MajecSTIC 2010
Bordeaux, France, du 13 au 15 octobre 2010

Aevol : un modèle individu-centré pour l'étude de la structuration des génomes

David P. Parsons[1,3], Carole Knibbe[2,3] et Guillaume Beslon[1,3]

1 : Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France.
2 : Université de Lyon, CNRS, Université Lyon 1, LIRIS, UMR5205, F-69622, France.
3 : IXXI, Institut Rhône-Alpin des Systèmes Complexes, Lyon, F-69007, France.

Contact : david.parsons@liris.cnrs.fr

Résumé
L'organisation des génomes présente de grandes disparités entre les différentes formes de vie, allant de génomes très courts et denses chez les virus à des génomes très longs et majoritairement non-codants chez les organismes multicellulaires. Dans cet article, nous présentons le modèle de génétique digitale Aevol, développé pour permettre l'étude de l'évolution de la structure des génomes d'organismes virtuels, ainsi qu'un aperçu des résultats que ce modèle a permis d'obtenir. Aevol permet de reproduire les différences observées chez les organismes réels en termes de structuration des génomes en faisant varier les conditions expérimentales. Les différentes campagnes expérimentales que nous avons menées ont permis de mettre en avant un phénomène d'*error threshold* comme une des principales explications de ces différences structurelles.

Abstract
Genome organization shows great variations throughout the different life forms, going from very short and dense genomes in viruses to very long and mostly non-coding genomes in multicellular organisms. In this paper, we present the Aevol model, a digital genetics model developed to study the evolution of genome structure in virtual organisms, and provide an overview of the results obtained with this model. The diversity of genome organizations can be accurately reproduced by the model when varying experimental conditions. The experiments we have conducted point us to a phenomenon of error threshold as one of the main explanations for these structural differences.

Mots-clés : Simulation individu-centrée, évolution, organisation des génomes, error threshold, mutations

Keywords: Individual-centered simulation, evolution, genome organization, error threshold, mutations

1. Introduction

S'il occupe une position centrale dans la biologie moderne, le processus d'évolution des espèces reste difficile à étudier. En effet, les mécanismes qui en sont responsables agissent sur des échelles de temps très longues, rendant difficile l'expérimentation directe sur l'objet d'étude[1]. À cet obstacle logistique, viennent s'ajouter des difficultés d'analyse des résultats : comment associer de façon claire un caractère observé à une cause évolutive précise lorsque l'on ne maîtrise pas l'ensemble des paramètres pouvant entrer en jeu ?
Les approches de génomiques comparatives permettent de s'affranchir de ces contraintes de temps, cependant, elles sont basées sur un cliché instantané des séquences (l'ADN des espèces actuelles) et doivent *inférer* leur passé évolutif.

[1] Même pour des espèces à reproduction rapide comme les bactéries, une expérience directe d'évolution prend des dizaines d'années [5].

La génétique digitale est un nouveau champ de recherche qui a émergé dans les années 90. Des structures de données représentant des systèmes biologiques abstraits (les « organismes digitaux ») dans une « chimie artificielle » [6] sont interprétées par des programmes dédiés qui leur permettent de se dupliquer, de subir des mutations, d'évoluer et de s'adapter à leur environnement. D'un point de vue algorithmique, la génétique digitale est très proche des algorithmes évolutionnaires, cependant, le but n'est plus de trouver une solution à un problème d'optimisation spécifique mais d'étudier le processus évolutif lui-même dans une perspective de vie artificielle.
Les approches de génétique digitale permettent de réaliser facilement et en un temps raisonnable des expériences d'évolution contrôlées et reproductibles, donnant accès à un enregistrement fossile complet de l'intégralité du processus évolutif. Il est alors possible de soulever des questions essentielles quant aux fondements génétiques de l'évolution de la complexité, de l'organisation des génomes ou de la robustesse et de l'évolvabilité [1].
Dans cet article, nous présentons le modèle d'évolution expérimentale *in silico* Aevol ainsi qu'un aperçu des résultats obtenus grâce à ce modèle.

2. Le modèle Aevol

Aevol est un modèle d'évolution artificielle développé au LIRIS pour permettre l'étude de la structuration des génomes dans le processus évolutif. Largement inspiré de la génétique bactérienne, sa principale particularité vis-à-vis des algorithmes génétiques classiques est qu'il met en œuvre explicitement des processus de transcription et de traduction du génome, faisant ainsi apparaitre un niveau intermédiaire (le protéome) entre le génotype et le phénotype. Cette couche intermédiaire autorise une grande flexibilité au niveau du génome, permettant à la structure du génome elle-même d'évoluer. Nous pouvons alors interroger le modèle sur les causes possibles des différences observées chez les organismes réels en termes de structuration et d'organisation du génome.

2.1. Aevol - Principes généraux

Aevol est un modèle individu-centré qui permet de simuler l'évolution d'une population de N individus. Chaque individu possède un génome dont la structure est inspirée des génomes bactériens. Il s'agit d'une séquence binaire circulaire double-brin contenant un nombre variable de gènes séparés par des séquences non-codantes (figure 1). Au début d'une simulation, tous les individus possèdent le même génome aléatoire, puis, à chaque génération, le génome de chaque individu est décodé pour déterminer son phénotype qui permettra à son tour de calculer la valeur d'adaptation (ou fitness) de l'individu.

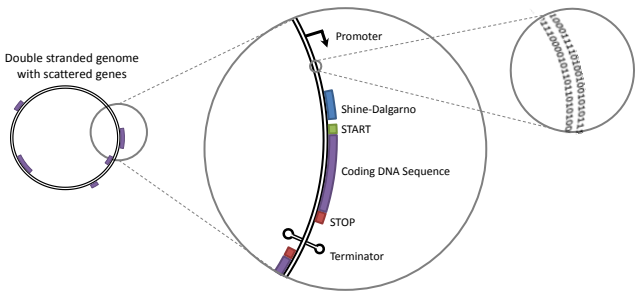


FIG. 1 – Dans Aevol, chaque organisme possède un génome binaire circulaire double-brin le long duquel des séquences codantes sont identifiées grâce à des séquences de signal prédéfinies. Les promoteurs et les terminateurs délimitent des séquences transcrites au sein desquelles des séquences codantes peuvent à leur tour être présentes entre un signal START et un codon STOP en phase (voir figure 2 pour le code génétique).

Plus un individu est adapté à son environnement, plus il a de chances de se reproduire. À chaque génération, la population est intégralement renouvelée. Une roulette biaisée selon la fitness des individus permet de déterminer le « parent » de chaque nouvel individu.
Pendant le processus de réplication, le génome peut subir différents types de mutations : des mutations ponctuelles (substitutions, petites insertions ou délétions) mais aussi des réarrangements à l'échelle du chromosome (duplications, délétions, translocations, inversions). La structure du génome est donc libre d'évoluer (nombre de gènes, taille du génome, ...) et on peut étudier l'émergence de différentes structures génétiques.

2.2. Du génotype au phénotype

Dans Aevol, le décodage du génotype est directement inspiré des processus de transcription et de traduction bactériens. Nous avons défini un ensemble de signaux qui, lorsqu'ils sont présents sur l'ADN, nous permettent d'identifier les séquences qui seront transcrites en ARNs et, sur celles-ci, les sous-séquences qui seront traduites en protéines. Ces protéines seront ensuite interprétées en termes de « fonctions biologiques » réalisées ou inhibées par la protéine.

2.2.1. Transcription du génome

Chez les bactéries, l'initiation de la transcription s'effectue en des sites particuliers, appelés promoteurs, où les ARN-polymérases reconnaissent une séquence consensus et commencent la synthèse de l'ARN. Dans Aevol, un promoteur est une séquence dont la distance de Hamming d avec une séquence consensus prédéfinie, est inférieure ou égale à d_{max}. La séquence que nous utilisons typiquement dans nos expériences comporte 22 bases : 0101011001110010010110 et on autorise jusqu'à $d_{max} = 4$ différences. Cette séquence est suffisamment longue pour que des séquences non-codantes n'aient qu'une faible probabilité de devenir codantes à la suite d'une mutation.
Le niveau d'expression e d'un ARN dépend de sa séquence promotrice. Plus le promoteur est proche de la séquence consensus, plus le niveau d'expression est élevé : $e = 1 - \frac{d}{d_{max}+1}$. Cette modulation de l'expression des gènes modélise de façon simple l'interaction entre l'ARN-polymérase et le promoteur, sans introduire de réseau de régulation[2].
Lorsqu'un promoteur est identifié, la séquence est transcrite jusqu'à ce qu'un terminateur soit rencontré. Les terminateurs doivent être plus fréquents que les promoteurs pour limiter le chevauchement des séquences transcrites. Nous avons donc défini les terminateurs comme des séquences capables de former des structures en tige-boucle, similaires aux terminateurs ρ-indépendants bactériens[3]. Dans nos expériences, les tailles typiquement utilisées sont de 4 pour la tige et de 3 pour la boucle, ainsi les terminateurs ont la structure $abcd***\overline{d}\overline{c}\overline{b}\overline{a}$, où $a, b, c, d = 0$ ou 1.

2.2.2. Traduction des ARNs

Les séquences transcrites (ARNs) ne conduisent pas systématiquement à la production d'une protéine. Comme pour la transcription, le processus de traduction débute et se termine lorsque le signal correspondant est rencontré. Ici, un signal de début de traduction est composé d'une séquence dite de Shine-Dalgarno, suivie, quelques bases plus loin, d'un codon START (voir le code génétique figure 2). Lorsque ce signal est rencontré, la séquence est lue codon par codon jusqu'à ce qu'un codon STOP soit trouvé dans le même cadre de lecture que le codon START. Le processus de traduction associe alors à chaque codon (ou triplet de bases), un « acide aminé » abstrait grâce à un code génétique et la séquence d'acides aminés forme la séquence primaire de la protéine (figure 2).
Comme chez les organismes réels, notre séquence génétique peut être lue suivant six cadres de lecture différents (trois sur chaque brin), ce qui permet aux organismes de présenter des gènes chevauchant (correspondant à des protéines différentes puisque lus sur des cadres de lecture différents).

2.2.3. Repliement des protéines et calcul du phénotype

Pour modéliser l'activité des protéines et le phénotype correspondant, nous avons défini une « chimie artificielle » simple [6] qui décrit le métabolisme d'un organisme dans un langage mathématique. Nous considérons qu'il existe un espace abstrait Ω de l'ensemble des processus méta-

[2] Une extension du modèle (RAevol) intègre un mécanisme explicite de régulation de l'expression des gènes [2, 3].
[3] Remarquablement, cette structure dite de « hairpin » permet de coder des terminateurs à la fois longs *et* fréquents.

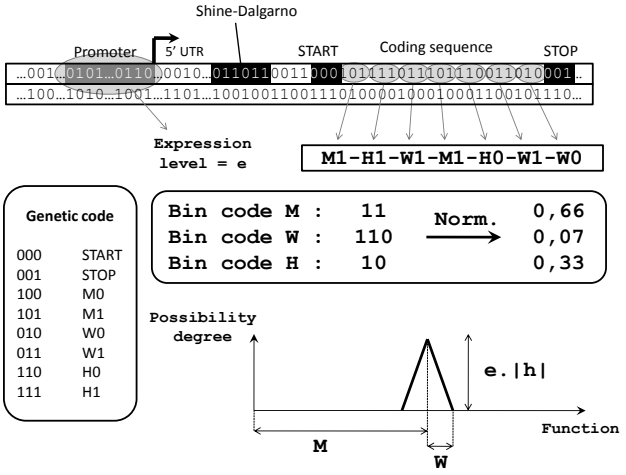


FIG. 2 – Schéma du processus de transcription-traduction-repliement dans Aevol. Les séquences transcrites sont celles qui commencent par un promoteur (séquence consensus) et finissent par un terminateur (structure tige-boucle), qui n'est pas sur la figure. Les séquences codantes (gènes) sont recherchées dans les séquences transcrites ; elles commencent par une séquence Shine-Dalgarno-START et se terminent par un codon STOP. Un code génétique artificiel (à gauche) est utilisé pour obtenir la séquence primaire de la protéine codée par un gène et un processus de « repliement » nous permet de calculer l'activité métabolique de la protéine (capacités fonctionnelles).

boliques possibles. Dans le modèle, $\Omega = [0, 1]$, un processus métabolique est alors un simple réel. Dans cet « espace métabolique », chaque protéine est impliquée dans un ensemble de processus, soit en contribuant à leur réalisation, soit en les inhibant. Cette contribution est décrite grâce à un formalisme de logique floue : une protéine peut activer ou inhiber un processus biologique avec un degré de possibilité compris entre 0 et 1 (positif pour une activation, négatif pour une inhibition). Une protéine est donc caractérisée par une fonction qui associe un degré de possibilité à chaque processus biologique. Pour des raisons de simplicité, nous utilisons des fonctions linéaires par parties ayant la forme de triangles isocèles (voir figure 2). Ainsi, trois nombres suffisent pour caractériser l'activité métabolique d'une protéine : la position m ($m \in \Omega$) du triangle sur l'axe fonctionnel, sa demi-largeur w et sa hauteur h (positive quand la fonction est réalisée par la protéine, négative quand elle est inhibée). La protéine contribue donc à la plage de processus métaboliques $[m - w, m + w]$, avec une préférence pour les processus les plus proches de m (pour lequel la plus grande efficacité h est atteinte). Ainsi, plusieurs types de protéines peuvent co-exister, allant de protéines très spécialisées et efficaces (faible w, fort h) à des protéines beaucoup plus polyvalentes et moins efficaces (fort w, faible h).

Le calcul de ces trois paramètres à partir de la séquence primaire de la protéine est l'étape qui correspondrait dans les vraies cellules au repliement de le protéine. Ici la séquence primaire de chaque protéine est décomposée en trois sous-séquences binaires entrelacées codant les valeurs des trois paramètres m, w et h. Par exemple le codon 010 (resp. 011) est traduit en l'Acide Aminé *W*0 (resp. *W*1), ce qui signifie qu'il contribue au paramètre W en ajoutant un bit 0 (resp. 1) à son code binaire. La séquence binaire correspondant à chaque paramètre est finalement interprétée comme un réel normalisé selon la longueur de la séquence et les valeurs possibles du paramètre.

Une fois toutes les protéines d'un organisme identifiées et caractérisées, leurs activités respectives sont combinées en utilisant les opérateurs de Lucasiewicz. L'ensemble flou qui en résulte représente le phénotype P de l'individu, il indique le degré avec lequel cet individu réalise chaque fonction biologique de Ω.

2.3. Environnement, adaptation et sélection

Dans Aevol, l'environnement est représenté par une cible phénotypique : l'ensemble flou E défini sur Ω qui représente le degré de possibilité optimal pour chaque fonction biologique. Pour évaluer un individu, on compare son phénotype P à la cible E. L'aire géométrique g entre ces deux ensembles représente l'« erreur métabolique » de l'individu (figure 3). Plus l'erreur métabolique est petite, meilleur est l'individu. Cette mesure pénalise aussi bien la sur- que la sous-réalisation de chaque fonction.

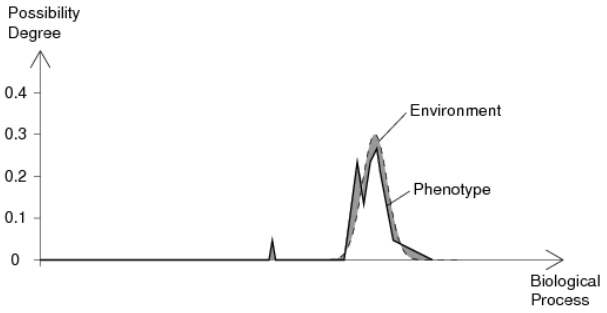


FIG. 3 – Mesure de l'adaptation d'un individu. Courbe pointillée : cible environnementale E. Courbe trait plein : phénotype P (profil métabolique obtenu en combinant toutes les protéines). Zone grisée : erreur métabolique g.

Chaque individu se voit attribuer une probabilité de reproduction en fonction de son erreur métabolique g et un tirage multinomial détermine le nombre de descendants effectif de chacun d'entre eux. Différentes méthodes de sélection sont disponibles, basés sur le rang de l'individu dans la population ou directement sur sa valeur d'adaptation [4]. Toutes les expériences mentionnées ici ont été réalisées avec des sélections sur le rang, sans croisement entre les individus.

2.4. Opérateurs génétiques

Pendant leur réplication, les génomes peuvent subir sept types de mutations génétiques, parmi lesquels trois sont locaux (substitution d'une base, insertion ou délétion de quelques bases) et quatre sont des réarrangements chromosomiques affectant des segments potentiellement longs du génomes (duplication, délétion, translocation et inversion). Les points de rupture de ces réarrangements sont tirés aléatoirement sur le chromosome selon une loi uniforme.

Les mutations affectent le génome mais n'ont pas nécessairement un effet phénotypique. Ainsi, une mutation ayant lieu dans une région non transcrite sera complètement neutre (sauf si elle crée un nouveau gène, ce qui est très rare). Les taux de mutations μ_i sont des paramètres du modèle, ils sont définis comme la probabilité par base et par réplication qu'un évènement de type i ait lieu.

Aevol est donc un modèle de génétique digitale dans lequel la structure des génomes est libre d'évoluer. Il intègre les principaux mécanismes impliqués dans l'expression et la modification du génome, introduisant un niveau intermédiaire entre le génotype et le phénotype et autorisant non seulement des opérateurs de mutations ponctuelles, mais aussi les réarrangements chromosomiques.

Ces particularités font d'Aevol un modèle particulièrement adapté à l'étude de l'organisation des génomes. Il permet de réaliser des campagnes expérimentales complètes dans différentes conditions expérimentales (*e.g.* différents taux de mutations) et d'observer comment les paramètres structuraux des génomes évoluent en fonction de ces conditions. Il est alors possible de vérifier la cohérence des résultats obtenus avec les différentes hypothèses proposées dans la littérature et d'essayer de comprendre les mécanismes à l'origine des phénomènes observés.

3. Une évolution typique dans Aevol

Aevol permet de mener des campagnes d'évolution expérimentale sur plusieurs dizaines de milliers de générations et d'analyser l'allure des génomes obtenus en fonction des paramètres. Si les structures finales peuvent être très différentes, le processus évolutif est quant à lui relativement stable d'une expérience à l'autre.
On observe ainsi une amélioration rapide de la fitness des individus dans les premières générations puis un ralentissement progressif. On notera que la fitness n'est jamais totalement stable et que des mutations avantageuses se produisent régulièrement, même après 500 000 générations.
L'évolution des individus s'accompagne de profondes modifications dans la structure de leur génome (figure 4). Dans un premier temps, la taille du génome augmente fortement pour passer des 5000 paires de bases initiales (initialisation par défaut dans aevol) à plusieurs dizaines de milliers de paires de bases en quelques centaines de générations. Le nombre de gènes et la taille des séquences non codantes augmentent aussi fortement. La deuxième phase se caractérise par une décroissance rapide de la taille du génome et du nombre de gènes, tandis que la taille des gènes, elle, continue de croître. Enfin, au cours de la troisième phase, la taille des génomes est stable. Par contre, l'organisme recommence à acquérir des gènes (mais plus modérément) tandis que la taille des séquences codantes augmente continûment.
Ainsi, dans un premier temps, les organismes augmentent rapidement la taille de leur répertoire génique, le plus souvent par duplication-divergence de gènes pré-existants. Ils sélectionnent ensuite les gènes les plus adaptés avant d'affiner leur répertoire génique en améliorant progressivement chacune de leurs séquences codantes.

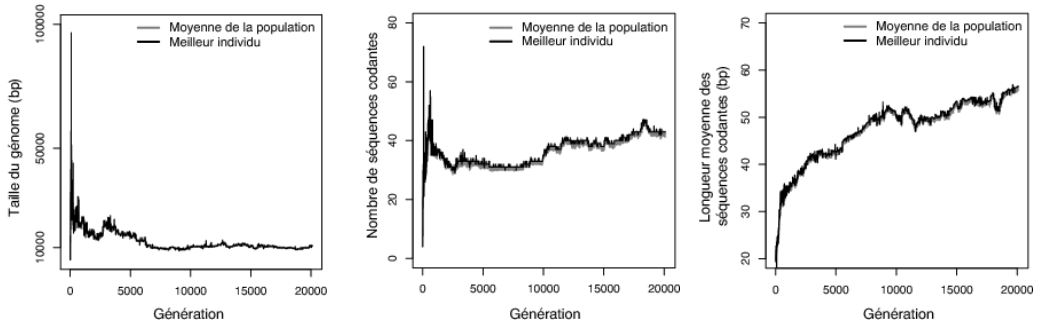
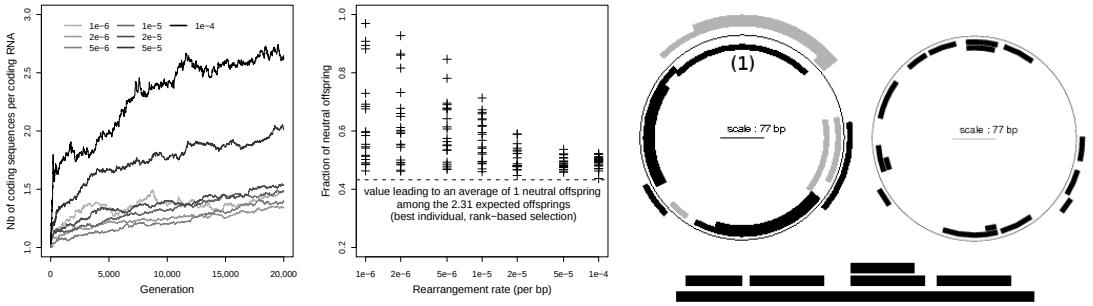


FIG. 4 – Évolution des paramètres structuraux des génomes pour une exécution « typique » de aevol. **(a)** Évolution de la taille du génome. **(b)** Évolution du nombre de séquences codantes (nombre de gènes). **(c)** Évolution de la taille moyenne des séquences codantes. D'après [8].

4. Résultats

Les différentes expériences que nous avons menées avec le modèle Aevol nous ont permis d'apporter des éléments de réponse à plusieurs questions ouvertes en biologie évolutive. En faisant varier les paramètres du modèle, nous avons observé de grandes variations dans l'organisation des génomes des individus. Nous avons ainsi constaté que, dans un environnement identique, une population évoluant avec un taux de mutations fort donnait naissance à des génomes beaucoup plus courts et compacts qu'une population sujette à des taux de mutations plus faibles [9]. Ce phénomène était déjà connu en ce qui concerne la quantité de séquences codantes sous la dénomination d'« error threshold » [7,13] ou de fardeau mutationnel [12], mais son extension à la quantité de non-codant constitue un résultat majeur du modèle.
De la même façon, nous avons pu établir un lien fort entre le caractère plus ou moins délétère des mutations et la compacité du génome [10].



(a) (b) (c)

FIG. 5 – **(a)** Nombre de gènes par ARN codant (contenant au moins un gène) au cours de l'évolution. Pour des raisons de clarté, chaque ligne correspond à la moyenne des valeurs de 21 simulations partageant le même taux de réarrangements. **(b)** Proportion de descendants non neutres du meilleur individu de chaque population après 20 000 générations. Les valeurs les plus hautes correspondent à des individus très robustes que l'on peut observer dans des populations qui n'ont pas suffisamment avancé dans le processus évolutif. **(c)** Génome du meilleur individu de la génération 20 000 d'une simulation typique avec des taux de mutations et de réarrangements élevés (1.10^{-4}). À gauche : ARNs (codants en noir, non-codants en gris). À droite : gènes. En bas : zoom sur l'opéron (1) avec ses 5 gènes.

Nous étudions à présent l'influence des taux de réarrangements sur l'organisation de la transcription. En effet, une analyse de la structure des ARNs montre que les variations de taille des génomes s'accompagnent de profondes différences dans la façon dont ils sont transcrits. Les génomes les plus longs présentent de très nombreux ARNs non-codants, leurs ARNs codants étant courts et ne codant généralement que pour une seule protéine. Les génomes courts, quant à eux, sont généralement transcrits en des ARNs beaucoup plus longs codant chacun pour plusieurs protéines, formant ainsi des *opérons* (figure 5). L'origine évolutive des opérons dans les génomes réels est une question ouverte en biologie [11].
Nous avons constaté qu'il existait un seuil de taux de réarrangements au-delà duquel les opérons deviennent la règle plutôt que l'exception. Cet effet de seuil est en fait le résultat de la combinaison de deux pressions antagonistes. Selon le phénomène d'*error threshold*, seuls les génomes courts peuvent être transmis fidèlement lorsque le niveau de variations génétiques est élevé. Par ailleurs, la sélection des individus les plus adaptés à l'environnement tend ici à favoriser ceux ayant beaucoup de gènes. La conjonction de ces deux pressions résulte ainsi en une pression de compaction des génomes et, *in fine*, en la formation d'opérons [14].

L'utilisation de Aevol nous permet ainsi de montrer les relations qui existent entre tous ces résultats. Ils traduisent en effet la nécessité, pour un organisme en évolution, de maintenir un équilibre entre robustesse et évolvabilité, entre capitalisation de l'acquis et exploration de nouvelles solutions. Ainsi, lorsque l'on analyse les meilleurs individus d'une population évoluée, on constate qu'ils partagent tous une caractéristique commune : leur nombre moyen de descendants neutres $F_\nu W$ (produit de la probabilité de reproduction neutre[4] F_ν et du nombre de descendants W – voir figure 5) est juste supérieur à 1.
Aevol a donc montré que les génomes ne sont pas façonnés uniquement par des pressions directes sur la fitness ou par des biais mutationnels. Ils sont aussi profondément structurés par des pressions indirectes (pressions de second ordre) dont celle pour atteindre un bon compromis entre exploration et exploitation. La compacité du génome est un levier d'ajustement de ce compromis car les génomes présentant plus de gènes et plus de non-codant subissent plus de réarrangements pouvant impacter la fitness [9].

[4] Cette probabilité peut être obtenue expérimentalement en effectuant 10 000 reproductions de l'individu et en comptant le nombre de descendants ayant la même fitness que le progéniteur.

5. Conclusion

En faisant évoluer, de façon réaliste, des organismes virtuels, Aevol nous permet d'étudier les mécanismes évolutifs. Aevol permet de retrouver de nombreuses caractéristiques structurelles des génomes d'organismes réels en faisant varier des paramètres tels que les taux de réarrangements ou la taille de la population. Il nous permet donc de proposer aux biologistes des hypothèses alternatives pouvant expliquer ces phénomènes. Le modèle Aevol peut donc être considéré comme un générateur d'hypothèses pour expliquer l'évolution de l'organisation des génomes.

D'un point de vue biologique, le modèle a vraisemblablement encore beaucoup à nous apprendre, nous projetons par exemple de mener des expériences parallèles sur le modèle et sur des organismes réels pour étudier l'évolution de la pathogénicité chez certaines bactéries. Une extension du modèle, RAevol, intègre un niveau explicite de régulation de l'expression des gènes. Cette extension nous a permis d'apporter des éléments expliquant la complexification des réseaux de régulation, montrant que celle-ci peut n'être qu'un simple effet de bord de la sélection d'un niveau adéquat de robustesse mutationnelle [3].

D'un point de vue informatique, l'identification de pressions de second ordre telles que nous avons pu les observer dans nos simulations pourrait ouvrir de nouvelles voies dans le domaine de l'optimisation par algorithmes génétiques. D'autre part, les données produites par le modèles peuvent servir de banc d'essai pour des algorithmes de découverte de connaissances [2].

Bibliographie

1. Christoph Adami. Digital genetics : unravelling the genetic basis of evolution. *Nat. Rev. Genet.*, 7(2) :109–118, February 2006.
2. Guillaume Beslon, David P. Parsons, Jose Maria Pena, Christophe Rigotti, et Yolanda Sanchez-Dehesa. From Digital Genetics to Knowledge Discovery : Perspectives in Genetic Network Understanding. *Intelligent Data Analysis Journal*, 14(2) :173–191, mars 2010.
3. Guillaume Beslon, David P. Parsons, Yolanda Sanchez-Dehesa, Jose Maria Pena, et Carole Knibbe. Scaling Laws in Bacterial Genomes : A Side-Effect of Selection of Mutational Robustness. In *BioSystems*, 2010. (À paraître).
4. Tobias Blickle et Lothar Thiele. A comparison of selection schemes used in evolutionary algorithms. *Evol. Comput.*, 4(4) :361–394, December 1996.
5. Zachary D. Blount, Christina Z. Borland, et Richard E. Lenski. Historical contingency and the evolution of a key innovation in an experimental population of escherichia coli. *Proceedings of the National Academy of Sciences*, 105(23) :7899–7906, June 2008.
6. P Dittrich, J Ziegler, et W Banzhaf. Artificial chemistries-a review. *Artif Life*, 7(3) :225–275, 2001.
7. M. Eigen. Self-organization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58 :456–523, 1971.
8. Carole Knibbe. *Structuration des génomes par sélection indirecte de la variabilité mutationnelle, une approche de modélisation et de simulation*. Thèse de doctorat, INSA-Lyon, 2006.
9. Carole Knibbe, Antoine Coulon, Olivier Mazet, Jean-Michel Fayard, et Guillaume Beslon. A long-term evolutionary pressure on the amount of noncoding DNA. *Mol. Biol. Evol.*, 24(10) :2344–2353, 2007.
10. Carole Knibbe, Olivier Mazet, Fabien Chaudier, Jean-Michel Fayard, et Guillaume Beslon. Evolutionary coupling between the deleteriousness of gene mutations and the amount of noncoding sequences. *J. Theor. Biol.*, 244(4) :621–630, 2007.
11. J. Lawrence. Selfish operons : the evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Curr. Opin. Genet. Dev.*, 9(6) :642–648, December 1999.
12. Michael Lynch. Streamlining and simplification of microbial genome architecture. *Annu. Rev. Microbiol.*, 60(1) :327–349, August 2006.
13. Gabriela Ochoa. Error thresholds in genetic algorithms. *Evolutionary Computation*, 14(2) :157–182, June 2006.
14. David P. Parsons, Carole Knibbe, et Guillaume Beslon. Importance of the rearrangement rates on the organization of genome transcription. In *Artif. Life*, août 2010. (À paraître).