Robustesse à la recombinaison dans un modèle de réseaux de gènes

Vincent Viguie

stage de 2ème année de master sous la direction du Pr. Olivier Martin

25 mars - 4 août 2006

Laboratoire de Physique Théorique et Modèles Statistiques
université Paris XI Orsay

Résumé

Au cours de ce stage, nous avons étudié en quoi la reproduction sexuée peut procurer un avantage sélectif par rapport à la reproduction asexuée, dans le cadre d'un certain modèle de réseau génétique. Nous avons ainsi déterminé numériquement que, dans ce modèle, et en présence de sélection naturelle, une population utilisant la reproduction sexuée devient plus résistante qu'une population identique mais n'évoluant qu'au gré de mutations ponctuelles. Nous avons relié ce phénomène à la corrélation existant entre robustesse aux mutations et probabilité que le descendant lors d'une reproduction sexuée soit viable. Nous exposons enfin une modélisation analytique de l'évolution d'une population sexuée.

mots-clef réseau de gène, évolution, reproduction sexuée

Table des matières

1 Evolution, gènes et réseaux de gènes **3**
1.1 L'évolution 3
1.2 Variabilité entre les individus et évolution des populations 4
1.3 Les réseaux de gènes 5

2 Modélisation des réseaux génétiques **6**
2.1 Modélisation des réseaux de gènes 6
2.2 Le modèle de A.Wagner 7
2.3 Comment modeliser la selection naturelle et quelles observables choisir pour suivre l'évolution d'une population ? 8
2.4 Le theoreme de E. Van Nimwegen, J.P. Crutchfield et M. Huynen 9

3 Population en évolution **10**
3.1 Généralités 10
3.2 Principe du programme 10
3.3 Effets de la recombinaison 11
3.4 A quoi est du l'accroissement de la robustesse ? 11

4 Propriétés intrinsèques du réseau des génotypes viables **12**
4.1 Principe du programme 13
4.2 Robustesse a la recombinaison 14
4.3 Etude suivant le nombre de lignes échangées 15
4.4 Corrélation entre la robustesse aux mutations et la robustesse à la recombinaison 15

5 Approximation de champ moyen **16**
5.1 Motivations 16
5.2 Dérivation 17
5.3 Commentaires sur la formule 19
5.4 Comparaison avec les simulations 19

A Augmentation de la robustesse d'une population soumise à des mutations ponctuelles seules **22**

B passer de la robustesse à la recombinaison en population à la robustesse à la recombinaison sur l'ensemble des génotypes viables **22**
B.1 Calcul de la probabilité que le descendant de 2 individus donnés soit viable . 22
B.2 cas d'une population 23

Introduction

De par ses implications sur l'origine de l'humanité notamment, la théorie de l'évolution est une des théories scientifiques ayant eu le plus d'impact sur les questions philosophiques et religieuses. Il n'est donc pas étonnant que cette théorie suscite encore l'engouement de nombreux scientifiques, et que, 150 ans après sa première formulation, elle soit toujours l'objet d'actives recherches. Sous sa forme actuelle, elle laisse de nombreuses questions sans réponses. En particulier, la question qui a motivé l'étude que j'ai réalisée au cours de mon stage est la question de l'origine du sexe : comment un tel procédé est-il apparu ? Dans quel cas présente-t-il un avantage par rapport à la simple reproduction asexuée ?

Pour étudier cette question, il faut utiliser un modèle faisant le lien entre les gènes et les caractéristiques de l'individu : nous avons choisi d'utiliser un modèle de réseau génétique. En effet, dans un organisme, les gènes ne s'expriment pas indépendamment les uns des autres. Ils s'activent ou s'inhibent suivant un réseau compliqué d'interactions. Nous avons, au cours de ce stage, étudié un modèle de réseau d'interactions, et essayé de voir en quoi cette structure en réseau peut influer sur la manière dont les espèces évoluent, et si elle peut aider à mieux comprendre l'apparition de la reproduction sexuée.

Dans une première partie, je vais exposer quelques idées générales sur la théorie de l'évolution et sur les réseaux de régulation des gènes. Je vais ensuite étudier comment modéliser de tels réseaux, et comment ces modèles peuvent permettre de répondre aux questions sur l'évolution. Dans un troisième temps, je vais étudier la manière dont l'évolution avec reproduction sexuée ou non, affecte une population de réseaux, avant dans une quatrième partie de présenter quelques résultats sur des propriétés statistiques de ces réseaux. Enfin, je vais terminer par une modélisation analytique de l'évolution d'une population sexuée.

1 Evolution, gènes et réseaux de gènes

1.1 L'évolution

Les théories de l'évolution En biologie, les théories de l'évolution cherchent à décrire le processus par lequel les espèces se modifient au cours du temps, et donnent naissance à de nouvelles espèces. Le fondateur de la théorie moderne de l'évolution est Charles Darwin, qui émit l'hypothèse de la sélection du plus apte (ou sélection naturelle) parmi des individus naturellement variants, et exposa cette théorie en 1859 dans son livre « l'Origine des espèces ». Pendant près d'un demi-siècle, les biologistes, mais aussi les paléontologues, s'affrontèrent sur la validité puis sur le fonctionnement de l'évolution. Depuis le milieu du XXe siècle, avec la Théorie synthétique de l'évolution, l'évolution fait l'objet d'un large consensus scientifique sur ses fondements et ses mécanismes.

Questions ouvertes, et énigme de la reproduction sexuée Cette théorie laisse cependant certaines questions ouvertes : par exemple sur la difficulté ou l'impossibilité de trouver en paléontologie des formes de transition entre espèces, ou encore la question de l'importance respective de la sélection naturelle et de la dérive génétique. Beaucoup d'autres questions sont liées à la reproduction sexuée : quelle est l'origine du sexe ? Comment un tel processus a-t-il pu être sélectionné ? Est-il véritablement avantageux ? en effet, l'obligation de trouver un partenaire, par exemple, est un désavantage par rapport à la reproduction asexuée. Certaines espèces peuvent « choisir »au cours de leur existence entre la reproduction sexuée et la reproduction asexuée : y a-t-il une stratégie optimale ? Une autre question posant problème

est la question de la spéciation : comment et dans quels cas deux populations ayant la même origine peuvent-elles diverger et finir par donner deux populations telles que les membres de l'une ne puissent plus se reproduire avec les membres de l'autre ?

Utilité de l'aproche théorique Pour répondre à ces questions, les approches expérimentales sont assez limitées. Deux types d'approches expérimentales sont en effet envisageables :

- les « cas d'étude », qui consistent à prendre une espèce comme modèle, et à l'étudier sur plusieurs générations. Il se pose alors le problème de la durée des expériences à mener pour pouvoir suivre le processus d'évolution
- « l'inférence historique », c'est-à-dire essayer de retracer l'évolution passée partir de traces fossiles. Selon que l'on essaie de retracer cette évolution à partir de traces fossiles ou bien en étudiant les variabilités moléculaires présentes aujourd'hui entre espèces, on se heurte à la faible quantité de traces fossiles à notre disposition ou à la grande difficulté de tirer des informations pertinentes des séquences de gènes ou de molécules.

Les approches purement théoriques et les simulations numériques basées sur des modèles bien définis tiennent pour ces raisons une place de plus en plus importante dans les efforts faits pour tenter de répondre aux questions sur l'évolution.

1.2 Variabilité entre les individus et évolution des populations

Avant de détailler les modèles et les simulations que j'ai mises en oeuvre, je vais faire un bref rappel des notions sur l'évolution que j'ai été amené à utiliser.

D'où provient la diversité des individus Les variations entre les individus d'une même espèce, phénomène qui tient une place centrale dans la théorie de l'évolution, ont principalement deux origines : les mutations et les échanges de matériel génétique.

Les mutations sont des modifications du patrimoine génétique des individus, qui apparaissent par exemple à la suite d'agressions physiques ou chimiques. Ces modifications ont lieu le plus souvent lors de la reproduction, comme erreurs lors de la réplication du génome. Les mutations peuvent être de plusieurs types : il peut s'agir bien sûr de mutations ponctuelles qui modifient une ou un petit nombre de bases de l'ADN, mais aussi également de duplications de gènes, de délétions de séquences génétiques, d'ajouts de nouvelles séquences au hasard, de déplacement des gènes d'un chromosome à l'autre, d'une modification du nombre de chromosomes... On pense que la plupart des mutations sont létales, certaines sont neutres (elles n'ont pas d'effet sur l'organisme) et petit nombre peuvent être bénéfiques.

A ces mutations peuvent s'ajouter des échanges de matériel génétique entre individus. Ces échanges de matériel génétique se produisent notamment lors de la reproduction sexuée, avec l'échange de chromosomes et les « recombinaisons », c'est-à-dire l'échange de matériel génétique entre 2 chromosomes, lorsque se produisent des crossing over.De tels échanges se produisent également chez les individus asexués, par « transfert horizontal de gènes » : de nombreuses bactéries sont capables d'intégrer et d'utiliser du matériel génétique présent dans le milieu, ou « donné »par une autre bactérie, par exemple par l'intermédiaire d'un virus. De nombreux gènes de résistance aux antibiotiques se diffusent par exemple de cette façon. De vives polémiques ont actuellement lieu parmi les biologistes à propos de l'ampleur de ce phénomène chez les micro-organismes.

Origine de la variation de fréquence des allèles au sein des populations En plus de ces phénomènes qui affectent les individus, divers mécanismes qui agissent à l'échelle des

populations conduisent à des différences génétiques de plus en plus importantes au cours du temps, en sélectionnant certains individus plutôt que d'autres

- le plus célèbre est la « sélection naturelle », qui fut le mécanisme proposé par Charles Darwin : dans un environnement donné, les chances de reproduction des organismes dépendent des allèles qu'ils possèdent. D'une génération à l'autre, les individus favorisés (mieux adaptés à leur environnement) se reproduiront plus que les défavorisés, et la proportion des allèles des individus favorisés augmentera au détriment de celle des allèles des défavorisés.
- il faut aussi prendre en compte le mécanisme de « dérive génétique ». Celui-ci consiste simplement en l'effet du hasard. Il s'agit de ce que l'on pourrait qualifier de « bruit de taille finie » : même au sein d'une population dont tous les individus sont également favorisés, la fréquence des différents allèles va varier du simple fait que certains individus vont mourir aléatoirement, ou, dans le cas de la reproduction sexuée, que seules certaines gamètes vont être fécondées. Cet effet est d'autant plus important que la population est petite.
- d'autres mécanismes d'origine géographique existent également, comme par exemple le processus lié aux migrations : les migrations sont l'occasion de transmission d'allèles d'une population à l'autre. Si la fréquence allélique d'un groupe migrant n'est pas représentative de la population dont il est issu, la migration va modifier la fréquence des allèles entre les populations concernées.

1.3 Les réseaux de gènes

Avant d'étudier la reproduction sexuée et les recombinaisons, il nous faut tout d'abord comprendre quel est le lien entre les parents et leurs descendants c'est-à-dire en particulier le lien entre les gènes et l'individu.

La découverte de la structure en double hélice de l'ADN au début des années 50 marque le début de la génétique moderne au sens large. Le déchiffrage du code génétique nous a ensuite révélé qu'en fait les séquences génétiques sont associées aux protéines présentes dans l'organisme. Rappelons le « dogme central »de la génétique dans sa version très simplifiée : l'ADN, support de l'information génétique, code les séquences d'acides aminés correspondant à des protéines données, véritables outils de base de la machinerie cellulaire. Cependant, l'ADN ne « produit »pas directement de protéines : la séquence d'ADN correspondant à un gène donné est d'abord copiée sous forme d'ARN messager (transcription), puis la séquence codée par cet ARN est lue par les ribosomes chargés de synthétiser la protéine (traduction). Les protéines vont ensuite se replier, adopter une structure tridimensionnelle bien définie et acquérir toutes leur fonctions physicichimiques. Dans le « dogme central », un gène est associé à une protéine, via la production d'un ARN messager.

Cependant, il est de nombreux problèmes que le « dogme »ne parvient pas à expliquer, et notamment, le problème de la différentiation cellulaire : au sein d'un individu, des cellules peuvent avoir des activités complètement différentes bien qu'elles aient les mêmes gènes. Ceci provient du simple fait que les différentes protéines ne sont pas produites de façon indépendante : des protéines peuvent interagir entre elles, activer ou réprimer la transcription d'autres gènes et donc la production d'autres protéines, interagir avec des ARN messagers ...Il est pertinent de décrire l'ensemble de ces interactions par un « réseau génétique ». Une représentation synthétique d'un tel réseau serait un graphe, dans lequel chaque noeud correspond en fait à un triplet gène,ARN,protéine et chaque lien à une interaction (cf figure 1). De tels réseaux peuvent en général avoir plusieurs états d'équilibre distincts : deux cellules ayant des gènes identiques mais ayant des activités différentes sont interprétées comme

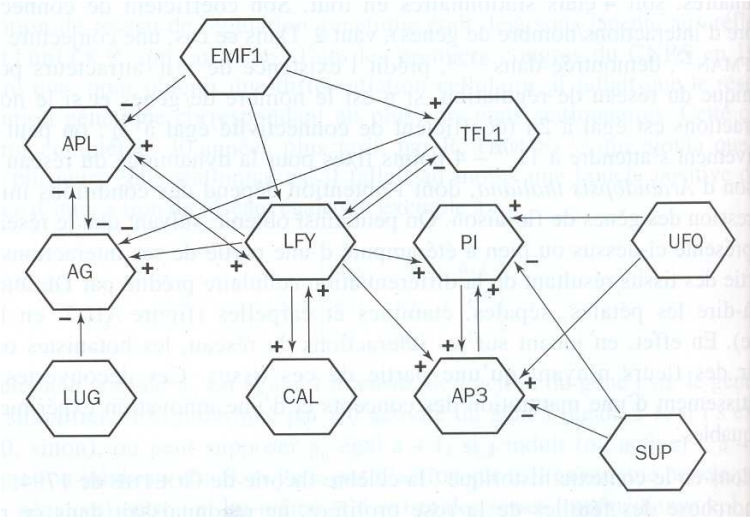

FIG. 1 – réseau de régulation de gènes de floraison d'*Arabidopsis thaliana* avec l'acronyme des différents gènes impliqués

deux cellules ayant le même réseau de gènes, mais chacun dans un état d'équilibre différent. C'est cette multiplicité d'états d'équilibre possibles, beaucoup plus que le nombre de gènes, qui peut expliquer la grande complexité des eukaryotes. Ces dernières années, beaucoup de données concernant les interactions entre protéines et gènes ont été accumulées, ce qui a permis d'analyser les circuits génétiques correspondant à de nombreux processus cellulaires (voir par exemple [15]).

Pour pouvoir, à partir des données expérimentales, construire et surtout étudier un réseau de gènes, il est utiles de disposer d'un formalisme général d'étude et de modélisation de ces réseaux, qui permette d'en comprendre le fonctionnement et les propriétés.

2 Modélisation des réseaux génétiques

2.1 Modélisation des réseaux de gènes

Premières modélisations Les premières modélisations des aspects généraux des réseaux de gènes datent de la fin des années 60 grâce aux travaux de Stuart Kauffman et René Thomas (voir par exemple pour S.Kauffman [6] [7] et pour R.Thomas [14]). En l'absence de résultats expérimentaux, S.Kauffman a considéré une représentation idéalisée d'un réseau de gène typique (c'est-à-dire aléatoire). Dans son modèle, les gènes sont équivalents, et leurs interactions forment un graphe orienté dans lequel un nombre fixe de liens arrive à chaque gène, les gènes à l'origine des liens étant des voisins aléatoires. L'état du gène est décrit par une variable binaire (allumé ou éteint), et le comportement dynamique de chaque gène, c'est-à-dire s'il va être allumé ou éteint au temps $t+1$, est gouverné par une fonction booléenne. Kauffman a ensuite proposé d'identifier les attracteurs dans l'espace des phases des réseaux, avec les différents types de cellules que l'on trouve dans un même organisme, et a réussi à estimer le nombre de tels attracteurs.

René Thomas a quant à lui étudié une description logique des mécanismes régulant l'expression des gènes. Son formalisme a notamment été appliqué avec succès à divers réseaux de régulations de gènes jouant un rôle dans la morphogénèse de la fleur d'*Arabidopsis thaliana*[11] et dans le développement de la mouche *Drosophilia melanogaster*[13][4].

Modèles actuels [1] Aujourd'hui plusieurs manières de modéliser les réseaux de gènes ont été développées et utilisées expérimentalement avec succès. Les modèles actuels peuvent en première approximation se diviser en deux catégories : ceux qui font une approche discrète et ceux qui font une approche continue. Chez ceux qui font une approche discrète, les gènes ne peuvent avoir qu'un nombre fini d'états, et les interactions entre gènes sont décrites par des fonctions logiques semblables à celles utilisées en programmation (par exemple [11] [2]). En général, le temps est aussi quantifié. Chez les modèles qui font une approche continue, au contraire, on suppose que l'activité des gènes est une fonction continue du temps, et leur évolution est modélisée par des équations différentielles avec typiquement des lois d'action de masse ou des lois de décroissances exponentielles (par exemple [5]).

L'analyse de ces réseaux nous est d'une grande utilité pour comprendre de manière générale le lien entre génotype et phénotype, et la manière dont une modification de l'un va entraîner une modification de l'autre. Pour répondre aux problèmes posés par la théorie de l'évolution l'idéal serait, maintenant que les génomes entiers de certaines espèces ont été séquencés, de déterminer la structure du réseau formé par l'ensemble des gènes d'un individu modèle. Cependant, nos connaissances actuelles sont encore loin de nous permettre une telle approche, car trouver des interactions entre gènes nécessite une connaissance extrêmement fine des données biochimiques des individus : dans ces réseaux d'interaction, le moindre composant oublié ou non détecté peut aboutir à changer complètement le comportement global. Pour étudier un individu dans sa globalité, nous en sommes donc réduit à essayer de déterminer des propriétés générales de classes de réseaux modèles dont la structure intègre le plus possible les données sur les réseaux déjà trouvés expérimentalement.

2.2 Le modèle de A.Wagner

Je vais maintenant présenter le modèle que j'ai utilisé au cours de mon stage. Il a été construit par A.Wagner [16] en 1994. Le développement et la structuration des organismes à partir d'une cellule souche fait intervenir des protéines qui régulent l'expression des gènes au niveau de la transcription, de manière à ce que les bons gènes s'expriment au bon moment et que la différentiation cellulaire de fasse dans de bonnes condition. Ces protéines régulent également leurs propres gènes : on peut donc construire un réseau génétique d'interactions déterminant les concentrations de ces protéines structurantes. Ce réseau joue donc un rôle central puisque ces protéines déterminent la structure globale de l'individu. C'est pourquoi, plutôt que de s'intéresser au réseau constitué par l'ensemble des gènes d'un individu, de nombreux biologistes du développement et de l'évolution ont choisi de s'intéresser à ces réseaux. Le modèle que j'ai utilisé au cours de mon stage a été construit par A.Wagner pour modéliser de tels réseaux.

Seule une fraction des gènes codant les régulateurs transcriptionnels sont susceptibles de s'exprimer dans une cellule donnée et à un instant donné. Pour permettre la différentiation cellulaire, l'expression de ces gènes varie d'une cellule à l'autre et au cours du temps. D'après les données expérimentales disponibles [10], ces réseaux reposent sur un faible nombre de gènes, entre 10 et 100. La manière dont les protéines interagissent entre elles et avec ces gènes n'est pas bien connue. Pour que notre modèle aboutisse à un formalisme où des calculs sont réalisables, un certain nombre d'hypothèses simplificatrices sont faites. On suppose ainsi :

1. que la régulation se fait uniquement au niveau de la transcription
2. que chaque gène produit un et un seule type de régulateur transcriptionnel
3. que l'effet d'un régulateur sur un gène agit indépendamment de l'effet des autres régulateurs sur ce même gène : tout est additif

Dans le modèle de A.Wagner, un réseau de gènes est représenté par un système dynamique dont la variable d'état est le vecteur des états d'expression des gènes du réseau :

$$(S_1(t), ..., S_N(t))$$

Par simplicité, on suppose que $S_i(t)$ ne peut avoir que deux valeurs : 1 et -1 correspondant aux situations où le gène est exprimé ou non exprimé, respectivement. L'état des gènes au temps 0 est appelé l'état d'expression initial. Partant de cet état, les interactions entre les gènes du réseau vont faire évoluer l'expression des gènes. Ces changement sont modélisés par l'équation matricielle :

$$S_i(t+\tau) = \sigma\left[\sum_{j=1}^{N} w_{ij} S_j(t)\right]$$

où σ est la fonction signe : $\sigma(x) = -1$ si $x < 0$ et $\sigma(x) = +1$ si $x \geq 0$. τ est une constante de temps caractéristique du phénomène que l'on considère. Sa valeur va dépendre de paramètres biochimiques, comme le taux de transcription ou le temps nécessaire pour exporter l'ARNm dans le cytoplasme. Les constantes réelles w_{ij} représentent la « force »de l'interaction du produit du gène j sur le gène i, et indiquent si cette interaction est répressive ($w_{ij} < 0$) ou activatrice ($w_{ij} > 0$). Un tel modèle a été utilisé avec succès dans la description et la prédiction d'interactions entre gènes expérimentalement [9]. Au cours de mon stage, nous avons étudié une version légèrement simplifiée de ce modèle : au lieu de prendre des coefficients w_{ij} réels, nous avons imposés que ces coefficients valent soit $+1$ soit -1 soit 0.

L'espace des états étant fini, la dynamique de ce modèle va mener à un état d'équilibre qui peut être un point fixe ou un cycle limite. L'étude des cycles limites peut avoir une certaine utilité dans l'étude du comportement cyclique de certains gènes, mais pour simplifier nous ne nous intéresserons ici qu'aux points fixes, et considérerrons, que les cycles ne correspondent pas à des systèmes biologiquement viables. L'état d'équilibre, s'il existe, vont constituer le « phénotype »de notre système.

Les mutations ponctuelles sont représentées par un changement d'un élément de la matrice choisi au hasard, avec la règle suivante : un -1 ou un 1 deviennent un 0, et un 0 devient aléatoirement un 1 ou un -1. Cette règle empêche qu'une interaction négative ne devienne positive en une seule mutation, ce qui ne serait pas très réaliste du point de vue biologique. Une recombinaison entre deux génomes est représentée par l'échange d'un certain nombre de lignes entre les matrices des génomes en question.

2.3 Comment modeliser la selection naturelle et quelles observables choisir pour suivre l'évolution d'une population ?

Les gènes que nous avons modélisés servent à réguler notamment l'expression des gènes structuraux, ou les gènes codant les protéines impliquées dans les processus de traduction[1]. Ces gènes sont fondamentaux chez un organisme, et, expérimentalement, la moindre altération dans l'expression de ces gènes cause des perturbations importantes qui sont en général létales [8][10]. Suite à cette observation, nous avons donc supposé qu'il existe un état d'équilibre optimal d'expression des gènes, c'est-à-dire un phénotype optimal. Si un réseau atteint un phénotype différent, nous considérerons donc que l'organisme correspondant sera beaucoup moins bien adapté à son environnement. Dans toute la suite, nous allons supposer que l'état d'expression initial des gènes et le phénotype optimal sont identiques pour tous les individus de la population.

[1] nous n'avons pas intégré ces derniers au réseau car eux n'agissent pas en retour sur les gènes de notre réseau

Comment quantifier l'adaptation d'un individu à son milieu ? Pour quantifier cette adaptation, la manière la plus simple consiste à créer une fonction fitness qui associe à chaque phénotype un réel entre 0 et 1 : plus cette fitness est proche de 1 plus l'individu est adapté. Il faut ensuite décider d'une règle déterminant la nombre de descendants individu en fonction de la fitness de son phénotype. Dans notre cas à nous, suite à la remarque que j'ai faite un peu plus haut, il semble raisonnable de supposer que cette fitness vaut soit 1 soit 0 suivant que le phénotype est le phénotype optimal ou non, et que les individus ayant une fitness de 1 vont se reproduire tandis que les autres vont mourir sans laisser de descendant. Nous allons à partir de maintenant qualifier ces individus de *viables* et de *non-viables*.

Quelle observable choisir ? Puisque tous les individus (viables) présentent la même adaptation à leur environnement, quelle observable choisir pour suivre l'évolution d'une population ? Dans le cas où tous les individus sont aussi bien adaptés, la seule chose qui va favoriser certains par rapport aux autres est la probabilité que les descendants de ces individus soient viables. En l'absence de mutations ou de reproduction sexuée, les descendants sont identiques aux parents, donc ce qui va nous intéresser est la probabilité qu'un individu reste viable lorsqu'on le soumet à des mutations ou lorsqu'on le fait se reproduire avec un autre individu de la population. C'est un changement de point de vue important : ce qui nous intéresse à partir de maintenant n'est plus l'adaptation d'un individu à son milieu, adaptation supposée ne varier que très faiblement d'un individu à l'autre, mais la robustesse de l'individu face aux mutations et mélange de gènes.

2.4 Le theoreme de E. Van Nimwegen, J.P. Crutchfield et M. Huynen

Le résultat dont je vais maintenant parler a été établi en 1999 par E. Van Nimwegen, J.P. Crutchfield et M. Huynen [12]. Je présente ce résultat ici, car une grande part de mon travail l'utilise. Il détermine la distribution des génotypes d'une population qui évolue par mutations ponctuelles et qui est soumise à une sélection de la part de son environnement.

Plus précisément : considérons le métagraphe, c'est-à-dire, le réseau de l'ensemble des génotypes possibles, avec un lien entre deux génotypes s'ils sont identiques à une mutation ponctuelles près. De ce graphe, on peut extraire un sous-graphe qui ne comporte que les génotypes viables. Maintenant, numérotons ces génotypes viables, et construisons la matrice d'adjacence G : l'entrée (i, j) est égale à 1 les génotypes i et j sont connectés, et 0 sinon. Soit enfin $\overrightarrow{P}(t)$ le vecteur de la population à l'instant t, c'est-à-dire le vecteur donnant la répartition de la population sur les différents génotypes.

Si on laisse la population évoluer par mutations ponctuelles, à chaque itération on a l'équation matricielle :

$$\langle Rmu \rangle_{POP} \overrightarrow{P}(t+1) = G.\overrightarrow{P}(t)$$

où $\langle Rmu \rangle_{POP}$ est un facteur égal à la probabilité moyenne qu'un individu de la population reste viable après une mutation ponctuelle.

Un calcul classique montre alors qu'en itérant un grand nombre de fois cette équation, le vecteur $\overrightarrow{P}$ va tendre vers le vecteur propre de G ayant la valeur propre la plus grande, et qu'un tel vecteur existe. Un autre calcul classique (que je détaille en annexe A) permet de montrer que la robustesse moyenne de la population va atteindre une valeur d'équilibre supérieure à la robustesse moyenne de l'ensemble des génotypes viables.

Cependant, il faut prendre garde au fait que cette formule ne s'applique qu'au cas d'une

population infinie. En effet, le vecteur $\overrightarrow{P}$ ne peut avoir que des composantes entières, alors que dans notre calcul nous avons supposé qu'il était à valeur réelle, et pouvait avoir des composantes non entières. Par conséquent lorsque la taille finie de la population est prise en compte, à chaque itération il va y avoir des fluctuations aléatoires qui vont causer une dérive par rapport au résultat prédit par la formule.

En résumé : l'état asymptotique d'une population qui n'évolue que par des mutations ponctuelles ne dépend que de la topologie réseau des génotypes viables, et non de la distribution initiale de la population[2].

3 Population en évolution

3.1 Généralités

L'évolution d'une population soumise à des mutations ponctuelles seules a déjà été bien étudiée. Comme prédit par l'article de E. Van Nimwegen, J.P. Crutchfield et M. Huynen [12], une telle population va atteindre un état d'équilibre, qui ne dépend pas de la population initiale, et dans lequel la robustesse moyenne aux mutations est sensiblement plus élevée que la robustesse moyenne prise sur l'ensemble du réseau neutre. Que se passe-t-il maintenant si on ajoute la possibilité d'évoluer par recombinaisons ? Pour répondre à cette question, nous avons écrit un programme permettant de faire évoluer une population par mutations et par recombinaison.

3.2 Principe du programme

On commence par créer, de la même manière que dans le code précédent, un individu viable. On copie cet individu un grand nombre de fois pour créer une population initiale d'individus identiques. Ensuite, à chaque itération, on provoque une mutation ponctuelle aléatoire chez chaque individu avec une certaine probabilité (le « taux de mutations »), et on effectue un échange d'un certain nombre de lignes de cet individu avec un autre pris au hasard dans la population.

Lorsqu'un individu n'est plus viable après une mutation ou une recombinaison, on l'efface de la population.

La population est donc en constante décroissance, et, pour se placer dans le cas de faible dérive génétique, c'est-à-dire pour éviter les effets de taille finie, on double la population en dupliquant chaque individu, dès que celle-ci est devenue inférieure à la moitié de la population initiale, ceci maintient la population près d'une taille cible donnée.

Nous avons mesuré à chaque itérations la robustesse moyenne de la population, c'est-à-dire la probabilité que le descendant d'un individu soit viable à la prochaine itération. Cette probabilité se mesure très simplement : c'est le rapport entre la population au temps $t+1$ et la population eu temps t (en faisant simplement attention aux cas où le programme duplique la population). Nous avons également mesuré à chaque itération la *robustesse aux mutations* moyenne de la population, c'est à dire la moyenne sur la population, de la probabilité qu'une mutation ponctuelle seule laisse un individu viable (c'est aussi le taux de décroissance de la population si on arrête les recombinaisons).

[2]en fait elle en dépend un peu : elle dépend rigoureusement de la topologie des *composantes connexes* du réseau des génotypes viables peuplés initialement

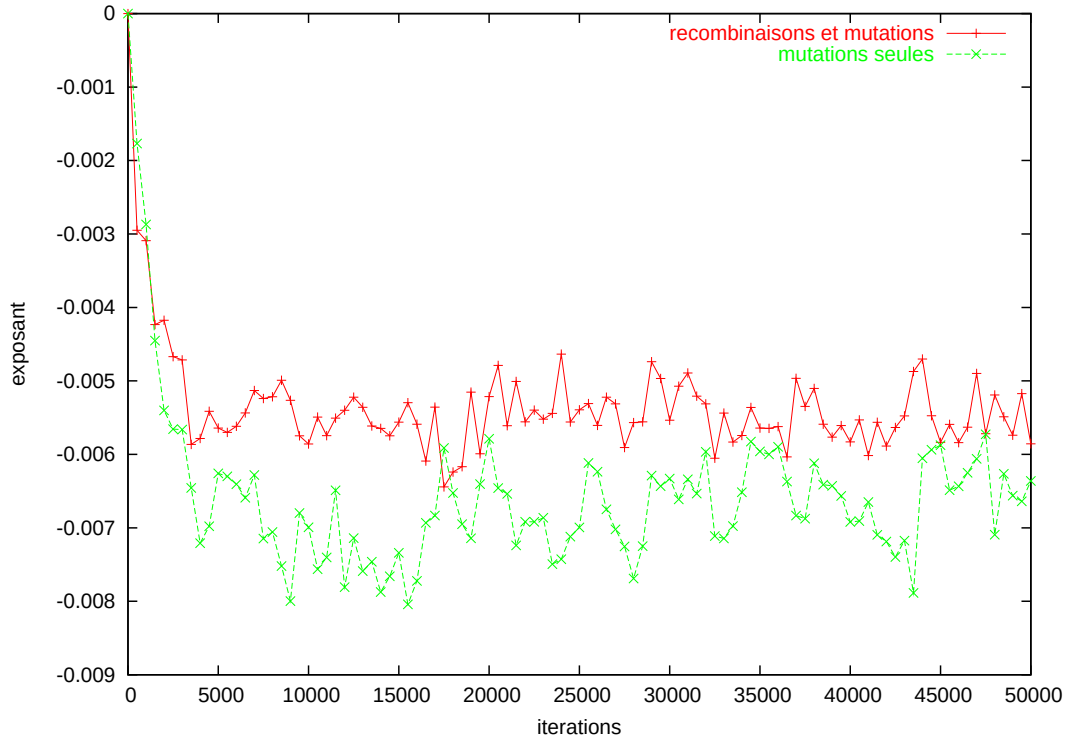

FIG. 2 – évolution du logarithme de la robustesse moyenne de la population en fonction du nombre d'itérations

3.3 Effets de la recombinaison

Tout d'abord, la première conclusion est, que, si la population n'est soumise qu'à des recombinaisons, l'état asymptotique dépend complètement de la population initiale. C'est pourquoi nous n'avons étudié que les cas d'évolution par recombinaison et par mutations, où au contraire, même si le taux de mutations est faible devant 1, l'état asymptotique semble indépendant de l'état initial.

La principale conclusion de notre étude, est que l'état stationnaire atteint par une population soumise à des mutations et à de la recombinaison présente une robustesse moyenne plus élevée que celui atteint par mutations seules (figure 2). La robustesse dans le cas (mutations + recombinaisons) augmente lorsque le taux de mutations diminue, tandis que dans le cas mutations seules la robustesse d'équilibre ne dépend bien sûr pas du taux de mutations (changer le taux de mutations dans ce cas correspond juste à faire une homothétie sur le temps, ce qui ne change pas l'état d'équilibre).

3.4 A quoi est du l'accroissement de la robustesse ?

Tout d'abord, nous nous sommes aperçu que lorsque l'état stationnaire est atteint, les recombinaisons ne jouent plus qu'un rôle négligeable dans la robustesse totale. En effet, nous nous sommes aperçu qu'une population soumise à uniquement à des recombinaisons, en général, diminue d'une part très lentement, et d'autre part ne disparaît pas complètement : elle tend vers une population plus faible que la population initiale, mais non nulle[3]. Au contraire, une population soumise uniquement à des mutations, va toujours s'annuler de manière exponentielle (figure 3). Même si elles sont rares, les mutations sont donc l'origine

[3]c'est assez logique : une population composée d'une seule matrice ne peut plus décroître si on la soumet à de la recombinaison, puisque recombiner une matrice avec elle-même ne modifie pas celle-ci

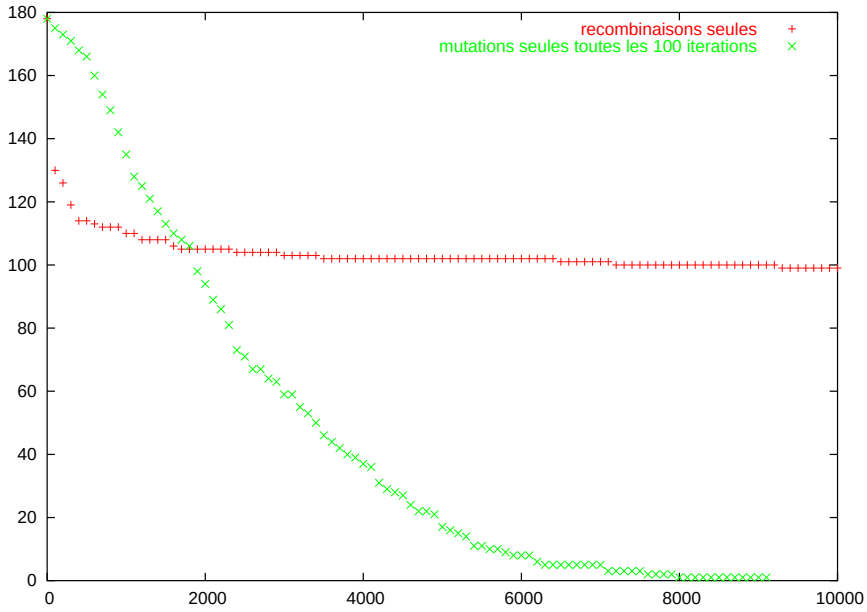

FIG. 3 – evolution du nombre d'individus d'une population soumise soit à des mutations ponctuelles, soit à des recombinaisons sans mutations

essentielle de décroissance de la population, et tout se joue au niveau de la robustesse aux mutations. En regardant l'évolution de la robustesse moyenne aux mutations (figure 4), nous avons pu mettre en évidence que la présence de la recombinaison augmente considérablement la robustesse aux mutations d'équilibre. En résumé, la présence de les recombinaisons ne font que très faiblement décroître les populations, mais augmentent considérablement la robustesse moyenne aux mutations, ce qui a pour effet d'augmenter la robustesse totale. Comment les recombinaisons peuvent-elles augmenter la robustesse aux mutations, alors que les deux phénomènes sont *a priori* très différents ? C'est ce que nous allons étudier dans la section 4.

Mais auparavant il me reste à signaler que les résultats de cette étude ne varient pas qualitativement, lorsque l'on choisit d'utiliser des réels distribués selon une gaussienne centrée en 0 au lieu d'utiliser simplement les trois entiers 1, -1 ou 0 comme éléments des matrices.Ceci est plutôt un signe positif par rapport aux simplifications que nous avons faites dans la construction du modèle.

4 Propriétés intrinsèques du réseau des génotypes viables

Quel est le lien entre recombinaisons et mutations ? Il semblerait naturel que, de même que les mutations ponctuelles font évoluer les populations vers une robustesse à la mutation moyenne plus grande que celle de départ, les recombinaisons devraient faire augmenter la *robustesse à la recombinaison* moyenne de la population. Y a-t-il alors une corrélation entre robustesse à la recombinaison et robustesse aux mutations ? La réponse à cette question nécessite d'étudier l'organisation de l'ensemble des génotypes viables. Nous avons donc écrit un autre programme échantillonnant de manière uniforme l'ensemble des génotypes viables, ce qui nous a permis d'étudier les corrélations entre les robustesses, mais aussi de mieux comprendre les valeurs de la robustesse à la recombinaison.

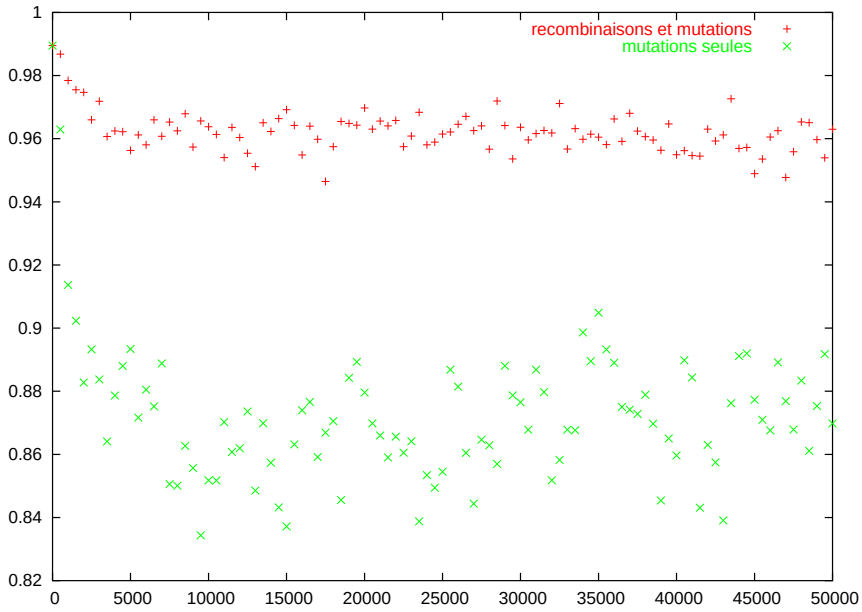

FIG. 4 – évolution de la robustesse aux mutations en fonction du nombre d'itérations

4.1 Principe du programme

Pour qu'un génotype soit considéré comme viable,il faut

1. que le vecteur de sortie soit un point fixe
2. que le système évolue spontanément du vecteur d'entrée au vecteur de sortie

Il est assez simple de quantifier le nombre de génotypes respectant la première condition : si les génotypes comportent N gènes, alors ils sont représentés par des matrices $N * N$ et chaque ligne de la matrice a une probabilité $1/2$ de donner la bonne composante du vecteur de sortie.

Seuls $\frac{1}{2^N}^{\text{ième}}$ de l'ensemble des génotypes satisfont donc la première condition.

La deuxième condition est beaucoup plus difficile à quantifier, mais il semble raisonnable que seule une petite fraction des génotypes vérifiant la première condition va vérifier cette seconde. Dans tous les cas, seule une très petite fraction des génotypes va être viable, fraction d'autant plus petite que le nombre N des gènes est grand.

A cause du très petit nombre de génotypes viables, ni l'énumération de l'ensemble des génotypes possibles, ni un échantillonnage aléatoire ne peuvent en pratique servir à trouver des génotypes viables efficacement dès que $N > 4$. Cependant même si ces génotypes représentent une très petite faction de l'ensemble des génotypes, il a été montré[3] que, si on relie entre eux tous les génotypes identiques à une mutation ponctuelle près, la quasi-totalité des génotypes viables constitue une grande composante connexe.

Ceci nous a donc incité à utiliser une stratégie de marche aléatoire : on part d'un génotype que l'on sait être viable, puis on effectue une mutation ponctuelle, pour creer un nouveau genome. Si celui-ci est viable, et que l'on ne l'a pas déjà, on le garde et on recommence, et si celui-ci n'est pas viable, on essaie en effectuant une autre mutation ponctuelle au lieu de la précédente. Pour échantillonner uniformément l'ensemble des génotypes possibles, nous avons dû faire attention de bien respecter le principe de balance détaillée : pour accepter un génotype B relié à R_B génotypes viables, après avoir accepté un génotype A relié à R_A génotypes viables, nous avons introduit la condition suivante :

– si $R_B < R_A$, on accepte toujours B
– sinon, on accepte B avec la probabilité R_A/R_B

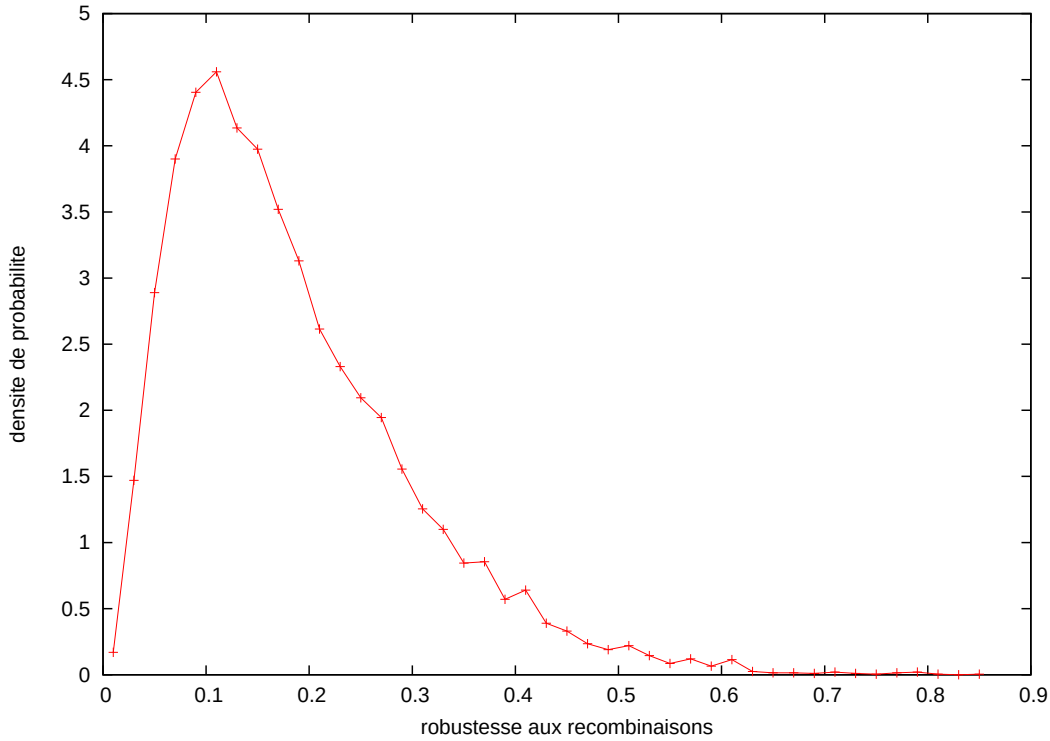

FIG. 5 – histogramme des robustesses des génotypes viables pour 10 gènes

Enfin, pour que l'échantillonnage ne dépende pas trop du génotype initial, nous avons introduit une thermalisation : on n'a commencé à enregistrer les génotypes viables qu'après en avoir déjà trouvé une centaine.

Après avoir trouvé un certain nombre de génotypes viables, notre programme a analysé les génotypes trouvés en mesurant pour chacun les robustesses aux mutations et à la recombinaison. La robustesse aux mutations, définie je le rappelle comme étant la probabilité qu'un génotype reste viable après une mutation ponctuelle, a été calculé comme dans le programme précédent en essayant tout simplement toutes les mutations ponctuelles possibles sur chaque matrice, et en comptant le nombre de viables et le nombre d'essais.

La robustesse à la recombinaison d'un génotype donné est définie de la même manière comme étant la probabilité que le génotype reste viable après recombinaison avec un autre génotype aléatoire, lorsque le nombre et la position des lignes échangées sont aléatoires elles aussi. Pour la mesurer avons fait pour chaque génotype des moyennes sur 100 recombinaisons aléatoires. Nous avons essayé aussi en faisant des moyennes sur plus de 100 essais, mais cela n'a rien changé qualitativement aux résultats que je vais présenter[4].

4.2 Robustesse a la recombinaison

Nous avons trouvé dans les simulations que, comme la robustesse aux mutations, la robustesse à la recombinaison peut varier beaucoup d'un genome a l'autre. Lorsque le nombre de gènes N augmente, la distribution n'a pas tendance à devenir piquée autour d'une valeur, ce qui fait que l'on ne peut pas, par exemple, se restreindre à étudier une valeur typique de la robustesse (cf fig 5).

[4]en particulier, cela n'a pas rendu les distribution plus étroites

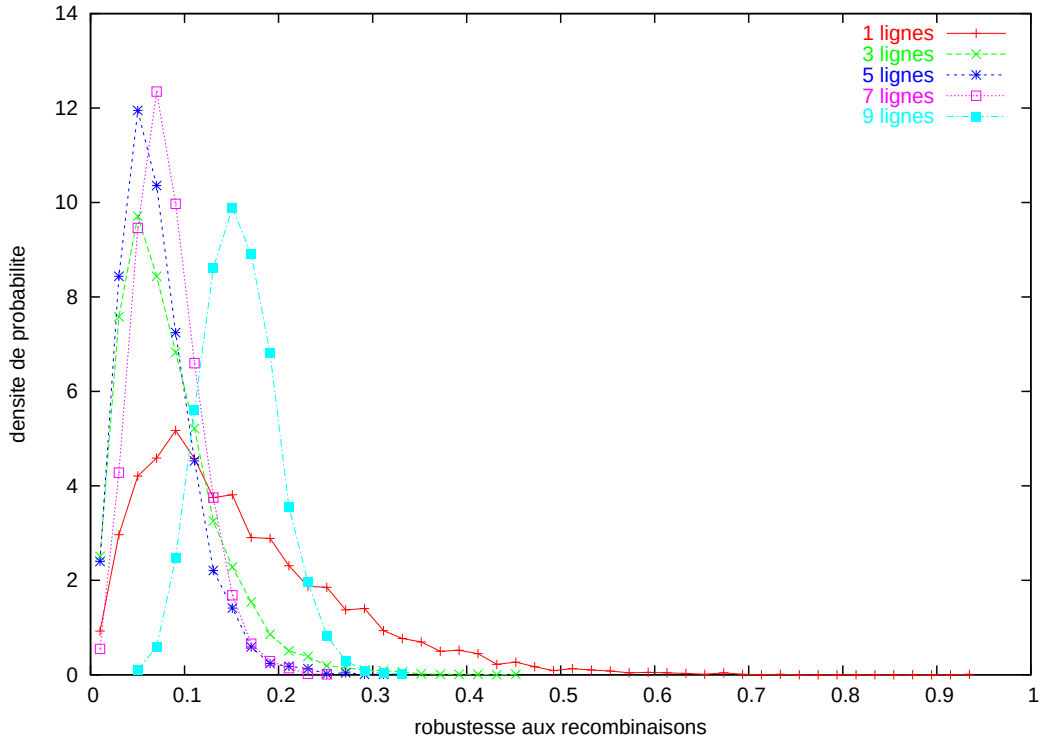

FIG. 6 – histogramme des robustesses pour un échantillonnage uniforme des génotypes viables, suivant le nombre de lignes échangées à chaque recombinaison (10 gènes)

4.3 Etude suivant le nombre de lignes échangées

Pour comprendre ces valeurs, nous avons calculé la distribution des robustesses en imposant cette fois le nombre de lignes échangées à chaque recombinaison. Nos conclusions furent les suivantes : à nombre N de gènes fixé, lorsque le nombre de lignes échangées augmente, le pic étalé représentant la distribution des robustesses se déplace vers les valeurs faibles, en devenant de plus en plus fin. Le minimum est atteint lorsque l'on échange la moitié des gènes à chaque recombinaison, puis le pic se redéplace à nouveau vers les valeurs plus grandes (cf figure 6).

4.4 Corrélation entre la robustesse aux mutations et la robustesse à la recombinaison

Nous avons trouvé qu'il existe une corrélation positive entre robustesse aux mutations et robustesse à la recombinaison (figure 7). Cette corrélation est d'autant plus forte que Rr ou Rmu sont grandes. Cette dernière conclusion paraît assez naturelle : en effet, pour qu'un génotype soit robuste face à la recombinaison, un bon critère semble être que la matrice représentant le génotype fasse passer de l'état d'expression initial à l'état d'expression final en un petit nombre d'itérations, critère qui rend robuste face à toutes les perturbations possibles du génotype, donc en particulier aussi aux mutations ponctuelles.

Cependant, il faut faire attention à notre définition de la robustesse à la recombinaison. Quand on considère l'évolution d'une population, on souhaite savoir la probabilité que le descendant d'un individu soit viable lorsque cet individu se reproduit uniquement avec d'autres individus *de la même population*, ce qui définit Rr_{pop} la « robustesse au sein de la population qui évolue ». Ce n'est pas cela que l'on a calculé ici, puisqu'on a pris la probabilité que le descendant soit viable lorsque l'individu se reproduit avec un autre individu viable

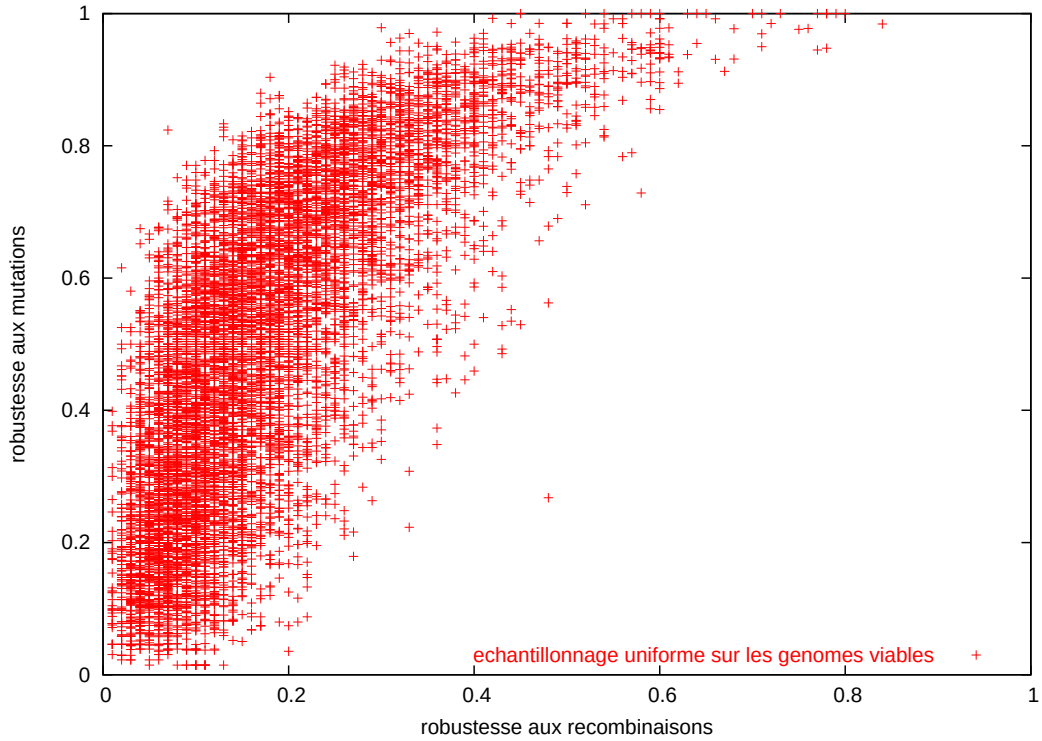

FIG. 7 – corrélation entre Rr et Rmu. Plus Rr ou rmu sont grandes, plus la corrélation semble importante

quelconque : Rr_{unif} la « robustesse au sein de l'ensemble des génotypes viables ».

Néanmoins, il existe un lien entre les deux : On peut en effet montrer qu'elles sont proportionnelles (cf annexe B.1) moyennant quelques approximations. Effectivement, en utilisant le facteur numérique donné par le calcul en annexe, on a pu superposer les courbes Rmu en fonction de Rr obtenues pour une population qui évolue par recombinaisons et mutations à celle obtenue pour l'ensemble des génotypes viables. Cette corrélation peut donc permettre d'expliquer au moins en partie les grandes robustesses aux mutations obtenues lorsque l'on fait évoluer une population par recombinaisons et mutations : lorsque la population évolue, les recombinaisons vont sélectionner les individus qui ont une forte robustesse à la recombinaison. Comme ces individus ont statistiquement une plus grande robustesse aux mutations que la moyenne, la recombinaison va augmenter la robustesse aux mutations de la population.

5 Approximation de champ moyen

5.1 Motivations

Pour une population soumise à des recombinaisons, il n'existe pas d'équivalent de la formule de l'article de E. Van Nimwegen, J.P. Crutchfield et M. Huynen, et qui permettrait de connaître simplement la distribution asymptotique de la population. Dans ce cas en effet, la grande difficulté vient du fait que l'évolution d'une population n'est plus régie par une équation linéaire : si, dans le cas de l'évolution par mutations ponctuelles, chaque individu évolue de manière indépendante des autres, dans le cas de l'évolution par recombinaison, il faut utiliser deux individus pour en créer un nouveau. La probabilité d'obtenir un descendant donné dépend donc maintenant du produit des probabilités de trouver chacun des parents.

Pour essayer de traiter l'équation d'évolution malgré la non-linéarité, nous avons tenté

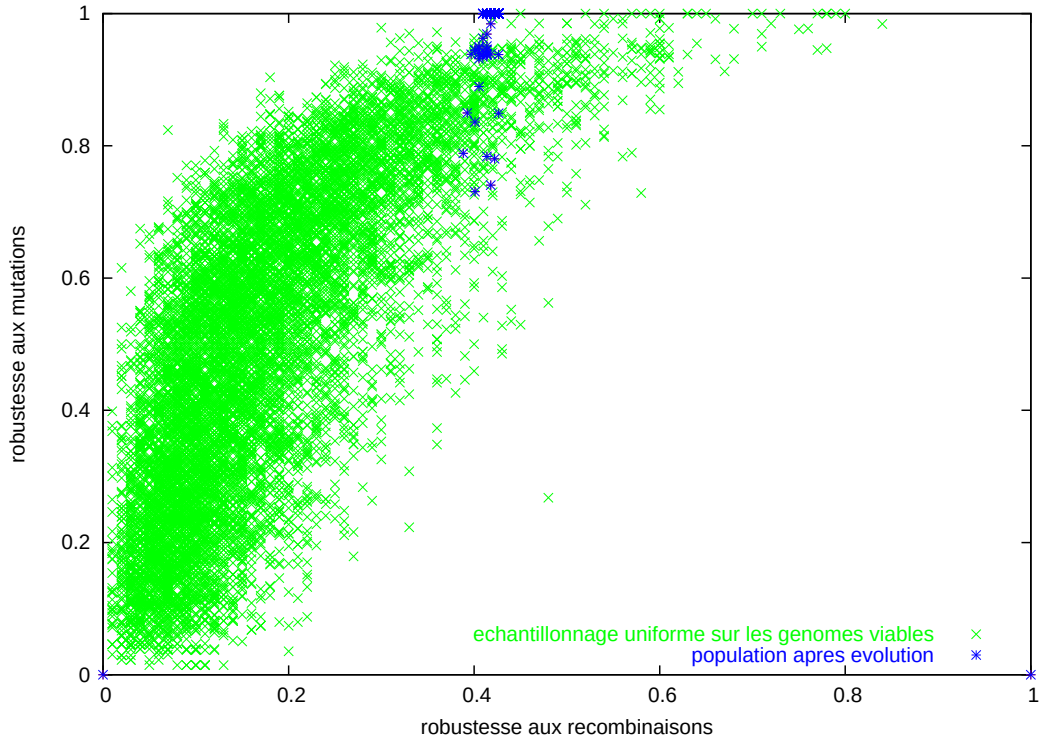

FIG. 8 – superposition de la courbe précédente, et des génotypes issus d'une population qui a évolué à cause de mutations et de recombinaisons

une approximation de champ moyen. Nous avons par la suite testé la validité de cette approximation par des simulations numériques. Malheureusement, cette étude s'est déroulée vers la fin du stage, et nous n'avons donc pu tester cette approximation que dans quelques cas bien particuliers, ce qui est expliqué dans la dernière partie de ce rapport.

5.2 Dérivation

Fixons un nombre N de gènes et considérons l'espace E des génotypes possibles que l'on peut créer avec ces N gènes. Dans le modèle de A.Wagner, cet espace contient 3^{N^2} génotypes. Dans cet espace, une fraction f des génotypes sont viables. Indexons ces derniers par un entier i. On cherche à étudier l'évolution d'une population comptant $n_{population\ totale}(t)$ individus. Tous ces individus sont censés être viables, et j'appelle $n_i(t)$ le nombre d'individus de cette population ayant le génotype i à l'instant t. Comment écrire l'équation d'évolution de $n_i(t)$?

Pour cela, imaginons tout d'abord qu'un individu de génotype i puisse être créé par la recombinaison entre un individu de génotype j et un individu de génotype k.

- Comment savoir si un couple de génotypes (j, k) peut donner i par recombinaison ? Dans le modèle de A.Wagner, le critère est très simple : il suffit que la matrice représentant j ait au moins une ligne en commun avec la matrice représentant i. La matrice de k quant à elle doit au moins avoir en commun avec la matrice de i les lignes pour lesquelles la matrice de j diffère de la matrice de i.
- La probabilité qu'un individu de génotype j se recombine avec un individu de type k, et non avec un individu d'un autre type, est :

$$\frac{n_j * n_k}{\sum_a n_a} = \frac{n_j * n_k}{n_{population\ totale}}$$

– Il faut multiplier ce nombre par la probabilité que, lors de la recombinaison, le bon nombre l de lignes soit échangé, et que ce soient les bonnes lignes, probabilté que j'appelle $\rho(l)$.
– il reste maintenant à sommer sur toutes les possibilités.

schéma ? On arrive ainsi, en faisant attention de ne compter qu'une seule fois chacun des couple (i, j) à :

$$n_i(t+1) = \frac{1}{2} \underbrace{\sum_{l=1}^{N-1}}_{\substack{\text{nombre de lignes échangées}}} \underbrace{\sum}_{\substack{\text{position des lignes}}} \underbrace{\sum_j n_j(t)}_{\substack{j \text{ a au moins } l \text{ lignes en commun avec } i}} \underbrace{\sum_k n_k(t) * \frac{\rho(l)}{n_{population\ totale}}}_{\substack{k \text{ permet à } j \text{ de donner } i \text{ par recombinaison}}}$$

ou encore, en imposant à j d'avoir plus de la moitié de ses lignes en commun avec i, c'est-à-dire en fixant $l \geq (N-1)/2$:

$$= \sum_{l=\frac{N-1}{2}}^{N-1} \sum \sum_j n_j(t) \sum_k n_k(t) * \frac{\rho(l)}{n_{population\ totale}}$$

Dans cette formule, la seule contrainte sur les j est qu'ils aient au moins l lignes en commun avec i, et la seule contrainte sur k ait qu'il ait les $N-l$ lignes restantes en commun avec i. Pour un j donné, il existe beaucoup de k qui vont convenir, et inversement. Cependant, dans la mesure où $l > (N-1)/2$, la contrainte sur k est beaucoup plus faible que celle sur j : k doit avoir moins de la moitié de ses lignes en commun avec i, alors que j doit en avoir plus de la moitié. L'hypothèse que j'ai faite a donc été de considérer que

– vu qu'à un génotype j correspond beaucoup de génotypes k sur lesquels il va falloir sommer
– et vu que ces génotypes k ne sont pas trop reliés au génotype i (ils ont moins de la moitié de leurs lignes en commun avec lui)

on peut supposer que l'ensemble des k correspondant à un j donné est un échantillonnage à peu près aléatoire des génotypes viables, et puisqu'on fait la somme sur un grand nombre de termes, même si la population n'est pas vraiment équirépartie, qu'en introduisant la valeur moyenne $< n >$ de la population par génotype, on peut remplacer la somme $\sum_k n_k$ par la somme $\sum_k < n >$. Cela conduit donc à la formule suivante :

$$n_i(t+1) = 2 * \sum_{l=\frac{N-1}{2}}^{N-1} \sum \sum_j n_j(t) \sum_k < n > (t) * \frac{\rho(l)}{n_{population\ totale}}$$

Or $n_{population\ totale} = 3^{N^2} * f * < n >$ puisqu'il y a $3^{N^2} * f$ génotypes viables occupés chacun en moyenne par $< n >$ individus :

$$= 2 * \sum_{l=\frac{N-1}{2}}^{N-1} \sum \sum_j n_j(t) \sum_k \frac{\rho(l)}{3^{N^2} * f}$$

Il nous faut maintenant connaître le *nombre* de partenaires k correspondant à un j donné. Ce nombre correspond au nombre de génotypes viables ayant $N-l$ lignes données en commun avec i. Sans regarder la condition de viabilité, il y a en tout 3^{N*l} génotypes qui ont $N-l$ lignes en commun avec i. Comme $l > N/2$, k a relativement peu de lignes en commun avec

i, et on peut donc supposer que le fait de savoir qu'un génotype a ces quelques lignes en commun avec i est à peu près indépendant du fait de savoir si ce génotype est viable ou non.

Si on fait cette hypothèse, la proportion f de viables dans l'espace total des génotypes est égale à la fraction des viables dans l'espace des génotypes ayant au moins $N-l$ lignes en commun avec i. Donc à un j donné correspondent $3^{N*l} * f$ partenaires k. Au final, on arrive donc à la formule :

$$n_i(t+1) = 2 * \sum_{l=\frac{N-1}{2}}^{N-1} \quad \sum_{\text{position des lignes}} \quad \sum_j n_j(t) \quad \frac{\rho(l)}{3^{N*(N-l)}}$$

5.3 Commentaires sur la formule

Cette formule est formellement l'analogue de la formule établie par dans l'article de E. Van Nimwegen, J.P. Crutchfield et M. Huynen : Elle relie la population du génotype i à la somme des populatons des génotypes *voisins* de i c'est-à-dire ici « qui ont un certain nombre de lignes en commun avec i ». [5]

Dans le cadre de l'approximation réalisée ici, on peut donc faire les mêmes conclusions que dans l'article de E. Van Nimwegen, J.P. Crutchfield et M. Huynen : il existe un unique état d'équilibre, qui dépend uniquement du réseau des génotypes, et la population d'équilibre présente une robustesse moyenne (à la recombinaison) beaucoup plus élevée que la robustesse moyenne de l'ensemble des génotypes viables. Rappelons les hypothèses de notre calcul :

1. si je prends un génotype viable, que je choisis moins de la moitié de ses lignes, et que je fais la moyenne des populations des génotypes qui ont ces lignes-ci en commun avec lui, j'obtiens un nombre identique à la moyenne sur l'ensemble des génotypes viables (hypothèse de champ moyen)
2. si je prends un génotype viable, et que je modifie aléatoirement plus de la moitié de ses lignes, la probabilité que le nouveau génotype soit viable est la même que la probabilité qu'un génotype quelconque soit viable
3. comme pour le calcul original de E. Van Nimwegen, J.P. Crutchfield et M. Huynen, on doit considérer une population infinie

a priori, les deux premières hypothèses ne sont valables que pour N grand, et la deuxième impose en outre que le nombre de génotypes représentés dans la populations soit grand. En effet, il faut qu'à peu près l'ensembles des génotypes viables possibles puisse être atteint par recombinaison de deux individus de la population[6].

5.4 Comparaison avec les simulations

Pour tester cette formule, plusieurs approches sont possibles :

– prendre une population, la faire évoluer par recombinaisons, et comparer l'évolution de cette population à l'évolution prédite par la formule précédente. Malheureusement, pour pouvoir utiliser cette formule, il faut avoir une liste de tous les génotypes viables, ce qu'il nous a été impossible de trouver pour $N > 4$. [7]

[5] En fait, comme ici suivant le nombre de lignes en commun il y a un facteur multiplicatif différent, ce n'est pas exactement une matrice d'adjacence que l'on utilise mais plutôt une combinaison linéaire de matrices d'adjacences, mais cela ne change rien au calcul déterminant la population asymptotique.

[6] par exemple si toute la population est concentrée en un génotype, les recombinaisons ne changeront rien, et notre formule ne s'appliquera pas...

[7] les simulations ont montré que le modèle n'est pas valable pour N=4, ce qui était prévisible puisqu'il a été dérivé en supposant N assez grand.

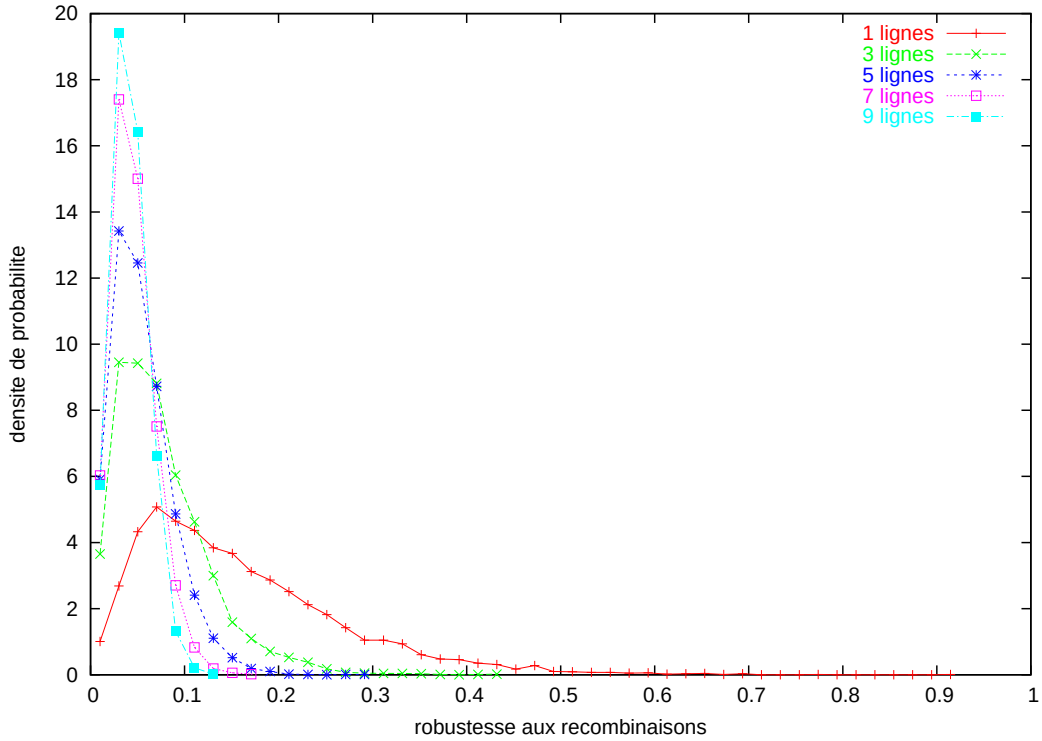

FIG. 9 – histogramme des robustesses pour un échantillonnage uniforme des génotypes viables, suivant le nombre de lignes remplacées par des lignes aléatoires (10 gènes)

– tester les approximations séparément. c'est cette démarche qui a donné le plus de résultats.

Nous avons ainsi cherché à tester l'hypothèse 2 : « si je prends un génotype viable, et que je modifie aléatoirement plus de la moitié de ses lignes, la probabilité que le nouveau génotype soit viable est la même que la probabilité qu'un génotype quelconque soit viable ». Pour cela, pour $N = 10$ gènes, nous avons fait un échantillonnage uniforme des génotypes viables, puis nous avons remplacé un nombre donné de lignes par des lignes aléatoires pour chaque génotype de l'échantillon, et mesuré les robustesses moyennes par rapport à cette manipulation. Nous nous sommes effectivement aperçu qu'au delà de plus de la moitié des lignes remplacées, l'histogramme des robustesses mesurées ne changeait plus, et devenait de plus en plus piqué à mesure que l'on évaluait les robustesses de plus en plus précisément. Ceci semble donc indiquer que notre seconde hypothèse est réaliste. Nous n'avons malheureusement pas eu le temps d'aller plus loin dans le test des hypothèses.

Conclusion

En conclusion, nous avons montré que dans le modèle de réseaux de gènes de A. Wagner, la présence de recombinaisons en plus des mutations permet à la population qui évolue d'avoir, une fois l'état stationnaire atteint, une robustesse plus grande. Ceci semble être dû au fait que les recombinaisons sont rarement létales, donc ne diminuent que peu la robustesse de la population, mais font évoluer le système vers des états qui sont robustes face aux recombinaisons, et que ces états sont eux-mêmes très robustes face aux mutations à cause d'une corrélation statistique. Le système devient donc plus robustes aux mutations que s'il n'évoluait que par des mutations ponctuelles. Nous avons enfin essayé de quantifier ces effets en développant un modèle de champ moyen, modèle qu'il faut à présent continuer de tester

pour en connaître la validité. Dans son domaine de validité, celui-ci pourrait notamment permettre d'étudier les questions liées à la spéciation, par exemple étudier l'écart qui apparaît entre deux populations voisines initialement mais qui évoluent séparément. La question reste posée de savoir si les conclusions présentées ici sont générales, et si d'autres modèles tout aussi bien fondés biologiquement peuvent donner des résultats qui contredisent celles-ci. De nombreux phénomènes auraient également pu être intégrés dans notre étude, comme les effets de diploïdie, ou encore l'existence au sein d'une population de deux types d'individus différents : mâles et femelles.

Remerciements

Je voudrais ici remercier très chaleureusement Olivier Martin de m'avoir accepté en stage, de m'avoir fait découvrir ce thème et surtout de m'avoir, toujours avec gentillesse, tant appris sur les simulations numériques, ainsi que tout le LPTMS d'Orsay pour son accueil chaleureux !

A Augmentation de la robustesse d'une population soumise à des mutations ponctuelles seules

Soit R la robustesse aux mutations d'un individu. Montrons que $\langle d \rangle_{POP}$ la robustesse aux mutations moyenne de la population est toujours supérieure à la robustesse moyenne de l'ensemble des génotypes viables $\langle d \rangle_{unif}$. La matrice G qui fait passer de la distribution de population au temps t à la distribution au temps $t+1$ est symétrique, et est donc diagonalisable de valeurs propres réelles. L'iteration un grand nombre de fois de cette matrice projète le vecteur initial sur le vecteur propre Ψ_{max} de valeur propre la plus grande λ_{max}.

L'expression

$$A = \frac{\sum_{ij} \phi_i G_{ij} \phi_j}{\sum_i \phi_i^2}$$

est maximisée quand $\phi = \Psi_{max}$, et ce maximum est λ_{max}. Un calcul direct montre que l'expression ci-dessus est égale à $A = \langle d \rangle_{\Psi_{max}}$ donc à $\langle d \rangle_{POP}$. Si on prend maintenant ϕ comme étant une distribution uniforme, on a $A = \langle d \rangle_{unif}$, donc $\langle d \rangle_{unif} \leq \langle d \rangle_{POP}$.

B passer de la robustesse à la recombinaison en population à la robustesse à la recombinaison sur l'ensemble des génotypes viables

B.1 Calcul de la probabilité que le descendant de 2 individus donnés soit viable

Fixons les notations : il y a n génotypes viables différents, chacun ayant une robustesse Rr à la recombinaison, définie comme étant la probabilité que le descendant d'un individu ayant ce génotype avec un autre individu viable quelconque, soit viable.

On peut penser aux génotypes comme étant des noeuds d'un réseau, avec une arrête joignant 2 noeuds lorsque des individus ayant les 2 génotypes en question peuvent avoir un descendant viable. Dans ce cas, il y a en tout $n * \sum Rr/2$ arrêtes.

Fixons 2 individus A et B. A est connecté à $Rr_A * n$ arrêtes et B à $Rr_B * n$. La probabilité que la première arrête de B soit connéctée à A est :

$$\frac{Rr_A * n}{(Rr_A * n + 2 * Rr_B * n)}$$

La probabilité que la première arrête de B ne soit pas connecté à A mais que la deuxième le soit est :

$$\frac{Rr_A * n}{(Rr_A * n + 2 * Rr_B * n - 2)} * \frac{2 * Rr_B * n}{(Rr_A * n + 2 * Rr_B * n)}$$

En continuant ainsi de suite, on arrive à la formule :

$$P = \sum_{i=1}^{n*Rr_B} \frac{Rr_A * n}{(Rr_A * n + 2(Rr_B * n - i))} * 2^i * \prod_{l=0}^{i} \frac{Rr_B * n - l}{Rr_A * n + 2(Rr_B * n - l)}$$

qui après quelques simplifications au premier ordre en $\frac{Rr_A}{\sum Rr}$ et $\frac{Rr_B}{\sum Rr}$ aboutit à :

$$P \approx \frac{n * Rr_A * Rr_B}{\sum Rr}$$

B.2 cas d'une population

La robustesse Rr_A^{pop} de l'individu A au sein d'une population donnée sera :

$$Rr_A^{pop} \approx \frac{n * Rr_A^{unif} * \left\langle Rr^{unif} \right\rangle_{POP}}{\sum Rr^{unif}} = \frac{Rr_A^{unif} * \left\langle Rr^{unif} \right\rangle_{POP}}{\left\langle Rr^{unif} \right\rangle_{unif}}$$

où Rr^{unif} désigne la robustesse au sein de l'ensemble des génotypes viables, et Rr^{pop} la robustesse au sein de la population. Comment calculer $\left\langle Rr^{unif} \right\rangle_{POP}$ sachant que par les simulations numériques on n'a accès qu'à $\left\langle Rr^{unif} \right\rangle_{unif}$ et à $\left\langle Rr^{pop} \right\rangle_{POP}$? Si on fait une moyenne de la formule sur tous les individus de la population, on obtient :

$$\left\langle Rr^{pop} \right\rangle_{POP} = \frac{\left\langle Rr^{unif} \right\rangle_{POP}^2}{\left\langle Rr^{unif} \right\rangle_{unif}}$$

et donc :

$$Rr_A^{pop} \approx Rr_A^{unif} * \frac{\sqrt{\left\langle Rr^{pop} \right\rangle_{POP} * \left\langle Rr^{unif} \right\rangle_{unif}}}{\left\langle Rr^{unif} \right\rangle_{unif}}$$

Références

[1] R. Albert. *Lect. Notes Phys.*, 650, 2004.

[2] R. Albert and H. G. Othmer. *J. Theor. Biol.*, 223, 2003.

[3] Stafano Ciliberti, Olivier C. Martin, and Andreas Wagner. Circuit topology and the evolution of robustness in complex regulatory gene networks. *soumis à Nature*, 2006.

[4] A. Ghysen and R. Thomas. *BioEssays*, 25, 2003.

[5] V. V. Gursky, J. Reinitz, and A. M. Samsonov. *Chaos*, 11, 2001.

[6] S. A. Kauffman. Modeling network dynamics : the lac operon, a case study. *J. Theor. Biol.*, 22 :437, 1973.

[7] S. A. Kauffman. *Tho origins of Order*. 1993.

[8] M. Kimura. *Jpn. J. Genet.*, 66, 1991.

[9] Fangting Li, Tao Long, Ying Lu, Qi Ouyang, and Chao Tang. The yeast cell-cycle network is robustly designed. *PNAS*, 101(14) :4781–4786, 2004.

[10] W. McGinnis and R. Krumlauf. *Cell*, 68, 1992.

[11] L. Mendoza, D. Thieffry, and E. R. Alvarez-Buylla. *Bioinformatics*, 15 :593, 1999.

[12] Erik Van Nimwegen, James P.Crutchfield, and Martin Huynen. Neutral evolution of mutational robustness. *Proc. Nat. Acad. Sci.*, 96(9716-9720), août 1999.

[13] L. Sánchez and D. Thieffry. *J. Theor. Biol.*, 188 :391, 1997.

[14] R. Thomas. *J. Theor. Biol.*, 42 :563, 1973.

[15] Jose M.G. Vilar, Calin C. Guet, and Stanislas Leibler. Modeling network dynamics : the lac operon, a case study. *J. Cell Biol.*, 161(3) :471–476, 2003.

[16] A. Wagner. *Evolution*, 50(3), 1996.