être de s'affranchir de tout ou partie de la *genotype-phenotype map* d'Aevol tout en conservant les principes généraux décrits ci-dessus, à savoir la notion de cible résiduelle et les mécanismes de réarrangements chromosomiques. Nous présentons ici une ébauche de ce que pourrait être un tel modèle simplifié :

Soit un génome de N bases binaires (dans les illustrations suivantes nous prendrons $N = 4$) et un *fitness landscape* défini par un unique optimum (par exemple 1111) et par une fonction de fitness $f()$ décroissante en fonction de la distance à l'optimum (donc $f(1111) > f(1011) > f(0011) > f(0010) > f(0000)$). Un processus évolutif utilisant des mutations ponctuelles (de type « switch ») produirait des dynamiques évolutives de type $0000 \rightarrow 0100 \rightarrow 0110 \rightarrow 1110 \rightarrow 1111$, où le symbole « $\rightarrow$ » correspond à la mutation

d'une base dans le génome (on néglige ici les effets de la dérive génétique).

Il est possible de construire un modèle alternatif comportant, en plus des mutations de type switch, des mutations de type duplication-délétion. Prenons comme règle de duplication-délétion la duplication intégrale de la séquence. Un organisme est alors constitué d'un ensemble $\mathcal{S}$ de séquences (une duplication consistant donc à copier intégralement une des séquences de $\mathcal{S}$, une délétion consistant en retour à supprimer une séquence dans l'ensemble $\mathcal{S}$). Il est alors nécessaire de se doter d'une règle de composition permettant de calculer le phénotype (une séquence de N bits) issu de $\mathcal{S}$. Une règle possible serait la somme binaire (si $\mathcal{S} = \{0011, 0110\}$, alors le phénotype de l'organisme est la somme binaire $0011 + 0110 = 1001$). Dans ce nouveau modèle, un chemin évolutif possible serait : $\{0000\} \Rightarrow \{0000, 0000\} \rightarrow \{0010, 0000\} \rightarrow \{0010, 0100\} \Rightarrow \{0010, 0100, 0100\} \rightarrow \{0011, 0100, 0100\} \Rightarrow \{0011, 0100, 0100, 0100\}$ où le symbole « $\Rightarrow$ » dénote une duplication ($\rightarrow$ correspondant toujours au switch). Ce chemin évolutif produit la séquence phénotypique suivante : $0000 \rightarrow 0000 \rightarrow 0010 \rightarrow 0110 \rightarrow 1010 \rightarrow 1011 \rightarrow 1111$ qui converge bien vers l'optimum.

Ce modèle, certes simplifié à l'extrême, illustre bien le mécanisme du cliquet de la complexité : en ajoutant un opérateur de duplication et une règle de composition, nous avons autorisé un très grand nombre de chemins évolutifs alternatifs dont certains (comme celui donné en exemple ci-dessus) entraînent un accroissement de complexité largement supérieur à la « complexité de l'environnement » (qui, elle, n'a pas changé par rapport au modèle ne comportant que des mutations ponctuelles). En outre, le second modèle accumule de l'information génétique (donc de la complexité) au cours de son histoire évolutive, ce qui n'est pas le cas du premier. On voit donc comment la modification des opérateurs de variation peut *éventuellement* autoriser l'action du cliquet de la complexité. En outre, même si celle-ci reste à faire, un tel modèle se prête parfaitement à une analyse formelle et pourrait donc permettre de dégager les lois sous-jacentes.

2.3 Vers d'autres horizons ?

2.3.1 Introduction

De la même façon que les populations d'organismes étudiés dans cette thèse sont des systèmes intrinsèquement complexes, les systèmes sociaux ne se laissent pas réduire à une somme de lois indépendantes et se déploient à des échelles multiples. Dans ce domaine, bien des systèmes mériteraient qu'on s'y arrête pour les rapprocher de nos travaux. La loi, par exemple, que nul n'est censé ignorer mais dont la subtilité essentielle se trouve redoublée par l'inflation des textes normatifs, rappelle la profusion de complexité que connaissent les organismes des populations que nous avons étudiées dans les chapitres précédents. Sa complexité est-elle inéluctable ? L'édification du Code civil des Français sous Napoléon, intervenant après une longue guerre civile qui laisse la France en ruines, prête à repartir de pas grand-chose fait figure d'exemple d'une simplification. Sans que ceci ne constitue une preuve quelconque, comment ne pas être intrigué par le parallèle que trace cet exemple avec la faramineuse augmentation des taux de mutation que doivent surmonter les organismes de nos expériences (chapitre V, section 3.2) avant d'avoir une chance de changer de trajectoire et de se simplifier. Dans le même ordre d'idée, tout doctorant se doit de maîtriser rapidement les méandres de l'inscription ou de la réinscription

en thèse et l'expérience quotidienne ne permet à personne d'échapper à la pensée que l'administration tout entière est un autre exemple de cette loi fondamentale de l'accroissement contingent de la complexité (on pourra déplorer que dans ce registre on ne trouve pas d'exemple analogue à la simplification que nous évoquions à l'instant pour la loi...).

D'autres exemples, aux effets tout à fait concrets et palpables dans la vie quotidienne, peuvent venir à l'esprit, comme l'urbanisme et ses méandres ou la sophistication croissante et vertigineuse des produits financiers (les *produits dérivés*, en particulier). Un autre exemple encore, cher à l'auteur : le code source des programmes informatiques, qui procède généralement d'ajouts successifs. Même s'il peut connaître des phases de réorganisation et de rationalisation au travers d'étapes de *refactoring* (comme le dit l'expression consacrée), ce n'est dans la plupart des cas que lorsqu'une réécriture complète du logiciel est entreprise qu'une simplification significative peut voir le jour... Pour autant que cette renaissance ne se solde pas par un échec.

Mais, de façon plus spécifique, c'est vers les domaines de l'économie et de la cliométrie (l'histoire de l'économie) que notre curiosité s'est en particulier dirigée car les idées de Darwin y ont connu un certain succès et ont conduit à des élaborations qu'il nous faut évoquer avec les concepts de « *path dependence* », « d'*increasing returns* » et de « *lock-in* ». Ce n'est cependant pas une question triviale que de déterminer le degré de parenté qu'entretiennent ces concepts avec les nôtres ou, pour poser tout de suite la question sans ambages : le cliquet de la complexité et la *path dependence* parlent-ils de la même chose ?

Nous allons dans les paragraphes qui suivent décrire le concept de *path dependence* tel que le présentent les auteurs qui nous ont semblé être les plus emblématiques du domaine (en particulier, W. B. Arthur et A. David). Nous montrerons ensuite comment ce champ de recherche, en particulier représenté par la branche de l'« *evolutionary economics* », quoiqu'il trouve sa source dans les idées darwiniennes, a perdu le contact avec les recherches actuelles sur l'évolution darwinienne. Ce constat nous conduira à poser la question du lien qu'entretiennent les concepts économiques de *path dependence*, d'*increasing returns* et de *lock-in* avec ceux, proprement biologiques, liés au cliquet de la complexité.

2.3.2 Exemples introductifs

L'histoire est souvent présentée comme une succession d'événements qui procèdent les uns des autres comme s'ils entretenaient entre eux des relations aussi nécessaires que les lemmes d'une démonstration mathématique. Le concept de *path dependence*, sans qu'il exclue la validité de cette représentation de l'histoire dans de nombreux cas (même s'il aurait tendance à la qualifier d'*anhistorique* dans la mesure où, dans de tels cas, le temps n'est finalement qu'un facteur subalterne), met l'accent sur l'importance d'événements qu'on pourrait, si l'on n'y prenait garde, considérer comme anecdotiques. Paul A. David a proposé en 2007 un article (David, 2007) dans lequel il récapitule les nombreux travaux qu'il a menés autour de ce concept de *path dependence* (David, 2007). Il y propose notamment des exemples puisés dans l'Histoire, relevant des grands conflits qui ont marqué le temps, mais il rapporte surtout des exemples spécifiquement économiques, dans une perspective technologique, qui illustrent bien le phénomène et la pertinence de son étude.

L'exemple incontournable de la *path dependence* indique d'emblée l'affinité des auteurs pour les questions économiques en rapport avec l'évolution de la technique. Il s'agit du clavier QWERTY. Au contraire de l'intuition, il a été suggéré (Yamada, 1980) que cette

disposition visait à *ralentir* les dactylographes pour limiter les problèmes mécaniques que rencontraient les machines à écrire, dont les marteaux pouvaient se bloquer. Cette hypothèse a connu un immense succès. Et malgré une controverse sur la logique qui présida réellement à l'agencement des touches, certes ravivée de façon convaincante par Yasuoka et Yasuoka (2011), l'idée demeure au moins que cette disposition n'était pas nécessairement la meilleure des points de vue de l'ergonomie et de l'efficacité dactylographique. Les tenants des dispositions alternatives, en particulier du clavier « Dvorak », affirment en tout cas que leur disposition favorite permet aux dactylographes entraînés d'atteindre des vitesses de frappe supérieures. Il n'en demeure pas moins qu'aux variantes régionales près, et même en l'absence de statistiques précises à ce sujet, la disposition QWERTY reste si majoritaire que la plupart des utilisateurs ignorent qu'il en existe d'autres. C'est justement l'élément clé en ce qui concerne la problématique qui nous intéresse : un événement, relativement anodin au moment où il se produit – en l'occurrence : le choix de la disposition du clavier – connaît de telles conséquences qu'il devient un état de fait. En ce qui concerne le clavier QWERTY, il a acquis une telle popularité, qu'il est devenu économiquement absurde d'envisager une véritable alternative concurrente, fut-elle meilleure du point de vue de l'efficacité dactylographique (et donc économique !). Il s'agit-là d'un exemple typique de *lock-in* où un événement économique et les événements qui en découlent entérinent un choix historique. En ce sens, il s'agit aussi d'un exemple du concept plus général de *path dependence* qui, dans son principe, ne se borne pas aux questions de conquêtes de marchés économiques, même si c'est surtout dans ce registre que se placent P. A. David et W. B. Arthur. Nous proposerons dans la section suivante une définition explicite de ces notions.

L'exemple du clavier QWERTY, bien qu'on le trouve de façon récurrente dans les articles sur la *path dependence*, peut sembler anecdotique. Il ne s'agit en effet que d'un exemple, marquant dans la mesure où il parle d'un des outils les plus ubiquitaires de nos vies modernes et pour lequel nos mémoires musculaires soulignent la stabilité de l'état de fait induit. Avant d'en proposer quelques autres, aux conséquences plus significatives, il faut remarquer que la notion *d'anecdotique* est ici quelque peu fragile dans la mesure où elle peut conduire à négliger le phénomène de *path dependence* lui-même puisqu'il peut être généralement occulté par l'effet du biais du survivant : au fond, tout ce qui a survécu à une compétition en sort auréolé d'un halo qui fait croire à sa nécessité historique. C'est particulièrement le cas dans la présentation hagiographique de l'histoire, où la narration pare les vainqueurs des vertus qui leur assurent par avance la victoire. Cette façon de voir risque de faire négliger tout ce qui contreviendrait à une histoire envisagée comme une succession de nécessités (au même titre d'ailleurs que la complexité des systèmes biologiques est « naturellement » interprétée comme une succession de nécessités). En ce sens, David (2007) reprend chez l'historien G. Mattingly l'exemple éclairant de l'Invincible Armada, affrétée en 1588 par le roi d'Espagne pour établir sa domination sur l'Angleterre. Selon (Mattingly, 1987), celle-ci n'aurait échoué dans son entreprise qu'à cause d'un défaut *fortuit* d'approvisionnement en eau potable. Le Duc de Medina Sidonia eût-il supervisé cet approvisionnement avec plus de soin, l'issue de cette entreprise militaire aurait, selon l'auteur, été toute différente et cela aurait, éventuellement, changé la face du monde que nous connaissons.

David (2007) cite d'autres exemples historiques comme la séparation entre Hollande et Belgique, la façon dont la religion dominante de telle ou telle région se trouve déterminée

et ces exemples lui donnent l'occasion de souligner la possibilité qu'existent en effet différents états d'équilibre, vers lesquels pourrait se porter l'histoire, et de poser la question des mécanismes sous-jacents. Il propose aussi plusieurs exemples industriels tels que les techniques de production d'énergie nucléaire, le type de courant, alternatif ou continu, utilisé dans les réseaux de distribution d'électricité ou les techniques retenues pour les moteurs automobiles ; autant de choix, dont il suggère qu'ils n'étaient pas les meilleurs même si ce jugement n'est possible que rétrospectivement.

2.3.3 *Path dependence*, *increasing returns* et *lock-in* : définitions

Si l'on suit Cecere *et al.* (2014), c'est l'étude de la compétition entre technologies rivales qui a d'abord conduit Arthur (1989) à élaborer le concept de dépendance au chemin pour expliquer comment des retours sur investissement qui augmentent au gré de l'adoption d'une technologie (*increasing returns*) conduisent à des situations de dominance (*lock-in*) qui découlent essentiellement d'éléments contingents, c'est à dire du hasard (*path dependence*).

Le chemin que suit l'histoire rencontre des points de bifurcation qui peuvent mener à des états distincts. La notion de *path dependence* insiste, d'une part, sur le fait que la direction retenue à ces carrefours peut n'être qu'un accident et, d'autre part, elle indique que, dans les situations où les retours sur investissements augmentent avec l'adoption de plus en plus massive de ce chemin, ces états sont rendus stables du fait que leur emprise sur le marché se renforce automatiquement. Les coûts qu'induirait l'adoption d'une alternative en deuxième intention sont en effet prohibitifs. Ce phénomène conduit l'économie à entériner des solutions techniques qui ne sont pas nécessairement les meilleures.

Pour souligner la nuance entre *lock-in* et *path dependence*, la *path dependence* met l'accent sur les événements contingents qui se trouvent au niveau des bifurcations dans les processus historiques alors que le *lock-in* désigne les maxima locaux qui caractérisent les situation auxquelles ces événements contingents ont mené, rendus stables par les *increasing returns*.

2.3.4 Des liens entre économie évolutive et biologie évolutive

Les idées de Darwin ont trouvé leur place en économie, elles y ont même fondé de nouvelles branches d'étude. On peut penser à l'importance qu'a la notion d'évolution dans la pensée de Friedrich Hayek ; on peut aussi constater qu'il existe des journaux scientifiques actifs et reconnus – tel le *Journal of Evolutionary Economics* – qui revendiquent leur lien avec la théorie darwinienne. Il semble cependant que cette ligne de pensée économique se soit détachée des efforts de la recherche en évolution pour n'en garder que les intuitions fondatrices. Il faut aussi noter que pendant que le courant de la *path dependence* relève d'une économie inspirée par la théorie de l'évolution, on trouve également un mouvement en sens inverse qui va de la recherche en évolution vers l'économie, ce que représentent les riches articles de G. J. Vermeij (Vermeij, 1995, 1999). Il semble cependant que ces deux courants ne se soient pas encore rencontrés : d'un côté, même si Vermeij cite Arthur, ce n'est qu'en passant ; de l'autre, la façon dont David et Arthur s'appuient sur l'évolution relève de la métaphore, du moins ne prennent-ils pas le soin d'expliciter, par exemple, comment des opérateurs de mutation agiraient dans leur cadre d'étude. De fait, les cartes suivantes (VI.1 et VI.2), extraites du Web of Science indiquent les croisements entre les

disciplines. Spécifiquement, la figure VI.1 recense les domaines (selon les catégories du *Web of Science*) dont des publications ont cité l'article séminal d'Arthur (1989). On y constate que seules 290 citations relèvent du champ des *Environmental Sciences [and] Ecology*. La figure VI.2 reprend ces résultats en se focalisant sur ces 290 articles, ce qui permet de voir que, décidément, quasiment aucune de ces publications ne se rattache ni à la biologie ni, à fortiori, à la biologie évolutive. Seuls trois de ces 290 articles relèvent du champs disciplinaire *Evolutionary Biology* mais deux d'entre eux sont ceux que nous avons déjà évoqués : (Vermeij, 1995), (Vermeij, 1999) tandis que le troisième est un travail théorique de Eörs Szathmáry (1991).

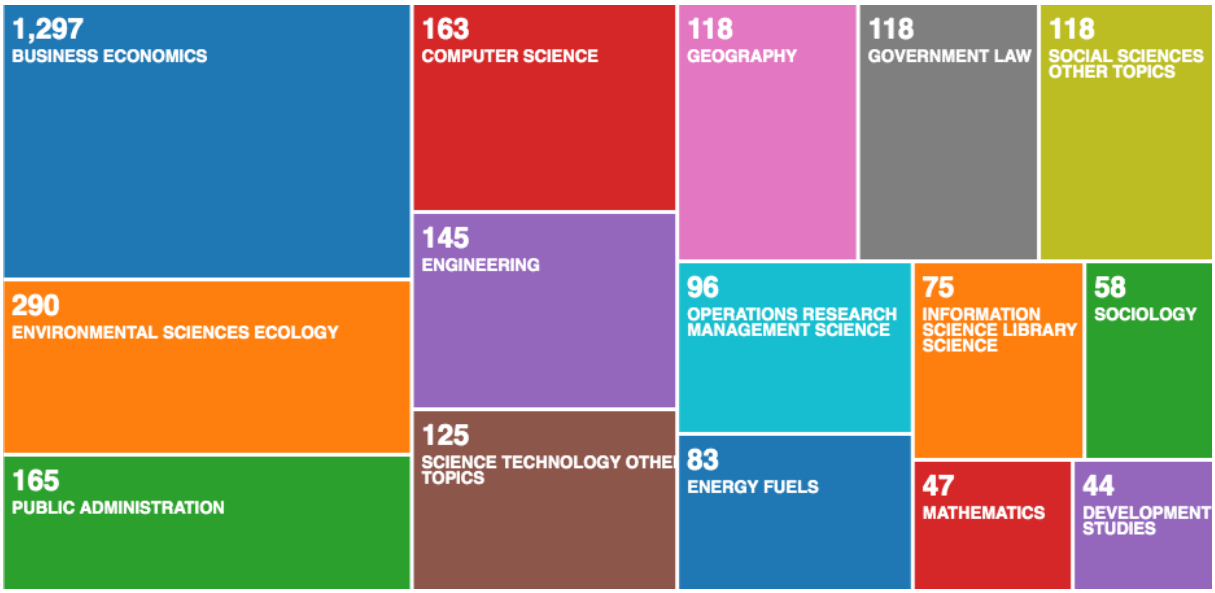


Figure VI.1 – Selon *Web of Science*, (Arthur, 1989) a été cité 2 608 fois (en date du 16 juillet 2020). Cette carte est une classification de ces 2 608 citations dans les domaines scientifiques répertoriés par *Web of Science*. On constate qu'à l'exception des 290 citations qui relèvent des champs *Environmental Sciences [and] Ecology*, aucune de ces citations ne relève de la biologie et, à fortiori, de la biologie évolutive.

2.3.5 Est-ce la même chose ?

Comme il a été décrit précédemment, la conceptualisation de la *path dependence* s'articule autour de trois idées : la contingence de l'élément qui détermine la bifurcation des événements, la stabilité qu'assurent à l'état qui est ainsi atteint, les *increasing returns* et la non optimalité de cet état, qui distingue foncièrement le résultat du processus de celui auquel aurait conduit un processus ergodique. Ainsi résumé, il est frappant de constater la gémellité de ces notions avec le cliquet de la complexité : contingence car rien ne semble déterminer de façon nécessaire ce qui conduit les organismes sur la voie de la simplicité ou de la complexité ; stabilité car nous avons vu comme il est difficile de dévier la trajectoire de ces organismes une fois qu'elle est engagée et seule une multiplication par cent des taux de mutation nous a permis d'observer une déviation significative ; non optimalité car, comme il a été souligné à plusieurs reprises, le chemin de la simplicité,

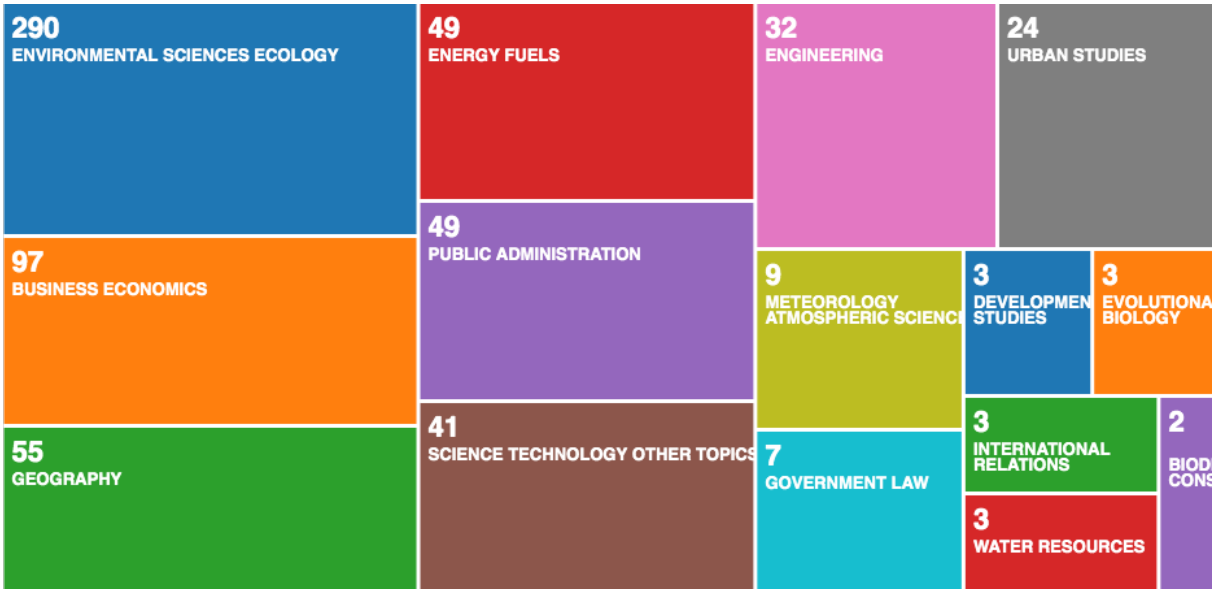
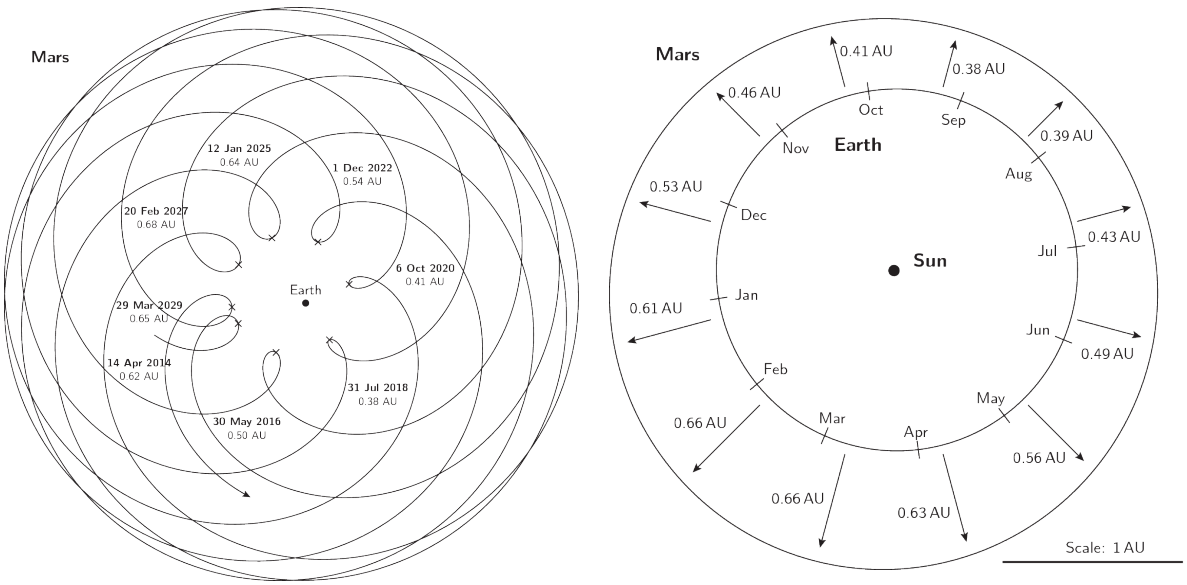


Figure VI.2 – Cette carte s'intéresse spécifiquement aux 290 citations de (Arthur, 1989) qui relèvent des champs *Environmental Sciences [and] Ecology* (cf. VI.1). Il y apparaît notamment que seuls trois articles se rattachent à l'*evolutionnary biology*. Ce sont (Szathmáry, 1991) et (Vermeij, 1995, 1999).

le plus souvent « raté » par les organismes, les conduirait pourtant à des fitness remarquablement supérieures à celui de la complexité. La convergence des idées et des cadres d'interprétation est donc patente. On notera d'ailleurs que le terme de *lock-in* comme celui de cliquet convoquent tous les deux des représentations mécaniques pour parler de processus irréversibles dont chaque nouvelle étape renforce graduellement les effets.

Enfin, de nombreux exemples montrent que les processus de *path dependence* pourraient être liés à la complexification. Nous avons déjà évoqué la construction de la loi. Un autre exemple pourrait être la complexification des théories scientifiques. Ainsi, on ne peut qu'être étonné devant la complexité à laquelle avait abouti le système géocentrique de Ptolémée pour rendre compte des gains de précision qu'apportait le progrès des instruments d'observation (figure VI.3a). La question de la nature de la « révolution », au sens de Khun, qui a permis de dépasser un système géocentrique « complexe » (on peut ici penser la complexité comme la longueur de la description mathématique du système) vers un système héliocentrique « simple » (figure VI.3b) serait d'ailleurs à comparer avec les contraintes de robustesse que nous avons mises en évidence. Le fait que l'étude des astres et de phénomènes tels que le mouvement rétrograde de mars, jadis réservé aux plus grands esprits de l'époque, soit devenu un des exercices classiques proposés aux élèves découvrant tout juste la mécanique, montre l'ampleur du bénéfice intellectuel que représentent pourtant de tels changements de paradigmes.



(a) Référentiel géocentrique (b) Référentiel héliocentrique

Figure VI.3 – Trajectoire de la planète mars dans les référentiels géocentrique (a) et héliocentrique (b). À gauche, dans le référentiel géocentrique, la terre est représentée par le point au centre de la figure et la courbe représente la trajectoire de la planète mars entre 2014 et 2029. À droite, le point central représente le soleil, le cercle intérieur représente l'orbite de la terre, le cercle extérieur l'orbite de le planète mars. C'est à Copernic (1473-1563) et Kepler (1571-1630) que l'on doit d'avoir dépassé le système ptolémaïque qui plaçait la terre au centre de l'univers et astreignait les représentations astronomiques à la complexité de la figure (a) alors que le système géocentrique permet de décomposer l'étude des trajectoires en termes de trajectoires circulaires ou elliptiques. Crédit : In-The-Sky.org

Un des objectifs historiques des « sciences de la complexité » est de rechercher des concepts ou des mécanismes unificateurs dans les différents domaines scientifiques ou applicatifs. Les rapprochements potentiels entre, d'un côté, les concepts d'économie ou de sciences sociales que sont la « path-dependence » et le « lock-in » et, de l'autre, le cliquet de la complexité que nous avons identifié au cours de cette thèse – par exemple via des formalisations théoriques telles que celles que nous avons évoquées précédemment – ouvrent sur ce plan des perspectives enthousiasmantes.

Bibliographie

Adami, C. (1994). Evolutionary learning in the 2d artificial life system" avida". *In Artificial Life IV*, pages 377–381. The MIT Press.

Adami, C. (1998). *Introduction to artificial life*. Springer Science & Business Media.

Adami, C. (2002a). Sequence complexity in darwinian evolution. *Complexity*, 8(2):49–56.

Adami, C. (2002b). What is complexity ? *BioEssays*, 24(12):1085–1094.

Adami, C. (2006). Digital genetics : unravelling the genetic basis of evolution. *Nat. Rev. Genet.*, 7(2):109–118.

Adami, C. et Brown, C. T. (1994). Evolutionary learning in the 2D artificial life systems avida. *In* Brooks, R. et Maes, P., éditeurs : *Artificial Life IV*, pages 377–381. MIT Press.

Adami, C., Ofria, C. et Collier, T. C. (2000). Evolution of biological complexity. *PNAS*, 97(9):4463–4468.

Albantakis, L., Hintze, A., Koch, C., Adami, C. et Tononi, G. (2014). Evolution of integrated causal structures in animats exposed to environments of increasing complexity. *PLoS Comput Biol*, 10(12):e1003966.

Arendt, D., Hausen, H. et Purschke, G. (2009). The 'division of labour'model of eye evolution. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 364(1531):2809–2817.

Arthur, W. B. (1989). Competing technologies, increasing returns, and lock-in by historical events. *The economic journal*, 99(394):116–131.

Auerbach, J. E. et Bongard, J. C. (2014). Environmental influence on the evolution of morphological complexity in machines. *PLoS computational biology*, 10(1).

Banzhaf, W., Baumgaertner, B., Beslon, G., Doursat, R., Foster, J. A., McMullin, B., De Melo, V. V., Miconi, T., Spector, L., Stepney, S. *et al.* (2016). Defining and simulating open-ended novelty : requirements, guidelines, and challenges. *Theory in Biosciences*, 135(3):131–161.

Bapteste, É. (2017). *Tous entrelacés ! : des gènes aux super-organismes : les réseaux de l'évolution*. Belin.

Barabási, A.-L. et Oltvai, Z. N. (2004). Network biology : understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113.

Batut, B., Knibbe, C., Marais, G. et Daubin, V. (2014). Reductive genome evolution at both ends of the bacterial population size spectrum. *Nature Reviews Microbiology*, 12(12):841.

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

Benkirane, R., Morin, E., Prigogine, I. et Varela, F. (2002). *La complexité, vertiges et promesses : 18 histoires de sciences*. Le pommier.

Bennett, C. H. (1995). Logical depth and physical complexity. *In The universal Turing machine (2nd ed.) a half-century survey*, pages 207–235. Springer.

Bertin, E., Gandrillon, O., Beslon, G., Grauwin, S., Jensen, P. et Schabanel, N. (2011). Les complexités : point de vue d'un institut des systèmes complexes. *Hermès, La Revue*, 60(2):145–150.

Beslon, G., Liard, V., Parsons, D. P. et Rouzaud-Cornabas, J. (2020). Of evolution, systems and complexity. *In* Crombach, A., éditeur : *Evolutionary Systems Biology, vol. 2*, page 17. Springer.

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

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

Biebricher, C. et Eigen, M. (2005). The error threshold. *Virus Research*, 107(2):117–127.

Biller, P., Guéguen, L., Knibbe, C. et Tannier, E. (2016a). Breaking good : accounting for fragility of genomic regions in rearrangement distance estimation. *Genome biology and evolution*, 8(5):1427–1439.

Biller, P., Knibbe, C., Beslon, G. et Tannier, E. (2016b). Comparative genomics on artificial life. *In Conference on Computability in Europe*, pages 35–44. Springer.

Bongard, J. (2010). The utility of evolving simulated robot morphology increases with task complexity for object manipulation. *Artificial Life*, 16(3):201–223.

Bongard, J. et Paul, C. (2001). Making evolution an offer it can't refuse : Morphology and the extradimensional bypass advances in artificial life. *In* Kelemen, J. et Sosík, P., éditeurs : *Advances in Artificial Life*, volume 2159 de *Lecture Notes in Computer Science*, chapitre 43, pages 401–412. Springer Berlin / Heidelberg, Berlin, Heidelberg.

Bonnici, V. et Manca, V. (2016). Informational laws of genome structures. *Scientific reports*, 6:28840.

Brunet, T. et Doolittle, W. F. (2018). The generality of constructive neutral evolution. *Biology & Philosophy*, 33(1-2):2.

Cairns-Smith, A. (1995). The complexity ratchet. *In Progress in the Search for Extraterrestrial Life.*, volume 74, page 31.

Carde, Q., Foley, M., Knibbe, C., Parsons, D. P., Rouzaud-Cornabas, J. et Beslon, G. (2019). How to reduce a genome ? alife as a tool to teach the scientific method to school pupils. *In The 2018 Conference on Artificial Life : A Hybrid of the European Conference on Artificial Life (ECAL) and the International Conference on the Synthesis and Simulation of Living Systems (ALIFE)*, pages 497–504. MIT Press.

Cecere, G., Corrocher, N., Gossart, C. et Ozman, M. (2014). Lock-in and path dependence : an evolutionary approach to eco-innovations. *Journal of Evolutionary Economics*, 24(5):1037–1065.

Chow, S. S., Wilke, C. O., Ofria, C., Lenski, R. E. et Adami, C. (2004). Adaptive radiation from resource competition in digital organisms. *Science (New York, N.Y.)*, 305(5680):84–86.

Clune, J., Mouret, J.-B. et Lipson, H. (2013). The evolutionary origins of modularity. *Proc. Roy. Soc. B*, 280(1755):20122863.

Crick, F. H. (1958). On protein synthesis. *In Symp Soc Exp Biol*, numéro 138-63 de n/a, page 8.

Crick, F. H. (1968). The origin of the genetic code. *Journal of molecular biology*, 38(3):367–379.

Crombach, A. et Hogeweg, P. (2007). Chromosome rearrangements and the evolution of genome structuring and adaptability. *Molecular Biology and Evolution*, 24(5):1130–1139.

Crombach, A. et Hogeweg, P. (2008). Evolution of evolvability in gene regulatory networks. *PLoS Comp Biol*, 4(7).

Crombach, A. et Hogeweg, P. (2009). Evolution of resource cycling in ecosystems and individuals. *BMC Evolutionary Biology*, 9(1):122+.

Darwin, C. (1859). *On the Origin of Species by means of Natural Selection or the preservation of Favoured Races in the Struggle for Life*. Murray, John.

David, P. A. (2007). Path dependence : a foundational concept for historical social science. *Cliometrica*, 1(2):91–114.

Dawkins, R. (1997). Human chauvinism. *Evolution*, 51(3):1015–1020.

Dawkins, R. et Krebs, J. R. (1979). Arms races between and within species. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 205(1161):489–511.

de Back, W., Wiering, M. et de Jong, E. (2006). Red queen dynamics in a predator-prey ecosystem. *In Proceedings of the 8th annual conference on Genetic and evolutionary computation*, pages 381–382.

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

Delahaye, J.-P. *et al.* (1994). *Information, complexité et hasard*. Hermes Paris.

Desbenoit, B., Galin, E. et Akkouche, S. (2005). Modeling cracks and fractures. *The Visual Computer*, 21(8-10):717–726.

Devert, A., Bredeche, N. et Schoenauer, M. (2006). Blindbuilder : A new encoding to evolve Lego-Like structures genetic programming. *In* Collet, P., Tomassini, M., Ebner, M., Gustafson, S. et Ekárt, A., éditeurs : *Genetic Programming*, volume 3905 de *Lecture Notes in Computer Science*, chapitre 6, pages 61–72. Springer Berlin / Heidelberg, Berlin, Heidelberg.

Dittrich, P., Ziegler, J. et Banzhaf, W. (2001). Artificial chemistries-a review. *Artif Life*, 7(3):225–275.

Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. *The american biology teacher*, 35(3):125–129.

Doolittle, W. F. (2012). A ratchet for protein complexity. *Nature*, 481(7381):270–271.

Doolittle, W. F., Lukeš, J., Archibald, J. M., Keeling, P. J. et Gray, M. W. (2011). Comment on "does constructive neutral evolution play an important role in the origin of cellular complexity ?" doi 10.1002/bies. 201100010. *Bioessays*, 33(6):427–429.

Drake, J. W. (1991). A constant rate of spontaneous mutation in dna-based microbes. *Proc Natl Acad Sci USA*, 88(16):7160–7164.

Edlund, J. A., Chaumont, N., Hintze, A., Koch, C., Tononi, G. et Adami, C. (2011). Integrated information increases with fitness in the evolution of animats. *PLoS Comput Biol*, 7(10):e1002236.

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

Eigen, M. et Schuster, P. (1977). A principle of natural self-organization. *Naturwissenschaften*, 64(11):541–565.

Elena, S. et Sanjuan, R. (2008). The effect of genetic robustness on evolvability in digital organisms. *BMC Evolutionary Biology*, 8(1):284+.

Elena, S. F., Wilke, C. O., Ofria, C. et Lenski, R. E. (2007). Effects of population size and mutation rate on the evolution of mutational robustness. *Evolution*, 61(3):666–674.

Elliott, T. A. et Gregory, T. R. (2015). What's in a genome ? the c-value enigma and the evolution of eukaryotic genome content. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 370(1678):20140331.

Espinosa-Soto, C. et Wagner, A. (2010). Specialization can drive the evolution of modularity. *PLoS Comp. Biol.*, 6(3).

Félix, M.-A. et Wagner, A. (2008). Robustness and evolution : concepts, insights and challenges from a developmental model system. *Heredity*, 100(2):132–140.

Feyerabend, P. (1979). Contre la méthode : l'anarchisme méthodologique. *Ed. du Seuil, Paris*.

Finlay, B. J. et Esteban, G. F. (2009). Can biological complexity be rationalized ? *BioScience*, 59(4):333–340.

Fischer, S., Bernard, S., Beslon, G. et Knibbe, C. (2014). A model for genome size evolution. *Bull. Math. Biol.*, 76(9):2249–2291.

Gago, S., Elena, S. F., Flores, R. et Sanjuan, R. (2009). Extremely high mutation rate of a hammerhead viroid. *Science*, 323(5919):1308.

Gilbert, W. (1986). Origin of life : The rna world. *nature*, 319(6055):618–618.

Giovannoni, S. J., Tripp, H. J., Givan, S., Podar, M., Vergin, K. L., Baptista, D., Bibbs, L., Eads, J., Richardson, T. H., Noordewier, M. *et al.* (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *science*, 309(5738):1242–1245.

Gould, S. J. (1996). *Full House : The Spread of Joy from Plato to Darwin*. Harmony Books.

Gray, M. W., Lukeš, J., Archibald, J. M., Keeling, P. J. et Doolittle, W. F. (2010). Irremediable complexity ? *Science*, 330(6006):920–921.

Hahn, M. W., Wray, G. A. *et al.* (2002). The g-value paradox. *Evolution and Development*, 4(2):73–75.

Hansen, T. F. (2003). Is modularity necessary for evolvability ? : Remarks on the relationship between pleiotropy and evolvability. *Biosystems*, 69(2-3):83–94.

Heylighen, F. (1999). The growth of structural and functional complexity during evolution. *The evolution of complexity*, pages 17–44.

Heylighen, F. (2007). Five questions on complexity. *arXiv preprint nlin/0702016*.

Higgins, N. (2005). *The bacterial chromosome*. ASM Press.

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

Hintze, A. et Adami, C. (2008). Evolution of complex modular biological networks. *PLoS computational biology*, 4(2).

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K. *et al.* (2016). A new view of the tree of life. *Nature microbiology*, 1(5):16048.

Ideker, T., Galitski, T. et Hood, L. (2001). A new approach to decoding life : systems biology. *Annual review of genomics and human genetics*, 2(1):343–372.

Kashtan, N. et Alon, U. (2005). Spontaneous evolution of modularity and network motifs. *Proc Natl Acad Sci USA*, 102(39):13773–13778.

Kashtan, N., Parter, M., Dekel, E., Mayo, A. E. et Alon, U. (2009). Extinctions in heterogeneous environments and the evolution of modularity. *Evolution*, 63(8):1964–1975.

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

Kitano, H. (2002). Systems biology : a brief overview. *science*, 295(5560):1662–1664.

Knibbe, C. (2006). *Structuration des génomes par sélection indirecte de la variabilité mutationnelle, une approche de modélisation et de simulation.* Thèse de doctorat, INSA-Lyon.

Knibbe, C., Coulon, A., Mazet, O., Fayard, J.-M. et Beslon, G. (2007a). A long-term evolutionary pressure on the amount of noncoding DNA. *Mol. Biol. Evol.*, 24(10):2344–2353.

Knibbe, C., Fayard, J.-M. et Beslon, G. (2008). The topology of the protein network influences the dynamics of gene order : From systems biology to a systemic understanding of evolution. *Artificial life*, 14(1):149–156.

Knibbe, C., Mazet, O., Chaudier, F., Fayard, J.-M. et Beslon, G. (2007b). Evolutionary coupling between the deleteriousness of gene mutations and the amount of non-coding sequences. *J. Theor. Biol.*, 244(4):621–630.

Knibbe, C., Parsons, D. P. et Beslon, G. (2011). Parsimonious modeling of scaling laws in genomes and transcriptomes. *In Proceedings of the Eleventh European Conference on the Synthesis and Simulation of Living Systems (ECAL 11)*, pages 414–415. MIT Press.

Kolmogorov, A. N. (1963). On tables of random numbers. *Sankhyā : The Indian Journal of Statistics, Series A*, pages 369–376.

Komosiński, M. et Ulatowski, S. (1999). Framsticks : Towards a simulation of a Nature-Like world, creatures and evolution advances in artificial life. *In* Floreano, D., Nicoud, J.-D. et Mondada, F., éditeurs : *Advances in Artificial Life*, volume 1674 de *Lecture Notes in Computer Science*, chapitre 33, pages 261–265. Springer Berlin / Heidelberg.

Krakauer, D. C. (2011). Darwinian demons, evolutionary complexity, and information maximization. *Chaos : An Interdisciplinary Journal of Nonlinear Science*, 21(3):037110.

Lenski, R. E., Barrick, J. E. et Ofria, C. (2006). Balancing robustness and evolvability. *PLoS biology*, 4(12).

Lenski, R. E., Ofria, C., Collier, T. C. et Adami, C. (1999). Genome complexity, robustness and genetic interactions in digital organisms. *Nature*, 400(6745):661–664.

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

Lewin, B. (2007). *Genes IX*. Jones and Bartlett.

Liard, V., Parsons, D., Rouzaud-Cornabas, J. et Beslon, G. (2018). The complexity ratchet : Stronger than selection, weaker than robustness. *In Artificial Life Conference Proceedings*, pages 250–257. MIT Press.

Liard, V., Parsons, D. P., Rouzaud-Cornabas, J. et Beslon, G. (2020). The complexity ratchet : Stronger than selection, stronger than evolvability, weaker than robustness. *Artificial Life*, 26(1):38–57.

Lukeš, J., Archibald, J. M., Keeling, P. J., Doolittle, W. F. et Gray, M. W. (2011). How a neutral evolutionary ratchet can build cellular complexity. *IUBMB life*, 63(7):528–537.

Lynch, M. et Conery, J. S. (2003). The origins of genome complexity. *Science*, 302(5649):1401–1404.

Lynch, M. et Walsh, B. (2007). *The origins of genome architecture*, volume 98. Sinauer Associates Sunderland, MA.

Maslov, S., Krishna, S., Pang, T. Y. et Sneppen, K. (2009). Toolbox model of evolution of prokaryotic metabolic networks and their regulation. *Proceedings of the National Academy of Sciences*, 106(24):9743–9748.

Mattick, J. S., Taft, R. J. et Faulkner, G. J. (2010). A global view of genomic information–moving beyond the gene and the master regulator. *Trends in genetics*, 26(1):21–28.

Mattingly, G. (1987). *The Armada*, volume 17. Houghton Mifflin Harcourt.

Maynard-Smith, J. (1970). Time in the evolutionary process. *Studium generale ; Zeitschrift fur die Einheit der Wissenschaften im Zusammenhang ihrer Begriffsbildungen und Forschungsmethoden*, 23(3):266–272.

Maynard-Smith, J. et Szathmary, E. (1997). *The major transitions in evolution*. Oxford University Press.

Mayr, E. (1961). Cause and effect in biology : Kinds of causes, predictability, and teleology are viewed by a practicing biologist. *Science*, 134(3489):1501–1506.

McShea, D. W. (1993). Evolutionary change in the morphological complexity of the mammalian vertebral column. *Evolution*, 47(3):730–740.

McShea, D. W. (1996). Metazoan complexity and evolution : Is there a trend ? *Evolution*, 50(2):477–492.

McShea, D. W. (2017). Evolution of complexity. *Evolutionary Developmental Biology : A Reference Guide*, pages 1–11.

McShea, D. W. et Brandon, R. N. (2010). *Biology's first law : the tendency for diversity and complexity to increase in evolutionary systems*. University of Chicago Press.

McShea, D. W., Wang, S. C. et Brandon, R. N. (2019). A quantitative formulation of biology's first law. *Evolution*, 73(6):1101–1115.

Metropolis, N. et Ulam, S. (1949). The monte carlo method. *Journal of the American statistical association*, 44(247):335–341.

Miconi, T. (2008). Evolution and complexity : The double-edged sword. *Artificial life*, 14(3):325–344.

Mileti, D. S., Gillespie, D. F. et Haas, J. E. (1977). Size and structure in complex organizations. *Social Forces*, 56(1):208–217.

Misevic, D., Ofria, C. et Lenski, R. E. (2006). Sexual reproduction reshapes the genetic architecture of digital organisms. *Proc. R. Soc. B.*, 273(1585):457–464.

Mitchell, M. (2009). *Complexity : A guided tour*. Oxford University Press.

Molina, N. et van Nimwegen, E. (2008). The evolution of domain-content in bacterial genomes. *Biology Direct*, 3:51.

Molina, N. et van Nimwegen, E. (2009). Scaling laws in functional genome content across prokaryotic clades and lifestyles. *Trends in genetics*, 25(6):243–247.

Moran, N. A. (2007). Symbiosis as an adaptive process and source of phenotypic complexity. *Proceedings of the National Academy of Sciences*, 104(suppl 1):8627–8633.

Moran, N. A. et Mira, A. (2001). The process of genome shrinkage in the obligate symbiont buchnera aphidicola. *Genome biology*, 2(12):research0054–1.

Moya, A., Oliver, J. L., Verdú, M., Delaye, L., Arnau, V., Bernaola-Galván, P., de la Fuente, R., Díaz, W., Gómez-Martín, C., González, F. M. *et al.* (2020). Tracking evolutionary trends towards increasing complexity : a case study in cyanobacteria. *bioRxiv*.

Musso, F. et Feverati, G. (2012). Mutation–selection dynamics and error threshold in an evolutionary model for turing machines. *Biosystems*, 107(1):18 – 33.

Nghe, P., Kogenaru, M. et Tans, S. J. (2018). Sign epistasis caused by hierarchy within signalling cascades. *Nat. Comm.*, 9(1):1–9.

Nilsson, D.-E. (2013). Eye evolution and its functional basis. *Visual neuroscience*, 30(1-2):5–20.

Nuñez, P. A., Romero, H., Farber, M. D. et Rocha, E. P. (2013). Natural selection for operons depends on genome size. *Genome biology and evolution*, 5(11):2242–2254.

Ofria, C. et Wilke, C. O. (2004). Avida : A software platform for research in computational evolutionary biology. *Artificial life*, 10(2):191–229.

Ohno, S. (2013). *Evolution by gene duplication*. Springer Science & Business Media.

O'Neill, B. (2003). Digital evolution. *PLoS Biol*, 1(1):e18+.

Orr, H. A. (2000). Adaptation and the cost of complexity. *Evolution*, 54(1):13–20.

Packard, N., Bedau, M. A., Channon, A., Ikegami, T., Rasmussen, S., Stanley, K. et Taylor, T. (2019a). Open-ended evolution and open-endedness : Editorial introduction to the open-ended evolution i special issue.

Packard, N., Bedau, M. A., Channon, A., Ikegami, T., Rasmussen, S., Stanley, K. O. et Taylor, T. (2019b). An overview of open-ended evolution : Editorial introduction to the open-ended evolution ii special issue. *Artificial life*, 25(2):93–103.

Pajot, P. (2018). La naissance d'une théorie au carrefour des disciplines. *la Recherche*, n/a(537 (juillet-août 2018)):453–470.

Pallen, M. et Gophna, U. (2007). Bacterial flagella and type iii secretion : case studies in the evolution of complexity. *In Gene and Protein Evolution*, volume 3, pages 30–47. Karger Publishers.

Pallen, M. J. et Matzke, N. J. (2006). From the origin of species to the origin of bacterial flagella. *Nature Reviews Microbiology*, 4(10):784–790.

Parsons, D. (2011). *Sélection indirecte en évolution darwinienne : mécanismes et implications*. Thèse de doctorat, Lyon, INSA.

Parsons, D. P., Knibbe, C. et Beslon, G. (2010). Importance of the rearrangement rates on the organization of transcription. *In Proceedings of Artificial Life XII*, pages 479–486.

Parsons, D. P., Knibbe, C. et Beslon, G. (2011). Homologous and nonhomologous rearrangements : Interactions and effects on evolvability. *In ECAL*, pages 622–629.

Pavlicev, M. et Wagner, G. P. (2012). A model of developmental evolution : selection, pleiotropy and compensation. *Trends in Ecology & Evolution*, 27(6):316–322.

Peignier, S., Rigotti, C. et Beslon, G. (2015). Subspace clustering using evolvable genome structure. *In Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*, pages 575–582.

Peignier, S., Rigotti, C. et Beslon, G. (2020). Evolutionary subspace clustering using variable genome length. *Computational Intelligence*, 36(2):574–612.

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

Phillips, P. (2008). Epistasis – the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11):855–867.

Pigliucci, M. (2008). Is evolvability evolvable ? *Nature Reviews Genetics*, 9(1):75.

Pollack, J. B. et Lipson, H. (2000). The GOLEM project : Evolving hardware bodies and brains. *Evolvable Hardware, NASA/DoD Conference on*, 0:37+.

Raeside, C., Gaffé, J., Deatherage, D. E., Tenaillon, O., Briska, A. M., Ptashkin, R. N., Cruveiller, S., Médigue, C., Lenski, R. E., Barrick, J. E. *et al.* (2014). Large chromosomal rearrangements during a long-term evolution experiment with escherichia coli. *MBio*, 5(5):e01377–14.

Ray, T. S. (1991). An approach to the synthesis of life. *Artificial Life II, Santa Fe Institute Studies in the Sciences of Complexity, vol. XI*.

Ray, T. S. (1992). Evolution, ecology and optimization of digital organisms. *Santa Fe Institute working paper 92-08-04*.

Rocabert, C., Knibbe, C., Consuegra, J., Schneider, D. et Beslon, G. (2017). Beware batch culture : Seasonality and niche construction predicted to favor bacterial adaptive diversification. *PLoS computational biology*, 13(3):e1005459.

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

Sanchez-Dehesa, Y. (2009). *RÆvol : un modèle de génétique digitale pour étudier l'évolution des réseaux de régulation génétiques*. Thèse de doctorat, INSA-Lyon.

Sanjuán, R. et Elena, S. F. (2006). Epistasis correlates to genomic complexity. *PNAS*, 103(39):14402–14405.

Saunders, P. T. et Ho, M.-W. (1976). On the increase in complexity in evolution. *Journal of Theoretical Biology*, 63(2):375–384.

Saunders, W., Work, D. et Nikolaeva, S. (1999). Evolution of complexity in paleozoic ammonoid sutures. *Science*, 286(5440):760–763.

Schopf, T. J., Raup, D. M., Gould, S. J. et Simberloff, D. S. (1975). Genomic versus morphologic rates of evolution : influence of morphologic complexity. *Paleobiology*, 1(1):63–70.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Shen, A. (2000). Algorithmic information theory and kolmogorov complexity. *Lecture notes of an introductory course. Uppsala University Technical Report*, 34:2000–034.

Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, 106(6):467–482.

Sims, K. (1994a). Evolving 3d morphology and behavior by competition. *Artif. Life*, 1:353–372.

Sims, K. (1994b). Evolving virtual creatures. *In Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, SIGGRAPH '94, pages 15–22. ACM.

Snyder, L. A., Loman, N. J., Fütterer, K. et Pallen, M. J. (2009). Bacterial flagellar diversity and evolution : seek simplicity and distrust it ? *Trends in microbiology*, 17(1):1–5.

Soyer, O. S. et Bonhoeffer, S. (2006). Evolution of complexity in signaling pathways. *PNAS*, 103(44):16337–16342.

Soyer, O. S. et O'Malley, M. A. (2013). Evolutionary systems biology : what it is and why it matters. *BioEssays*, 35(8):696–705.

Spielberg, S., Kennedy, K. et Molen, G. R. (1993). Jurassic park. *Universal Pictures*.

Stanley, K. O. et Miikkulainen, R. (2002). Efficient reinforcement learning through evolving neural network topologies. *In Proceedings of the 4th Annual Conference on Genetic and Evolutionary Computation*, pages 569–577.

Stoltzfus, A. (1999). On the possibility of constructive neutral evolution. *Journal of Molecular Evolution*, 49(2):169–181.

Szathmáry, E. (1991). Simple growth laws and selection consequences. *Trends in Ecology & Evolution*, 6(11):366–370.

Szathmáry, E. (2015). Toward major evolutionary transitions theory 2.0. *Proceedings of the National Academy of Sciences*, 112(33):10104–10111.

Taylor, T., Bedau, M., Channon, A., Ackley, D., Banzhaf, W., Beslon, G., Dolson, E., Froese, T., Hickinbotham, S., Ikegami, T. *et al.* (2016). Open-ended evolution : Perspectives from the oee workshop in york. *Artificial life*, 22(3):408–423.

Ten Tusscher, K. H. et Hogeweg, P. (2009). The role of genome and gene regulatory network canalization in the evolution of multi-trait polymorphisms and sympatric speciation. *BMC Evolutionary Biology*, 9(1):159+.

Ten Tusscher, K. H. et Hogeweg, P. (2011). Evolution of networks for body plan patterning ; interplay of modularity, robustness and evolvability. *PLoS Computational Biology*, 7(10).

Thomas, C. A. J. (1971). The genetic organization of chromosomes. *Annual review of genetics*, 5(1):237–256.

Vadée-Le-Brun, Y., Rouzaud-Cornabas, J. et Beslon, G. (2016). In silico experimental evolution suggests a complex intertwining of selection, robustness and drift in the evolution of genetic networks complexity. *In ALife XIV*, pages 180–188.

Valentine, J. W. (2000). Two genomic paths to the evolution of complexity in bodyplans. *Paleobiology*, 26(3):513–519.

Van Nimwegen, E., Crutchfield, J. P. et Huynen, M. (1999). Neutral evolution of mutational robustness. *Proceedings of the National Academy of Sciences*, 96(17):9716–9720.

Vermeij, G. J. (1995). Economics, volcanoes, and phanerozoic revolutions. *Paleobiology*, pages 125–152.

Vermeij, G. J. (1999). Inequality and the directionality of history. *The American Naturalist*, 153(3):243–253.

Von Bertalanffy, L., Chabrol, J.-B., László, E. et Paulre, B. (1973). *Théorie générale des systèmes*. Dunod Paris.

Waddington, C. H. (1969). Paradigm for an evolutionary process. *Towards a theoretical biology*, 2:106–128.

Wagner, A. (2005). Distributed robustness versus redundancy as causes of mutational robustness. *Bioessays*, 27(2):176–188.

Wagner, A. (2007). *Robustness and evolvability in living systems*. Princeton university press.

Wagner, A. (2008). Robustness and evolvability : a paradox resolved. *Proceedings of the Royal Society B : Biological Sciences*, 275(1630):91–100.

Wagner, G. P. et Zhang, J. (2011). The pleiotropic structure of the genotype–phenotype map : the evolvability of complex organisms. *Nature Reviews Genetics*, 12(3):204–213.

Wagner, P. J. (1996). Contrasting the underlying patterns of active trends in morphologic evolution. *Evolution*, 50(3):990–1007.

Wang, Z., Liao, B.-Y. et Zhang, J. (2010). Genomic patterns of pleiotropy and the evolution of complexity. *Proceedings of the National Academy of Sciences*, 107(42): 18034–18039.

Weaver, W. (1948). Science and complexity. *American Scientist*, 36(4):536–544.

Weinreich, D. M., Watson, R. A. et Chao, L. (2005). Perspective : sign epistasis and genetic costraint on evolutionary trajectories. *Evolution*, 59(6):1165–1174.

Whitesides, G. M. et Ismagilov, R. F. (1999). Complexity in chemistry. *science*, 284(5411):89–92.

Wiener, N. (1948). *Cybernetics or control and communication in the animal and the machine*. Technology Press.

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

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

Wilson, A. C. et Duncan, R. P. (2015). Signatures of host/symbiont genome coevolution in insect nutritional endosymbioses. *Proceedings of the National Academy of Sciences*, 112(33):10255–10261.

Woods, R. J., Barrick, J. E., Cooper, T. F., Shrestha, U., Kauth, M. R. et Lenski, R. E. (2011). Second-order selection for evolvability in a large escherichia coli population. *Science*, 331(6023):1433–1436.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. *In Proceedings of the sixth international congress of Genetics*, volume 1, pages 356–366.

Yaeger, L. (1994). Computational genetics, physiology, metabolism, neural systems, learning, vision, and behavior or poly world : Life in a new context. *In Santa Fe Institute Studies in the Sciences of Complexity*, volume 17, pages 263–263. Addison-Wesley Publishing Co.

Yaeger, L., Griffith, V. et Sporns, O. (2008). Passive and driven trends in the evolution of complexity. *In ALife XI*, pages 725–732.

Yamada, H. (1980). *A historical study of typewriters and typing methods, from the position of planning Japanese parallels*. Journal of Information Processing.

Yasuoka, K. et Yasuoka, M. (2011). On the prehistory of qwerty. *Zinbun*, n/a(42):161–174.

Zaman, L., Meyer, J. R., Devangam, S., Bryson, D. M., Lenski, R. E. et Ofria, C. (2014). Coevolution drives the emergence of complex traits and promotes evolvability. *PLoS Biol*, 12(12):e1002023.

FOLIO ADMINISTRATIF

THÈSE DE L'UNIVERSITÉ DE LYON OPÉRÉE AU SEIN DE L'INSA LYON

NOM : Liard

DATE de SOUTENANCE : 19/10/2010

Prénoms : Vincent, Marie

TITRE : M.

NATURE : Doctorat

Numéro d'ordre : 2020LYSEI085

Ecole doctorale : Infomaths (ED 512)

Spécialité : Informatique et applications

RÉSUMÉ :
L'origine évolutive de la complexité des systèmes biologiques interroge les sciences du vivant depuis de nombreuses années. Dans cette thèse nous avons utilisé la plateforme d'évolution expérimentale in silico « Aevol » pour tester l'existence d'un « cliquet de la complexité », c'est à dire d'un processus historique qui fait augmenter la complexité même dans des conditions où celle-ci n'est pas requise. Pour ce faire nous avons fait évoluer des populations d'organismes numériques dans des conditions environnementales telles que des organismes simples puissent se reproduire et prospérer. Malgré cela nous observons que, dans une large majorité des simulations, la complexité des organismes augmente continûment. L'étude à posteriori des simulations montre pourtant que ces organismes complexes sont beaucoup moins adaptés que les organismes simples et qu'ils ne présentent aucun avantage de robustesse ou d'évolvabilité. Ceci exclut la sélection de l'ensemble des explications possibles pour l'évolution de la complexité. Par ailleurs, des expériences complémentaires ont montré que la sélection est néanmoins nécessaire pour que la complexité évolue, ce qui exclut également les effets non sélectifs. En analysant le devenir à long terme des organismes complexes, nous avons enfin montré que ces organismes ne reviennent presque jamais à la simplicité malgré le bénéfice potentiel que cela représenterait en termes de fitness. Cela suggère l'existence d'un cliquet de complexité alimenté par une épistasie négative : les mutations qui conduiraient à des solutions simples, favorables en début de simulation, deviennent délétères après la fixation d'autres mutations. Nos résultats suggèrent également que ce cliquet de la complexité serait plus puissant que la sélection, mais qu'il peut être inversé par la robustesse en raison des contraintes qu'elle impose sur la capacité de codage du génome.

MOTS-CLÉS :
Évolution, complexité, épistasie, évolvabilité, évolution expérimentale in silico, robustesse

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

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

Président de jury :
Xxx

Composition du jury :			
Schneider, Dominique	Professeur des Universités	UGA	Rapporteur
Muller, Jean-Pierre	Directeur de Recherche	CIRAD	Rapporteur
Dillmann, Christine	Professeure des Universités	Université Paris Saclay	Examinatrice
Lopez, Philippe	Professeur des Universités	UPMC	Examinateur
Beslon, Guillaume	Professeur des Universités	INSA-LYON	Directeur de thèse
Rouzaud-Cornabas, Jonathan	Maître de conférences	INSA-LYON	Co-encadrant