Modèles computationnels de cellules biologiques pour l'évolution moléculaire et la nutrition

Carole Knibbe

N° Identificateur : 2021 HDR 001 Année 2021

HABILITATION A DIRIGER DES RECHERCHES

présentée devant

l'Institut National des Sciences Appliquées de Lyon
et l'Université Claude Bernard LYON I

Modèles computationnels de cellules biologiques pour l'évolution moléculaire et la nutrition

SPECIALITE : Biologie computationnelle

par
Carole KNIBBE

Soutenue le 8 janvier 2021 devant la Commission d'examen

(par ordre alphabétique)

Nicolas Bredèche, PU, Sorbonne Université, rapporteur

Hubert Charles, PU, INSA de Lyon, examinateur

Hidde De Jong, DR Inria Rhône-Alpes, rapporteur

Ingrid Lafontaine, MCU HDR, Sorbonne Université, rapporteur

Marie Caroline Michalski, DR INRAE, examinatrice

Hédi Soula, PU, Sorbonne Université, examinateur

Christophe Soulage, PU, Université Claude Bernard Lyon 1, examinateur

Jean-Daniel Zucker, DR IRD, examinateur

Laboratoire CarMeN (Cardiovasculaire, Métabolisme, Diabétologie et Nutrition, UMR INSERM U.1060, INRAE U1397, Université Lyon 1, INSA de Lyon)

Equipe Inria Beagle, Centre Inria Rhône-Alpes

Résumé

Ce mémoire présente les approches de modélisation et de simulation que j'ai développées, avec mes étudiantes et mes étudiants, mes collaborateurs et collaboratrices, pour étudier certaines propriétés des cellules vivantes. Il peut s'agit soit de propriétés relativement générales et difficilement accessibles à l'expérimentation directe (comme l'évolvabilité et/ou la robustesse mutationnelle que pourrait apporter un génome particulièrement dense en gènes, en opérons, ou en séquences répétées, toutes choses égales par ailleurs), soit de propriétés spécifiques pour lesquelles la modélisation doit guider la conception des expériences (comme l'existence ou non d'une étape limitante dans le trafic des triglycérides à l'intérieur des cellules de la muqueuse intestinale).

Selon les questions abordées, les modèles développés sont des modèles individu-centrés, des chaînes de Markov ou des systèmes d'équations différentielles simulés numériquement, ou encore des modèles hybrides, dans lesquels chaque individu de la population d'agents possède un réseau de réactions intracellulaires décrit par un système d'équations différentielles.

Les deux premiers chapitres présentent les contributions que mes étudiants, mes collaborateurs et moi avons apportées durant les 13 dernières années : en évolution moléculaire (au laboratoire LIRIS et dans l'axe "Models for Molecular Evolution" de l'équipe Inria Beagle) dans le chapitre I, et en nutrition (laboratoire CarMeN et dans l'axe "Computational cellular biochemistry" de l'équipe Inria Beagle) dans le chapitre II. Enfin, le troisième chapitre présente le projet de recherche que je souhaite animer autour de la modélisation mécanistique et quantitative des processus moléculaires, cellulaires et physiologiques qui gouvernent le destin des lipides alimentaires dans l'organisme.

Remerciements

Ce mémoire n'aurait simplement pas existé sans les étudiantes et étudiants qui ont brillamment réalisé les travaux présentés ici. Mes tout premiers remerciements vont donc à David, Stephan, Bérénice, Charles, Julie, qui m'ont fait assez confiance pour se lancer dans une thèse sous ma co-direction, qui se sont attaqués à des questions difficiles, qui ont fait preuve d'inventivité et de ténacité pour dépasser les difficultés techniques, qui n'ont pas ménagé leurs efforts pour obtenir des résultats solides, le tout avec curiosité scientifique, enthousiasme et bonne humeur. C'était un privilège de travailler avec des doctorants comme vous.

Je remercie également les étudiants de licence et de master qui ont pris le risque de partir avec moi sur des sujets exploratoires, et dont le travail a permis dégager des pistes prometteuses : Mathias Millet, Alice Joffard, Nicolas Comte, Ella Beaumann, Mathilde Dumond, Houleymatou Baldé, Tom Dusséaux, Juliette Geoffray, Damien Agopian, et Justine Antoine. Un grand merci également à Vincent Liard pour son précieux et rigoureux travail de développement pendant l'ADT aevol.

Je remercie très sincèrement mes rapporteurs, Ingrid Lafontaine, Hidde De Jong et Nicolas Bredèche, d'avoir accepté d'évaluer ce mémoire en dépit leurs emplois du temps très chargés, et ce d'autant plus que le mémoire couvre deux domaines d'application en biologie (l'évolution moléculaire et la nutrition). Un grand merci également à Hédi Soula, Jean-Daniel Zucker, Marie-Caroline Michalski, Christophe Soulage et Hubert Charles d'avoir accepté de participer au jury comme examinateurs, là aussi malgré une période bien chargée.

Je remercie Attila Baskurt puis Mohand-Saïd Hacid de m'avoir accueillie au laboratoire LIRIS de 2007 à 2017, et de soutenir les activités de recherche en biologie computationnelle au sein de ce laboratoire. De la même façon, je remercie Hubert Vidal de m'accueillir depuis 2017 au laboratoire CarMeN et de soutenir également les approches de biologie *in silico*. Je remercie également François Sillon, Patrick Gros puis Frédéric Desprez, directeurs successifs du centre Inria Rhône-Alpes, ainsi que l'Inria dans son ensemble, pour leur soutien à la biologie computationnelle, et à l'équipe Inria Beagle en particulier.

Je remercie également les institutions qui ont soutenu financièrement les projets présentés ici, à savoir le CNRS (projet exploratoire pluridisciplinaire inter-instituts), l'Inria (notamment pour ma délégation et pour l'ADT aevol), le programme européen FP7 "Evolving Technologies", l'Institut Rhône-Alpin des Systèmes Complexes (IXXI), l'INSA de Lyon (BQR) et BioSyl, la fédération de recherche lyonnaise en biologie des systèmes.

Tous ces projets sont nés de collaborations avec des chercheurs et chercheuses passionnés dont j'ai eu la chance de croiser le chemin, et que je remercie chaleureusement pour les interactions stimulantes : Dominique Schneider, Thomas Hindré, Dusan Misevic, François Taddéi, Charles Ofria, Françoise Odin, Sara Franceschelli, Eric Tannier, Gabriel Marais, Vincent Daubin, Laurent Guéguen, Priscilla Biller, Matthieu Boulesteix, Hugues Berry, Julie-Anne Nazare, Frédéric Carrière, Yves Jorand, Samuel Bernard, Kirsty Spalding et Peter Arner. Merci aussi à Fabien Crauste pour toutes les discussions et les perspectives, scientifiques et humaines, qu'elles m'ont apporté.

Je remercie tous les membres du LIRIS, de CarMeN et de l'équipe Inria Beagle pour leur aide, leurs perspectives et leurs conseils. Mention spéciale aux assistantes pour leur précieuse aide sur le plan administratif, et pour avoir fait en sorte que mes étudiants et moi-

même nous sentions bien accueillis : Catherine Lombardi, Brigitte Guyader, Caroline Ferri, Mabrouka Gheraissa, Marina Verges, Sokunthea Lim, Caroline Lothe, Florence Maillard, Claire Sauer et Laetitia Lécot-Gauthé.

Au LIRIS, je voudrais notamment remercier Alain Mille, Jean-François Boulicaut, et les membres de feu les équipes Silex et Turing : le bout de chemin que nous avons fait ensemble m'a beaucoup appris, et je vous suis reconnaissante de m'avoir laissé saisir l'opportunité d'embarquer dans le navire Inria Beagle.

Dans ce navire-là, je remercie particulièrement Hugues Berry, pas seulement pour la joie qu'il apporte dans la salle café, mais aussi pour l'aide apportée dans mon virage thématique vers la biochimie computationnelle.

A CarMeN, merci à Isabelle Delton, Céline Costaz, Françoise Hullin-Matsuda, Fabienne Laugerette, Nathalie Bernoud-Hubac, Marion Létisse, Chloé Robert, Marine Vincent et Armelle Penhoat de m'avoir associée à leurs travaux ou à leurs réflexions. Armelle, j'en profite pour souligner la chance que j'ai de partager des projets et un bureau avec une personne aussi bienveillante et avisée ! Merci aussi à Adina Lazar pour tous les bons moments passés dans le bureau du deuxième. Un grand merci à Annie Durand pour son aide très précieuse sur le plan technique, ainsi qu'à Maxence Rabia, Monique Estienne, Patrick Molière, Vanessa Euthine, Marie Aoustet et Valérie Large pour leur aide à mon arrivée à CarMeN. Merci aux lutins festifs, Marie-Michèle Boulet et Charline Buisson notamment, pour l'organisation des moments conviviaux et la bonne ambiance qu'ils ont apporté.

Je voudrais remercier tout particulièrement Marie-Caroline Michalski et Guillaume Beslon, qui m'accueillent dans leurs équipes respectives à CarMeN et à l'Inria.

Marie-Caroline, ton enthousiasme scientifique, ta bonne humeur et ton optimisme sont communicatifs ! A tes côtés, on se sent capable de déplacer des montagnes, ce qui explique peut-être que le caractère ambitieux des projets de recherche présentés ici ;-)

Guillaume, quel chemin parcouru ensemble depuis mon stage de fin d'études en 2003... En tant que chef d'équipe Inria, tu as réussi à créer un environnement scientifique et humain exceptionnel, où l'on se sent sereins, confiants, unis, curieux et enthousiastes, des scientifiques heureux en somme. Ta bienveillance et tes conseils avisés m'ont accompagnée et guidée, et dans mon nouveau rôle de directrice de département, il m'arrive fréquemment de me demander ce que toi, mon père, Barack Obama et Albus Dumbledore feraient à ma place ;-) Tu m'as encouragée et soutenue dans mon virage thématique, alors que cela aurait sûrement plus simple pour toi que je continue dans l'évolution *in silico*. Mais bien au-delà de ton rôle de chef d'équipe, tu es aussi un de mes plus proches amis, qui a été présent dans les bons comme dans les mauvais moments. Merci pour tout...

Je voudrais aussi remercier mes proches, collègues, famille et amis, avec qui j'ai partagé les hauts et les bas de la rédaction de ce manuscrit : Maman, mes soeurs Marion et Julie, mais aussi Guille et Emma Orsi, Christophe Soulage, Fabien Chaudier, Arnaud Trollé, Cyril Fayard, Laurent Soulère, Stéphane Chambert et Francesca Rebasti : merci pour vos encouragements, ils m'ont donné du coeur à l'ouvrage !

Dora, Ella et Alex, à 14, 9 et 5 ans, vous savez déjà ce que veut dire HDR : "Horriblement Difficile à Rédiger" ! Merci pour votre patience pendant le processus de rédaction, et surtout merci pour la joie que vous apportez dans la maison et dans mon coeur. Vous êtes des rayons de soleil !

Samuel, ce mémoire n'aurait probablement jamais été terminé sans toi. Non seulement tu as directement contribué à plusieurs des travaux et projets présentés ici, mais tu m'as aussi permis de rédiger ce mémoire en m'offrant ce qui me manquait le plus : du temps (ainsi que des tasses de thé, des carrés de chocolat et des bons dîners). Je ne pense pas pouvoir te remercier assez pour tout ce que tu as fait ! Plus généralement, merci de partager ma vie et d'y apporter autant de bonheur.

Table des matières

Curriculum vitae et liste des publications

Carole KNIBBE

40 ans
Née le 5 décembre 1980 à Hoorn (Pays-Bas)
Nationalité française

Maître de conférences (section CNU : 27)
Institut National des Sciences Appliquées (INSA) de Lyon
Département Biosciences
Laboratoire CarMeN (INSERM U1060/ INRA U1397)
Equipe Inria Beagle - Centre Inria Rhône-Alpes

carole.knibbe@insa-lyon.fr

Cursus universitaire et fonctions occupées

Institut National des Sciences Appliquées (INSA) de Lyon, France

Depuis sept. 2020	Directrice du département Biosciences de l'INSA de Lyon.
2017 - 2020	Directrice des études du parcours "Bioinformatique et Modélisation" (L3, M1, M2) de l'INSA de Lyon.
Depuis 2017	Maître de conférences au département Biosciences de l'INSA de Lyon, au laboratoire CarMeN (INSERM U1060/ INRA U1397) et dans l'équipe Inria Beagle (Centre Inria Rhône-Alpes).

Université Claude Bernard Lyon 1, Lyon, France

2015-2017	Responsable de l'équipe Beagle du laboratoire LIRIS.
2013-2015	En délégation Inria dans l'équipe Beagle du centre Inria Rhône-Alpes.
Depuis 2012	Bénéficiaire de la PES / PEDR.
Depuis 2010	Membre de l'équipe Beagle du centre Inria Rhône-Alpes.
2007-2017	Maître de conférences au département Informatique de l'Université C. Bernard Lyon 1 et au laboratoire LIRIS (UMR CNRS 5205).

Inserm, Paris, France

2006-2007	Post-doctorante au laboratoire de Génétique Moléculaire Evolutive et Médicale (Inserm U571, Faculté de Médecine Necker - Enfants Malades, Université Paris 5), dirigé par Miroslav Radman. Supervision: François Taddéi.

Institut National des Sciences Appliquées (INSA) de Lyon, France

2003-2006	Doctorat en Bioinformatique et Modélisation, laboratoires BF2I et PRISMa de l'INSA de Lyon. Titre du mémoire : *Structuration des génomes par sélection indirecte de la variabilité mutationnelle – une approche de modélisation et de simulation*. Directeurs de thèse : Jean-Michel Fayard et Guillaume Beslon. + Monitrice au département Biologie de l'Université Lyon 1.
2003	DEA en Analyse et Modélisation des Systèmes Biologiques, stage aux laboratoires BF2I et PRISMa de l'INSA de Lyon. Mention très bien, major de promotion.
2003	Ingénieur en Bioinformatique et Modélisation, Institut National des Sciences Appliquées (INSA) de Lyon, félicitations du jury. Formation pluridisciplinaire en mathématiques, en informatique et en biologie.

Activités d'enseignement

Depuis sept. 2020	Directrice du département Biosciences de l'INSA de Lyon.
2017-2020	Directrice des études du parcours "Bioinformatique et Modélisation" (niveaux L3, M1, M2) de l'INSA de Lyon.
2014-2017	Correspondante du département Informatique pour le montage de la nouvelle mention de Master de Bioinfomatique de l'Université Lyon 1 (départements partenaires : Biologie et Chimie-Biochimie).

Niveau	Années[1]	Type intervention	Enseignement	Formation	Volume horaire
L1	De 2003-2004 à 2005-2006	TD et TP	Biomathématiques et Biostatistiques	Licence de Biologie de l'Université Lyon 1	~45 h eqTD par an
L1	De 2007-2008 à 2009-2010	TP	Mise à niveau en informatique	Toutes les licences de Sciences et Technologies de l'Université Lyon 1	~48 h eqTD par an
L1	2007-2008	TD et TP.	Algorithmique et programmation en C	Licence d'Informatique de l'Université Lyon 1	~38 h eqTD

[1] Années particulières : 2011-2012: congé maternité, 2013-2014: délégation Inria, 2014-2015: demi-délégation Inria et congé maternité.

L1	2016-2017	CM et TP.	Applications en mathématiques et informatique (en C)	Licence d'Informatique de l'Université Lyon 1	~38 h eqTD par an
L2	De 2007-2008 à 2015-2016	CM, TD, TP. Co-responsable UE depuis 2008. Environ 180 étudiants à l'automne, 40 au printemps.	Algorithmique et Programmation en C	Licence d'Informatique de l'Université Lyon 1	~16 h eqTD la première année, puis ~120 h eqTD par an
L2	2007-2008	TP	Pratique d'Unix	Licence d'Informatique de l'Université Lyon 1	20 h eq TD
L2	2016-2017	TD et TP	Programmation fonctionnelle pour le web	Licence d'Informatique de l'Université Lyon 1	20 h eq TD
L3	De 2008-2009 à 2016-2017		Suivi de stages en entreprise	Licence d'Informatique de l'Université Lyon 1	3 h eq TD par an
3BIM (L3)	Depuis 2017-2018	CM et TP. Responsable de l'UE.	Algorithmique et programmation en Python	Filière Bioinformatique et Modélisation de l'INSA de Lyon	56 h eq TD par an
3BIM (L3)	Depuis 2017-2018	TP	Développement logiciel en C++	Filière Bioinformatique et Modélisation de l'INSA de Lyon	24 h eq TD par an
3BIM (L3)	2017-2018	CM et TP. Responsable de l'UE.	Bases de données	Filière Bioinformatique et Modélisation de l'INSA de Lyon	27 h eq TD par an
3BIM (L3)	Depuis 2017-2018	TP. Responsable de l'UE.	Architecture des ordinateurs et systèmes d'exploitation	Filière Bioinformatique et Modélisation de l'INSA de Lyon	10 h eq TD par an
4BIM (M1)	De 2003-2004 à 2005-2006	TP	Phylogénie moléculaire	Filière Bioinformatique et Modélisation de l'INSA de Lyon	~15 h eqTD par an
M1	2007-2008	TP	Intelligence artificielle	Master d'Informatique de l'Université Lyon 1	12 h eq TD par an

M1	2015-2016	TP	Programmation web	Master d'informatique de l'Université Lyon 1	~40 h eqTD
M1	2016-2017	TP Co-responsable UE.	Mise à niveau en informatique pour biologistes et biochimiste	Master de Bioinformatique de l'Université Lyon 1	~20 h eqTD
M1	2016-2017	CM et TP Co-responsable UE.	Programmation orientée-objet pour la bioinformatique	Master de Bioinformatique de l'Université Lyon 1	~50 h eqTD
4BIM (M1)	Depuis 2017-2018	Organisation de conférences-métiers	Carrières en bioinformatique	Filière Bioinformatique et Modélisation de l'INSA de Lyon	35 h eq TD par an
M1 et M2	2016-2017	Suivi de projets	Projets en bioinformatique	Master de Bioinformatique de l'Université Lyon 1	~15 h eqTD
M2	De 2007-2008 à 2016-2017	CM et TP	Informatique bio-inspirée	Master d'Informatique de l'Université Lyon 1	12 à 22 h eq TD par an
M2 et 5BIM	De 2008-2009 à 2012-2013	CM. Responsable UE. 15 à 20 étudiants chaque année.	Modélisation et simulation en biologie et médecine	Master Systèmes Complexes de l'ENS Lyon et Filière Bioinformatique et Modélisation de l'INSA de Lyon	15h eqTD par an
M2	De 2008-2009 à 2010-2011	CM et TP	Algorithmes évolutionnaires	Master "Conception et Intégration Multimédia, parcours programmation et développement de jeux vidéo" de l'Université Lyon 2	12h eq TD par an
M2	2009-2010	TP	Résolution de problèmes combinatoires	Master d'Informatique de l'Université Lyon 1	10h eqTD par an
M2	De 2011-2012 à 2016-2017	CM et TP. Co-responsable UE. Environ 50 étudiants par an.	Méthodologie scientifique et préparation à la recherche	Master d'Informatique de l'Université Lyon 1	~25 h eq TD par an

Publications

Edition d'ouvrages (2)

C. Knibbe, G. Beslon, D. Misevic D., editors (2019). ECAL 2017 Special Issue, *Artificial Life* 25(4).

C. Knibbe et al., editors (2017). *Proceedings of ECAL 2017, the 14th European Conference on Artificial Life*, Lyon, 4-8 Sept 2017, MIT Press.

Revues internationales (19)

C. Rocabert, G. Beslon, **C. Knibbe**, S. Bernard. (2020). Phenotypic Noise and the Cost of Complexity. *Evolution* 74 (10): 2221-2237.

C. Robert, L. Couëdelo, **C. Knibbe**, L. Fonseca, C. Buisson, E. Errazuriz-Cerda, E. Meugnier, E. Loizon, C. Vaysse and M.-C. Michalski. (2020). Rapeseed lecithin increases lymphatic lipid output and alpha-linolenic acid bioavailability in rats. *Journal of Nutrition*, in press.

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

M. Vincent, O. Ménard, J. Etienne, J. Ossemond, A. Durand, R. Buffin, E. Loizon, E. Meugnier, A. Deglaire, D. Dupont, J.-C. Picaud, **C. Knibbe**, M.-C. Michalski, A. Penhoat. (2020). Human milk pasteurisation reduces pre-lipolysis but not digestive lipolysis and moderately decreases intestinal lipid uptake in a combination of preterm infant in vitro models. *Food Chemistry* 329: 126927.

F. Laugerette, C. Vors, M. Alligier, G. Pineau, J. Drai, **C. Knibbe**, B. Morio, S. Lambert-Porcheron, M. Laville, H. Vidal, M.-C. Michalski. (2020). Postprandial Endotoxin Transporters LBP and sCD14 Differ in Obese vs. Overweight and Normal Weight Men during Fat-Rich Meal Digestion. *Nutrients* 12(6): 1820.

S. Danthine, C. Vors, D. Agopian, A. Durand, R. Guyon, F. Carriere, **C. Knibbe**, M. Létisse, M-C. Michalski. (2019). Homogeneous triacylglycerol tracers have an impact on the thermal and structural properties of dietary fat and its lipolysis rate under simulated physiological conditions. *Chem Phys Lipids* 225: 104815.

Y. Chena, M. Arnal-Levron, F. Hullin-Matsuda, **C. Knibbe**, P. Moulin, C. Luquain-Costaz, I. Delton. (2018). In vitro oxidized HDL and HDL from type 2 diabetes patients have reduced ability to efflux oxysterols from THP-1 macrophages. *Biochimie* 153:232-237.

C. Knibbe. Souvenirs from ECAL 2017: create, play, experiment, discover – revealing the experimental power of virtual worlds. (2018). *Adaptive Behavior* 26 (1):37-40.

C. Rocabert, **C. Knibbe**, J. Consuegra, D. Schneider, G. Beslon. (2017). Beware Batch Culture: Seasonality and Niche Construction Predicted to Favor Bacterial Adaptive Diversification. *PLoS Computational Biology* 13(3): e1005459.

P. Biller, L. Gueguen, **C. Knibbe**, E. Tannier. (2016). Breaking good: accounting for fragility of genomic regions in rearrangement distance estimation. *Genome Biology and Evolution* 8(5): 1427-1439.

S. Fischer, S. Bernard, G. Beslon, **C. Knibbe**. (2014). A model for genome size evolution. *Bulletin of Mathematical Biology* 76(9):2249-91.

B. Batut, **C. Knibbe**, G. Marais, V. Daubin. (2014). Reductive genome evolution at both ends of bacterial population size spectrum. *Nature Reviews Microbiology* 12(12): 841-850.

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

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

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

C. Knibbe, J.M. Fayard, G. Beslon. (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.

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

C. Knibbe, A. Coulon, O. Mazet, J.-M. Fayard, G. Beslon (2007). A long-term evolutionary pressure on the amount of non-coding DNA. *Molecular Biology and Evolution* 24(10):2344–2353.

Conférences internationales avec comité de lecture et actes (11)

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

C. Rocabert, **C. Knibbe**, J. Consuegra, D. Schneider, G. Beslon. (2017). Environmental seasonality drives digital populations towards stable cross-feeding. In C. Knibbe et al., editors, *Proceedings of ECAL 2017, the 14th European Conference on Artificial Life*, Lyon, 4-8 Sept 2017, MIT Press, pp. 368-369.

P. Biller, **C. Knibbe**, G. Beslon, E. Tannier. (2016). Comparative genomics of artificial life. In L. Bienvenu and N. Jonoska (ed.), *CiE 2016 (Computability in Europe)*, Paris, France, volume 9709 of *Lecture Notes in Computer Science*, Springer, pp. 35-44.

C. Knibbe, D.P. Parsons. (2014). What happened to my genes? Insights on gene family dynamics from digital genetics experiments. In *ALIFE 14 (14th Intl. Conf. on the Synthesis and Simulation of Living Systems)*, Sayama, H. et al. (ed.) New York, NY. MIT Press, Cambridge, Massachusetts, pp. 33-40.

G. Beslon, B. Batut, D. Parsons, D. Schneider, **C. Knibbe**. (2013). An alife game to teach evolution of antibiotic resistance. In *European Conference on Artificial Life*, Taormina. pp. 43-50. MIT Press.

D.P. Parsons, **C. Knibbe**, G. Beslon. (2012). The Paradoxical Effects of Allelic Recombination on Fitness. In *Artificial Life XIII*, East Lansing, Michigan (USA). ISBN 978-0-262-31050-5. 2012.

D.P. Parsons, **C. Knibbe**, G. Beslon. (2011). Homologous and nonhomologous rearrangements: Interactions and effects on evolvability. In *European Conference on Artificial Life (ECAL)*, T. Lenaerts, M. Giacobini, H. Bersini, P. Bourgine, M. Dorigo, R. Doursat ed. Paris. pp. 622-629. MIT Press.

C. Knibbe, D.P. Parsons, G. Beslon. (2011). Parsimonious modeling of scaling laws in genomes and transcriptomes. In *European Conference on Artificial Life (ECAL),* T. Lenaerts, M. Giacobini, H. Bersini, P. Bourgine, M. Dorigo, R. Doursat ed. Paris. pp. 414-415. MIT Press .

D.P. Parsons, **C. Knibbe**, G. Beslon. (2010) Importance of the rearrangement rates on the organization of transcription. Dans *Artificial Life*, MIT Press ed. Odense, Danemark. pp. 479-486. ISBN 978-0-262-29075-3.

G. Beslon, Y. Sanchez-Dehesa, D.P. Parsons, J.M. Pena, **C. Knibbe**. (2009). Scaling Laws in Digital Organisms. Dans *Proc. Information Processing in Cells and Tissues (IPCAT'09)*, Ascona, Switzerland. pp. 111-114.

C. Knibbe, G. Beslon, V. Lefort, F. Chaudier and J.-M. Fayard (2005). Self-adaptation of genome size in artificial organisms. In *Proceedings of ECAL 2005 (European Conference on Artificial Life)*, volume 3630 of *Lecture Notes in Artificial Intelligence*, Springer, pp. 423-432.

Autres revues, chapitres d'ouvrages (2)

C. Knibbe. (2013). L'évolution expérimentale in silico. In *Modéliser et simuler - Epistémologies et pratiques de la modélisation et de la simulation - Tome 1*, édité by F. Varenne et M. Silberstein, Editions Matériologiques, ISBN: 978-2-919694-19-8, pp. 581–610.

G. Beslon et **C. Knibbe**. (2010). Petits bricolages en évolution. In *Des mondes bricolés? Arts et sciences à l'épreuve de la notion de bricolage*, édité par F. Odin and C. Thuderoz, Presses Polytechniques et Universitaires Romandes, ISBN: 978-2-88074-901-9.

Workshops, posters, autres conférences

C. Robert, C. Buisson, L. Couëdelo, C. Knibbe, E. Meugnier, E. Loizon, L. Fonseca, F. Laugerette, C. Vaysse, M.-C. Michalski. (2020). "Impact of vegetable lecithin on lipid metabolism and the bioavailability of alpha-linolenic acid in rodents". Poster, *2020 online AOCS Annual Meeting*.

S. Danthine, C. Vors, A. Durand, F. Carrière, C. Knibbe, M. Létisse & M.-C. Michalski. (2020). Homogeneous triacylglycerol tracers can impact fat crystallization and its lipolysis rate under simulated physiological conditions. Poster, *2020 online AOCS Annual Meeting*.

C. Robert, C. Buisson, L. Couëdelo, C. Knibbe, E. Meugnier, E. Loizon, L. Fonseca, F. Laugerette, C. Vaysse, M.-C. Michalski. (2020). "Differential metabolic impact of natural food-grade emulsifiers rich in alpha-linolenic acid". *Nutrition 2020*, Seattle.

M. Vincent, O. Menard, J. Etienne, C. Knibbe, J. Ossemond, A. Durand, R. Buffin, E. Loizon, E. Meugnier, A. Deglaire, D. Dupont, J-C. Picaud, M-C. Michalski, Armelle Penhoat. (2019). "Impact of human milk pasteurisation on lipolysis kinetics and intestinal lipid uptake using a combination of in vitro models for preterm infants". Poster, 5th international congress of the European Milk Bank Association (EMBA 2019), Turin, Italie.

C. Rocabert, C. Knibbe, J. Consuegra, D. Schneider & G. Beslon (2016). « In Silico Experimental Evolution Highlights the Influence of Environmental Seasonality on Bacterial Diversification ». *2nd EvoEvo Workshop, (Satellite workshop of CCS2016)*, Septembre 2016, Amsterdam (Pays-Bas).

P. Biller, E. Tannier, G. Beslon, C. Knibbe (2016). In silico experimental evolution provides independent and challenging benchmarks for comparative genomics (titre court: Comparative genomics on artificial life). In *Journées ouvertes en Biologie, Informatique et Mathématiques (JOBIM) 2016*, Juin 2016, Lyon, France.

B. Batut, G. Beslon & C. Knibbe (2016). « Unexpected genome inflation and streamlining in variable environments ». In *Journées ouvertes de Biologie Informatique & Mathématiques (JOBIM) 2016*, Juin 2016, Lyon, France.

C. Rocabert, C. Knibbe, G. Beslon. (2015). Towards a Integrated Evolutionary Model to Study Evolution of Evolution. In *EvoEvo Workshop (Satellite workshop of ECAL 2015)*, 24 July 2015, York (UK).

S. Fischer, S. Bernard, G. Beslon, C. Knibbe (2013). Genome size evolution: challenging intuition with modelling". Dans *International Conference on Stochastic Models in Ecology, Evolution and Genetics*, Angers, France, 9-13 décembre 2013.

S. Fischer, C. Knibbe, S. Bernard & G. Beslon. (2013). "What is mutational bias? Comparing average DNA losses and gains is not sufficient for assessing spontaneous shrinkage". Poster, *SMBE 2013 : Society for Molecular Biology and Evolution,* Chicago, USA.

B. Batut, V. Daubin, G. Beslon C. Knibbe & G. Marais. (2013). "The evolutionary forces at work in the genome reduction of *Prochlorococcus marinus*". Poster, *SMBE 2013 : Society for Molecular Biology and Evolution,* Chicago, USA.

S. Fischer, C. Knibbe, S. Bernard & G. Beslon. "Studying the indirect impact of intrachromosomal rearrangements on genome structure... or how to prevent genomes from growing indefinitely without direct selective cost". Poster, *SMBE 2012 : Society for Molecular Biology and Evolution*, Juin 2012, Dublin, Irlande.

B. Batut, M. Dumond, G. Marais, G. Beslon & C. Knibbe (2012). « Simulating evolutionary scenarios to test whether they can induce reductive evolution ». Poster, *SMBE 2012 : Society for Molecular Biology and Evolution*, Juin 2012, Dublin, Irlande.

S. Fischer, C. Knibbe, S. Bernard, G. Beslon (2011). Unravelling laws of genome evolution with both mathematical and individual-based models. *8th European Conference on Mathematical and Theoretical Biology (ECMTB 2011)*, Krakow, Pologne, 28 juin - 2 juillet 2011.

D.P. Parsons, C. Knibbe, G. Beslon (2010). Influence of the rearrangement rates on the organization of genome transcription, Dans Integrative Post-Genomics 2010, Lyon.

D.P. Parsons, C. Knibbe, G. Beslon. (2010). Influence of the rearrangement rates on the organization of genome transcription. Dans JOBIM 2010, Montpellier.

D.P. Parsons, C. Knibbe, G. Beslon. (2010) Aevol : un modèle individu-centré pour l'étude de la structuration des génomes. Dans MajecSTIC 2010, Bordeaux.

D.P. Parsons, G. Beslon, C. Knibbe, Y. Sanchez-Dehesa, J.M. Pena. (2009). Evolution of scaling laws in artificial regulation networks. Dans Integrative Post-Genomics, Lyon. p. 22.

G. Beslon, S. Franceschelli, C. Knibbe. (2008). Petits bricolages en évolution. Dans "Génies de la Bricole et du Bricolage", colloque en hommage à Claude Lévy-Strauss, Lyon.

G. Kaneko, C. Knibbe, G. Beslon. (2007). Effect of bottlenecks on genome size : investigations by digital genetics. Dans Integrative Post-Genomics, Lyon.

Logiciels

- **Aevol (Artificial EVOLution)** : simulateur individu-centré d'évolution de génomes virtuels. Environ 50 000 lignes de code C++. Disponible soit sous forme de paquet Debian, soit via le site http://aevol.fr/. Participation au développement de 2003 à 2017.
- **Tewep (Transposable Elements Within Expanding Populations)** : simulateur individu-centré de la dynamique des éléments transposables dans le génome d'une population en expansion géographique. Environ 4000 lignes de code C++. Collaboration avec le généticien des populations Matthieu Boulesteix (Laboratoire de Biométrie et Biologie Evolutive UMR CNRS 5558).
- **Evo^{2}Sim** : simulateur multi-échelles développé avec Charles Rocabert et Guillaume Beslon durant le projet européen EvoEvo, prenant en compte la dynamique (ultra-rapide) des réseaux métaboliques, la dynamique (rapide) des réseaux de régulation génique, la dynamique (moyenne à lente) de partage des ressources dans l'écosystème et la dynamique (lente) de l'évolution des gènes et de la structure des génomes.
- **Simuscale** : simulateur multi-échelles de populations de cellules dont la division, différenciation ou mort dépend de la dynamique de leurs réseaux moléculaires internes. Développement du premier prototype avec Samuel Bernard (Institut Camille Jordan et équipe Inria "Dracula"). Actuellement développé par l'équipe Inria Dracula.

Encadrement

Thèses (5)

2008 – 2011 Co-direction à 50%, avec Guillaume Beslon (LIRIS) de la thèse de David Parsons, sur le thème "Indirect selection in Darwinian Evolution: Mechanisms and Implications". Thèse soutenue le 8 décembre 2011 à l'INSA de Lyon.

2010 – 2013 Co-direction à 50%, avec Guillaume Beslon (LIRIS) et Samuel Bernard (ICJ), de la thèse de Stephan Fischer, sur le thème "Modélisation de l'évolution de la taille des génomes et de leur densité en gènes par mutations locales et grands réarrangements chromosomiques". Thèse soutenue le 2 décembre 2013 à l'INSA de Lyon.

2011 - 2014 Co-direction à 30%, avec Gabriel Marais (LBBE) et Guillaume Beslon (LIRIS), de la thèse de Melle Bérénice Batut, sur le thème "Etude de l'évolution réductive des génomes bactériens par expériences d'évolution in silico et analyses bioinformatiques". Thèse soutenue le 21 novembre 2014 à l'INSA de Lyon.

2013 - 2017 Co-direction à 50%, avec Guillaume Beslon (LIRIS), de la thèse de M. Charles Rocabert, sur le thème "Studying evolution of bacterial micro-organisms by computer simulation approaches". Thèse soutenue le 17 novembre 2017 à l'INSA de Lyon.

2019 - présent Co-direction à 70%, avec Marie-Caroline Michalski (CarMeN), de la thèse de Mme Julie Etienne, sur le thème "Modélisation et simulation du flux de triglycérides alimentaires, de l'absorption entérocytaire à la sécrétion des chylomicrons".

Stages de M2 (6)

2011 Bérénice Batut: "Quand l'évolution darwinienne réduit la complexité: Etude de l'évolution réductive de certains génomes bactériens par expériences d'évolution *in silico*". Stage de M2 du parcours « Modélisation des systèmes complexes » du master Informatique Fondamentale de l'ENS de Lyon, encadré à 100%.

2014 Mathias Millet: "Artificial chemistries : on the influence of parameters on evolution of populations". Stage de M2 encadré à 50% avec Hédi Soula (50%, laboratoire CarMeN).

2015 Alice Joffard: "Modélisation et simulation de la dynamique des éléments transposables lors d'une expansion géographique". Stage de M2 encadré à 50%, avec Matthieu Boulesteix (50%, Laboratoire de Biométrie et Biologie Evolutive UMR CNRS 5558).

2016 Nicolas Comte: "Evolution expérimentale in silico : production de benchmarks pour les méthodes d'inférence phylogénétique". Stage de M2 encadré à 50% avec Eric Tannier (25%, Laboratoire de Biométrie et Biologie Evolutive UMR CNRS 5558) et Guillaume Beslon (25%).

2019 Ella Beaumann: "Modélisation compartimentale de cinétiques postprandiales cliniques". Stage de M2 encadré à 70%, avec Samuel Bernard (15%, Institut Camille Jordan et équipe Inria Dracula) et Marie-Caroline Michalski (laboratoire CarMeN).

2019 Julie Etienne: "Analysing and modelling the traffic of triglycerides through enterocytes". Stage de M2 encadré à 80%, avec Hugues Berry (20%, laboratoire LIRIS et équipe Inria Beagle).

Autres encadrements de stages (licence, M1) (6)

2011 Mathilde Dumond (ENS Lyon L3 biologie ENS Lyon), « Modélisation de l'évolution structurelle de génomes bactériens dans différents contextes environnementaux ».

2012 Houleymatou Baldé (L2 Informatique, Univ. Lyon 1), « Conception et réalisation d'un formulaire en ligne pour la saisie des paramètres du simulateur aevol ».

2016 Tom Dusséaux (M1 Informatique, Univ. Lyon 1): "Un nouveau mapping du génotype au phénotype pour aevol". Stage co-encadré à 50% avec Guillaume Beslon (50%).

2017 Juliette Geoffray (L3 Mathématiques et Informatique du Vivant, Univ. Lyon 1): "Modelling in vitro lipolysis of milk triglycerides".

2018 Damien Agopian (M1 INSA Lyon): "Influence des traceurs au 13C sur la digestion de matière grasse laitière : étude par lipolyse in vitro et modélisation". Co-encadrement à 50% avec Marion Létisse (CarMeN).

2019 Justine Antoine (M1 INSA Lyon): " Modelling lipid physiology and energy imbalance during weight loss". Co-encadrement à 20%, avec Samuel Bernard (80%, ICJ, Inria Dracula).

Coordination de projets de recherche et participation à des projets internationaux

2011-2012 Porteuse du projet exploratoire pluridisciplinaire inter-instituts (PEPII) CNRS: "Analyser, simuler and expérimenter l'évolution des génomes bactériens". Cinq laboratoires impliqués, en microbiologie, mathématiques, bioinformatique et informatique. Budget total de 50 000 euros.

2014-2016 Porteuse de l'Aide au Développement Technologique (ADT) Inria « aevol » (2014-2016). Financement d'un ingénieur confirmé sur deux ans et affectation de 40% d'un ingénieur senior du service SED de l'Inria.

2013-2016 Membre du projet européen Evoevo (http://www.evoevo.eu/, appel "Evolving Technologies" du programme FP7). Projet porté par Guillaume Beslon (équipe Inria Beagle, Lyon), avec des partenaires microbiologistes et informaticiens de l'Université d'Utrecht (Pays-Bas), l'Université de Valencia (Espagne), l'Université Joseph Fourier (Grenoble) et l'Université de York (Royaume-Uni).

2017-2019 Porteuse du projet "Lipuscale: Simulation hybride de la digestion et de l'absorption des triglycérides par l'intestin" soutenu à hauteur de 5000 euros sur deux ans par l'Institut Rhône-Alpin des Systèmes Complexes (IXXI). Deux laboratoires impliqués, en biologie de la nutrition et en mathématiques.

2018-2020 Porteuse du BQR "SiMoLip : Impact de l'organisation spatiale des matières grasses sur leur digestion et sur leur absorption : couplage entre approches expérimentales et simulations" soutenu par l'INSA de Lyon à hauteur de 24000 euros sur deux ans. Trois laboratoires impliqués, en biologie de la nutrition, en informatique et en matériaux.

2019 Co-porteuse, avec Samuel Bernard (ICJ et Inria Dracula) du projet "Modelling lipid physiology and energy imbalance during weight loss", soutenu à hauteur de 4000 euros sur un an par BioSyl, la fédération de recherche lyonnaise en biologie des systèmes. Deux laboratoires impliqués, en biologie de la nutrition et en mathématiques.

Implication dans la communauté scientifique, rayonnement

Responsabilités au sein du laboratoire LIRIS

2015-2017 Membre du conseil de laboratoire du LIRIS

2015-2017 Membre de la commission ZRR du LIRIS

2015-2017 Responsable de l'équipe Beagle du LIRIS.

Jurys de thèse

Examinatrice pour la thèse de M. Anton Crombach, 2009, Utrecht University (Pays-Bas).

Examinatrice pour la thèse de M. Antoine Frénoy, 2014, Université Paris Descartes, Paris.

External examiner pour la thèse de M. Gael Jalowicki, 2016, University College Dublin, Irlande.

Comités de sélection, comités d'attribution de financement

2007-2010 Présidente du Comité de pilotage de l'Institut Rhône-Alpin des Systèmes complexes (IXXI). Comité constitué de 12 membres élus. Il participe aux décisions et orientations scientifiques, et évalue les demandes de financement émanant des chercheurs. Environ 100 000 euros par an sont attribués en soutien à des projets de recherche ou à des événements scientifiques, en fonction des recommandations du CP.

2010-2012 Vice-Présidente du Comité de pilotage de l'IXXI.

2014 Membre du comité de sélection pour un poste de maître de conférences au département Biosciences de l'INSA de Lyon, printemps 2014.

2016 Membre du comité de sélection pour un poste de maître de conférences en bioinformatique au sein de l'UFR de médecine de l'Université Paris Diderot.

2016-2018 Membre de la Commission de Développement Technologique du centre Inria Rhône-Alpes. Evaluation de demandes de financement pour des projets de développement logiciel issus des activités de recherche des équipes Inria (CDD d'ingénieurs).

2017-2019 Membre du comité des études doctorales du centre Inria Rhône-Alpes. Evaluation de demandes de financement de thèses dans les équipes Inria du centre.

2018 Membre du comité de sélection pour un poste de maître de conférences au département Biosciences de l'INSA de Lyon.

Organisation de conférences internationales

Scientific chair de la conférence ECAL 2017 (European Conference on Artificial Life), Lyon, 4-8 septembre 2017.

Comités de programmes de conférences internationales

Membre du comité de programme de ECCB 2012 (European Conference on Computational Biology)

Membre du comité de programme de ECCB 2014 (European Conference on Computational Biology)

Membre du comité de programme de ALIFE 2014 (International Conference on the Synthesis and Simulation of Living Systems)

Membre du comité de programme de ECAL 2015 (European Conference on Artificial Life)

Membre du comité de programme de ALIFE 2016 (International Conference on the Synthesis and Simulation of Living Systems)

Reviewer pour ECCB 2016 (European Conference on Computational Biology)

Conférences et séminaires invités

Séminaire de l'équipe Inria TAO, "Lessons from the evolution of artificial genomes", Orsay, le 6 février 2007

Séminaire du laboratoire Matière et Systemes Complexes, "Lessons from the evolution of artificial genomes", Paris, le 2 avril 2007

Séminaire de Modélisation du Vivant (Semovi), "Evolvability, robustness and genome structure: where is the link?", Lyon, le 27 juin 2007

Santa Fe Institute (USA), How the topology of the protein network influences the evolution of genome structure in artificial organisms", Santa Fe, USA**,** le 12 juillet 2007.

Ecole avancée Biodiversité et ressources biologiques, "Evolution of evolution : molecular mechanisms, models and virtual experiments", Lyon, juin 2008.

Séminaire du « Bioinformatics group », Utrecht University, "Evolution in action: robustness and evolvability in digital genomes and gene networks", Utrecht (Pays-Bas), le 22 avril 2009.

York Centre for Complex Systems, "Evolution in action: robustness and evolvability in digital genomes and gene networks", York, UK, le 27 novembre 2009.

Conférence EvoLyon 2009, "Robustesse et évolvabilité des génomes", Lyon, le 25 novembre 2009.

Ecole thématique interdisciplinaire d'échanges et de formation en biologie de Berder 2012, "Evolution réelle de génomes artificiels, ou évolution artificielle de génomes réels ?", Berder, 2012.

Journée de bilan du RTRA Finovi, "PEACE: Parallel Experimental and Computational Evolution of virulence in Legionella pneumophila", juin 2013.

Conférence EvoLyon 2013, "L'évolution expérimentale in silico", Lyon, 21 novembre 2013.

Séminaire du laboratoire Laboratoire Interdisciplinaire de Physique, "Computational and mathematical models for the evolution of genomic architecture… or: How really bad our intuition is, when genome evolution is concerned", Grenoble, septembre 2014.

International conference on Theoretical Approaches for the Genome and the Proteome, "What happened to my genes? Insights on gene family dynamics from digital genetics experiments", Chambéry, 4-5 décembre 2014.

Center for Research and Interdisciplinarity, Université Paris Descartes, "Genome size evolution: challenging intuition with modelling", Paris, novembre 2014.

Workshop du Laboratoire International Associé EvoAct (Evolution in action) "Modeling regulation and metabolism in MISEEM (Multi-scale In Silico Experimental Evolution Model)", Autrans, avril 2016.

Séminaire de Modélisation du Vivant (Semovi), "Insights on genome dynamics from in silico experimental evolution and mathematical modelling", Lyon, septembre 2016.

Conférence Jacques Monod "Evolutionary Genomics and Systems Biology: Bringing Together Theoretical and Experimental Approaches", "Genome evolution: challenging intuition with modelling and simulation", Roscoff, 10-14 octobre 2016.

Symposium de Biologie des Systèmes de Sorbonne Université, "Beware batch culture: Seasonality and niche construction predicted to favor bacterial adaptive diversification", 1er décembre 2017.
Gordon Conference 2019 "Organismal, Cellular, Molecular and Theoretical Approaches to Understanding Evolution". "The dynamics of innovation in digital viral-like genomes". Easton, MA, USA, 9-14 juin 2019.

Introduction

Ce mémoire présente mes activités de recherche depuis 2007, année de mon recrutement comme maître de conférences, ainsi que le projet de recherche que je souhaite animer pour les dix à quinze prochaines années. L'exercice n'est pas tout à fait trivial, tant pour moi que pour mes lecteurs, du fait du virage thématique que j'ai opéré en 2017 à l'occasion de ma mutation de l'Université Lyon 1 à l'INSA de Lyon. L'opportunité de prendre la direction du parcours "Bioinformatique et Modélisation" de l'INSA s'est accompagnée de celle de passer d'un laboratoire d'informatique, le LIRIS, à un laboratoire de recherche biomédicale, le laboratoire CarMeN. Si les questions biologiques ont beaucoup changé — de l'évolution moléculaire à la nutrition —, l'approche méthodologique est restée celle de la modélisation et de la simulation de systèmes biologiques, compétence centrale de l'équipe Inria Beagle, dont je faisais partie avant 2017 et dont je suis restée membre. Cette équipe travaille sur deux axes : l'évolution expérimentale *in silico* (simulations d'évolution moléculaire) et la biologie cellulaire computationnelle (simulations de processus moléculaires à l'intérieur et entre les cellules). Lors de mon virage thématique, je suis passée du premier au second axe en interne de l'équipe Inria.

Mon domaine de recherche est donc la biologie computationnelle, définie comme la conception de modèles et de simulations visant à mieux comprendre les systèmes biologiques et les interactions complexes au sein de ceux-ci. Méthodologiquement, les modèles que je développe sont, selon les cas, des modèles individus-centrés, des chaînes de Markov ou des systèmes d'équations différentielles. Quelque soit le type de formalisme, ils sont en général :

- mécanistiques – c'est-à-dire construits à partir des connaissances des mécanismes bio-physico-chimiques qui régissent le système (contrairement aux modèles dits empiriques ou phénoménologiques qui décrivent des corrélations indépendamment des mécanismes sous-jacents),
- dynamiques – c'est-à-dire caractérisés par un état qui change en fonction du temps, de façon déterministe ou stochastique selon le formalisme choisi,
- pour certains d'entre eux, ils sont également multi-échelles – c'est-à-dire mettant simultanément en jeu plusieurs échelles de temps et d'espace, par exemple une échelle moléculaire rapide pour le métabolisme à l'intérieur d'une cellule, une échelle moléculaire plus lente pour l'expression des gènes, et l'échelle de la population de cellules pour les processus de division, de mort cellulaire et d'évolution.

Ces modèles peuvent avoir plusieurs usages, et l'usage visé peut influencer la structure du modèle lors de sa conception. Par exemple, lorsque le système biologique est régi par des mécanismes de contrôle multiples, rendant le raisonnement intuitif difficile, on peut avoir besoin de connaître l'effet d'un mécanisme isolé, ou de déterminer lesquels sont les plus critiques. On va dans ce cas construire un ou plusieurs modèles minimaux, chacun comportant un seul mécanisme ou un petit nombre de mécanismes. Ces modèles minimaux fournissent alors des cadres conceptuels calculables, qui permettent de tester si le raisonnement intuitif prédit correctement ou non les conséquences de l'ensemble d'hypothèses. Les modèles que j'ai développé ou contribué à développer en évolution moléculaire étaient de ce type.

Un autre usage visé peut être la prédiction quantitative, auquel cas on peut parfois se contenter d'un modèle phénoménologique plutôt que mécanistique. Ce sont des modèles quantitatifs que je propose de développer en nutrition dans mon projet de recherche. Ces modèles quantitatifs sont en général applicables à un ensemble plus restreint de systèmes biologiques que les modèles minimaux (une voie métabolique particulière chez une espèce en particulier par exemple), mais ils peuvent permettre de guider le design des expériences futures, ou de façon plus ambitieuse, de réaliser des essais précliniques *in silico*. (Kovatchev et al. 2009) est par exemple un premier pas dans cette direction pour le système glucose-insuline. Quand les modèles quantitatifs sont aussi mécanistiques, ils permettent de surcroît d'identifier des processus physiologiques candidats pour expliquer les effets observés.

Un troisième usage possible, applicable tant pour les modèles minimaux que pour les modèles quantitatifs, est la réalisation (virtuelle) d'expériences impossibles, pour des raisons techniques ou éthiques. En évolution moléculaire, nous avons par exemple simulé l'évolution de génomes en l'absence totale de mutations ponctuelles, pour déterminer si l'adaptation à un nouvel environnement est possible uniquement avec des réarrangements chromosomiques. Nous avons aussi simulé l'évolution de réseaux de régulation en présence mais aussi en l'absence de coût énergétique à la synthèse des protéines. En nutrition, nous avons simulé l'effet de scénarios extrêmes concernant les taux de stockage et de mobilisation des lipides adipeux.

Ce mémoire est organisé en trois chapitres, qui décrivent plus en détail comment ces différents usages des modèles se déclinent dans mes travaux et projets de recherche, et quels éclairages ils ont permis ou permettront d'apporter en évolution moléculaire et en nutrition. Les deux premiers chapitres présentent les contributions que mes étudiants, mes collaborateurs et moi avons apportées durant les 13 dernières années : en évolution moléculaire dans le chapitre I, et en nutrition dans le chapitre II. Le premier chapitre est naturellement significativement plus long que le second, puisqu'il couvre une période plus longue (2007-2017). Le troisième chapitre présente le projet de recherche que je souhaite animer autour de la modélisation mécanistique et quantitative des processus moléculaires, cellulaires et physiologiques qui gouvernent le destin des lipides alimentaires dans l'organisme.

Chapitre I : Contributions en évolution moléculaire

> *Proof-of-concept models can both bring to light hidden assumptions present in verbal models and generate counterintuitive predictions. When a verbal model is converted into a mathematical one, casual or implicit assumptions must be made explicit; in doing so, any unintended assumptions are revealed. Once these hidden assumptions are altered or removed, the predicted outcomes and resulting inferences of the formal model may differ from, or even contradict, those of the verbal model. This benefit of mathematical models has brought clarity and transparency to virtually all fields of evolutionary biology. Additionally, in spite of their abstract simplicity, proof-of-concept models, much like simple, elegant experiments, have the capacity to surprise. Even formalizations of seemingly straightforward verbal models can yield outcomes that are unanticipated using a verbal chain of logic.*
>
> (Servedio et al. 2014)

Les organismes vivants sont contraints dans leur organisation moléculaire par les lois de la physique, mais aussi par le fait qu'ils sont engagés dans un processus évolutif. D'une part, tous les états physiquement possibles ne donnent pas les mêmes chances de survie et de reproduction. D'autre part, partant d'un état donné, seuls certains états sont accessibles par une étape de mutation. Ces contraintes évolutives délimitent des trajectoires possibles parmi l'ensemble des états physiquement possibles.

Mes travaux de recherche en évolution moléculaire ont visé à identifier des contraintes évolutives qui façonnent les grandes propriétés d'organisation des génomes (nombre de gènes, quantité d'ADN non codant, existence d'opérons...) et des réseaux de gènes (réseaux métaboliques et réseaux de régulation). Comment ces propriétés ont-elles émergé ? Confèrent-elles un avantage sélectif, ou ne sont-elles que des effets secondaires fortuits d'événements contingents ? A quels paramètres du processus évolutif sont-elles sensibles ?

Pour contribuer à ces questions, mes étudiantes et étudiants, mes collaborateurs et moi avons utilisé une approche encore originale, appelée évolution expérimentale *in silico*. Cette approche consiste à développer des modèles computationnels de l'évolution de type "proof-of-concept" au sens de (Servedio et al. 2014). Ces modèles n'ont pas pour but de faire des prédictions quantitatives pour un système biologique particulier, mais de formaliser et de rendre calculable un modèle conceptuel, verbal, pour tester si ce modèle verbal se comporte réellement comme l'expérience de pensée voudrait le croire. Ainsi, il s'agit de formaliser l'ensemble d'hypothèses du modèle verbal, et de générer l'univers des possibles dans le cadre de cet ensemble d'hypothèses : "life as it could be", comme on dit dans le domaine de la vie artificielle (Langton et al. 1991). On peut alors déterminer quelles sont les hypothèses qui sont cruciales pour l'émergence ou le maintien de telle ou telle propriété d'organisation à l'échelle du temps évolutif.

Le chapitre présente tout d'abord plus en détail l'approche d'évolution expérimentale *in silico*, puis décrit les éclairages de l'évolution de l'organisation des génomes et des réseaux, obtenus

par cette approche avec mes collaborateurs, mais aussi et surtout avec les doctorant·e·s et les stagiaires que j'ai co-encadrés durant mes 10 ans au laboratoire LIRIS et dans l'équipe Inria Beagle.

I.1 L'évolution expérimentale *in silico*

Cette section (I.1) est extraite de C. Knibbe. (2013). L'évolution expérimentale in silico. In Modéliser et simuler - Epistémologies et pratiques de la modélisation et de la simulation - Tome 1, édité par F. Varenne et M. Silberstein, Editions Matériologiques, ISBN: 978-2-919694-19-8, pp. 581–610.

Comme la préservation des tissus mous est rare dans les fossiles, la paléontologie fournit une connaissance précieuse mais limitée du passé du monde vivant. La phylogénie moléculaire est une méthode alternative qui consiste à comparer des séquences d'ADN, d'ARN ou de protéines issues d'espèces actuelles, et à utiliser des méthodes informatiques pour construire une histoire évolutive plausible, prenant la forme d'un arbre phylogénétique. Il est alors possible d'inférer les séquences ancestrales et les mutations qui ont été fixées[2] dans les différentes branches de l'arbre. Cependant, avec ces méthodes de reconstruction phylogénétique, on ne sait pas encore dire si une mutation donnée a été fixée parce qu'elle apportait un avantage sélectif, parce qu'elle était située à proximité d'une autre mutation avantageuse dans la séquence d'ADN, ou encore par pur hasard. Autrement dit, ces méthodes ne permettent pas de discriminer, dans l'histoire évolutive, ce qui relève de la nécessité de ce qui relève de la contingence. De plus, ces méthodes reposent sur des hypothèses, très spéculatives et difficiles à tester, concernant les forces évolutives qui gouvernent l'apparition spontanée des mutations et leur fixation dans les populations. Par exemple, certaines méthodes reposent sur le principe de parcimonie, c'est-à-dire qu'elles cherchent une histoire évolutive qui minimise le nombre de mutations dans l'arbre. Le scénario le plus simple n'est pourtant pas forcément le chemin réellement suivi par l'évolution, qui procède par succession d'essais-erreurs (mutation-sélection) et non par planification — ou, pour reprendre la métaphore de François Jacob, comme un bricoleur et non comme un ingénieur (Jacob 1977).

Le biologiste de l'évolution Stephen Jay Gould suggérait une approche plus directe pour comprendre les propriétés des systèmes vivants, qu'il appelait "rejouer le film de la vie" : "You press the rewind button and, making sure you thoroughly erase everything that actually happened, go back to any time and place in the past — say, to the seas of the Burgess Shale. Then let the tape run again and see if the repetition looks at all like the original." (Gould & Lewontin 1979). L'idée d'observer directement l'évolution en action, avec des répétitions pour distinguer la nécessité de la contingence, a donné naissance à un nouveau champ en biologie, appelé évolution expérimentale. De nombreuses équipes mènent des expériences d'évolution en laboratoire sur diverses espèces de plantes, de vertébrés, mais aussi et surtout de micro-organismes, du fait de leur reproduction rapide. Des techniques moléculaires avancées permettent à ces biologistes d'analyser finement l'évolution de

[2] Une mutation est dite fixée dans une population quand tous les individus de la population portent cette mutation. Le processus de fixation prend du temps car une mutation apparaît en général au départ dans un seul individu. La fréquence de la mutation dans la population peut augmenter sous l'effet de la sélection darwinienne ou simplement sous l'effet de fluctuations aléatoires (dérive génétique).

propriétés cellulaires spécifiques, comme la résistance à un stress thermique (Lenski & Bennett 1993) ou la capacité d'utiliser le citrate en présence d'oxygène (Blount et al. 2008; Blount et al. 2012).

Cependant, l'évolution expérimentale est une technique coûteuse qui ne peut pas être menée sur des centaines de souches, et il n'est pas toujours aisé de généraliser les résultats d'une expérience à d'autres contextes ou d'autres espèces. C'est pourquoi, dès 1992, le biologiste de l'évolution John Maynard Smith complétait la proposition de Stephen Jay Gould de la manière suivante : "So far, we have been able to study only one evolving system [la vie sur Terre] and we cannot wait for interstellar flight to provide us with a second. If we want to discover generalizations about evolving systems, we will have to look at artificial ones." (Smith 1992). Il est donc utile de développer des simulations informatiques, ou une "évolution expérimentale *in silico*" (EEIS), où des organismes artificiels sont créés, mutent et se reproduisent dans l'ordinateur. Dans ces expériences *in silico*, la connaissance des événements évolutifs est exhaustive (y compris celle des mutations qui n'auraient pas été fixées), les expériences peuvent être répliquées de nombreuses fois, et les différents paramètres susceptibles d'influencer l'évolution peuvent être contrôlés, apportant ainsi un éclairage complémentaire à celui des expériences à la paillasse (Hindré et al. 2012).

I.1.1 Positionnement de l'approche

I.1.1.a EEIS et vie artificielle

D'un point de vue épistémologique, la démarche de l'évolution expérimentale *in silico* s'inscrit dans celle de la vie artificielle. Ce champ a pris forme à la fin des années 80 lorsqu'une centaine de biologistes, chimistes, physiciens et informaticiens, travaillant indépendamment sur des sujets proches, se sont retrouvés au premier "Atelier interdisciplinaire sur la synthèse et la simulation des systèmes vivants", sous l'impulsion de Chris Langton, avec l'aide du Centre d'études des phénomènes non linéaires de Los Alamos et du Santa Fe Institute. La vie artificielle vise à inférer des principes universels sous-jacents à tout système vivant, en créant des systèmes artificiels qui capturent partiellement la complexité du vivant pour la rendre accessible à de nouvelles formes d'expérimentation. Ces systèmes artificiels permettent en effet de contrôler précisément les paramètres, de répliquer facilement les expériences, et d'accéder à toutes les données pertinentes pour analyser les résultats (Miller 1995). L'évolution est une thématique centrale dans ce contexte : (Bedau et al. 2001), dans un article intitulé "Open problems in artificial life", fixent l'objectif de "déterminer ce qui est inévitable" dans l'évolution d'un système vivant, c'est-à-dire de dégager ce qui est général, reproductible, de ce qui est contingent. Il s'agit de "réaliser les expériences que la méthode scientifique nous dicte, mais que nous ne pouvons pas faire avec les échelles de temps et d'espace de structures matérielles comme les cellules elles-mêmes" (Forbes 2005).

Dans la communauté de la vie artificielle, le statut de ces expériences *in silico* est très débattu. Certains y voient des nouvelles formes de vie (strong Alife), alors que d'autres les considèrent comme des simulations du réel (weak Alife). J'adopte pour ma part une position pragmatique. Selon moi, ces expériences ne sont pas tant réalisées pour *prouver* (que tel mécanisme est à l'origine de telle caractéristique des cellules vivantes), que pour "stimuler l'intuition" (Rennard 2002) et dégager des mécanismes potentiels qu'il faudra ensuite confronter au réel. De ce point de vue, les expériences réalisées dans l'ordinateur ne sont pas une fin en soi mais simplement une étape permettant de *générer des hypothèses à la fois originales et plausibles*. Dès lors, il importe peu que l'on considère les objets informatiques

manipulés comme des organismes virtuels, des modèles d'organismes ou des simulations d'organismes. Comme le suggère Volker Grimm, ces objets informatiques ont pleinement rempli leur rôle lorsqu'on peut les oublier en eux-mêmes (et donc évacuer la question de leur statut) pour ne plus parler que des mécanismes plus généraux qu'ils ont permis de mettre à jour : "The decisive thing with modelling is not the model per se, but what the model and working with the model does to our mind. It could even be argued that a criterion to determine good models is that they are no longer needed afterwards. If the whole process of modelling has succeeded, something will have happened in our head, namely that an understanding of relationships has emerged. We should then be in a position to communicate our insights to others without referring to the model." (Grimm 1999).

I.1.1.b EEIS et modélisation individu-centrée, émergence, immergence

Les modèles utilisés en évolution expérimentale *in silico* sont de type "individu-centrés", c'est-à-dire que tous les individus de la population sont explicitement représentés et que chaque individu est caractérisé par les valeurs que prennent ses attributs (par exemple, la longueur de son génome, sa probabilité de reproduction). Le concept d'émergence, en général résumé par la maxime "le tout est plus que la somme des parties", est un concept central des approches individu-centrées et de la vie artificielle. Cependant, lorsque ces approches sont utilisées pour modéliser l'évolution, ce n'est pas tant le tout, le collectif, qui est l'objet d'intérêt. Ce sont davantage les parties, c'est-à-dire les individus et leurs propriétés après plusieurs milliers de générations. Plus exactement, il s'agit de comprendre comment le fait d'appartenir à une population qui évolue (et donc, de subir la sélection naturelle) façonne les individus. On s'intéresse donc à la façon dont le tout agit sur les parties, en renversant le concept d'émergence. Si les parties acquièrent des propriétés inattendues dans le contexte du tout — des propriétés "immergentes", "micro-émergentes" ou "localement émergentes" selon les auteurs — , alors les parties sont plus que simplement des parties. Comme celui de l'émergence, le concept d'immergence requiert la présence explicite du niveau micro en plus du niveau macro. C'est pour cela que les approches individu-centrées sont particulièrement adaptées à l'étude de ces phénomènes.

I.1.1.c EEIS et algorithmes évolutionnaires

Les modèles d'évolution expérimentale *in silico* sont algorithmiquement très proches des algorithmes évolutionnaires utilisés en optimisation combinatoire. En effet, le principe consiste à définir une tâche que les individus devront effectuer, à les sélectionner selon leur performance, et à encoder dans un génome les paramètres qui définissent la forme ou le comportement des individus. Les mutations se produisent lors de la reproduction des individus : lorsque le génome est recopié, il peut subir des mutations qui vont se traduire par des modifications des paramètres régissant la forme ou le comportement du nouvel individu. Une biochimie artificielle doit être définie pour donner les règles de décodage du génome. L'Algorithme 1 donne un exemple d'algorithme dit "générationnel" pour une population d'organismes asexués. Comme pour les algorithmes évolutionnaires, de nombreuses variantes sont possibles : la reproduction peut être sexuée (un individu va alors avoir deux parents), seule une partie de la population peut être renouvelée à chaque pas de temps (on parle alors d'algorithme "steady-state"), la taille de la population peut être variable au cours du temps, etc.

Ainsi, la différence entre l'évolution expérimentale et les algorithmes évolutionnaires ne se trouve pas tant dans les algorithmes employés que dans les objectifs scientifiques qui guident leurs développements respectifs. Les algorithmes évolutionnaires sont développés

pour trouver des solutions à des problèmes combinatoires. Ce qui importe est la qualité de la solution obtenue, bien plus que le réalisme biologique des mécanismes de mutation et de sélection. Bien qu'historiquement, les algorithmes génétiques aient été proposés avec un encodage des solutions sous forme de séquences de "nucléotides" binaires (séquences de 0 et de 1, par analogie avec la séquence de A, C, G, T de l'ADN), d'autres encodages sont maintenant couramment employés, par exemple sous forme de vecteurs de réels ou même sous forme de structures arborescentes, bien loin de l'analogie avec la séquence d'ADN. Avec des encodages aussi différents, il faut également repenser les mécanismes de mutation, et souvent s'éloigner des mécanismes biologiques pour que ces opérateurs de variation aient du sens par rapport au codage choisi. Par exemple, dans le cas de l'encodage des paramètres de la solution sous forme d'un vecteur de réels, la "mutation" d'un paramètre consiste en général à tirer au hasard un ε selon une loi normale centrée sur 0 et à ajouter cet ε à l'ancienne valeur du paramètre.

Inversement, en évolution expérimentale *in silico*, la performance finale des individus importe peu. Ce sont les mécanismes de l'évolution qui sont l'objet de la recherche. Une simulation dans laquelle la performance se dégrade peut même s'avérer plus intéressante qu'une simulation dans laquelle la performance s'améliore, parce qu'elle indique qu'une pression évolutive — à découvrir — s'est opposée à la sélection immédiate du plus apte. En l'occurrence, il peut s'agir de la dérive génétique (lorsque la population est petite notamment), ou encore de la sélection indirecte de la robustesse mutationnelle (lorsque les taux de mutations sont très élevés). L'évolution expérimentale *in silico* est vraiment, dans ce sens, expérimentale, par opposition à la démarche des algorithmes évolutionnaires qui s'apparenterait davantage à une démarche d'ingénierie. Alors que les algorithmes évolutionnaires sont utilisés de façon pragmatique pour trouver la meilleure solution obtenue parmi tous les individus testés, l'évolution expérimentale *in silico* n'est pas tendue vers un objectif d'optimisation. Il s'agit d'étudier non seulement le meilleur individu, mais aussi et surtout la variabilité de la performance qui peut exister dans le temps, entre les différents individus d'une population, et entre les différentes répétitions d'une simulation évolutive. De même, alors que les algorithmes évolutionnaires partent en général d'une population de génomes aléatoires, il est courant en évolution expérimentale de démarrer une expérience à partir de génomes obtenus lors d'un ou plusieurs runs évolutifs préliminaires. On peut par exemple construire des expériences de compétition entre deux sous-populations A et B, en démarrant avec 50% d'individus de type A et 50% d'individus de type B, comme dans l'étude de (Wilke et al. 2001). Il est également possible de réaliser des expériences d'invasion, en démarrant avec un seul individu de type B dans une population de type A, afin de déterminer dans quelles conditions le type B peut envahir la population, comme dans l'étude de (Crombach & Hogeweg 2007). Il s'agit de réaliser dans l'ordinateur des expériences similaires à celles que les biologistes mènent en laboratoire, ou même d'aller plus loin en collectant des données impossibles à collecter en laboratoire. Cela transparaît dans l'Algorithme 1, avec la sauvegarde systématique des relations de parenté et des événements de mutation, avantageux ou non. Cette sauvegarde permet par exemple de retracer la lignée ancestrale d'un individu final, de détecter les mutations qui ont eu lieu sur cette lignée, de savoir si d'autres lignées ont été en compétition avec elle et pendant combien de temps, ou encore de "rejouer l'évolution" à partir de n'importe quel moment du passé, comme le suggérait Gould.

Entrées :
- nombre T_{max} de générations à simuler,
- nombre N d'individus dans la population (constant dans cet exemple),
- un environnement et une tâche à réaliser dedans,
- une biochimie artificielle pour déterminer la forme et/ou le comportement d'un individu à partir de son génome,
- un fichier *FicGenomesInitiaux* contenant N génomes pour initialiser les individus de la première génération,
- trois noms de fichiers *FicPerf*, *FicMutations* et *FicGenomesFinaux*.

Sorties :
- un fichier *FicPerf* contenant pour chaque génération la performance moyenne et maximale des individus,
- un fichier *FicMutations* indiquant, pour chaque individu de chaque génération, quel individu était son parent et quelles mutations il a subies par rapport à ce parent,
- un fichier *FicGenomesFinaux* contenant les N génomes de la génération finale.

Début

Créer N individus à partir des N génomes du fichier *FicGenomesInitiaux*

Pour chaque génération t **allant de** 1 **à** T_{max} **Faire** :

Pour chaque individu i **allant de** 1 **à** N **Faire :**
Décoder son génome selon la biochimie artificielle
Evaluer sa performance p_i à réaliser la tâche dans l'environnement
Ecrire les performances moyenne et maximale dans le fichier *FicPerf*

Si $t = T_{max}$ **Alors :**
Ecrire dans *FicGenomesFinaux* les N génomes
Sinon :
{sélection darwinienne des plus aptes}
Pour chaque individu i **allant de** 1 **à** N **Faire :**
Comparer sa performance p_i à celle des autres individus
En déduire sa probabilité de reproduction w_i
Tirer les nombres W_1, W_2, ..., W_N de descendants des individus
selon une loi multinomiale de paramètres (N, w_1, w_2, ..., w_N)
Pour chaque individu i **allant de** 1 **à** N **Faire :**
Pour chaque descendant j **allant de** 1 **à** W_i **Faire :**
{reproduction asexuée avec mutations}
Recopier le génome de i dans le descendant j
Faire subir au génome de j des mutations aléatoires
Sauvegarder dans *FicMutations* la relation de parenté entre i et j et les mutations effectuées
Remplacer les N anciens individus par les N nouveaux individus

Fin

Algorithme 1 : Exemple d'algorithme générationnel pour simuler l'évolution d'une population de N organismes asexués.

I.1.1.d EEIS et simulateurs d'évolution de séquences biologiques

En 2012, deux articles de revue en apparence très similaires ont été publiés chez le même éditeur par des groupes géographiquement voisins. L'un, intitulé "Computer simulations: tools for population and evolutionary genetics", écrit par les chercheurs grenoblois et italiens (Hoban et al. 2012), est paru dans *Nature Reviews Genetics*. L'autre, intitulé "New insights into bacterial adaptation through *in vivo* and *in silico* experimental evolution", écrit par notre groupe lyonnais et nos collaborateurs grenoblois (Hindré et al. 2012), est paru dans *Nature Reviews Microbiology*. Chacun des articles cite plus de 100 références. Pourtant, aucune référence n'est commune aux deux articles. Dysfonctionnement des processus de publication ? Non, car en réalité les deux articles ne parlent pas de la même chose. En effet, à côté de l'évolution expérimentale *in silico*, il existe deux autres catégories de simulateurs d'évolution de séquences biologiques :

- les simulateurs utilisés en phylogénie moléculaire pour générer des benchmarks pour les méthodes de reconstruction phylogénétique (voir par exemple (Beiko & Charlebois 2007; B. G. Hall 2008; Strope et al. 2009; Dalquen et al. 2012)). Dans ces simulations, on ne simule pas tous les individus de la population. Une seule séquence supposée représentative de toute l'espèce est simulée, la sélection étant implicitement intégrée dans les processus mutationnels. Ainsi, seules les mutations supposées *a priori* neutres ou favorables sont simulées. Par exemple, on interdit les mutations qui transformeraient un codon d'un gène en un codon stop, parce qu'on fait l'hypothèse qu'une telle mutation serait délétère. La vitesse d'évolution des différents gènes ou domaines de gènes est aussi fixée *a priori* par l'utilisateur.

- les simulateurs utilisés en génétique des populations pour tester l'effet de scénarios (démographiques par exemple) sur la diversité moléculaire au sein d'une population. Ce sont ces simulateurs qui faisaient l'objet de la revue de (Hoban et al. 2012). Dans ces simulateurs, tous les individus de la population sont explicitement simulés, et les mutations peuvent être délétères, neutres ou favorables, mais on définit *a priori* la distribution de l'effet des mutations. Par exemple, on paramètrera le simulateur pour que, dans un gène, 70% des mutations soient délétères avec un coefficient de sélection[3] $s = -0.005$, 20% soient neutres avec $s = 0$, et 10% soient avantageuses avec $s = 0.005$. Ainsi, c'est l'utilisateur qui définit la force de la sélection qui s'appliquera sur les différents gènes ou domaines de gènes.

L'évolution expérimentale *in silico* se distingue de ces deux types d'approches par le fait qu'aucune hypothèse n'est imposée sur la force de la sélection qui va s'exercer sur les différents gènes ou portions de gènes. On définit une tâche que les organismes artificiels doivent réaliser, et une biochimie artificielle pour décoder la séquence génomique et déterminer la capacité de l'organisme à réaliser la tâche. C'est ce calcul d'un phénotype à partir du génotype qui distingue l'évolution expérimentale *in silico* des deux autres approches. Comme chez les organismes réels, le génotype subit les mutations mais c'est le phénotype qui est sélectionné. On ne détermine pas *a priori* sur quels gènes la sélection va opérer. Cela permet de laisser l'évolution agir le plus librement possible au niveau

[3] En génétique des populations, le coefficient de sélection d'un mutant est une mesure de sa valeur adaptative relative, comparée à celle du type sauvage.

moléculaire ; on fait le moins d'hypothèses possibles sur la façon dont "devrait", à notre avis, se passer l'évolution au niveau des séquences. En d'autres termes, l'évolution expérimentale *in silico* modélise le phénomène évolutif et observe ses effets sur les séquences, quand les simulateurs "classiques" ne modélisent que les effets présupposés de l'évolution sur les séquences.

I.1.2 Les principales familles de formalismes en EEIS

Pour les organismes unicellulaires, il existe à l'heure actuelle cinq types de formalismes permettant de faire de l'évolution expérimentale *in silico*, c'est-à-dire des modèles basés sur le principe de calcul d'un phénotype à l'aide d'une biochimie artificielle, ayant pour objectif de complémenter les expériences humides d'évolution expérimentale[4]. On peut les distinguer par la façon dont le génome est représenté, selon la nomenclature proposée par (Hindré et al. 2012):

- Le génome-programme : le génome est une séquence d'instructions dans un langage de programmation de bas niveau, proche de l'assembleur. Les individus sont évalués selon leur capacité à se répliquer dans la mémoire de l'ordinateur et à effectuer des opérations logiques ou arithmétiques. Tierra (Ray 1991), la première plate-forme d'EEIS, et Avida (Adami 2006), le standard international actuel, sont fondés sur cette approche. Ces plates-formes ont notamment permis d'étudier l'émergence de parasites (Ray 1991), la radiation adaptative[5] (Chow et al. 2004), l'évolution de la complexité (Lenski et al. 2003), de la modularité (Misevic et al. 2006), de la robustesse (Wilke et al. 2001) et de l'évolvabilité (Elena & Sanjuán 2008).
- Le génome-graphe : les individus sont caractérisés par un graphe représentant un réseau de régulation génique, un réseau de neurones ou même un circuit logique. Il n'y a pas de notion de séquence d'ADN dans ce formalisme, les mutations ont directement lieu sur le graphe. Elles consistent à changer les poids des connexions ou la topologie du graphe, en ajoutant ou en enlevant des nœuds. Ce formalisme a notamment permis d'étudier l'évolution de la modularité (Kashtan & Alon 2005; Espinosa-Soto & Wagner 2010), la relation entre la robustesse aux mutations et la robustesse au bruit des processus de développement (Kaneko 2011), et l'évolution de la communication et de l'altruisme (Floreano et al. 2007; Waibel et al. 2011).
- Le génome-collection-d'allèles : le génome est constitué d'un nombre fixe de gènes, n. Chaque gène peut exister sous un nombre fini ou infini de variants appelés allèles. Ces allèles sont représentés par des numéros ou des caractères. Un individu est donc caractérisé par ses n allèles. Avant de lancer la simulation, on définit l'impact de chaque allèle sur la valeur sélective[6] de l'individu, et la façon dont les différents gènes interagissent (de façon additive ou multiplicative par exemple). Ce type de formalisme

---

[4] Seules sont listées ici les approches comparables aux expériences humides, c'est-à-dire les approches dans lesquelles le génome en évolution contient plusieurs gènes.

[5] Une radiation évolutive ou radiation adaptative est une évolution rapide, à partir d'un ancêtre commun, d'un ensemble d'espèces caractérisées par une grande diversité écologique et morphologique. Chaque nouvelle espèce est adaptée à une niche particulière.

[6] La valeur sélective (ou fitness, succès reproducteur ou valeur adaptative) décrit la capacité d'un individu d'un certain génotype à se reproduire. On peut évaluer la valeur sélective d'un individu par son nombre de descendants à la génération suivante.

a permis notamment permis d'étudier l'évolution de lignées hyper-mutatrices (Taddei et al. 1997; Tenaillon et al. 1999) et la spéciation bactérienne en l'absence de sélection (Hanage et al. 2006).

- Le génome-collier-de-perles : le génome est une chaîne comportant un nombre variable d'éléments fonctionnels — les "perles" — de différents types (gènes contribuant au métabolisme, gènes codant pour des facteurs de transcription, sites de fixation pour les facteurs de transcription, éléments transposables, etc.). Un facteur de transcription va influencer l'expression des gènes qui se trouvent en aval des sites de fixation de ce facteur. La chaîne d'éléments peut donc encoder un réseau de régulation génique. Chaque type d'élément peut exister sous un nombre prédéfini de variants appelés allèles. Les mutations peuvent changer l'allèle d'un élément, mais aussi changer le nombre et l'ordre des éléments. Un sous-ensemble de gènes cibles est prédéfini comme étant nécessaire à la survie dans un environnement donné, et la tâche des individus consiste à posséder et exprimer les gènes cibles correspondant à l'environnement courant. Ce formalisme a notamment permis d'étudier l'évolvabilité des génomes (Crombach & Hogeweg 2007) et des réseaux de gènes (Crombach & Hogeweg 2008), la transformation des ressources dans un écosystème (Crombach & Hogeweg 2009), ou encore la spéciation en l'absence de barrière géographique (Tusscher & Hogeweg 2009; Rocabert et al. 2017).
- Le génome-séquence-de-nucléotides : le génome est une chaîne de longueur variable de "nucléotides" ({A, C, G, T}, ou bien {0, 1}), dans laquelle des séquences-signal bio-inspirées marquent le début et la fin des gènes. Par exemple, dans le cas d'un génome binaire, on peut décider que la séquence 000 indiquera le début d'un gène, tandis que 001 marquera la fin d'un gène. Les gènes sont donc de longueur variable et les séquences intergéniques (ADN non codant) également. Chaque gène code une fonction mathématique ou un automate élémentaire et la tâche à réaliser est généralement une identification de fonction. Les mutations peuvent modifier la séquence des gènes, mais aussi les séquences intergéniques, le nombre et l'ordre des gènes. Le simulateur que nous développons dans l'équipe, *ævol*, appartient à cette catégorie, de même que les modèles AGE et ARN développés en Suisse par C. Mattiussi et au Canada par W. Banzhaf. Ces modèles ont permis d'étudier notamment l'évolution du nombre de gènes et de la quantité d'ADN non codant (Knibbe et al. 2007), l'évolution de l'organisation des gènes en opérons (Parsons et al. 2010) et l'évolution de la taille et de la topologie des réseaux de régulation génique (P. D. Kuo et al. 2006; Mattiussi & Floreano 2007; Beslon et al. 2010).

Chaque formalisme a ses forces et ses faiblesses. Le formalisme le plus approprié dépend fortement de la question scientifique d'intérêt. Par exemple, le modèle que j'ai développé durant ma thèse, *ævol*, a été conçu pour étudier l'évolution de l'organisation fonctionnelle des génomes et des réseaux de gènes bactériens (le modèle est décrit plus en détail dans la section suivante). La structure des génomes virtuels est donc volontairement très proche de l'organisation d'un génome bactérien, avec une notion d'ADN codant ou non codant, des gènes en nombre variable et de longueur variable, pouvant s'organiser en opérons (c'est-à-dire être sous le contrôle d'un même promoteur et être par conséquent transcrits ensemble). En revanche, le génome de ces organismes ne code pas pour la machinerie de réplication : un génome sans aucun gène est tout de même réplicable, comme si la machinerie de réplication était donnée à côté du génome, déjà fonctionnelle et sans qu'elle puisse elle-même évoluer. Au contraire, dans Tierra ou Avida, les génomes-programmes doivent contenir les instructions nécessaires à la réplication, et certaines mutations peuvent donc rendre un

génome-programme incapable de se répliquer. De ce point de vue, les organismes virtuels de Tierra et d'Avida sont plus "vivants" que ceux d'*ævol*, et sont plus pertinents pour étudier des questions comme celle de l'origine de la vie par exemple. Par contre, leurs génomes, qui sont des programmes en pseudo-assembleur, ne peuvent pas être structurellement comparés à des génomes réels : par exemple, il n'est pas possible avec Tierra ou Avida d'étudier l'évolution d'une organisation en opérons.

Ainsi, l'évolution expérimentale *in silico* est une méthode atypique d'étude de l'évolution. Elle consiste, dans l'esprit de la vie artificielle, à analyser une évolution réelle d'organismes artificiels, alors que les approches plus classiques de simulation d'évolution de séquences tendraient plutôt à projeter une évolution artificielle sur des séquences réelles. Elle permet de mettre à jour des phénomènes inattendus, comme les phénomènes de sélection indirecte, montrant en cela les limites de la simple expérience de pensée, pourtant si tentante tant la sélection darwinienne semble être simple. Stimulant l'intuition du chercheur, l'incitant à tester des scénarios alternatifs, elle s'avère un complément précieux aux expériences humides sur des organismes réels. Ces expériences humides restent cependant indispensables pour valider les mécanismes hypothétiques décelés grâce aux simulations, car ceux-ci peuvent potentiellement dépendre des choix *ad hoc* faits dans la biochimie artificielle construite pour calculer la valeur adaptative des individus à partir de leur génome.

I.2 Le modèle *ævol*

Travail réalisé entre autres grâce à l'obtention d'une Aide au Développement Technologique (ADT) Inria « aevol » (2014-2016). Financement d'un ingénieur confirmé (Vincent Liard) pendant deux ans, et affectation de 40% d'un ingénieur senior du service SED de l'Inria. CK porteuse du projet.

Le modèle *ævol*, initialement développé pendant ma thèse puis étendu durant les thèses de David Parsons et Bérénice Batut[7] et durant l'ADT Inria *aevol*, est un modèle d'évolution expérimentale *in silico* appartenant à la famille "génome-séquence-de-nucléotides". Il permet, lorsqu'il est déployé sur un cluster de calcul, de mener des expériences d'évolution artificielle en parallèle sur des centaines de milliers de générations, dans une démarche analogue à celle de l'évolution expérimentale *in vitro*. Les mécanismes d'« évolvabilité » ainsi identifiés permettent de mieux comprendre l'évolution des systèmes vivants, et peuvent également être une source d'inspiration pour la conception de systèmes artificiels auto-adaptatifs (voir par exemple (Peignier et al. 2015)).

aevol est un modèle individu-centré permettant de simuler l'évolution d'une population de "bactéries" virtuelles, en représentant explicitement la séquence génomique de chaque "bactérie" (un seul chromosome circulaire composé de "nucléotides" binaires) dans la mémoire de l'ordinateur, et en simulant explicitement également les mutations spontanées qui peuvent arriver lorsqu'un individu se reproduit. Des mutations locales, mais aussi des réarrangements chromosomiques peuvent se produire n'importe où dans le génome, à l'intérieur des gènes ou entre les gènes, et peuvent changer le nombre de gènes, leur ordre ou encore la séquence des protéines encodées. Le modèle inclut aussi un processus de sélection darwinienne : les individus doivent réaliser une tâche computationnelle — en l'occurrence une approximation de fonction — grâce à l'information contenue dans leurs

[7] Puis par bien d'autres doctorants, mais que je n'ai pas supervisés personnellement.

séquences codantes (Figure 1). La population, de taille constante N, est entièrement renouvelée à chaque génération selon un schéma de type Wright-Fisher asexué.

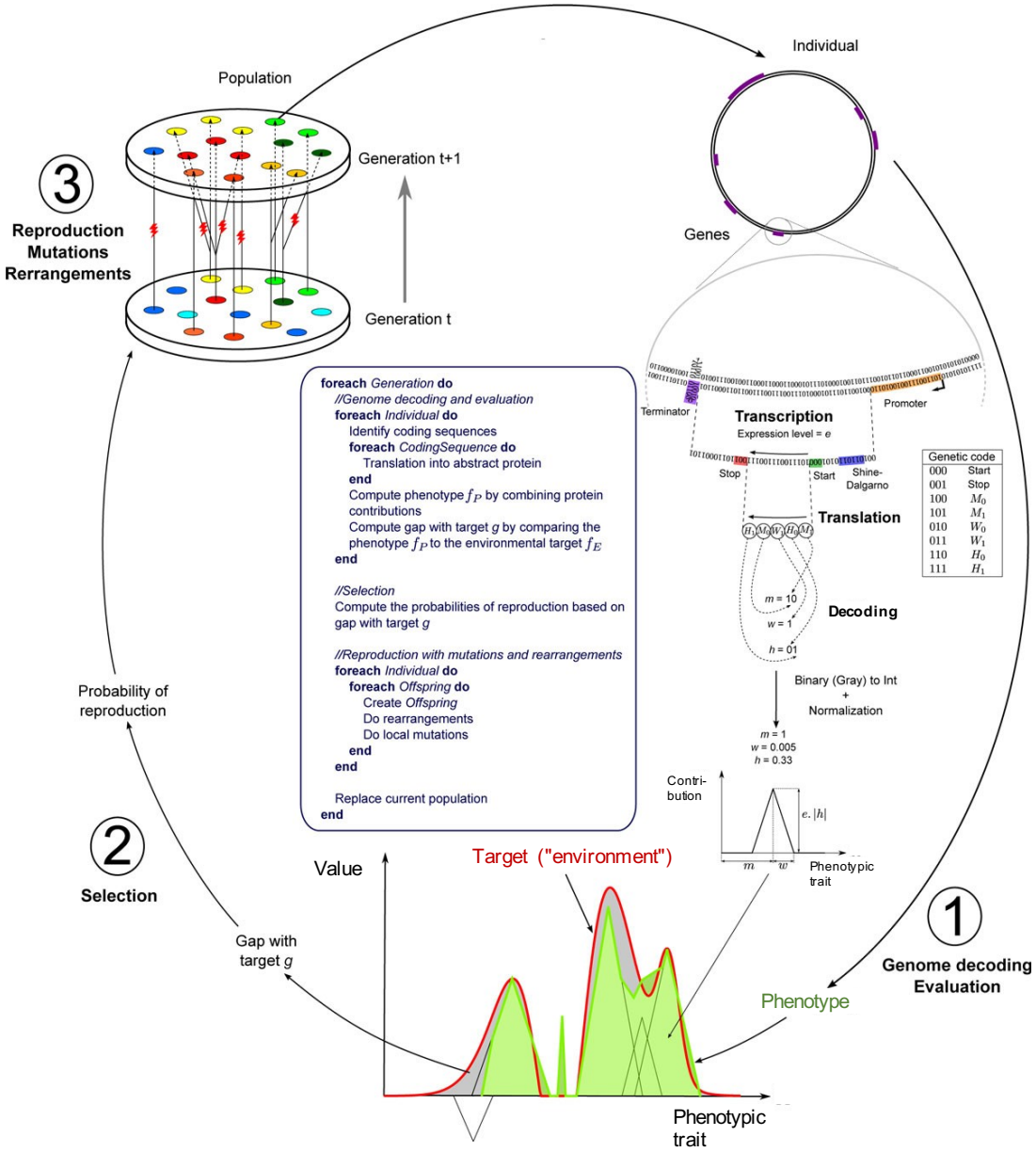


Figure 1 : Illustration du modèle aevol. Adapté de (Batut et al. 2013). Les individus sont haploïdes asexués et possèdent un seul chromosome circulaire, double brin, constitué de "nucléotides" binaires. Des séquences-signal prédéfinies (promoteurs, terminateurs, séquence "Shine-Dalgarno-like", codons start et stop) délimitent les séquences transcrites et, en leur sein, les séquences codantes. Un code génétique artificiel est utilisé pour décoder chaque séquence codante en un profil triangulaire. Le phénotype est un profil abstrait résultant de la somme de profils triangulaires encodés par les gènes. La fitness de l'individu dépend de la distance entre son phénotype et un profil-cible, possiblement variable dans le temps. La population, de taille constante, est entièrement renouvelée à chaque génération selon un schéma de type Wright-Fisher asexué. A chaque reproduction d'un individu, des mutations peuvent se produire : mutations ponctuelles, petites insertions, petites délétions, mais aussi duplications, délétions et inversions de grands segments chromosomiques.

Dans la version la plus simple du modèle, les points de cassure des réarrangements chromosomiques sont choisis au hasard, uniformément le long du chromosome, et peuvent

donc tomber ou non au milieu d'un gène. Nous verrons que durant sa thèse, David Parsons a développé une variante du modèle dans laquelle les points de cassure des réarrangements sont préférentiellement choisis dans les zones de forte similarité de séquence (Parsons et al. 2011). La thèse de Stephan Fischer a par la suite permis d'analyser mathématiquement l'impact des événements segmentaux sur l'évolution de la taille du génome (Fischer et al. 2014).

Selon l'objectif scientifique recherché, on peut soit démarrer avec un génome aléatoire cloné N fois, soit avec un génome issu d'une précédente simulation cloné N fois, soit avec une copie d'une population polymorphe issue d'une précédente simulation. Dans la thèse de Bérénice Batut, nous avons utilisé cette dernière approche pour tester la capacité de différents scénarios à provoquer une évolution réductive sur un génome déjà "constitué" (Batut et al. 2013).

La Figure 2 montre une capture d'écran du simulateur.

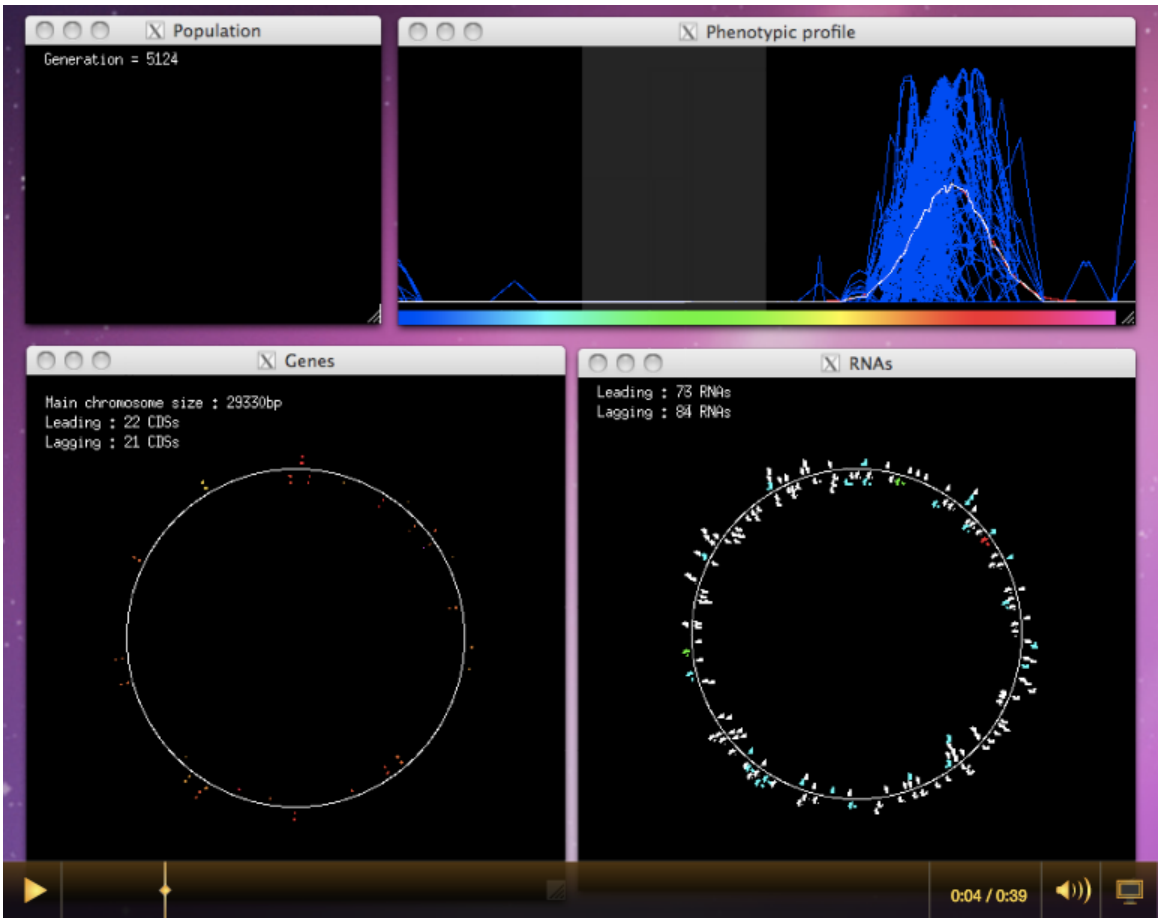


Figure 2 : Capture d'écran d'une simulation réalisée avec le simulateur *ævol*. Les deux fenêtres du bas sont deux façons de visualiser le génome (circulaire) de la meilleure bactérie dans la population courante de bactéries virtuelles. A gauche, on visualise les séquences codantes, séparées par des régions non codantes, et à droite les séquences transcrites. Une séquence transcrite grise est un ARN non codant. Les autres séquences transcrites sont colorées en fonction de leur niveau d'expression. La fenêtre en haut à droite représente le phénotype de la meilleure bactérie courante (courbe blanche), les phénotypes des autres bactéries de la population (courbes bleues), et la cible à atteindre (courbe rouge).

I.3 Le rôle clé des réarrangements chromosomiques

Thèse : David Parsons (2008-2011). Co-dirigée à 50%, avec Guillaume Beslon (LIRIS et Inria Beagle)

Le but de ce simulateur est de pouvoir réaliser (virtuellement) des expériences impossibles : le contrôle absolu des paramètres permet d'étudier certains mécanismes isolément, en allant au-delà de la simple expérience de pensée, qui est parfois trompeuse. En évolution moléculaire, l'expérience de pensée est souvent focalisée sur les mutations ponctuelles et tend à faire abstraction des réarrangements chromosomiques, peut-être plus difficiles à penser. Pourtant, le simulateur nous montre que les populations d'organismes artificiels arrivent très bien à évoluer sans mutations ponctuelles (Figure 3).

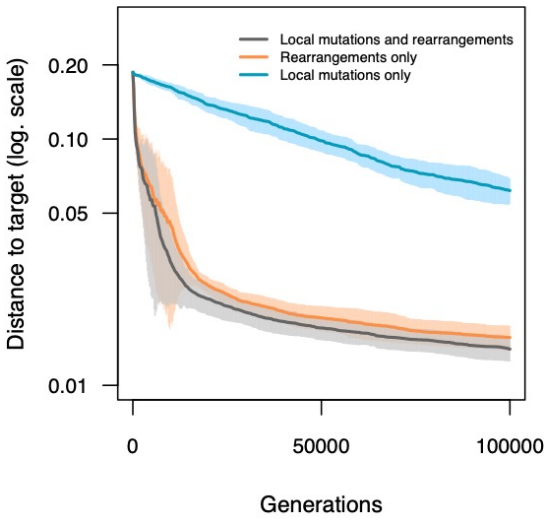


Figure 3 : Trajectoires évolutives de populations asexuées simulées avec *aevol*, évoluant soit uniquement avec des mutations locales (courbe bleue), soit uniquement avec des réarrangements chromosomiques (courbe orange), soit avec les deux types de mutations (courbe grise). L'axe des ordonnées représente la distance entre le phénotype et la cible environnementale : plus cette distance est petite, plus la valeur sélective (fitness) du phénotype est grande). Résultats non publiés.

Grâce aux simulations réalisées avec ce modèle, nous avons pu montrer que le nombre de gènes et, de façon plus surprenante, la quantité d'ADN non codant, s'ajustent en fonction du taux spontané de mutation (réarrangements inclus), sous l'effet de la sélection indirecte d'un niveau intermédiaire de variabilité mutationnelle. Ainsi, dans les simulations, un taux de réarrangement spontané très élevé est compensé à l'échelle du temps évolutif par un génome très court et très dense en gènes, alors qu'un taux de réarrangement très bas permet l'évolution de génomes beaucoup plus grands et beaucoup moins denses (Figure 4A et B) (Knibbe et al. 2007; Parsons et al. 2010).

Plus spécifiquement, nous avons observé une relation quasi-linéaire entre le logarithme du taux spontané de réarrangements et le logarithme de la taille du génome (Figure 4C), qui se retrouve à la fois sur le nombre de gènes (Figure 4D) et la quantité d'ADN non codant (Figure 4E). Ces relations linéaires en échelles logarithmiques correspondent à des lois de puissance en échelles normales. Toutefois, il convient de rester prudents car la plage de taux de réarrangements testés ne couvre que 2 ordres de grandeur, ce qui est peu pour décrire un comportement en loi de puissance. Jusqu'à récemment, il était difficile de tester des taux de réarrangement plus faibles, car cela conduisait à des génomes si grands que la population saturait la mémoire de l'ordinateur. La parallélisation du code d'*aevol*, réalisée récemment par l'équipe Inria Beagle, va permettre d'étendre la plage des taux de réarrangements testés.

Quoi qu'il en soit, nous savons déjà qu'il est possible d'obtenir des génomes longs et très peu denses en gènes sans avoir à faire appel ni à des différences d'environnement ou de style de vie, ni à des éléments transposables, ni à un biais mutationnel vers les petites insertions.

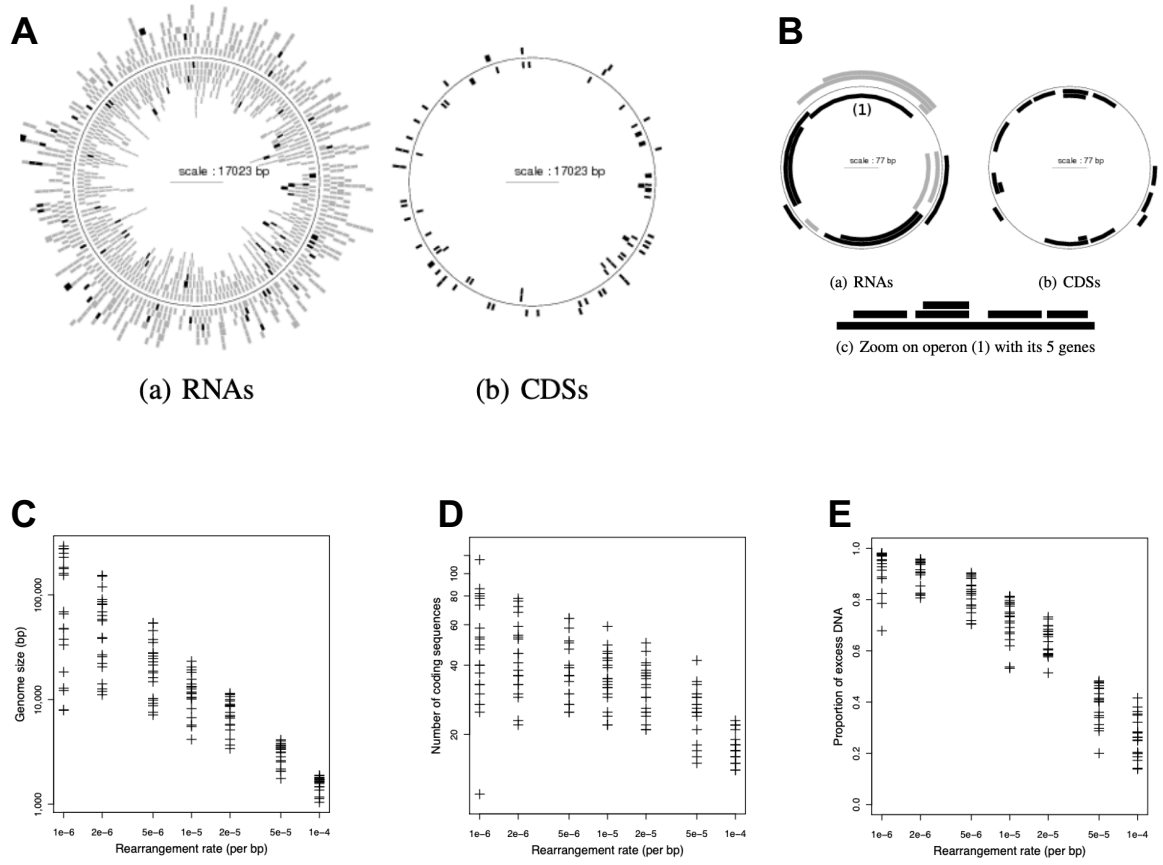


Figure 4 : Plus le taux spontané de réarrangements chromosomiques est élevé, plus les génomes obtenus sont compacts. A. Exemple de génome obtenu sous un taux de réarrangement spontané de 4.10^{-6} événements par paire de base par réplication : génome du meilleur individu de la population après 20 000 générations d'évolution asexuée, représenté d'une part avec ses séquences transcrites (a) et d'autre part avec ses séquences traduites (b). Dans le panel (a), les séquences grises sont des ARN non codants. **B.** Exemple de génome obtenu sous un taux de réarrangement spontané de 4.10^{-4} événements par paire de base par réplication. Certains ARN portent plusieurs séquences codantes (opérons). **C, D, E.** Caractéristiques des génomes obtenus après 20 000 générations d'évolution asexuée, en fonction du taux spontané de réarrangements. Les points de cassure des réarrangements sont ici choisis au hasard, uniformément le long du chromosome. Pour chaque taux de réarrangements, 7 taux de mutations locales ont été testés, allant de 3.10^{-6} à 3.10^{-4} mutations locales par paire de base par réplication, et 3 répétitions ont été effectuées pour chaque combinaison de paramètres. D'après (Parsons et al. 2010).

L'hypothèse la plus fréquemment évoquée pour expliquer la grande quantité d'ADN apparemment non fonctionnel dans les génomes eucaryotes est que ces séquences sont passivement produites par des mécanismes mutationnels biaisés vers la croissance du génome (comme la prolifération des éléments transposables), puis fixées par dérive génétique (Lynch & Conery 2003). Or dans les simulations réalisées avec *aevol*, il n'y a pas d'élément transposable, et les petites délétions sont aussi fréquentes que les petites insertions[8]. Le

[8] Nous verrons un peu plus loin que s'il y a un biais dans le modèle *aevol*, c'est plutôt un biais à la délétion et qu'il est dû à une asymétrie (non triviale) entre les grandes duplications et les grandes

modèle inclut bien de la dérive génétique puisque la population est de taille finie, mais il ne suppose pas que c'est la seule force à l'œuvre : le modèle se place dans des conditions d'évolution adaptative, c'est-à-dire d'adaptation à un nouvel environnement. Les individus sont initialement mal adaptés et des mutations favorables sont donc possibles.

Des simulations menées avec *R-aevol*, une variante d'*aevol* où les protéines peuvent réguler le niveau d'expression d'autres gènes (**Figure 5**A), suggèrent que le taux spontané de réarrangements chromosomiques impacte aussi la complexité des réseaux de régulation génétique (Beslon et al. 2010) (**Figure 5**B, C). De plus, nous avons observé dans les génomes vainqueurs que le nombre de gènes avec une activité métabolique augmente de façon quasi-linéaire avec le nombre total de gènes, tandis que le nombre de facteurs de transcription augmente, lui, de façon quasi-quadratique (**Figure 5**E). Ces lois de puissance sont qualitativement similaires à celles décrites par (Molina & Van Nimwegen 2008) pour 630 génomes bactériens (**Figure 5**F), alors même que les simulations correspondent à un "modèle nul" pour l'évolution de la régulation.

En effet, premièrement, dans les simulations, l'environnement était constant dans le temps. Les bactéries virtuelles n'avaient pas besoin de rendre leur phénotype dynamique ni de répondre à un quelconque signal. La régulation transcriptionnelle n'était, en fait, pas strictement nécessaire. Dans ces simulations, elle n'était utile que pour ajuster plus finement le niveau d'expression des gènes, mais ce niveau devait rester stable à l'échelle de la vie de l'individu : l'environnement étant constant dans le temps, les réseaux sélectionnés étaient ceux qui convergeaient rapidement vers un état stationnaire.

Deuxièmement, l'environnement était identique pour toutes les simulations. Seul le taux de mutation (incluant les réarrangements chromosomiques) différait d'une simulation à l'autre. Les différences de complexité observées dans les réseaux obtenus ne provenaient donc pas de différences de style de vie.

Dans *R-aevol*, un réseau génétique plus complexe peut être simplement obtenu en diminuant le taux spontané de réarrangements chromosomiques. Cela permet un génome plus grand avec davantage de gènes, ce qui augmente statistiquement le nombre d'associations possibles entre séquences promotrices et protéines régulatrices.

Ainsi, que ce soit dans *aevol* ou dans *R-aevol*, le taux spontané de réarrangements chromosomiques ressort comme un paramètre clé qui détermine la taille et la complexité des génomes obtenus, sans avoir à faire appel à des différences d'environnement ou de style de vie, ni à des biais mutationnels explicites.

---

délétions. Ce biais avait été mis en évidence par Antoine Coulon dans une note de juillet 2007, puis fut analysé en détail dans la mathématisation du modèle réalisée par Stephan Fischer dans le cadre de sa thèse.

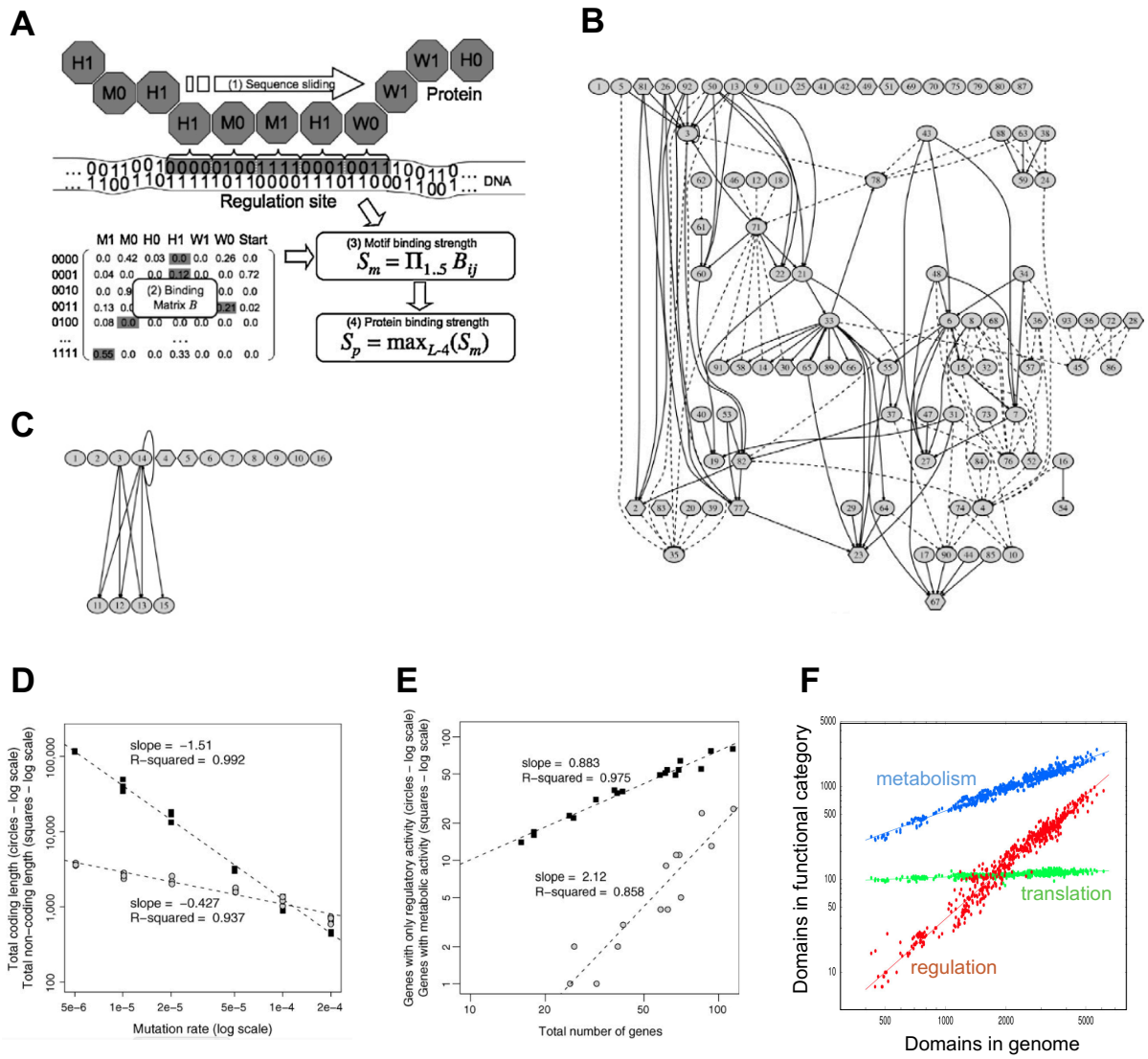


Figure 5 : Lorsque la régulation transcriptionnelle est rendue possible dans le modèle, le taux spontané de réarrangements chromosomiques influence aussi la complexité du réseau de régulation. D'après (Beslon et al. 2010). **A.** Principe du calcul de l'affinité entre une protéine et la région promotrice d'un gène dans *R-aevol*. Selon que l'affinité (éventuelle) se trouve juste en amont ou juste en aval de la première base transcrite, la protéine aura une action activatrice ou inhibitrice sur l'expression du gène. **B.** Exemple de réseau de régulation génétique obtenu après 15000 générations dans *R-aevol*, en environnement constant et sous un taux spontané de réarrangement de 2.10^{-5} par bp par réplication. Lignes pleines : liens d'activation ; lignes pointillées : liens d'inhibition ; ellipses : gènes ayant une activité métabolique ; hexagones : gènes ayant uniquement une activité régulatrice. **C.** Exemple de réseau obtenu sous un taux spontané de réarrangement de 8.10^{-4} par bp par réplication. **D.** "Lois de puissance" obtenues dans *R-aevol* entre le taux spontané de réarrangements d'une part, et le nombre de bases codantes (cercles) ou non codantes (carrés) d'autre part. **E.** "Lois de puissance" également obtenues dans R-aevol entre le nombre total de gènes d'une part, et le nombre de gènes ayant une activité métabolique (carrés) ou le nombre de gènes ayant uniquement une activité de régulation (cercles) d'autre part. **F.** Adapté de (Molina & Van Nimwegen 2008) : lois de puissance observées dans 630 génomes bactériens entre le nombre total de gènes d'une part, et le nombre de gènes ayant une activité métabolique (bleu) ou le nombre de gènes ayant une activité de régulation (rouge) d'autre part.

Cette relation entre le taux de réarrangements d'une part, et la taille et la densité en gènes du génome d'autre part, est robuste à la façon dont les réarrangements chromosomiques sont modélisés, que ce soit en choisissant les points de cassure au hasard dans le chromosome (Knibbe et al. 2007; Parsons et al. 2010), ou en les basant sur les similarités de séquences (Parsons et al. 2011) : comparer la Figure 4C et la Figure 6A. Lorsque les points de cassure des réarrangements sont préférentiellement choisis dans les zones de forte similarité de séquence, on observe dans les génomes vainqueurs qu'ils contiennent des séquences répétées directes ou inversées susceptibles de promouvoir des réarrangements, et que la plupart de ces séquences répétées se trouvent dans des zones intergéniques (Figure 6B).

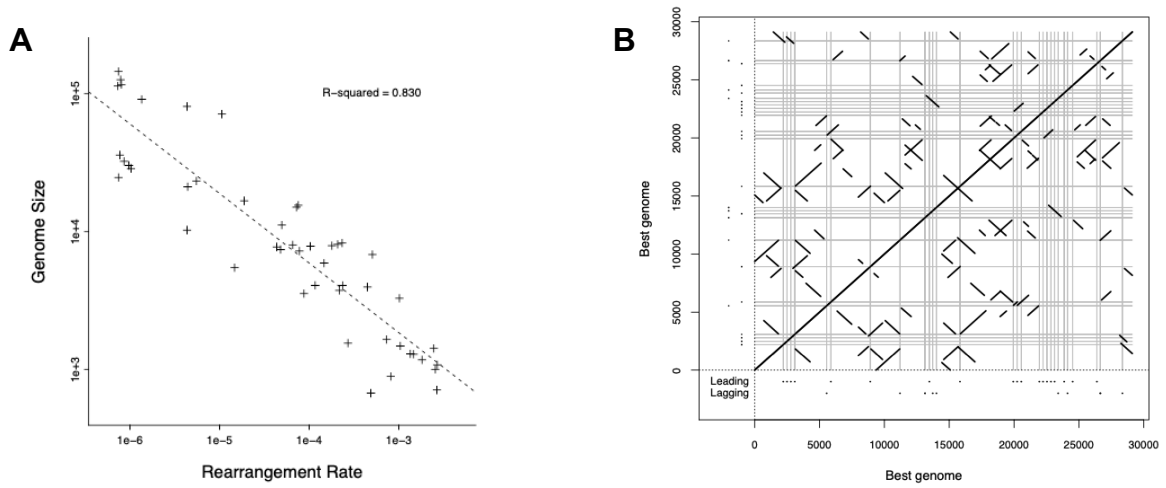


Figure 6 : Caractéristiques de génomes obtenus après 20 000 générations d'évolution en choisissant les points de cassure des réarrangements dans des zones présentant une forte similarité de séquence. (A) On retrouve la relation entre la taille du génome et le taux spontané de réarrangements. Ce taux dépend ici d'une part de la présence de séquences similaires dans le génome, et d'autre part d'un "taux de voisinage" représentant le degré de repliement du chromosome et donc le nombre de paires de séquences candidates à promouvoir un réarrangement. (B) "Dot plot" des séquences similaires trouvées dans l'un des génomes vainqueurs, produit en utilisant l'outil de bioinformatique Mummer. Les lignes grises sont les positions des gènes. D'après (Parsons et al. 2011).

L'algorithme de réarrangements basés sur les similarités de séquences et la version d'*aevol* qui l'implémente ont été réalisés par David Parsons dans le cadre de sa thèse, co-dirigée par Guillaume Beslon et moi-même. Dans le cas de l'évolution expérimentale *in silico*, il faut rechercher des paires de séquences similaires dans des millions de génomes (typiquement 1000 génomes par génération pendant des dizaines de milliers de générations dans *aevol*). Même si les génomes sont courts comparés à des génomes réels, leur nombre fait qu'une approche exhaustive est hors de portée, et même une approche heuristique de type BLAST serait trop coûteuse en temps de calcul. Cependant, nous n'avons pas besoin de rechercher *toutes* les paires de séquences similaires : pour réaliser des réarrangements, il nous suffit d'en trouver *assez*. David a donc mis au point un algorithme de recherche intermittente (Bénichou et al. 2005), décrit dans l'Algorithme 2 et la Figure 7. Le principe de cet algorithme est de tirer des paires de positions au hasard uniformément dans le chromosome, puis d'effectuer une recherche locale pour chaque paire de positions, afin de déterminer si les voisinages des deux positions contiennent des séquences avec un score de similarité suffisant. L'idée sous-

jacente est que les paires de positions représentent des séquences qui seraient spatialement proches du fait du repliement (non modélisé explicitement) du chromosome, et qu'un réarrangement peut se produire si les deux séquences spatialement proches sont suffisamment similaires.

David a implémenté deux options pour la recherche locale : avec ou sans gap, c'est-à-dire en autorisant ou non des petites insertions ou délétions en plus des mutations ponctuelles (mismatches). Si les gaps ne sont pas autorisés, le score est simplement calculé en ajoutant +1 pour un match et -2 pour un mismatch, et l'espace local de recherche est suffisamment petit pour qu'une recherche exhaustive soit effectuée. Le réarrangement se produit avec une probabilité p_{rear} qui dépend du meilleur score trouvé dans la zone locale de recherche (Figure 7c). Si les gaps sont autorisés, alors la recherche locale est effectuée comme dans PSI-BLAST et Gapped BLAST (Altschul et al. 1997), en recherchant des paires de hits sur une même diagonale ("*two-hit*"), en les étendant et en autorisant des gaps pour joindre deux *two-hits* sur des diagonales différentes. Le calcul du score incorpore alors une pénalité pour l'ouverture et l'extension d'un gap.

$initial\_nb\_pairs \leftarrow L * \mu_n$
$nb\_pairs \leftarrow initial\_nb\_pairs$
while $nb\_pairs > 0$ **do**
  Draw 2 random positions $pos1$ and $pos2$
  Draw type of rearrangement
  if *Inversion* **then** $sense \leftarrow indirect$
  else $sense \leftarrow direct$
  Draw minimal alignment score for a rearrangement to occur
  Search Alignment($pos1$, $pos2$, $sense$, min_score)
  if *Alignment found* **then**
    Proceed to Rearrangement
    Update L
  end
  $nb\_pairs \leftarrow nb\_pairs - 1$
  $nb\_pairs \leftarrow \frac{nb\_pairs}{initial\_nb\_pairs} * L * \mu_n$
end

Algorithme 2 : Algorithme de réarrangement basé sur les similarités de séquence dans le modèle *aevol*. D'après (Parsons et al. 2011). L est la longueur du chromosome en bp et μ_n est un "taux de voisinage" qui représente le degré de repliement du chromosome.

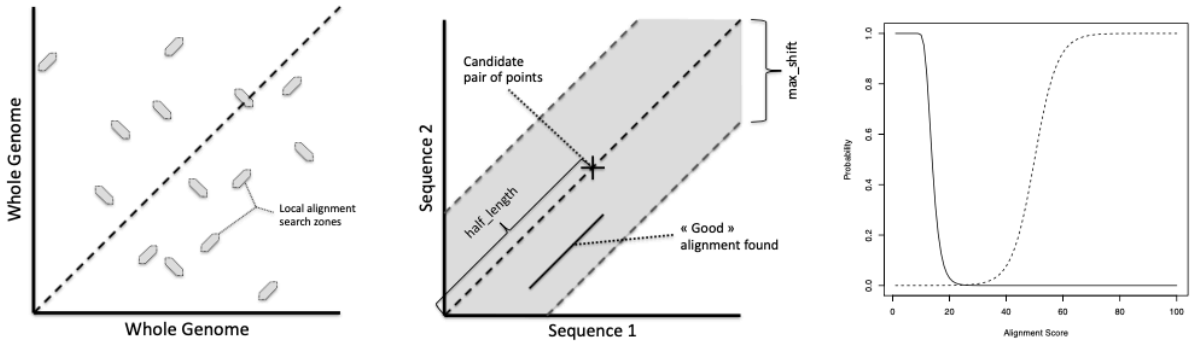


Figure 7 : Principe de la recherche intermittente utilisée dans *aevol* pour trouver des paires de séquences similaires pouvant servir de points de cassure pour les réarrangements. D'après (Parsons et al. 2011). **(a)** Des paires de positions sont tirées au hasard uniformément dans le chromosome. **(b)** Pour chaque paire de positions, une recherche locale est effectuée pour déterminer si les voisinages des deux positions contiennent des séquences avec un score de similarité suffisant. **(c)** Courbe pleine : Probabilité de trouver un alignement de score donné dans une séquence aléatoire. Courbe pointillée : Probabilité p_{rear} de réaliser un réarrangement selon le meilleur score trouvé dans la zone de recherche locale.

Ainsi, dans cette version d'*aevol*, on ne contrôle pas directement le taux de réarrangements. A la place, on définit un "taux de voisinage" μ_n qui représente le degré de repliement du chromosome. Le nombre de paires tirées sera alors $\mu_n L$, où L est la longueur du chromosome. Chaque paire est candidate à promouvoir un réarrangement, mais ne peut le faire que si elle contient des séquences suffisamment similaires. Le nombre de réarrangements effectivement réalisés dépend donc à la fois de $\mu_n L$ et de la présence ou non de séquences similaires dans le chromosome. En pratique, on observe que des séquences similaires sont effectivement maintenues dans les génomes (Figure 6B), que le taux de réarrangements réalisés est approximativement proportionnel à $\mu_n^{1.75}$ ($\mu_{\text{rear}} \cong 10^{-2.25}\mu_n^{1.75}$, $R^2 = 0.83$) et que le génome est d'autant plus court et dense en gènes que μ_{rear} est élevé (Figure 6A), comme dans la version initiale d'*aevol* (Parsons et al. 2011).

I.4 Identification des mécanismes sous-jacents

Projet exploratoire pluridisciplinaire inter-instituts (PEPII) CNRS: "Analyser, simuler and expérimenter l'évolution des génomes bactériens". Cinq laboratoires impliqués, en microbiologie, mathématiques, bioinformatique et informatique. Budget total de 50 000 euros. CK porteuse du projet.

I.4.1 Un modèle mathématique pour la dynamique mutationnelle sans sélection

Thèse : Stephan Fischer (2010-2013). Co-dirigée à 50%, avec Guillaume Beslon (LIRIS et Inria Beagle) et Samuel Bernard (Institut Camille Jordan et Inria Dracula).

Comme nous l'avons vu, les simulations individu-centrées menées avec *aevol* et *R-aevol* ont révélé l'influence du taux spontané de réarrangements chromosomiques sur la taille et de la densité en gènes des génomes. Pour isoler le ou les mécanismes sous-jacents, nous avons construit un modèle encore plus minimaliste de l'évolution des génomes, sacrifiant la notion d'information génétique pour ne prendre en compte que la taille du génome et sa densité en gènes, et leurs variations sous l'effet des réarrangements chromosomiques. Ce travail a été effectué dans le cadre de la thèse de Stephan Fischer, sous la direction conjointe de Guillaume Beslon et moi-même, et en co-encadrement avec Samuel Bernard.

Le modèle conçu avec Stephan est un modèle mathématique en population infinie et asexuée, et à générations discrètes. Dans la version la plus simple, les génomes de la population sont simplement caractérisés par leur taille s, un entier positif sans contrainte de valeur maximale ($s \in \mathbb{N}^*$). L'état de la population à la génération t est représenté par un vecteur $\boldsymbol{\nu}_t$ de taille infinie, dans lequel $\boldsymbol{\nu}_t(s)$ est la fraction de génomes de la population ayant une taille s. On suppose que qu'un génome est constitué d'un seul chromosome et que quatre types de mutations peuvent modifier sa taille :

- Petite délétion : 1 à l_{sdel} bases sont retirées au génome. La taille après la délétion appartient donc à $\{s_0 - l_{\text{sdel}}, ..., s_0 - 1\}$. Nous supposons que le nombre de bases retirées suit une loi arbitraire mais indépendante de s_0. Une matrice stochastique $\boldsymbol{M}_{\text{sdel}}$ donne les probabilités de transitions entre les différentes tailles de génomes après exactement une petite délétion : $(\boldsymbol{M}_{\text{sdel}})_{ij}$ est la probabilité qu'un génome de taille i arrive à une taille j après une petite délétion.

- Petite insertion : 1 à l_{ins} paires de bases sont ajoutées au génome. Pour un génome de taille initiale s_0, la taille après l'insertion appartient donc à $\{s_0 + 1, ..., s_0 + l_{\text{ins}}\}$. Nous supposons que le nombre de bases gagnées suit une loi arbitraire mais indépendante de s_0. Notons que dans ce cadre, la transposition réplicative d'un élément transposable peut être vue comme une de ces "petites" insertions, dans le sens où le nombre de bases ajoutées par cet événement ne dépend pas de la taille initiale du génome. Une matrice stochastique $\boldsymbol{M}_{\text{ins}}$ donne les probabilités de transitions entre les différentes tailles de génomes après exactement une petite insertion.

- Grande délétion : le nombre de bases supprimées varie de 1 à s_0. La taille finale appartient à $\{0, ..., s_0 - 1\}$. Le scénario le plus simple consiste à supposer que chaque taille finale possible peut être atteinte avec probabilité uniforme $1/s_0$, ce qui correspond au cas où les extrémités du segment délété sont choisies au hasard uniformément le long du génome, comme par défaut dans *aevol*. Dans ce cas, la taille moyenne des réarrangements augmente linéairement avec la taille du génome. D'autres distributions sont cependant possibles. Une matrice stochastique $\boldsymbol{M}_{\text{ldel}}$ donne les probabilités de transitions entre les différentes tailles de génomes après exactement une grande délétion.

- Duplication : le nombre de bases copiées varie de 1 à s_0. La taille après la duplication appartient à $\{s_0 + 1, ..., 2s_0\}$, plusieurs distributions étant possibles sur ce support, comme pour les grandes délétions. Une matrice stochastique $\boldsymbol{M}_{\text{dup}}$ donne les probabilités de transitions entre les différentes tailles de génomes après exactement une duplication.

On définit quatre taux μ_{sdel}, μ_{ins}, μ_{ldel} et μ_{dup}, par paire de base et par génération. Le taux de mutation total μ est la somme de ces quatre taux. La matrice stochastique $\boldsymbol{M}_1 = \frac{\mu_{\text{sdel}}}{\mu}\boldsymbol{M}_{\text{sdel}} +$

$\frac{\mu_{\text{ins}}}{\mu}\boldsymbol{M}_{\text{ins}} + \frac{\mu_{\text{ldel}}}{\mu}\boldsymbol{M}_{\text{ldel}} + \frac{\mu_{\text{dup}}}{\mu}\boldsymbol{M}_{\text{dup}}$ donne les probabilités de transitions entre les différentes tailles de génomes *après exactement une mutation*. En supposant que le nombre de mutations subies pendant la génération suit une loi de Poisson de paramètre μs_0, on peut définir[9] la matrice stochastique $\boldsymbol{M}_G$ qui donne les probabilités de transition entre les différentes tailles de génome *après une génération* : $(\boldsymbol{M}_{\text{G}})_{ij} = \sum_{n=0}^{+\infty} \frac{e^{-\mu i}(\mu i)^n}{n!}(\boldsymbol{M}_1^n)_{ij}$.

En l'absence de sélection, la dynamique du système est donnée par l'équation $\boldsymbol{\nu}_{t+1} = \boldsymbol{\nu}_t \boldsymbol{M}_G$. Cette équation peut être indifféremment interprétée comme l'évolution de la densité d'une population asexuée infinie dans l'espace des tailles de génomes, ou comme l'évolution de la chaîne de Markov homogène $(\mathbb{N}^*, \boldsymbol{M}_G)$.

L'analyse de ce modèle a permis de démontrer le théorème suivant[10] (Fischer et al. 2014) :

> **Théorème 1** (Fischer et al. 2014). Supposons que les distributions de taille des duplications, grandes délétions et indels soient telles qu'il existe une fonction d'échelle f positive et croissante vérifiant les conditions suivantes :
>
> Pour $\Delta(s) = \mathbb{E}[f(S_{n+1}) - f(S_n) | S_n = s]$ avec S_n la taille du génome après n mutations,
> - si la $(n+1)^{\text{ème}}$ mutation est une grande délétion, $\Delta(s) \xrightarrow[s\to+\infty]{} \delta_{\text{ldel}}$,
> - si la $(n+1)^{\text{ème}}$ mutation est une duplication, $\Delta(s) \xrightarrow[s\to+\infty]{} \delta_{\text{dup}}$,
> - si la $(n+1)^{\text{ème}}$ mutation est une petite délétion, $\Delta(s) \xrightarrow[s\to+\infty]{} \delta_{\text{sdel}}$,
> - si la $(n+1)^{\text{ème}}$ mutation est une petite insertion, $\Delta(s) \xrightarrow[s\to+\infty]{} \delta_{\text{ins}}$,
>
> où $\delta_{\text{ldel}} \leq 0$, $\delta_{\text{dup}} \geq 0$, $\delta_{\text{sdel}} \leq 0$ et $\delta_{\text{ins}} \geq 0$ sont des constantes dont au moins une n'est pas nulle,
>
> Alors la chaîne de Markov $(\mathbb{N}^*, \boldsymbol{M}_G)$ admet un vecteur de probabilité stationnaire asymptotique $\boldsymbol{\nu}_\infty$ si $\mu_{\text{ldel}}\delta_{\text{ldel}} + \mu_{\text{dup}}\delta_{\text{dup}} + \mu_{\text{sdel}}\delta_{\text{sdel}} + \mu_{\text{ins}}\delta_{\text{ins}} < 0$.
>
> Si la taille des duplications et grandes délétions grandit significativement plus vite avec la taille du génome que celle des petits indels, c'est-à-dire si $\delta_{\text{ins}} = \delta_{\text{sdel}} = 0$, la condition devient simplement $\mu_{\text{dup}}\delta_{\text{dup}} < \mu_{\text{ldel}}|\delta_{\text{ldel}}|$.

L'existence d'une distribution stationnaire $\boldsymbol{\nu}_\infty$ signifie que le génome ne croît pas à l'infini. Les conditions données pour cela par le théorème sont assez larges, car elles incluent des cas où l'intuition (l'expérience de pensée...) suggèrerait une croissance infinie.

Par exemple, lorsqu'on tire les points de cassure des réarrangements au hasard uniformément dans le génome comme dans la version la plus simple d'*aevol*, alors on peut prendre le logarithme pour la fonction f du théorème. On a alors $\delta_{\text{ldel}} = -1$, $\delta_{\text{dup}} = 2\ln 2 - 1$, $\delta_{\text{sdel}} = 0$ et $\delta_{\text{ins}} = 0$. D'après le Théorème 1, le taux de duplications peut être jusqu'à

---

[9] Même si les matrices sont infinies, leur produit matriciel est bien défini et associatif car elles sont stochastiques. Cela permet de garantir l'existence des puissances des matrices et en particulier l'existence de la matrice $\boldsymbol{M}_1^n$.

[10] Comme la matrice $\boldsymbol{M}_G$ est de taille infinie, beaucoup de théorèmes classiques (Perron-Frobenius par exemple) ne s'appliquent pas. La preuve de ce théorème repose sur l'utilisation de la condition de minoration de Doeblin en 2 pas de temps (Stroock 2006).

$1/(2\ln 2-1) \cong 2.6$ fois supérieur au taux de grandes délétions, et le taux de petites insertions peut être également très supérieur au taux de petites délétions, et pourtant le génome ne grandira pas infiniment. A l'aide d'une approximation continue valable pour les grands génomes et de l'inégalité de Chebyshev, nous avons pu montrer l'existence de bornes supérieures pour les quantiles de la distribution $\boldsymbol{\nu}_t$, valables quelque soit t (sauf peut-être à t=0). Ainsi, si $\mu_{\text{dup}} < 2.6\mu_{\text{ldel}}$, on peut garantir que quelque soit la génération, 50% au moins des génomes de la population seront plus petits que $Q_1(s^{\max,(1)})$, 80% des génomes au moins seront plus petits que $Q_2(s^{\max,(2)})$, et plus généralement $(100-100/(1+k^2))$% des génomes au moins seront plus petits que $Q_k(s^{\max,(k)})$, avec :

$$\begin{cases} Q_k(s) \;=\; e^{\ln s \,-\, As + kB\sqrt{s}} \\ s^{\max,(k)} = \dfrac{1}{A} + k^2\dfrac{B^2}{8A^2}\left(1+\sqrt{1+\dfrac{16A}{B^2}}\right) \\ A \;=\; \mu_{\text{ldel}} - (2\ln 2-1)\mu_{\text{dup}} \\ B \;=\; \sqrt{2(\ln 2-1)^2\mu_{\text{dup}} + 2\mu_{\text{ldel}}} \end{cases}$$

La Figure 8A montre l'allure de $Q_1(s)$, la borne supérieure pour la médiane de la taille du génome après une réplication, quand on part de la taille s. Le maximum de cette courbe, $Q_1(s^{\max,(1)})$ donne une borne supérieure valable pour tout génome de départ, et donc pour toute distribution $\boldsymbol{\nu}_{t-1}$ de génomes de départ. C'est donc une borne supérieure pour la médiane de la population à toute génération. Notons que cette borne resterait valide en présence de sélection : en éliminant certains individus et en sélectionnant plusieurs fois certains autres comme parents, l'opérateur de sélection modifierait la distribution des tailles de départ, mais la borne est valide pour toute distribution de départ.

Par ailleurs, la Figure 8A montre aussi qu'il existe une taille à partir de laquelle la médiane après réplication sera inférieure à la taille de départ, ce qui signifie que plus de la moitié des descendants vont rétrécir. Stephan a démontré que cette taille critique vaut B^2/A^2, soit environ $5.81/\mu_{\text{dupdel}}$ si $\mu_{\text{dup}} = \mu_{\text{ldel}} = \mu_{\text{dupdel}}$. Sous les hypothèses du modèle, cette taille critique est une limite supérieure de stabilité pour les génomes. Les génomes réels pour lesquels nous avons pu obtenir une estimation de μ_{dupdel} se trouvent effectivement bien en-deçà de cette taille critique (Figure 8B) : sous l'hypothèse (forte) que la taille d'une duplication ou d'une délétion suit une loi uniforme entre 1 et la taille du génome, les génomes réels considérés sont tels que leur probabilité de rétrécir en une réplication n'excède pas 0.01.

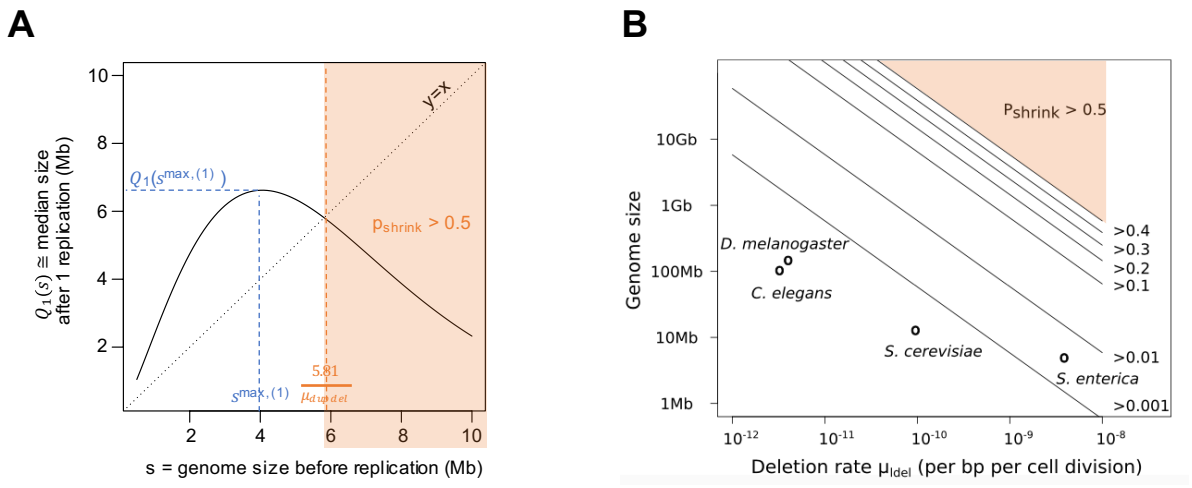


Figure 8 : Bornes supérieures prédites par le modèle mathématique, et position de génomes réels par rapport à ces bornes. D'après (Fischer et al. 2014). **A.** Allure de la fonction $Q_1(s)$ quand $\boldsymbol{\mu_{dup} = \mu_{ldel} = \mu_{dupdel} = 10^{-6}}$ par bp par réplication et quand la taille des segments dupliqués ou délétés est uniformément distribuée entre 0 et la taille s du génome avant réplication. $Q_1(s)$ est une borne supérieure pour la médiane de la taille du génome après une réplication. Le maximum de cette courbe, $Q_1(s^{max,(1)})$ donne une borne supérieure pour la médiane de la population à tout instant. Si $\boldsymbol{s > 5.81/\mu_{dupdel}}$, la médiane après réplication est inférieure à la taille avant réplication, ce qui signifie que plus de la moitié des descendants vont rétrécir. **B.** Position de génomes réels par rapport aux bornes de stabilité. Les taux spontanés de grandes délétions ont été calculés à partir de données expérimentales d'accumulation de mutations pour la bactérie *Salmonella enterica* (Nilsson et al. 2005), la levure *Saccharomyces cerevisiae* (Lynch et al. 2008), le ver *Caenorhabditis elegans* (Lipinski et al. 2011) et la mouche *Drosophila melanogaster* (Schrider et al. 2013). p_{shrink} est la probabilité qu'un génome rétrécisse lors d'une réplication si $\boldsymbol{\mu_{dup} = \mu_{ldel}}$ et si la taille des segments dupliqués ou délétés est uniformément distribuée entre 1 et la taille du génome.

Ces propriétés s'expliquent par deux éléments dans les hypothèses du modèle :

a) Si la taille d'une duplication ou d'une délétion suit une loi uniforme entre 1 et la taille du génome, le processus de duplications/grandes délétions semble symétrique mais il ne l'est pas. Même si on prend $\mu_{dup} = \mu_{ldel} = \mu_{dupdel}$ comme par défaut dans *aevol*, alors un grand génome diminue en moyenne de $100(1 - e^{\ln 2 - 1}) \cong 26.4\%$ après un événement mutationnel. Ainsi, bien que les taux de grandes délétions et de duplications soient égaux, et que la distribution de taille des segments délétés et dupliqués soient identiques, on a en fait un biais mutationnel vers une diminution du génome. En effet, comme les duplications et grandes délétions sont des processus multiplicatifs et non additifs, leurs effets absolus dépendent du point de départ. Dans un génome de taille s_0, la taille moyenne d'un segment dupliqué ou délété est $s_0/2$. Si on commence par une délétion, on arrivera à une taille $s_1 < s_0$, et le prochain événement aura une taille moyenne de $s_1/2$ seulement. On dit que le processus de duplication/délétion n'est pas invariant par translation. Il faut en moyenne 2.6 duplications pour compenser une délétion.

b) Les grands génomes subissent plus d'événements mutationnels que les petits génomes à chaque réplication, ce qui les rend plus instables.

Le modèle mathématique prédit que si $\mu_{\text{dup}} = \mu_{\text{ldel}}$ et qu'on relâche complètement la sélection, alors la taille du génome va s'effondrer, ce qui est effectivement ce qu'on avait observé dans *aevol* (Knibbe et al. 2007).

Toutefois, bien qu'une loi uniforme pour la taille moyenne des réarrangements soit pratique mathématiquement, elle n'est pas très réaliste. Nous avons donc utilisé des données de réarrangements réellement observées chez des bactéries pour simuler l'évolution spontanée d'un génome avec une distribution réaliste pour la taille des réarrangements (Figure 9A), tronquée à la taille courante du génome. Nous avons alors confirmé qu'à taux de duplication et de délétion égaux, en l'absence de sélection, le génome rétrécit (Figure 9C). Des simulations additionnelles montrent que le changement médian de taille du génome reste négatif après 1000 générations jusqu'à $\mu_{\text{dup}} = 1.2\mu_{\text{ldel}}$ (contre $\mu_{\text{dup}} = 2.6\mu_{\text{ldel}}$ pour une distribution uniforme). En effet, avec la distribution ajustée sur données réelles et tronquée à la taille du génome, la taille moyenne des segments augmente avec la taille du génome, mais beaucoup moins qu'avec la loi uniforme (Figure 9B).

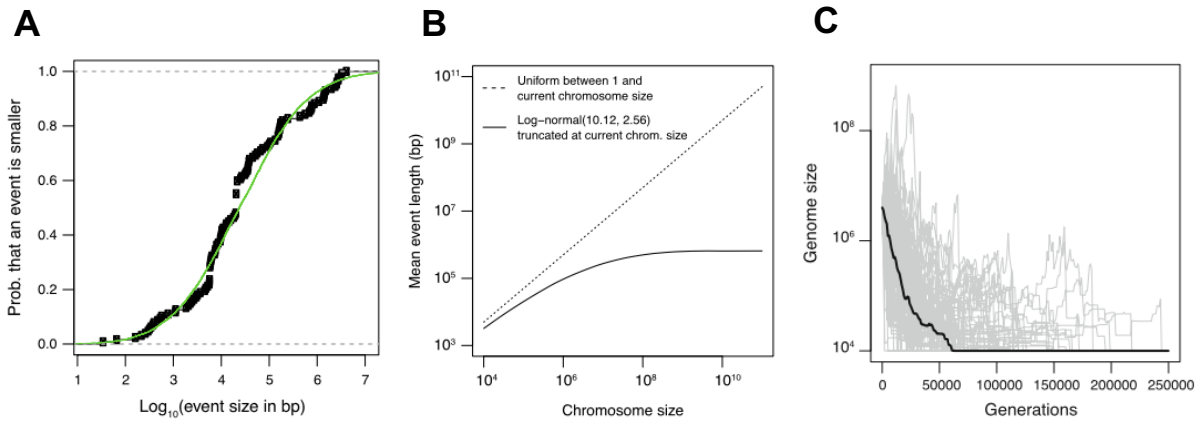


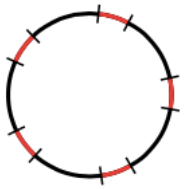
Figure 9 : Dynamique spontanée de rétrécissement des génomes, d'après Fischer et al. (2014). **A.** Fonction de répartition empirique de la taille (en bp) des réarrangements chromosomiques, basée sur 107 réarrangements vérifiés expérimentalement, observés dans des cultures d'*Escherichia coli* (D. Schneider, communication personnelle) et *Salmonella enterica* (Nilsson et al. 2005; Sun et al. 2012). La courbe verte est la distribution log-normale ajustée, $\ln \mathcal{N}(10.1214, 2.5602)$. **B.** Si on utilise cette loi log-normale tronquée à la taille du génome pour la taille des segments dupliqués et délétés, la taille moyenne des segments augmente avec la taille du génome mais moins vite qu'avec la loi uniforme, et finit par plafonner. **C.** Évolution spontanée de la taille d'un génome qui évoluerait sans sélection avec un taux égal de duplications et de délétions, dont la taille suivrait la distribution log-normale ajustée sur les données réelles mais tronquée à la taille du génome. Les courbes grises représentent 100 réalisations indépendantes du processus stochastique, et la courbe noire leur médiane. Les simulations ont été arrêtées lorsque la taille du génome devenait inférieure à 10000 bp, ce qui correspondrait à moins de 10 gènes dans un génome bactérien typique.

I.4.2 Un modèle numérique intermédiaire, avec sélection

Thèse : Stephan Fischer.

Pour déterminer si la dynamique spontanée de rétrécissement pouvait à elle seule expliquer les résultats obtenus avec *aevol*, il nous fallait d'abord comprendre comment elle interagissait avec une force de sélection simple. Pour cela, nous avons développé un modèle numérique intermédiaire. Il s'agit du modèle mathématique présenté dans la section précédente, avec les modifications suivantes (Fischer 2013):

- On décrit un génome par deux variables : son nombre n de régions codantes et son nombre L de paires de bases non codantes, et on suppose que toutes les régions codantes font la même taille l_{gene}. Alors la taille du génome est $s = L + nl_{gene}$. On suppose de plus que le génome est constitué d'un seul chromosome circulaire, et que les régions codantes sont régulièrement disposées le long de ce chromosome, c'est-à-dire que toutes les régions intergéniques ont la même longueur $l_{intergenic} = L/n$.



- Une petite insertion ou petite délétion ajoute ou enlève 1 à 6 paires de bases. Si elle se produit dans une région codante, celle-ci devient non codante (pseudogénisation).

- Pour les grandes délétions, on tire une position au hasard uniformément dans le chromosome et une longueur au hasard uniformément entre 1 et la taille du chromosome. Si une région codante est partiellement supprimée, la partie restante devient non codante.

- Pour les duplications, on tire une position au hasard uniformément dans le chromosome et une longueur au hasard uniformément entre 1 et la taille du chromosome. On tire ensuite une position au hasard uniformément pour insérer la copie du segment. Des régions codantes partiellement dupliquées deviennent des bases non codantes sur le segment dupliqué. Si le segment est inséré dans une région codante, celle-ci devient non codante.

- On ajoute un opérateur de sélection lors du changement de génération : les génomes sont sélectionnés selon leur fitness relative $\frac{\boldsymbol{F}}{\|\boldsymbol{\nu}_t\boldsymbol{F}\|}$ puis mutés, soit $\boldsymbol{\nu}_{t+1} = \frac{\boldsymbol{\nu}_t\boldsymbol{F}\boldsymbol{M}_G}{\|\boldsymbol{\nu}_t\boldsymbol{F}\|}$. Dans les simulations présentées ci-après, on suppose simplement que F=log(n). Ainsi, la sélection favorise les génomes ayant le plus de gènes, tandis que le nombre de bases non codantes n'a pas d'impact sur la fitness.

L'implémentation numérique de ce modèle est décrite en détail dans la thèse de Stephan Fischer (Fischer 2013). Les simulations numériques montrent une relation linéaire en échelles logarithmiques entre le taux de duplications et de grandes délétions et la taille totale du génome, exactement comme dans *aevol* (Figure 10A). Par contre, pour la densité en gènes, on obtient le comportement inverse à celui observé dans *aevol* : dans ce modèle numérique, plus de taux de réarrangements est élevé, *moins* le génome est dense en gènes (Figure 10B).

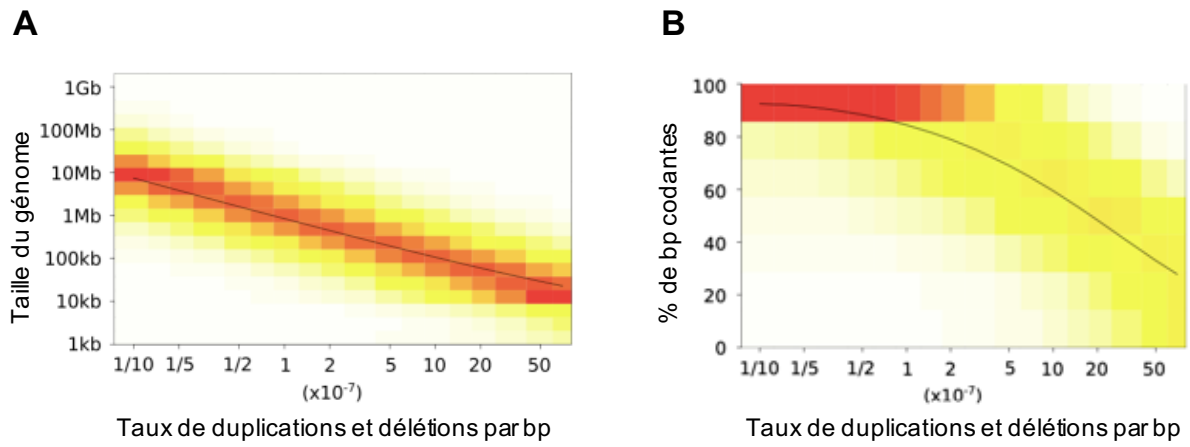


Figure 10 : Impact du taux de duplication et de délétions μ_{dupdel} sur la taille totale du chromosome (A) et sur le pourcentage de bases codantes (B) dans le modèle numérique intermédiaire, en présence d'une sélection favorisant les génomes ayant le plus de gènes. La courbe noire représente la valeur moyenne dans la population. Les niveaux de couleurs représentent la densité de la population. D'après (Fischer 2013).

Il semble donc que la dynamique mutationnelle biaisée vers le rétrécissement crée une borne supérieure pour la taille totale du génome, mais que la répartition entre codant et non-codant vers laquelle on converge sous cette borne dépende de l'action de la sélection. Une sélection simpliste comme celle du modèle numérique donne des résultats opposés à la sélection implémentée dans *aevol*, qui, elle, prend en compte la notion d'information génétique. Dans le modèle numérique, si un gène est perdu par délétion, la fitness peut être complètement restaurée en dupliquant n'importe quel autre gène. Dans *aevol*, la *séquence* des gènes est importante : un gène perdu implique une perte quasi-irrémédiable de fitness. Ainsi, comme nous allons le voir dans la section suivante, pour expliquer le pattern obtenu dans *aevol* pour le non codant, il faut analyser la dynamique mutationnelle en relation avec la perte de fitness qu'elle occasionne.

I.4.3 Un invariant dans les simulations et le rôle de la sélection indirecte

Thèses : David Parsons et Bérénice Batut (voir section I.5).

Nous avons observé un invariant dans toutes les simulations *aevol* : dans la lignée gagnante, le génome est tel que $p_{\text{neutral}} W \cong 1$, où W est le nombre de descendants que peut espérer le meilleur individu de la population et p_{neutral} la proportion de descendants dits neutres qu'il peut espérer générer (Figure 11). On appelle "descendant neutre" un descendant soit sans mutations, soit avec uniquement des mutations qui n'affectent pas le phénotype. Ainsi, dans la lignée gagnante, le génome est structuré de telle façon qu'à chaque génération, il produit en moyenne 1 descendant au phénotype identique, et $W - 1$ descendants de phénotypes différents.

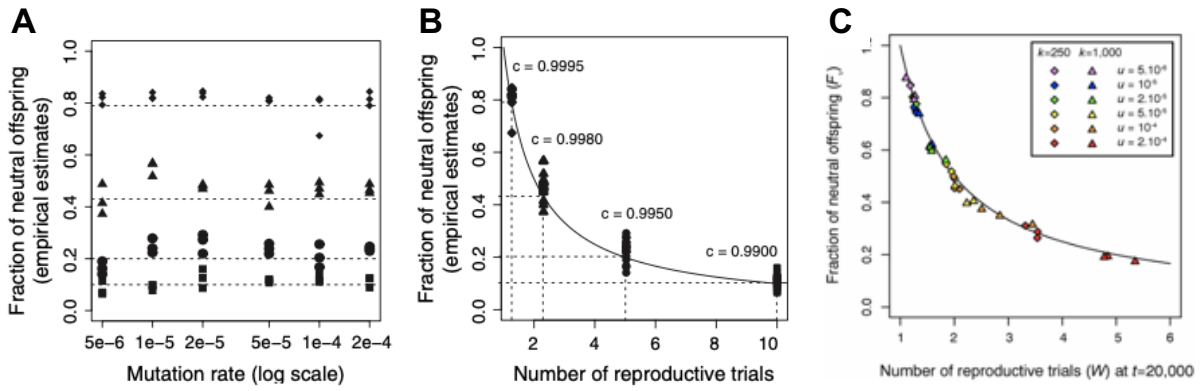


Figure 11 : Invariant observé dans les simulations *aevol* : le génome gagnant est tel qu'il produit en moyenne un descendant neutre par génération. D'après (Knibbe et al. 2007). **A.** Proportion $\boldsymbol{p}_{\text{neutral}}$ de descendants neutres du meilleur individu final (estimée en simulant 1000 réplications) en fonction du taux de mutation total, lorsque le nombre W de reproductions d'un individu dépend de son *rang* dans la population. Les différents symboles correspondent à différentes intensités de sélection (différents rapports $\boldsymbol{W}_{\text{best indiv}}/\boldsymbol{W}_{\text{worst indiv}}$). Les lignes pointillées correspondent à la valeur $1/W_{\text{best indiv}}$. **B.** Pour les mêmes simulations, proportion $\boldsymbol{p}_{\text{neutral}}$ de descendants neutres du meilleur individu final en fonction de son nombre moyen de reproductions $W_{\text{best indiv}}$. La courbe correspond à $1/W_{\text{best indiv}}$. **C.** Proportion $\boldsymbol{p}_{\text{neutral}}$ de descendants neutres du meilleur individu final en fonction de $W_{\text{best indiv}}$, pour d'autres simulations où la sélection est basée sur la valeur brute de fitness plutôt que sur le rang. La courbe correspond à $1/W_{\text{best indiv}}$

Ainsi, dans *aevol*, l'évolution de la taille du génome et de sa densité en gènes résulte de la tension entre plusieurs forces évolutives :

1. la dynamique mutationnelle caractérisée dans la thèse de Stephan Fischer, poussant vers la diminution, et fixant une borne supérieure de stabilité mutationnelle pour la taille du génome ;
2. la sélection directe des phénotypes les plus proches de la cible, ce qui nécessite beaucoup de "triangles" donc beaucoup de gènes ;
3. la "sélection indirecte" (Johnson 1999; Palmer & Feldman 2011) ou "sélection de second ordre" (Tenaillon et al. 2001; Woods et al. 2011) d'un niveau intermédiaire de variabilité mutationnelle, c'est-à-dire la survie à long terme des lignées présentant un niveau intermédiaire de variabilité mutationnelle : ni trop, ni trop peu de mutations. En effet, dans des conditions d'évolution adaptative, des mutations trop fréquentes, généralement délétères, conduisent rapidement à l'extinction de la lignée faute de pouvoir préserver la fitness ancestrale. Inversement, des mutations trop rares conduisent à une impasse évolutive : à long terme, la lignée risque l'extinction du fait de la compétition avec une autre lignée qui aurait, elle, trouvé une mutation avantageuse.

Pour illustrer ce dernier concept, on peut prendre l'exemple d'un gène qui influence le taux de mutation des autres gènes parce qu'il code pour une protéine impliquée dans la réparation de l'ADN. Une mutation dans ce gène de réparation, qui diminuerait la qualité de la réparation, va potentiellement réduire son coût énergétique, ce qui peut donner lieu à une sélection *directe* de la mutation, du fait de ses effets immédiats sur la vitesse de reproduction du mutant. Cette mutation peut aussi avoir pour effet d'augmenter le taux de mutation des autres gènes du génome : si un descendant du mutant subit une mutation favorable dans un

gène du métabolisme par exemple, alors cet individu double-mutant va se reproduire plus rapidement, et les deux mutations vont se propager ensemble dans la population. La mutation du gène de réparation aura été dans ce cas sélectionnée *indirectement* grâce à la mutation favorable qu'elle a permis de générer.

Jusqu'à récemment, ce phénomène était supposé relativement marginal car applicable uniquement dans les populations asexuées. En effet, dans les populations sexuées, la recombinaison de l'ADN peut casser l'association entre la mutation favorable sélectionnée et celle du gène de réparation (Sniegowski et al. 2000).

Nous avons montré que le phénomène de sélection indirecte peut en fait être un phénomène général, car la variabilité mutationnelle n'est pas seulement sous le contrôle d'un gène en particulier, mais dépend aussi de propriétés distribuées sur l'ensemble du génome, comme la quantité d'ADN non codant. Pour comprendre cela, il est fondamental de prendre en compte le rôle des réarrangements chromosomiques, comme le fait *aevol*, au contraire des modèles traditionnellement utilisés en génétique des populations. Dans nos simulations, à jeux de gènes identiques et taux de mutations ponctuelles identiques, deux génomes contenant plus ou moins d'ADN non codant n'ont pas la même variabilité mutationnelle. Celui qui est le plus long a plus de chances de contenir des paires de séquences répétées, susceptibles de provoquer des réarrangements chromosomiques comme des grandes délétions. L'ADN non codant est ainsi mutagène pour les gènes avoisinants.

La variabilité mutationnelle du phénotype ne dépend donc pas seulement du taux de mutation par base, mais aussi de la structure du génome. Dans nos simulations, à taux de mutations fixe, la sélection indirecte d'un niveau intermédiaire de variabilité mutationnelle se traduit par la sélection indirecte d'une taille de génome et d'une densité en gènes qui permettent de "compenser" le taux de mutations : génome long et peu dense en gènes si le taux de mutations est bas, génome court et dense en gènes si le taux de mutations est élevé (Knibbe et al. 2007; Parsons et al. 2011). Durant la thèse de Bérénice Batut, nous avons mené des simulations avec un biais mutationnel favorisant les petites délétions par rapport aux petites insertions : dans ces conditions où l'on attendrait une érosion des séquences non codantes, on constate au contraire qu'elles peuvent être activement maintenues par la sélection si l'environnement est variable (Figure 12). Si l'environnement se stabilise, alors on observe bien l'érosion attendue, ce qui semble confirmer le rôle joué par l'ADN non codant dans l'évolution adaptative.

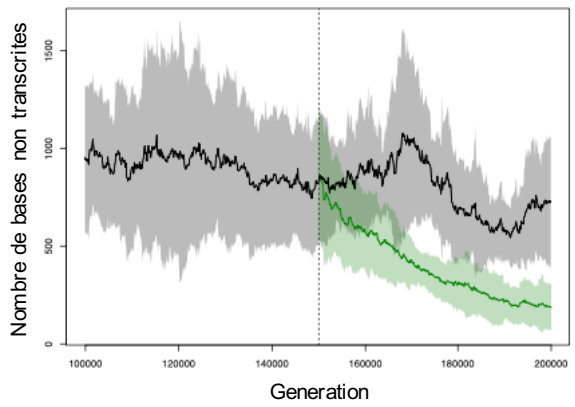


Figure 12 : Évolution de la quantité d'ADN non transcrit dans *aevol*, en présence d'un biais mutationnel favorisant les petites délétions, lorsque l'environnement est variable (courbe noire) ou rendu fixe à partir de la génération 150 000 (courbe verte). D'après (Batut 2014).

Durant la thèse de David Parsons, nous avons montré que ces résultats sont robustes à la présence d'échanges génétiques entre individus pouvant casser les associations entre mutations (Figure 13) : quelque soit le taux d'échanges génétiques entre les bactéries

virtuelles, on obtient des génomes d'autant plus courts et compacts que le taux de réarrangements est élevé. Cela vient du fait que la fréquence des réarrangements n'est pas contrôlée par un seul locus, mais dépend au contraire de la présence de séquences répétées réparties dans tout le génome. Ainsi, contrairement à ce que les modèles de génétique des populations laissaient croire, le phénomène de sélection indirecte ne concerne donc pas seulement les populations asexuées : il pourrait être donc à l'origine de contraintes évolutives sur l'organisation des génomes et des réseaux génétiques dans toutes les espèces.

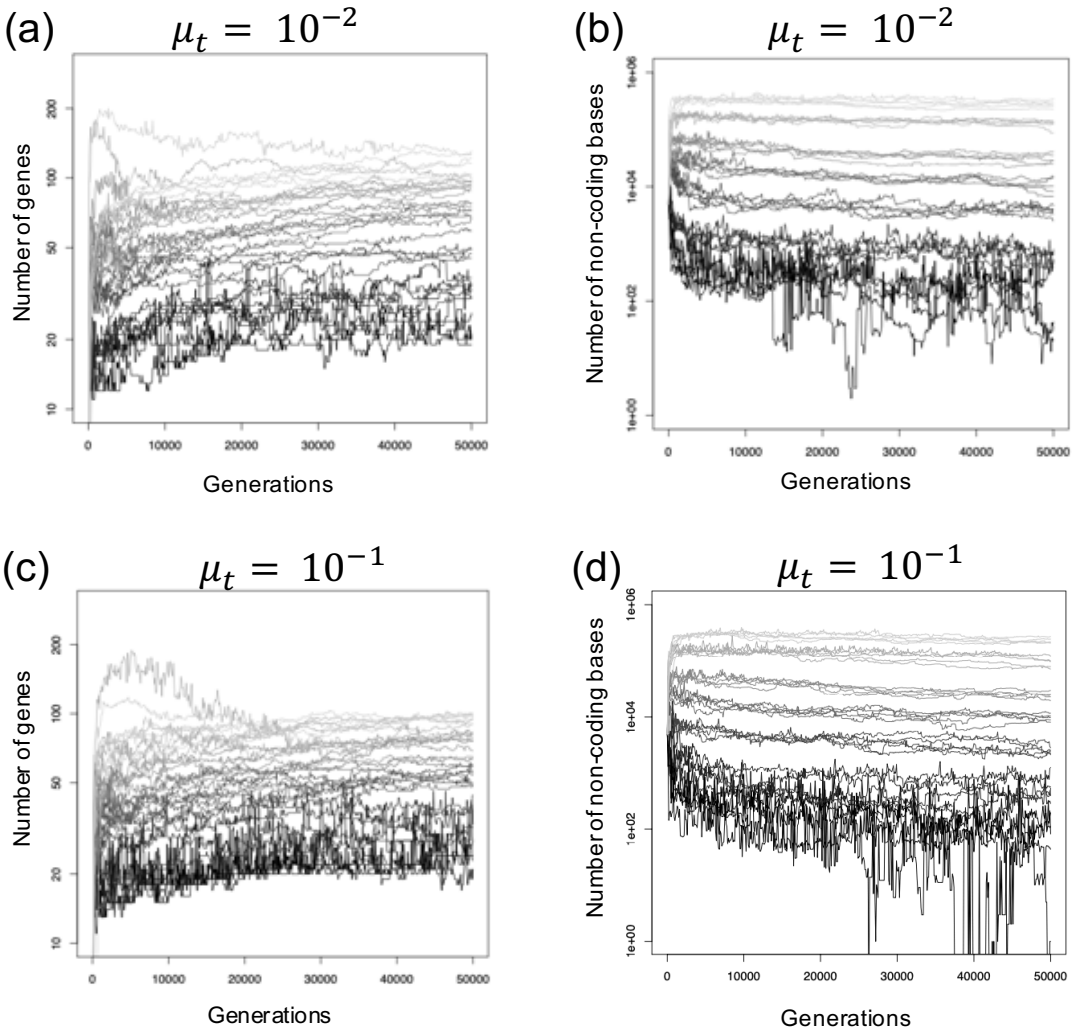


Figure 13 : Évolution du nombre de gènes (a, c) et de la quantité d'ADN non codant (b, d) dans *aevol*, en présence de transfert horizontal. D'après (Parsons 2011). Le niveau de gris correspond au taux de mutations incluant les réarrangements : de 7.10^{-6} (gris clair) à 7.10^{-4} (noir). $\boldsymbol{\mu_t}$ est la proportion de réplications durant lesquelles un transfert est tenté. Dans ce cas, un individu donneur est choisi au hasard dans la population. Un transfert a lieu si on trouve deux alignements A_1 et A_2 de scores suffisants entre le chromosome du donneur et celui du receveur. La séquence située entre A_1 et A_2 chez le donneur est alors copiée et vient remplacer la séquence située entre A_1 et A_2 chez le receveur, mimant ainsi une recombinaison post-conjugaison ou post-transduction.

I.5 Test de scénarios d'évolution réductive

PEPII CNRS: "Analyser, simuler and expérimenter l'évolution des génomes bactériens".

Stage de M2 : Bérénice Batut (2011). Stage encadré à 100%.

Thèse : Bérénice Batut (2011-2014). Co-encadrée à 30% avec Gabriel Marais (LBBE) et Guillaume Beslon (LIRIS et Inria Beagle).

Les simulations réalisées avec *ævol* peuvent apporter un éclairage complémentaire aux approches de génomique comparative et de phylogénie moléculaire sur les causes de l'évolution réductive observée dans certains génomes bactériens (Batut et al. 2013). En effet, chez certaines bactéries endosymbiotiques (comme *Buchnera aphidicola*, bactérie vivant dans certaines cellules des pucerons), mais aussi chez certaines lignées de cyanobactéries libres comme *Prochlorococcus*, le génome s'est significativement réduit en taille (Figure 14).

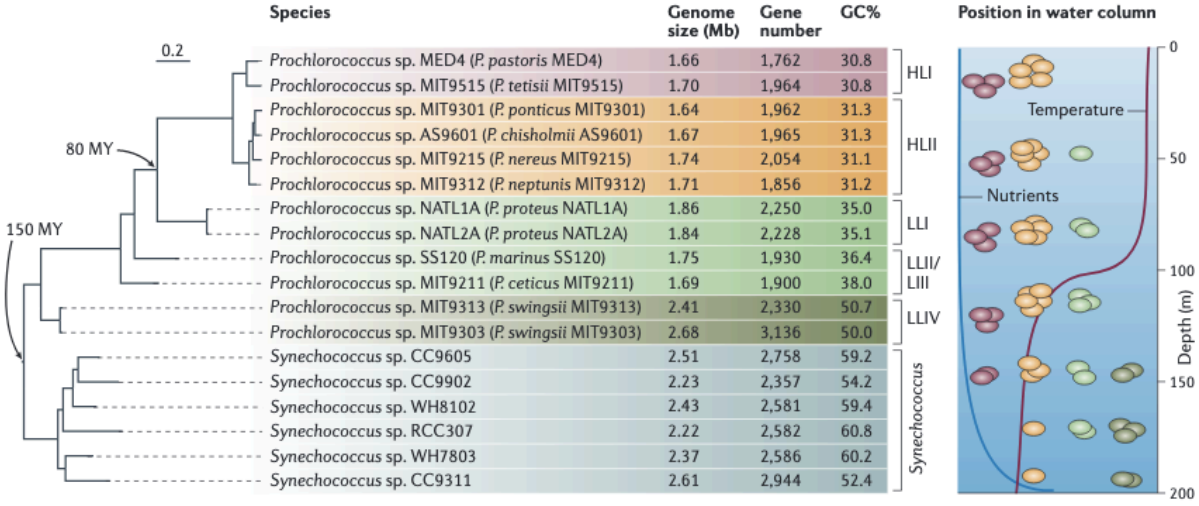


Figure 14 : Phylogénie, statistiques des génomes et préférences écologiques des écotypes de *Prochlorococcus*. D'après (Batut et al. 2014). L'arbre phylogénétique a été construit par Bérénice en utilisant PhyML (Guindon et al. 2010) sur un ensemble de gènes orthologues 1 à 1 (HOGENOM v6 (Penel et al. 2009)) alignés avec Prank (Löytynoja & Goldman 2005). Les temps de divergence sont issus de (Dufresne et al. 2005). La figure de droite est reproduite de (Partensky & Garczarek 2010). Le nombre de cellules symbolise l'abondance des écotypes aux différentes profondeurs de la colonne d'eau dans les océans. Rouge : "high-light" I (HLI), orange : "high-light" II (HLII), vert clair : "low-light" I (LLI), vert : "low-light" II et "low-light" III (LLII/LLIII), vert foncé : "low-light" IV (LLIV), noir *Synechococcus*. Des génomes réduits sont trouvés dans toutes les lignées des souches de *Prochlorococcus* sauf LLIV.

Pour les bactéries endosymbiotiques comme *Buchnera*, l'hypothèse largement admise pour expliquer l'évolution réductive du génome est celle du cliquet de Muller[11], qui affecte les

[11] Lors de la reproduction asexuée d'un organisme, son génome entier est copié et transmis à sa descendance, y compris les éventuelles mutations délétères comme des pertes de gènes. En supposant que les mutations de réversion sont improbables, la qualité génétique de la lignée se détériore, comme un cliquet dont les dents n'autorisent le mouvement que dans une seule direction. Ce processus dégénératif affecte principalement les populations non recombinantes à faible taille efficace de population. Le cliquet clique quand les génomes les moins chargés (avec $n - 1$ mutations

populations non recombinantes de petite taille efficace. Les endosymbiotes, transmis verticalement de la mère aux descendants via les œufs, sont supposés être fortement affectés par le cliquet de Muller, car leurs populations sont impactées par ces goulets d'étranglements à chaque reproduction de l'hôte et manquent d'opportunités de recombinaison avec d'autres bactéries. La sélection naturelle est donc peu efficace pour contrer l'accumulation de mutations délétères, principalement dans les gènes non essentiels, qui deviennent des pseudogènes et sont ensuite éliminés via le biais spontané vers les petites délétions, typique des bactéries (C.-H. Kuo & Ochman 2009; Mira et al. 2001). Ces gènes perdus ne peuvent être regagnés par manque de recombinaison. Même les gènes essentiels, pour lesquels la sélection est assez forte pour éviter une dégénérescence complète, peuvent toujours accumuler des substitutions non-synonymes délétères. Ainsi, sous l'effet du cliquet de Muller, les séquences d'un génome évoluent rapidement et les patrons mutationnels dominent la sélection.

Cependant, l'hypothèse du cliquet de Muller semble peu plausible pour des cyanobactéries libres comme *Procholorococcus*, dont l'abondance globale moyenne annuelle est estimée à $2.9 \cdot 10^{27}$ cellules (Flombaum et al. 2013). La taille efficace de population chez *Prochlorococcus* a été estimée à 10^9 (Kashtan et al. 2014), soit environ 20 fois celle *d'E. coli* ($5 \cdot 10^7$) (Charlesworth & Eyre-Walker 2006). De plus, les génomes de *Prochlorococcus* portent des gènes acquis par transfert horizontal (Lindell et al. 2004), et des analyses menées par Bérénice avec le logiciel PHI (Bruen et al. 2006) suggèrent qu'environ 12% des familles de gènes de *Prochlorococcus* subissent régulièrement de la recombinaison allélique (Batut 2014). Enfin, comme le montre le **Tableau 1**, toutes les caractéristiques génomiques observées chez *Prochlorococcus* ne sont pas en accord avec les prédictions de l'hypothèse du cliquet de Muller.

délétères) sont perdus par malchance lors de l'échantillonnage des gamètes d'une génération à l'autre, et que tout génome de la population possède au moins n mutations délétères. Cette accumulation de mutations délétères est rapide dans les petites populations et irréversible en l'absence de recombinaison, car seules des mutations de réversion improbables peuvent restaurer le type sauvage.

Characteristic	Patterns		Hypotheses for reductive genome evolution for *Prochlorococcus* spp.‡			
	Buchnera spp. versus *Escherichia coli*	Reduced *Prochlorococcus* spp. versus non-reduced *Prochlorococcus* spp.	MR	EA	HMR	BQH
Global genome characteristics						
Genome size	Reduction up to 80%[19,36,37]	Reduction up to 38%[38–40]	+	+	+	+
Proportion of coding DNA	Unchanged[5]	Slightly higher[40]	–	+	?	=
%GC	Decrease to 26%[5,6,31,35–37]	Decrease to 30.8–38%[38–40]	+	?	+	=
Gene repertoires						
Gene number	Reduction up to 80%[35,37]	Reduction up to 43%[39,40,45]	+	+	+	+
Gene family size	Smaller[37]	Smaller[45]	+	+	+	–
Pseudogenes	Higher proportion[6,37]	Possibly higher proportion[41]	+	–	+	+
Recombination genes	Many losses[5,6]	Some losses[28]	+	–	+	–
DNA replication and repair genes	Losses[5,6,35,37]	Losses[28,38,39,44]	+	–	+	–
		Gains[44]	–	+	+	–
Regulation genes	Losses[6]	Losses[38,40,98]	+	+	+	+
Metabolic genes	Losses[6,35]	Losses[38,44,98]	+	+	+	+
		Gains[44]	–	+	–	–
Sequence evolution						
Sequence evolution	Faster[5,6,30,31,35–37,49,99]	Faster[39,47]	+	–	+	–
K_a/K_s	Smaller N_e (REFS 31,37,46)	Larger N_e?[47]	–	+	+	?
Polymorphism level	Smaller N_e (REF. 72)	Not clear				
Change in amino acid composition	Deleterious changes[5,49,99]	Probably many adaptive[41] or neutral[39] changes	–	+	?	–
Genome architecture						
Genome architecture	Stable[6,19,37]	Not static[32]	+	+	+	=
HGT, genomic islands and bacteriophages	No[6,19,37]	Yes[32,44,45,100,101]	–	+	=	=
Codon usage						
Optimal codon preferences	Lower[49]	Lower[70]	+	?	+	=
tRNA gene patterns	Degenerate[71]	No information				

Tableau 1 : Caractéristiques comparées et hypothèses explicatives possibles de l'évolution réductive des génomes de *Buchnera aphidicola*, bactérie endosymbiotique du puceron, et de *Prochlorococcus marinus*, cyanobactérie marine libre. D'après (Batut et al. 2014). MR: Muller's ratchet (cliquet de Muller, accumulation de mutations délétères dans des populations non recombinantes de petite taille efficace) ; EA: environmental adaptation (changement de niche) ; HMR : high mutation rate (augmentation du taux de mutations suite à la perte de gènes de réparation de l'ADN) ; BQH : black queen hypothesis (hypothèse de la reine noire : perte des gènes associés à des tâches réalisées par d'autres espèces de la communauté bactérienne). Les symboles '+' et '–' indiquent si les observations confirment ou contredisent une hypothèse donnée pour l'évolution réductive des génomes de *Prochlorococcus*. '=' indique que l'hypothèse ne fait pas de prédiction pour cette caractéristique.

Dans le cadre de la thèse de Bérénice, nous avons utilisé le simulateur *aevol* pour tester l'impact sur le génome de différents scénarios évolutifs évoqués dans la littérature pour expliquer l'évolutive réductive de génomes comme ceux de *Buchnera* ou de *Prochlorococcus*. Pour cela, Bérénice a tout d'abord créé dix populations dites souches, chacune créée en partant d'un génome initial aléatoire et en laissant la population évoluer pendant 150 000 générations sous des conditions mimant un style de vie libre et des patrons mutationnels typiquement bactériens : environnement fluctuant aléatoirement autour d'un environnement moyen, recombinaison allélique possible (par transfert d'un segment d'ADN d'un donneur aléatoire puis remplacement de la séquence homologue dans le chromosome receveur), biais mutationnel favorisant les petites délétions par rapport aux petites insertions (C.-H. Kuo & Ochman 2009; Mira et al. 2001). Chaque population souche a ensuite subi chacun des

scénarios suivants (Figure 15) : aucun changement de paramètre (contrôle), changement de la taille de la population, changement de la force de la sélection, suppression du transfert, augmentation du taux de mutations locales, augmentation du taux de réarrangements intrachromosomiques, stabilisation de l'environnement, changement de forme de l'environnement moyen (mimant un changement de niche sans réduction du nombre de tâches métaboliques à effectuer), neutralisation ou suppression d'un lobe de l'environnement (mimant un changement de niche avec réduction du nombre de tâches métaboliques à effectuer, i. e. certains gènes deviennent inutiles ou délétères). Chacun des scénarios correspond à un changement de valeurs de paramètres, effectué à la génération 150 000, à la suite de quoi l'évolution se poursuit pendant 50 000 générations supplémentaires. Les génomes obtenus au final sont comparés à ceux obtenus dans des simulations de contrôle, où aucun changement de paramètre n'a été effectué.

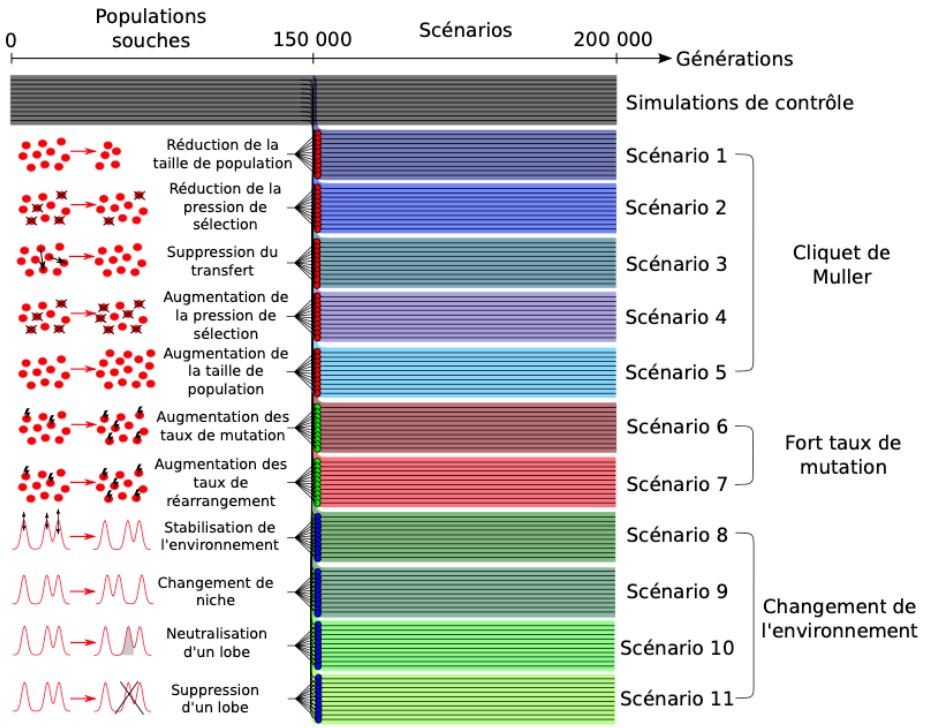


Figure 15 : Méthodologie d'étude de l'évolution réductive par la construction de populations souches et simulation de différents scénarios concernant les paramètres évolutifs. Les scénarios de l'évolution réduction sont rassemblés en trois catégories : cliquet de Muller (en bleu), fort taux de mutation (en rouge) et changement de l'environnement (en vert).

Sur les onze scénarios testés, seuls deux induisent une évolution réductive similaire à celle observée chez *Prochlorococcus*, c'est-à-dire à la fois une réduction du nombre de gènes, une augmentation de la densité en gènes et une accélération de la vitesse d'évolution des séquences. Ce sont les scénarios de réduction de la force de sélection et d'augmentation des taux de mutations locales (Figure 16). Les scénarios de neutralisation ou suppression d'un lobe de l'environnement induisent une réduction du nombre de gènes, ainsi qu'une réduction de la quantité d'ADN non codant mais dans une moindre proportion, aboutissant au final à des génomes qui ne sont pas plus denses en gènes que les génomes contrôles.

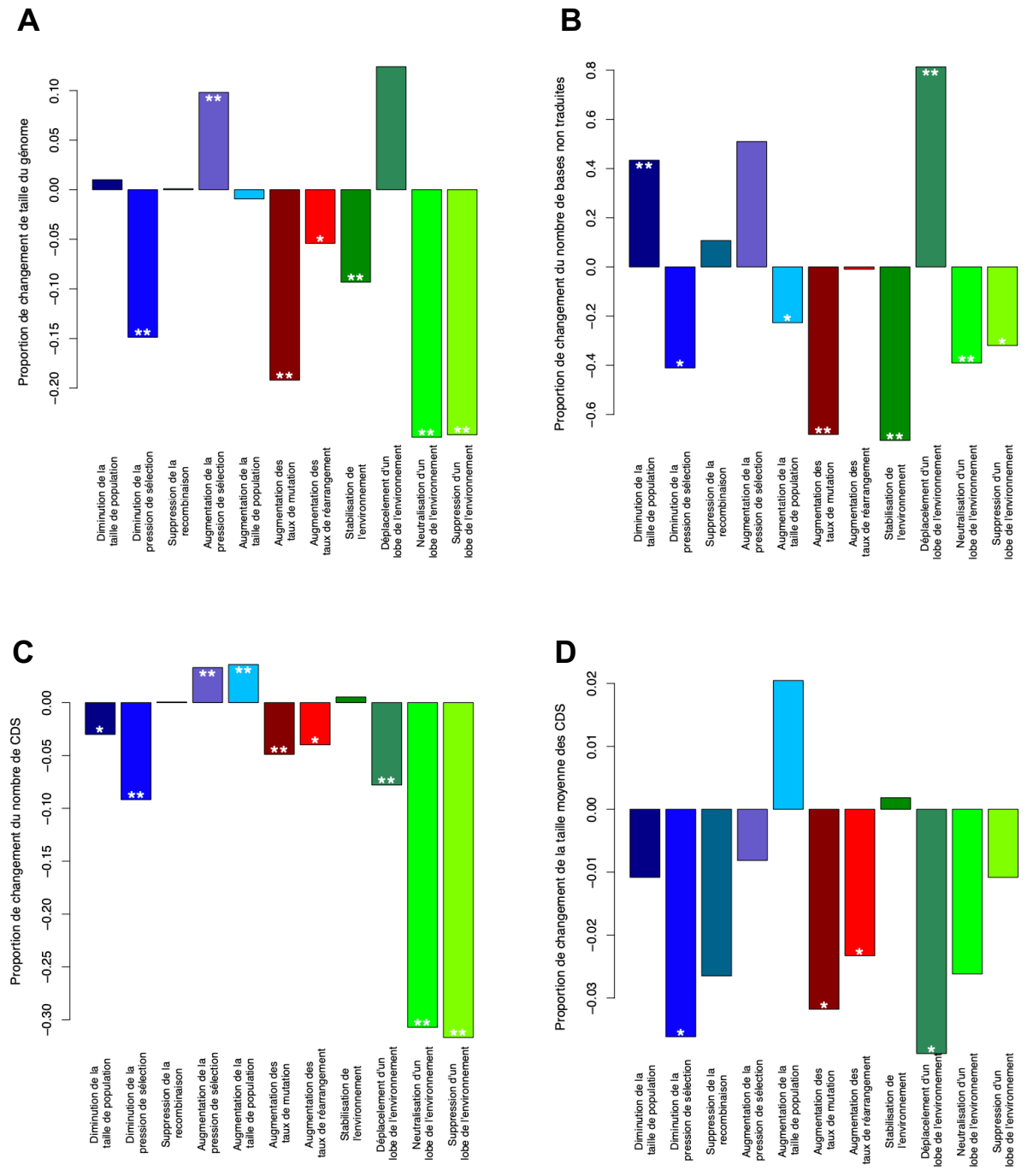


Figure 16 : Différences relatives entre les génomes finaux des simulations de contrôle et les génomes finaux des 11 scénarios, pour la taille du génome (A), le nombre de bases non traduites (B), le nombre de séquences codantes (C) et la taille moyenne des séquences codantes (D). Les scénarios sont représentés par les différentes couleurs, avec en bleu les scénarios liés au cliquet de Muller, en rouge les scénarios liés aux forts taux de mutation et en vert les scénarios de changement d'environnement. D'après (Batut 2014).

De façon assez remarquable, dans tous les scénarios, la dynamique évolutive des populations les ramène toujours vers l'invariant précédemment identifié, $p_{\text{neutral}} W_{\text{best}} \cong 1$ (Figure 17). En effectuant le changement de paramètres à la génération 150 000, on perturbe le système, soit en augmentant W_{best} (lorsqu'on change l'environnement), soit en baissant W_{best}

(lorsqu'on baisse la force de la sélection), soit en diminuant p_{neutral} (lorsqu'on augmente le taux de mutations ou de réarrangements). En 1 000 générations environ, la population revient à proximité de la courbe $p_{\text{neutral}} W_{\text{best}} = 1$, par sélection indirecte des génomes qui vérifient cet invariant dans les nouvelles conditions.

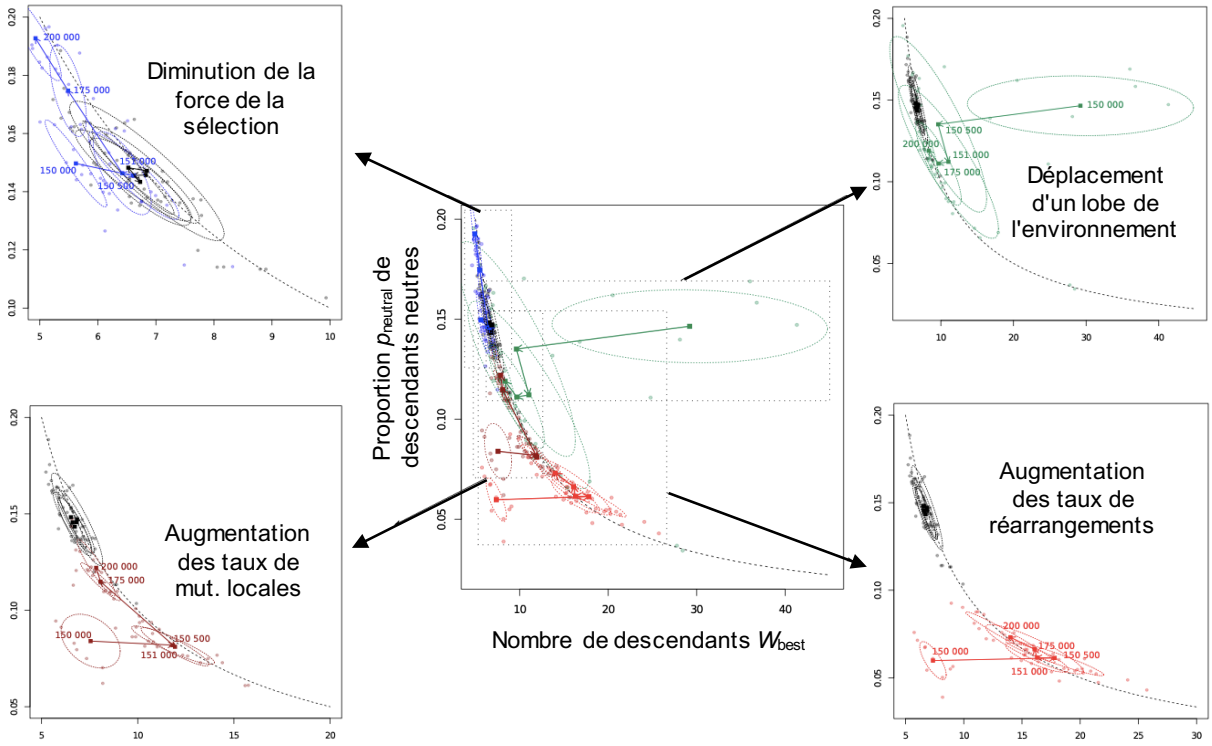


Figure 17 : Dynamique évolutives des populations simulées après changement d'un paramètre. Chaque point positionne le meilleur individu d'une population à une génération donnée. Chaque ellipse englobe 95% des répétitions de chaque scénario à une génération donnée (150 000, 150 500, 151 000, 175 000 et 200 000). Les flèches colorées relient les moyennes des répétitions dans l'ordre chronologique. D'après (Batut 2014).

En plus de cette campagne de simulations avec *aevol*, Bérénice a réalisé de nombreuses analyses bioinformatiques des génomes séquencés de *Prochloroccus*, non détaillées dans ce manuscrit mais dont les résultats sont synthétisés dans la Figure 18. En mettant en regard les résultats des simulations et des analyses bioinformatiques, Bérénice a pu proposer un scénario évolutif hypothétique pour ce clade, mettant en jeu deux épisodes mutateurs séparés temporellement par un changement de niche écologique (montée dans la colonne d'eau) (Figure 19). Une très récente analyse phylogénétique (Bourguignon et al. 2020; Marais et al. 2020) tend à confirmer que l'augmentation du taux de mutation, plus que la diminution de la taille efficace de la population, serait à l'origine de l'évolution réductive de certains génomes procaryotes.

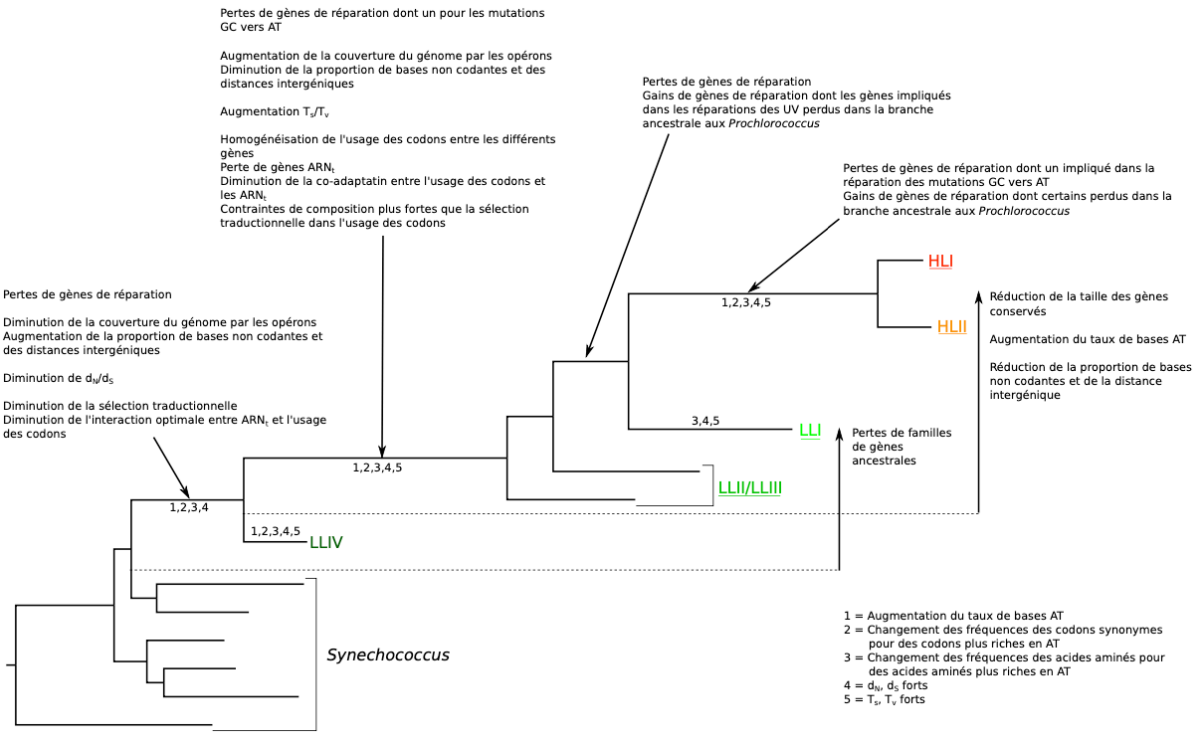


Figure 18 : Changements le long de la phylogénie de *Prochlorococcus* et de *Synechococcus* Seuls les différents écotypes sont représentés. Les écotypes dont le nom est souligné contiennent seulement des souches pour lesquelles le génome est réduit. D'après (Batut 2014).

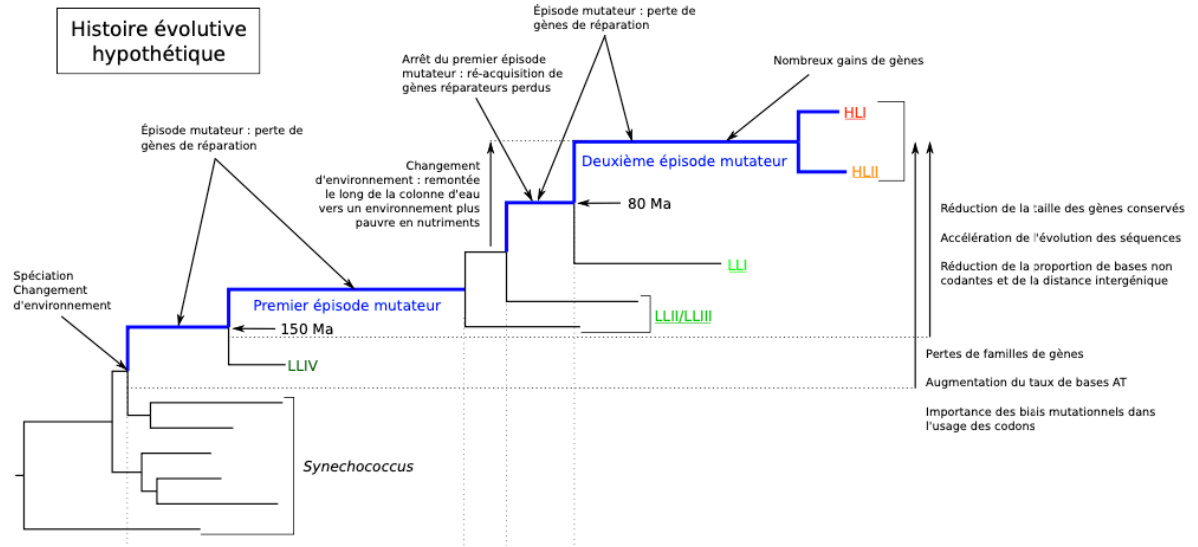


Figure 19 : Histoire évolutive hypothétique pour expliquer les changements génomiques le long de la phylogénie de *Prochlorococcus*. D'après (Batut 2014).

I.6 Production de benchmarks pour la génomique comparative

Stage de M2 : Nicolas Comte. Stage co-encadré à 50% avec Eric Tannier (LBBE et Inria Beagle) et Guillaume Beslon (LIRIS et Inria Beagle).

Post-doctorante : Priscila Biller.

Les travaux que j'ai menés avec Nicolas Comte, Eric Tannier (LBBE et Inria Beagle), Laurent Guéguen (LBBE) et Priscila Biller (post-doctorante Inria Beagle) ont permis de montrer que plus généralement, les données synthétiques produites avec *ævol* (voir par exemple la Figure 20) peuvent être utilisées comme benchmarks pour les méthodes de génomique comparative et de phylogénie moléculaire.

En effet, la préservation des tissus mous est rare dans les fossiles, et il n'existe donc que très peu de séquences d'ADN ancien. Ainsi, en général, on ne dispose pas des séquences ancestrales qui permettraient de valider directement un outil d'inférence phylogénétique. La méthode de validation la plus populaire consiste à développer, en parallèle de l'outil d'intérêt, un simulateur d'évolution de séquences, pour générer des séquences synthétiques dont les états ancestraux sont connus. Cependant, le simulateur est souvent programmé par la même équipe qui développe l'outil, et même lorsque ce n'est pas le cas, il fait en général les mêmes hypothèses simplificatrices que les outils d'inférence phylogénétique. Par exemple, seules les mutations fixées sont simulées parce que ce sont les seules visibles par les outils d'inférence ; la sélection est paramétrée pour correspondre à ce qui est visible par les outils d'inférence ; les gènes sont les unités évolutives parce que ce sont les unités utilisées par les outils d'inférence. Au final, ces simulateurs produisent des jeux de données plutôt faciles pour les outils à valider, mais pas nécessairement réalistes (Biller, Knibbe, et al. 2016). Les simulateurs d'évolution expérimentale *in silico* comme *aevol*, eux, n'ont pas été conçus avec un outil d'inférence phylogénétique en tête et permettent donc de tester ces outils d'inférence de façon plus indépendante.

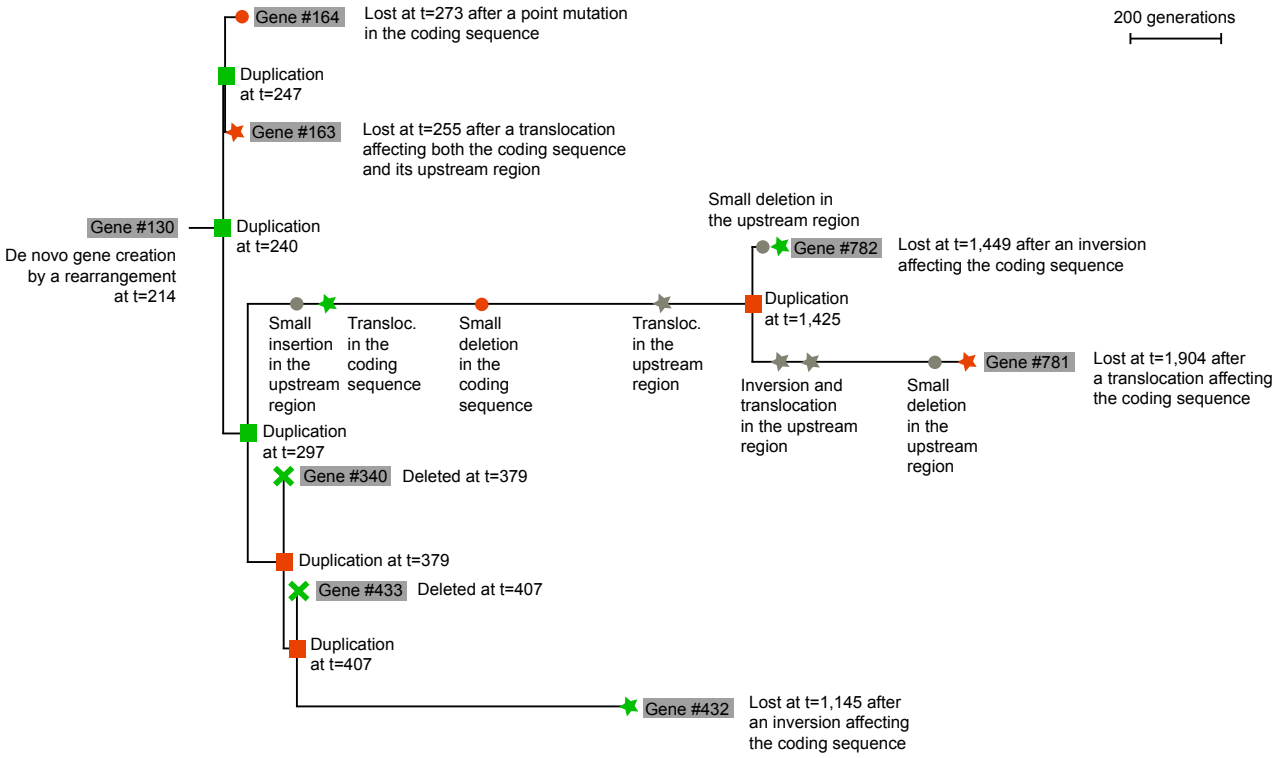


Figure 20 : Exemple de données synthétiques d'évolution moléculaires, produites avec le simulateur *aevol*, d'après (Knibbe & Parsons 2014). Il s'agit de l'histoire d'une famille de gènes. En évolution expérimentale *in silico*, tous les événements sont connus, on dispose d'un "enregistrement fossile" complet. Les événements gris étaient neutres, les verts avantageux et les rouges délétères.

Nous avons utilisé des jeux de données produits avec *aevol* pour tester des estimateurs de distance d'inversion entre deux génomes. La distance d'inversion entre deux génomes est le nombre d'inversions qui se sont produites dans les lignées évolutives qui les séparent. Comme le montre la Figure 21, tous les estimateurs sous-estiment le nombre réel d'inversions lorsque celui dépasse un certain seuil, de 20 inversions environ pour les estimateurs les moins performants à 200 inversions environ pour l'estimateur ER1, que nous avons proposé dans (Biller, Guéguen, et al. 2016).

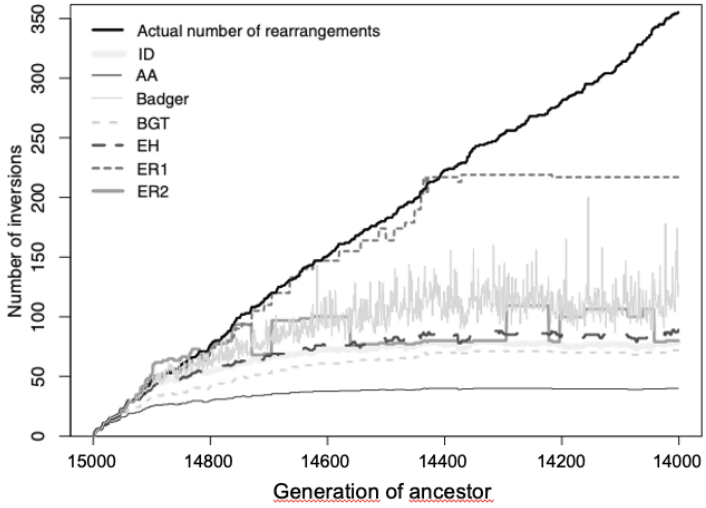


Figure 21 : Évaluation de 7 estimateurs de distance d'inversion sur des données générées avec aevol. La courbe noire représente la distance d'inversion réelle entre le meilleur génome de la génération 15 000, et son ancêtre à une génération donnée, en fonction de la génération de l'ancêtre. Les courbes grises représentent les distances estimées par 7 estimateurs de distance d'inversion. ID: inversion distance (Hannenhalli & Pevzner 1995), AA: Alexeev et Alekseyev (Alexeev & Alekseyev 2017), Badger (Larget et al. 2002), BGT: Biller, Guéguen et Tannier (Biller et al. 2015), EH: Eriksen et Hultman (Eriksen & Hultman 2004), ER1 et ER2: Erdös-Renyi 1 et 2 (Biller, Guéguen, et al. 2016). D'après (Biller, Knibbe, et al. 2016).

La différence entre ER1 et les autres estimateurs est qu'il prend en compte la distribution de taille des séquences intergéniques : au lieu de supposer que la probabilité de cassure est identique entre tous les gènes, ER1 suppose que cette probabilité est variable selon les régions du génome. Ainsi, pour cet estimateur, un génome est représenté par une permutation de gènes (ou plus généralement de "régions solides") et par un vecteur de probabilités de cassure (pour les séquences intergéniques, ou plus généralement les "régions fragiles"). Ces probabilités de cassure peuvent être choisies proportionnelles aux distances intergéniques : c'était le cas pour l'étude de la Figure 21. Les probabilités de cassure co-évoluent avec l'organisation du génome, c'est-à-dire que les inversions modifient à la fois l'ordre des gènes et le vecteur des probabilités de cassure (Figure 22). Par opposition, la plupart des estimateurs existants supposent que le vecteur de probabilités de cassure est constant dans le temps et que toutes les probabilités sont égales à $1/(n+1)$, où n est le nombre de gènes.

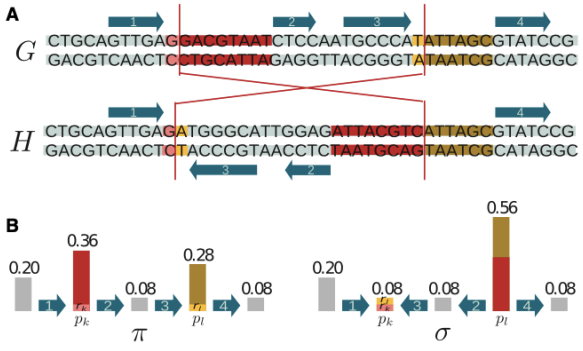


Figure 22 : Principe de l'estimateur ER1 de la distance d'inversion entre deux génomes. Les génomes sont représentés par des permutations de gènes et par un vecteur de probabilités de cassure, qui peuvent par exemple être considérées proportionnelles aux distances intergéniques. Les inversions modifient à la fois la permutation et le vecteur de probabilités. D'après (Biller, Guéguen, et al. 2016).

Ce sont les simulations *aevol* qui ont inspiré la conception de l'estimateur ER1, et, comme attendu, il fonctionne nettement mieux que les autres estimateurs sur les données génomiques simulées avec *aevol*. Dans (Biller, Guéguen, et al. 2016), nous avons testé les performances d'ER1 sur 21 paires de génomes amniotes, à différentes distances évolutives (humain, chimpanzé, macaque, souris, cheval, opossum et poulet). Les résultats montrent que même si ER1 donne des résultats plus plausibles que ceux obtenus sous des probabilités de cassure constantes, on obtient des résultats encore meilleurs si on n'identifie pas les régions fragiles avec les intergènes. C'est l'estimateur que nous avons appelé ER2 : avec ER2, on ne spécifie pas le nombre de régions fragiles, et on le co-estime en même temps que la distance d'inversion. Pour les génomes amniotes considérés, le nombre estimé de régions fragiles varie entre 600 pour les paires de génomes les plus proches phylogénétiquement, et 1800 pour les paires de génomes les plus éloignés.

I.7 Un modèle multi-échelles pour l'évolution des génomes et des réseaux

Projet européen Evoevo (http://www.evoevo.eu/, appel "Evolving Technologies" du programme FP7). Projet porté par Guillaume Beslon (équipe Inria Beagle, Lyon), avec des partenaires microbiologistes et informaticiens de l'Université d'Utrecht (Pays-Bas), l'Université de Valencia (Espagne), l'Université Joseph Fourier (Grenoble) et l'Université de York (Royaume-Uni). CK membre du projet.

Thèse : Charles Rocabert (2013-2017). Co-dirigée à 50% avec Guillaume Beslon (LIRIS et Inria Beagle).

Dans le cadre du projet européen "Evoevo" et de la thèse de Charles Rocabert, nous avons développé un nouveau simulateur multi-échelles, Evo²Sim (Rocabert 2017). Ce nouveau modèle prend en compte la dynamique (ultra-rapide) des réseaux métaboliques, la dynamique (rapide) des réseaux de régulation génique, la dynamique (moyenne à lente) de partage des ressources dans l'écosystème et la dynamique (lente) de l'évolution des gènes et de la structure des génomes. La Figure 23 présente les grands principes de ce modèle. Selon la classification proposée dans (Hindré et al. 2012) et reprise dans la section I.1.2, ce modèle appartient à la famille "génome-collier-de-perles" : la séquence nucléotidique n'est pas explicitement représentée, le génome étant représenté comme une séquence d'éléments à plus gros grain, de différents types (promoteur, site de fixation, séquence codante, séquence non codante).

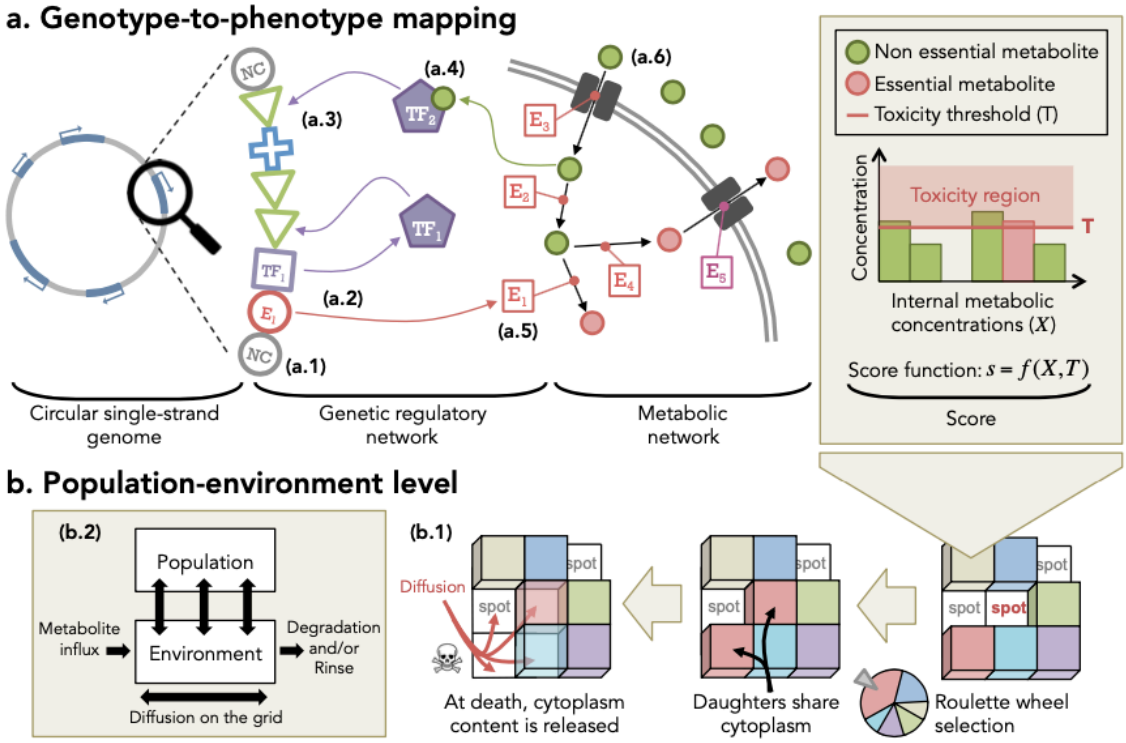


Figure 23 : Principe du modèle Evo²Sim, d'après (Rocabert 2017). (a) Représentation d'un organisme. Dans ce modèle, les organismes (unicellulaires et asexués) possèdent un chromosome circulaire constitué de "perles" de différents types (**a.1**), voir aussi la **Figure 24**). Toutes les séquences de perles ne sont pas fonctionnelles. Les régions fonctionnelles sont celles constituées d'un promoteur, éventuellement flanqué d'un ou plusieurs sites de fixations, suivi d'une ou plusieurs séquences codantes. Les séquences codantes peuvent coder pour des pompes, des enzymes ou des facteurs de transcription. Les pompes font entrer ou sortir des métabolites dans l'organisme (**a.6**). Les enzymes transforment un métabolite en un autre métabolite (**a.5**). Les métabolites peuvent se fixer aux facteurs de transcription et les activer ou les inhiber (**a.4**). Lorsqu'un facteur de transcription actif se lie à un site de fixation en aval (resp. en amont) d'un promoteur, il inhibe (resp. active) l'expression des séquences codantes sous le contrôle de ce promoteur (**a.3**). Le score d'un organisme est la somme des concentrations internes des métabolites prédéfinis comme essentiels, par exemple ceux dont l'index est un nombre premier. Un seuil létal de toxicité est prédéfini pour les concentrations internes de tous les métabolites. Un organisme a aussi une certaine probabilité de mourir par hasard à chaque pas de temps. **(b) Population et environnement**. Chaque organisme occupe un emplacement sur une grille 2D. Quand il meurt, ses métabolites internes sont relargués sur son emplacement et diffusent dans la grille. Les éventuels organismes voisins se disputent l'emplacement libéré : celui dont le score est le plus élevé se divise sur l'emplacement libre, si son score dépasse une valeur minimale prédéfinie. Lors de la division, le contenu cytoplasmique (métabolites et protéines) est réparti à parts égales dans les deux cellules filles. Un afflux externe de métabolites peut être programmé à un rythme régulier ou aléatoire. A chaque pas de temps, une fraction des métabolites de l'environnement est retirée.

La Figure 24 présente le détail des équations qui régissent les réseaux intracellulaires. Le réseau métabolique et le réseau de régulation sont couplés par le fait qu'un facteur de transcription peut être activé ou inhibé par l'association avec un métabolite.

Type of genetic unit	Attributes	Graphical symbol
Non coding unit (NC)	No attributes;	NC
Promoter unit (P)	Basal expression level β;	
Binding site unit (BS)	Transcription factor tag TF_{tag};	
Transcription factor coding unit (TF)	Binding site tag BS_{tag}; Co-enzyme tag CoE_{tag}; Co-enzyme constant k_{CoE}; Free activity A_{free}; Bound activity A_{bound}; Binding window W_{bind};	TF
Enzyme coding unit (E)	Substrate tag s; Product tag p; k_{cat} constant; k_{cat}/k_m constant;	E

Metabolic network

$$\frac{d[m_i]}{dt} = -\sum_p \frac{k_{cat}^{E_{i\to p}}[E_{i\to p}][m_i]}{K_M^{E_{i\to p}} + [m_i]}$$ enzymes using metabolite i as substrate

$$+\sum_s \frac{k_{cat}^{E_{s\to i}}[E_{s\to i}][m_s]}{K_M^{E_{s\to i}} + [m_s]}$$ enzymes producing metabolite i

In the special case where $s = p$, the enzyme is considered as a pump, actively pumping *in* the metabolite s if k_{cat} is positive, or *out* if k_{cat} is negative.

Regulatory network

$$\frac{d[E_{s\to p}]}{dt} = -\phi[E_{s\to p}] + \beta_{E_{s\to p}}\left(\frac{\theta^n}{(\sum_{d\in \text{down-stream binding sites}}\sum_{k\in \text{trans-cription factors}} a_{kd}[TF_k^+])^n + \theta^n}\right)\cdot\left(1+\left(\frac{1}{\beta_{E_{s\to p}}}-1\right)\left(\frac{(\sum\sum a_{ku}[TF_k^+])^n}{(\sum_{u\in \text{up-stream binding sites}}\sum_{k\in \text{trans-cription factors}} a_{ku}[TF_k^+])^n + \theta^n}\right)\right)$$

$$\frac{d[TF_j]}{dt} = -\phi[TF_j] + \beta_{TF_j}\left(\frac{\theta^n}{(\sum_{d\in \text{down-stream binding sites}}\sum_{k\in \text{trans-cription factors}} a_{kd}[TF_k^+])^n + \theta^n}\right)\cdot\left(1+\left(\frac{1}{\beta_{TF_j}}-1\right)\left(\frac{(\sum\sum a_{ku}[TF_k^+])^n}{(\sum_{u\in \text{up-stream binding sites}}\sum_{k\in \text{trans-cription factors}} a_{ku}[TF_k^+])^n + \theta^n}\right)\right)$$

with

$$[TF_j^+] = \begin{cases} \frac{[m_{CoE\,tag}][TF_j]}{k_{CoE}^{TF_j}+[m_{CoE\,tag}]} & \text{if } A_{free}^{TF_j}=0 \text{ and } A_{bound}^{TF_j}=1 \\ \frac{k_{CoE}[TF_j]}{k_{CoE}^{TF_j}+[m_{CoE\,tag}]} & \text{if } A_{free}^{TF_j}=1 \text{ and } A_{bound}^{TF_j}=0 \\ [TF_j] & \text{if } A_{free}^{TF_j}=1 \text{ and } A_{bound}^{TF_j}=1 \\ 0 & \text{if } A_{free}^{TF_j}=0 \text{ and } A_{bound}^{TF_j}=0 \end{cases}$$

$$a_{kd} = \begin{cases} 1 - \frac{|BStag^{TF_k} - BStag^{site_d}|}{W_{bind}^{TF_k}} & \text{if } |BStag^{TF_k} - BStag^{site_d}| < W_{bind}^{TF_k} \\ 0 & \text{else} \end{cases}$$

Energy

$$\frac{d\varepsilon}{dt} = c_{enz}\sum_{s,p}\left(\frac{k_{cat}^{E_{s\to p}}[E_{s\to p}][m_s]}{K_M^{E_{s\to p}}+[m_s]}(s-p)\right) - c_{pump}\left(\sum_s \frac{k_{cat}^{P_s^{in}}[P_s^{in}][m_{s_{out}}]}{K_M^{P_s^{in}}+[m_{s_{out}}]} + \sum_s \frac{k_{cat}^{P_s^{out}}[P_s^{out}][m_s]}{K_M^{P_s^{in}}+[m_s]}\right)$$

$$- c_{expr}\left(\sum_{s,p}\left(\text{production term of}\frac{d[E_{s\to p}]}{dt}\right) + \sum_s\left(\text{production term of}\frac{d[P_s^{in}]}{dt}\right) + \sum_s\left(\text{production term of}\frac{d[P_s^{out}]}{dt}\right)\right.$$

$$\left. + \sum_j\left(\text{production term of}\frac{d[TF_j]}{dt}\right)\right)$$

Figure 24 : Les "perles" qui constituent les génomes virtuels du modèle Evo²Sim contiennent les paramètres d'un système d'équations différentielles ordinaires donnant la dynamique d'un réseau métabolique et d'un réseau de régulation génique. Les deux réseaux sont couplés par le fait qu'un facteur de transcription peut être activé ou inhibé par l'association avec un métabolite. Dans les équations, les paramètres verts sont ceux encodés dans les "perles" du génome, ils peuvent évoluer par mutations.

Dans le modèle, les mutations ponctuelles peuvent changer le substrat, le produit ou les paramètres cinétiques d'une enzyme ou d'une pompe, le niveau d'expression basal d'un promoteur, l'affinité entre un facteur de transcription et un site de fixation, ou encore l'action d'un métabolite sur un facteur de transcription. Les duplications et les délétions peuvent modifier le nombre de gènes et donc d'enzymes, de pompes ou de facteurs de transcription. Les inversions peuvent changer l'ordre des gènes.

Dans ce modèle multi-échelles, il était important de bien calibrer les différentes échelles de temps les unes par rapport aux autres. Nous avons calibré les paramètres du modèle de façon aussi réaliste que possible, en nous basant sur des données disponibles pour *Escherichia coli* : volume cellulaire (1.5 à 4.4 μm^3) (Volkmer & Heinemann 2011), plage de concentrations protéiques typiques (5 à 50000 nM, avec 75% des protéines sous les 500 nM) (Ishihama et al. 2008), demi-vie des protéines (2 minutes à 70 heures) (Maurizi 1992), plage de concentrations typiques pour les métabolites intracellulaires (10^{-7} à 10^{-2} M). Pour les paramètres cinétiques des enzymes, nous avons utilisé les plages de valeurs observées par (Bar-Even et al. 2011) pour les milliers d'enzymes des bases de données Brenda et KEGG : 6 à 60000 min^{-1} pour k_{cat}, 10^{-7} à 10^{-1} M K_M, et 6.10^4 à 6.10^8 $min^{-1}M^{-1}$ pour le ratio k_{cat}/K_M.

Nous avons utilisé ce modèle pour répondre à deux questions : (1) quelles sont les conditions environnementales qui favorisent la diversification en différents écotypes stables, et (2) comment l'organisation du génome, le réseau métabolique et le réseau de régulation co-évoluent.

Pour la première question, nous avons reproduit *in silico* les différentes dynamiques des ressources lors d'une expérience d'évolution en laboratoire, selon que la culture est faite en chemostat (apport continu de milieu frais) ou en batch (renouvellement complet du milieu à intervalles de temps régulier). Les simulations montrent que seule la culture en batch permet l'évolution d'interactions de crossfeeding stables à long terme. Par crossfeeding, nous entendons l'émergence de deux souches, l'une qui consomme la ressource fournie dans le milieu, et l'autre qui se nourrit des sous-produits de la première (Figure 25). En chemostat, de telles interactions peuvent apparaître, mais elles ne sont pas stables à l'échelle du temps évolutif : un mutant généraliste finit toujours par envahir la population (Rocabert et al. 2017). C'est la saisonnalité introduite par les changements de milieu réguliers — une première saison riche en nourriture, suivie d'une seconde saison riche en sous-produits de "digestion" — , qui créée deux niches temporelles. L'existence de ces deux niches permet la spécialisation des deux souches, qui peut être une première étape vers une spéciation. Ainsi, la dynamique temporelle des ressources peut permettre une spéciation sympatrique, c'est-à-dire sans barrière géographique.

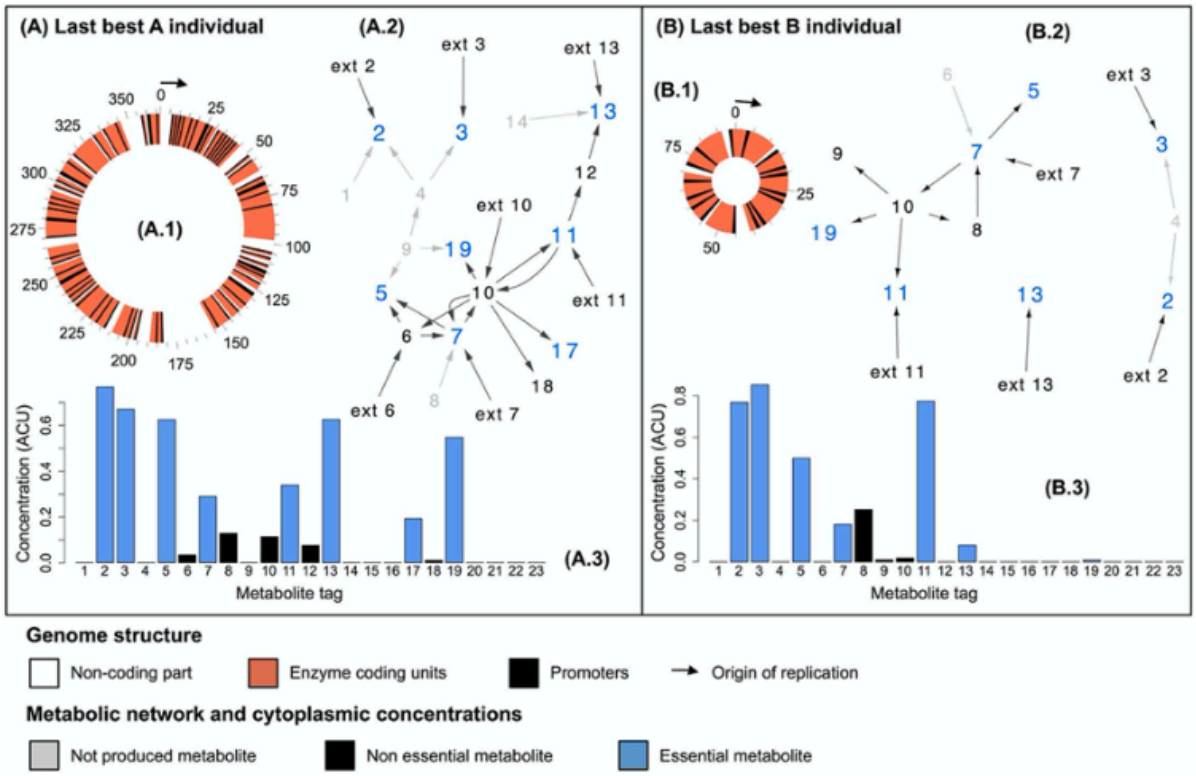


Figure 25 : Génomes et réseaux métaboliques de deux souches bactériennes virtuelles obtenues avec le simulateur Evo2Sim, d'après (Rocabert et al. 2017). Les métabolites sont identifiés par des "tags" arbitraires. La ressource fournie dans le milieu est le métabolite #10. Ici, deux souches ont émergé dans la population. La souche A consomme entre autres la ressource primaire (#10) tandis que la souche B ne la consomme pas, et consomme à la place des sous-produits de la souche A (#2, #3, #7, #11, #13).

Dans cette première série d'expériences, les populations étaient initialisées en tirant des génomes aléatoires jusqu'à ce qu'un génome viable soit trouvé, typiquement un génome contenant au moins une pompe pour le métabolite fourni dans le milieu, et une enzyme transformant ce métabolite en métabolite essentiel. Le réseau métabolique se complexifiait ensuite sous l'effet du processus évolutif. Par contre, bien que la régulation soit possible, elle n'a jamais émergé dans cette série d'expériences.

Ainsi, pour la seconde question, c'est-à-dire la façon dont le génome, le réseau métabolique et le réseau de régulation co-évoluent, nous avons initialisé les génomes avec un réseau de régulation minimal, spécialement conçu pour permettre la survie dans un environnement où les métabolites #20 et #22 sont fournis aléatoirement selon un processus de Poisson (Figure 26).

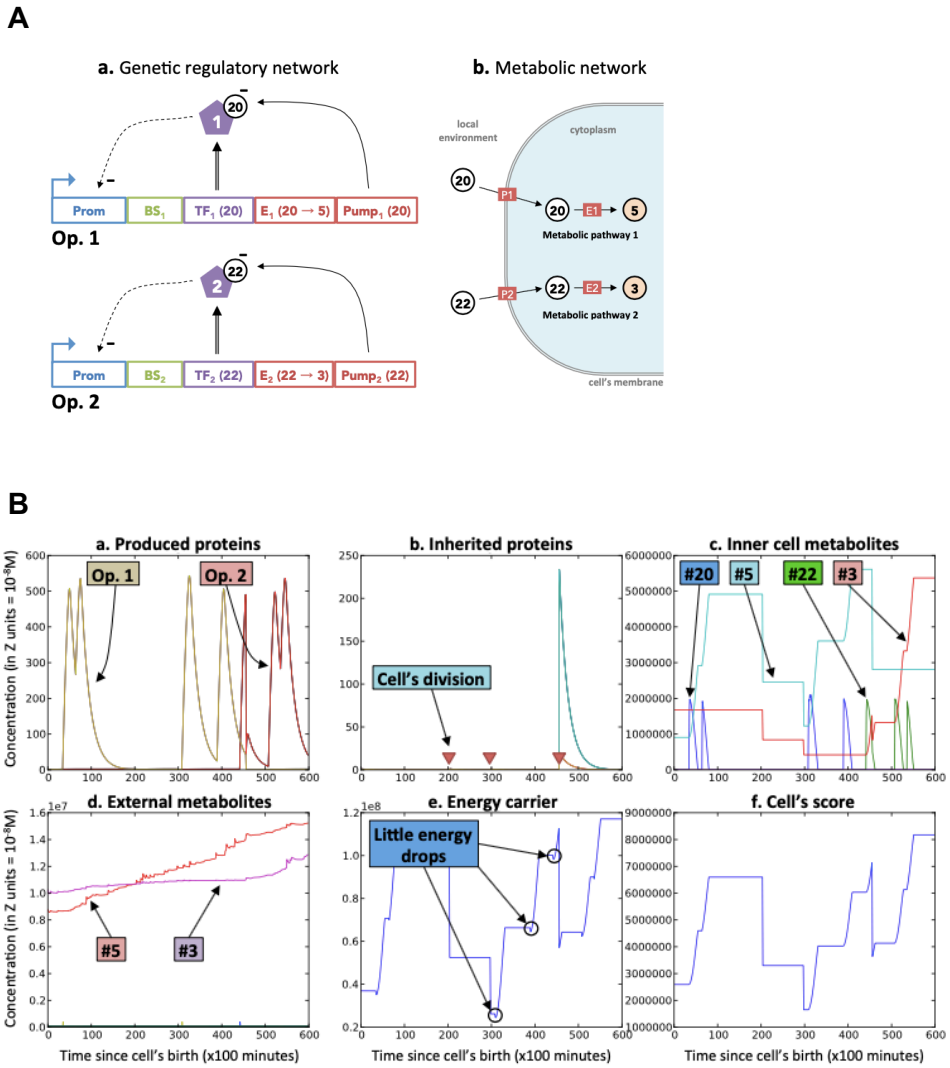


Figure 26 : A. Génome et réseaux initiaux, conçus à la main. Les réactions enzymatiques #20 → #5 et #22 → #3 sont cataboliques et fournissent de l'énergie à la cellule, mais les pompes pour les métabolites #20 and #22 nécessitent de l'énergie, ainsi que l'expression des protéines. Chacun des deux opérons s'auto-inhibe, sauf si son métabolite primaire est présent dans l'environnement. **B. Dynamiques des protéines, des métabolites et du score** dans un environnement où les métabolites #20 et #22 sont fournis aléatoirement selon un processus de Poisson**.** D'après (Rocabert 2017).

Ce génome conçu à la main n'était visiblement pas optimal, puisqu'en le laissant évoluer, il s'est considérablement modifié et est devenu plus robuste aux périodes de famine qui arrivent tôt ou tard dans le processus de Poisson, et qui mettent la population en danger d'extinction (Figure 27A). La Figure 27B montre les génomes finaux obtenus après évolution, pour deux simulations représentatives avec ou sans coût énergétique à l'expression des protéines. La taille et l'organisation du génome final dépend fortement de la présence ou non d'une contrainte énergétique sur l'expression des protéines. A taux de réarrangements chromosomiques égaux, les génomes obtenus sans ce coût sont en moyenne 4 fois plus longs qu'avec, contiennent 4 fois plus de non codant, 26 fois plus d'opérons mais ces opérons contiennent chacun 4.5 fois moins de séquences codantes (Rocabert 2017). Ainsi,

dans le modèle multi-échelles, les contraintes énergétiques ont une influence sur la taille du génome et sa densité en gènes. Une perspective immédiate de ce travail serait de quantifier l'impact relatif de ces contraintes énergétiques et des taux de réarrangements chromosomiques : le modèle mathématique de Stephan Fischer (section I.4.1) prédit que les taux de duplications et de grandes délétions fixent une borne supérieure $Q_1(s^{\max,(1)})$ en-dessous de laquelle la sélection (donc en particulier les contraintes énergétiques) peut jouer, vers le rétrécissement ou vers la croissance, mais sans dépasser la borne supérieure fixée par les taux de réarrangements.

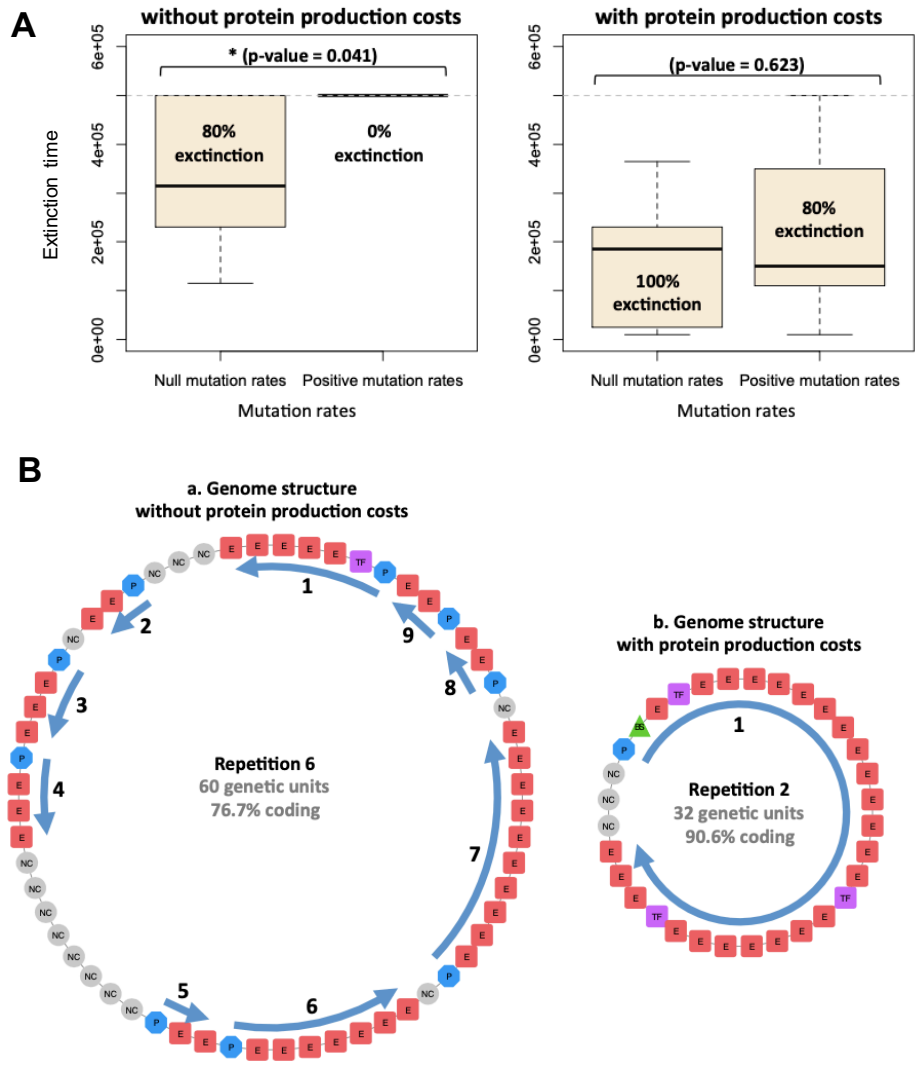


Figure 27 : A. L'évolution produit des organismes plus robustes que ceux conçus à la main vis-à-vis des périodes de "famine" induites par le processus poissonien d'apport de ressources. La durée maximale de simulation était de 500 000 pas de temps. **B. Exemples de génomes finaux obtenus après évolution.** Noter l'absence de sites de fixation (triangles verts), donc de réseau de régulation, dans le génome évolué en l'absence de coût énergétique à l'expression des protéines. D'après (Rocabert 2017).

La présence ou non d'un coût énergétique à l'expression des protéines semble aussi déterminer l'intérêt évolutif de la régulation. En effet, toutes les populations ayant évolué avec ce coût énergétique ont conservé un réseau de régulation, tandis qu'aucune n'en a conservé en l'absence de ce coût. Ainsi, en l'absence de coût, toutes les voies métaboliques

sont exprimées de façon constitutive dans tous les génomes finaux, malgré la dynamique fluctuante des ressources externes (Rocabert 2017). Ces résultats sont cohérents avec ceux obtenus par (Weiße et al. 2015) avec un modèle mathématique de croissance bactérienne et de compétition entre souches : l'intérêt évolutif de réguler l'expression des voies métaboliques ne serait pas tant de pouvoir exprimer les enzymes à des concentrations finement adaptées à celles des substrats, mais plutôt de pouvoir les exprimer à des concentrations compatibles avec les contraintes internes de la cellule.

Il faut cependant noter qu'en présence du coût énergétique à l'expression des protéines, un réseau de régulation est certes conservé, mais il est différent de celui conçu à la main pour l'initialisation. Dans l'exemple représentatif de la Figure 28, un seul facteur de transcription, inactivé par le métabolite #20, régule toutes les enzymes. Le réseau métabolique est lui aussi différent du réseau initial. Il est plus complexe et plus connecté, et laisserait bien des ingénieurs perplexes. Dans ces simulations, l'évolution a façonné des réseaux métaboliques complexes alors que la dynamique des ressources environnementales est apparemment relativement simple. Par ressources environnementales, nous, humains, pensons aux métabolites #20 et #22 introduits dans l'environnement, mais les organismes ont aussi accès aux métabolites internes libérés par leurs voisins lorsqu'ils meurent, des ressources que l'évolution n'a pas manqué d'exploiter. Il est en fait assez fréquent en évolution expérimentale *in silico* d'obtenir des organisations moléculaires complexes "gratuitement", c'est-à-dire sans avoir explicitement cherché à sélectionner une organisation complexe (voir par exemple (Liard et al. 2020)). La complexité résulte du fait que l'évolution procède essentiellement par "bricolage" (Jacob 1977), sans but, "réalisant ses « menus travaux » essentiellement « au petit bonheur » en fonction des conditions et des éléments rencontrés fortuitement au fil du temps et des rencontres (on retrouve d'ailleurs ici, étonnamment, l'étymologie ancienne du verbe bricoler : « aller par-ci par-là, en zig-zag[12] »)" (Beslon & Knibbe 2010).

[12] D'après le Dictionnaire historique de la langue Française, Ed. Le Robert, 2006

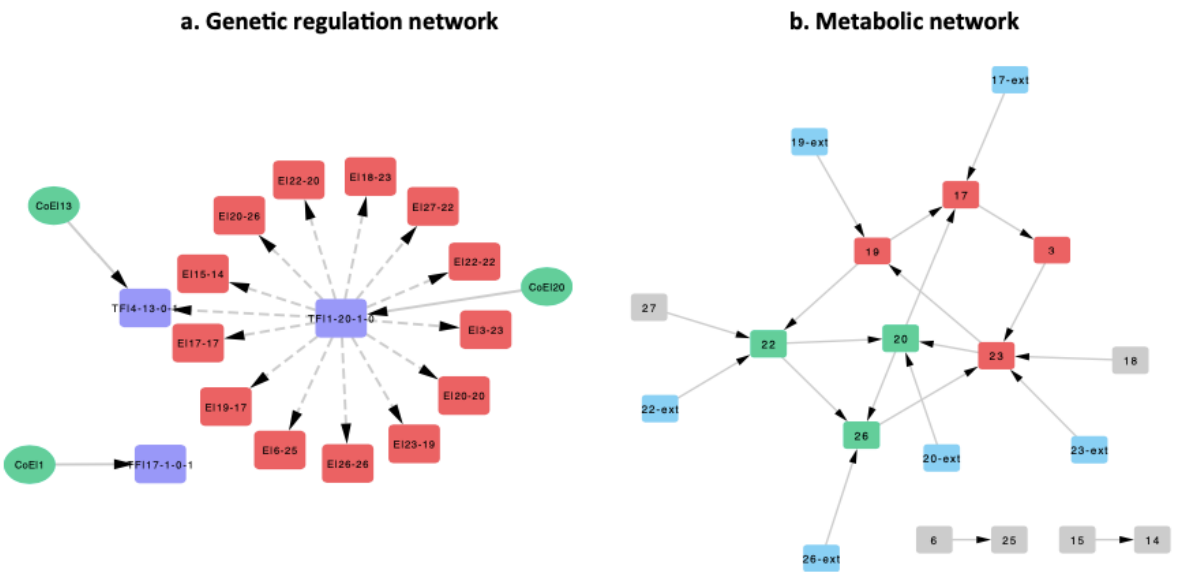


Figure 28 : Réseau de régulation et réseau métabolique d'un des organismes finaux après évolution en présence d'un coût énergétique à l'expression des protéines. D'après (Rocabert 2017).

I.8 Conclusion du chapitre

Dans son ensemble, le travail mené avec les doctorants que j'ai co-supervisé, avec mes étudiants de masters et avec mes collaborateurs montre que l'évolution *in silico* est source de surprises, dans le sens où elle nous révèle parfois que l'expérience de pensée peut être trompeuse. C'était le cas par exemple quand nous avons vu les génomes évoluer sans aucun problème en l'absence de mutations ponctuelles, ou *diminuer* en taille alors qu'on imposait un taux de duplications deux fois supérieur au taux de délétions. C'était le cas aussi lorsque nous avons vu des réseaux de régulation évoluer dans *aevol* alors que l'environnement était constant. C'était encore le cas quand l'évolution a considérablement modifié le beau réseau de régulation que nous avions construit à la main pour initialiser les individus dans Evo²Sim, alors que nous le pensions remarquablement bien adapté à la dynamique des ressources.

Toutefois, l'évolution *in silico* n'est pas complètement imprédictible non plus, dans la mesure où il nous a été possible de trouver un invariant fondamental indirectement sélectionné dans *aevol* (le nombre des descendants neutres), de mathématiser ce modèle computationnel, de prouver des conditions de convergence (le taux de duplications peut être jusqu'à 2.6 fois supérieur au taux de délétion) et des bornes supérieures pour la taille du génome, de comprendre que le réseau amélioré par l'évolution dans Evo²Sim correspond en fait mieux que le nôtre à la dynamique réelle des ressources.

Ce sont les possibilités offertes d' "ouvrir la boîte" et de réaliser des expériences impossibles qui rendent l'approche intéressante, et complémentaire des approches d'évolution expérimentale ou de génomique comparative. Complémentaire, car nous avons montré comment l'évolution *in silico* peut être utilisée pour tester des scénarios évolutifs relatifs à des génomes réels (bactéries endosymbiotiques et cyanobactérie *Prochlorococcus*) ou pour

produire des benchmarks pour des méthodes de génomique comparative (estimation de la distance d'inversion).

En passant d'aevol à Evo²Sim, j'ai amorcé une transition vers la modélisation des réseaux cellulaires et vers l'aspect quantitatif dans le paramétrage de ces réseaux. J'ai eu l'opportunité en 2017 de muter de l'Université Lyon 1 à l'INSA de Lyon, pour prendre la direction du parcours Bioinformatique et Modélisation de l'INSA, mais aussi pour passer d'un laboratoire d'informatique (LIRIS) à un laboratoire de recherche biomédicale dans le domaine des maladies métaboliques, de l'obésité, du diabète de type 2 et de leurs complications cardiovasculaires (CarMeN). Je suis restée membre de l'équipe Inria Beagle, mais j'ai basculé de son axe Evolution *in silico* vers son axe Biochimie computationnelle. J'ai pu, à cette occasion, continuer mon virage thématique vers la modélisation mécanistique et quantitative des réseaux cellulaires, mais cette fois dans le domaine de la nutrition.

Chapitre II : Contributions en nutrition

> *The nutritional evaluation of the relation between diet and health has traditionally focused on individual food constituents such as proteins, fats, carbohydrates, and micronutrients separately. This reductionist approach, which links one nutrient to one health effect, may partly explain some of the discrepancies between a food's predicted health effect on the basis of its nutrient content and its actual health effect when consumed as a whole food. A diet does not consist of single nutrients but of whole foods, either alone or alongside many other foods as part of a meal. Foods have complex structures both physically and nutritionally, which affect digestion and absorption and may generate interactions within the food matrix, thereby altering the bioactive properties of nutrients in ways that are not predictable from the nutrition-label information.*
>
> (Thorning et al. 2017)

À mon arrivée au laboratoire CarMeN, j'ai contribué, via des analyses statistiques, à différents projets de l'équipe de Marie-Caroline Michalski, visant à caractériser l'impact de différents facteurs sur la digestion et l'absorption des lipides. Cela m'a permis de me familiariser avec les différents types de données expérimentales disponibles au laboratoire : cinétiques enzymatiques *in vitro*, données d'absorption lipidique, de sécrétion et d'expression de gènes dans des cultures cellulaires, cinétiques lymphatiques chez l'animal après infusion intraduodénale de lipides, cinétiques plasmatiques chez des volontaires après un repas.

Ce sont ces données (ainsi que celles de la littérature) qui pourront être utilisées pour calibrer les modèles mécanistiques qui font l'objet de mon projet de recherche, en lien avec l'axe Biochimie computationnelle de l'équipe Inria Beagle. Le projet en question sera décrit dans le chapitre suivant. Le présent chapitre décrit les principaux projets CarMeN auxquels j'ai contribué, ainsi que le contexte scientifique dans lequel ils s'inscrivent.

II.1 Contexte scientifique

Durant les heures qui suivent un repas (phase dite "postprandiale"), les nutriments sont digérés et absorbés au niveau de l'intestin, puis leur concentration dans le sang augmente temporairement, et enfin diminue tandis qu'ils sont progressivement utilisés ou stockés par les différents tissus. On appelle *cinétique postprandiale* l'évolution au cours du temps de la concentration d'une molécule dans le plasma (ou plus rarement dans la lymphe), après l'absorption orale de nutriments.

Les cinétiques postprandiales consécutives à l'absorption de glucides sont très étudiées car elles constituent les pierres angulaires du diagnostic et du suivi des maladies nutritionnelles. En effet, l'un des examens clés pour le diagnostic du diabète consiste à mesurer la cinétique postprandiale du glucose et de l'insuline après l'absorption par voie orale d'une quantité

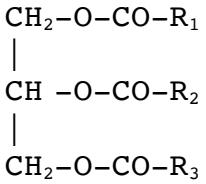
standard de glucose (OGTT, oral glucose tolerance test). L'absence de pic insulinémique après absorption, un pic insulinémique tardif, ou une forte augmentation de la glycémie sont des anomalies qui révèlent une insulino-résistance ou un diabète.

Les lipides[13] dans le sang sont, eux, traditionnellement mesurés de façon statique et à jeun : le bilan lipidique standard pour évaluer le risque cardio-vasculaire consiste à mesurer les taux de triglycérides et de cholestérol après 12h de jeûne. Cependant, un changement de paradigme s'opère depuis une dizaine d'années. De nouvelles recommandations préconisent de mesurer la lipémie en phase postprandiale plutôt qu'à jeun, car, pour les triglycérides au moins, la mesure postprandiale s'avère être un meilleur prédicteur du risque cardiovasculaire que la mesure à jeun (Bansal et al. 2007; Kolovou et al. 2011; Nordestgaard et al. 2016). Plus spécifiquement, une lipémie postprandiale de forte amplitude sur une longue durée constitue un facteur de risque élevé de développement de maladies cardio-vasculaires (Jackson et al. 2012). Des tests "OFTT" (oral fat tolerance tests), inspirés des tests OGTT, sont donc en cours de standardisation.

En plus de leur rôle dans le diagnostic, les cinétiques postprandiales peuvent aussi être des atouts en matière de prévention. En effet, elles dépendent non seulement de l'état physiologique du sujet (obèse ou non, diabétique ou non, par exemple), mais aussi de la nature des aliments absorbés. Or la quantification précise de l'impact des aliments sur la santé est un enjeu majeur pour la recherche en nutrition (van Ommen, Keijer, et al. 2008). Dans une optique de prévention, il s'agit de détecter cet impact avant qu'il ne se traduise par des maladies ou par des lésions dans les organes.

Plusieurs études ont montré que les cinétiques postprandiales sont sensibles aux interventions nutritionnelles et peuvent donc servir à quantifier l'impact des aliments sur la santé. Par exemple, l'addition de fibres solubles et visqueuses à un repas ralentit l'absorption des glucides et améliore la régulation de la glycémie, tandis que des régimes riches en acides gras polyinsaturés oméga-3 à longue chaîne peuvent diminuer la lipémie postprandiale (Vors et al. 2014). Plus étonnant, à composition identique en nutriments, les cinétiques postprandiales s'avèrent également sensibles à l'organisation spatiale des molécules. Ainsi, une même matière grasse conduira à un pic de lipémie plus précoce, plus prononcé, mais aussi plus rapidement éliminé si elle est émulsionnée que si elle est absorbée sous forme solide, en particulier chez des sujets obèses comme l'a montré le laboratoire CarMeN (Vors et al. 2013).

Il est donc important d'étudier les facteurs susceptibles de moduler le devenir des lipides alimentaires à toutes les étapes de la phase postprandiale. Comme les lipides alimentaires sont constitués d'environ 97% de triglycérides, environ 3% de phospholipides (Armand 2008) et moins de 0.5% de cholestérol (typiquement 300 mg/jour, (Xu et al. 2018)), je me focalise principalement ici sur les triglycérides et leur devenir en phase postprandiale. Leur formule générale est de la forme :



[13] Les lipides englobent les acides gras et leurs dérivés (triglycérides et phospholipides surtout), mais aussi les métabolites comprenant des stérols, comme le cholestérol.

où R_1, R_2 et R_3 sont des chaînes carbonées d'acides gras, comportant chacune entre 4 et 28 atomes de carbone dans la plupart des triglycérides naturels. Les triglycérides sont hydrophobes. Dans les aliments transformés, on les trouve sous différentes organisations supramoléculaires (Vors et al. 2016) :

- sous forme libre, en phase continue homogène : cas des huiles ou du saindoux ;
- sous forme de gouttelettes d'émulsion huile-dans-eau : cas du lait, des sauces vinaigrettes et mayonnaises, ou encore des yaourts qui sont des émulsions gélifiées. Ces émulsions sont stabilisées par des molécules tensioactives généralement amphiphiles : protéines (caséines, lactoferrine**,** protéines du lactosérum…), biopolymères non protéiques (alginate de propane-1,2-diol, pectine acétylée…), ou lipides polaires (lécithines, polysorbates…) ;
- sous forme de phase continue d'une émulsion eau-dans-huile : cas du beurre ou de la margarine ;
- sous forme d'inclusions lipidiques dans des matrices solides glucidiques ou protéiques : cas des produits élaborés comme les biscuits et les fromages.

Dans la cavité buccale, les triglycérides sont exposés à des forces mécaniques et à la salive, qui permettent une émulsification grossière des huiles et l'inversion des émulsions eau-dans-huile en émulsions huile-dans-eau. Une fois dans l'intestin, les sels biliaires permettent une émulsification plus fine grâce à leurs propriétés tensioactives. Toutefois, même émulsionnés, les triglycérides ne peuvent pas traverser la couche de mucus qui borde les cellules épithéliales de l'estomac et de l'intestin, ni la membrane de ces cellules (entérocytes).

Les lipases présentes dans l'estomac et l'intestin agissent à l'interface huile-eau, à la surface des gouttelettes de l'émulsion. Elles hydrolysent les triglycérides, produisant des monoglycérides et des acides gras libres, qui, eux, pourront être absorbés par les entérocytes après avoir été solubilisés dans des micelles mixtes par les sels biliaires (Figure 29). Dans les entérocytes, les acides gras et monoglycérides sont réassemblés en triglycérides puis empaquetés avec des lipides polaires, du cholestérol et des protéines dans des assemblages appelés chylomicrons. Leur taille, entre 20 et 1000 nm, ne leur permet pas de traverser la paroi des capillaires sanguins. Ils rejoignent plutôt les capillaires lymphatiques de l'intestin (chylifères, "lacteals" en anglais). Ils rejoignent ensuite la circulation sanguine au niveau du conduit thoracique, qui amène la lymphe dans la veine sous-clavière gauche. Dans le plasma, le pic de lipémie postprandiale dû à la présence de chylomicrons circulants s'observe entre 2h et 4h après le début du repas.

La clairance de la lipémie postprandiale est due à l'action de la lipoprotéine lipase (LPL), une enzyme synthétisée par les cellules musculaires et adipeuses. Elle sort des cellules et se fixe sur les cellules endothéliales des capillaires sanguins qui irriguent les tissus utilisant les acides gras libres comme nutriments, principalement au niveau du tissu adipeux, du cœur, des muscles squelettiques et des glandes mammaires. Les triglycérides transportés par les chylomicrons sont hydrolysés par la LPL. Les molécules d'acides gras et de glycérol ainsi libérées peuvent alors être absorbées par les cellules avoisinantes, en particulier celles des muscles pour l'énergie et celles du tissu adipeux pour le stockage.

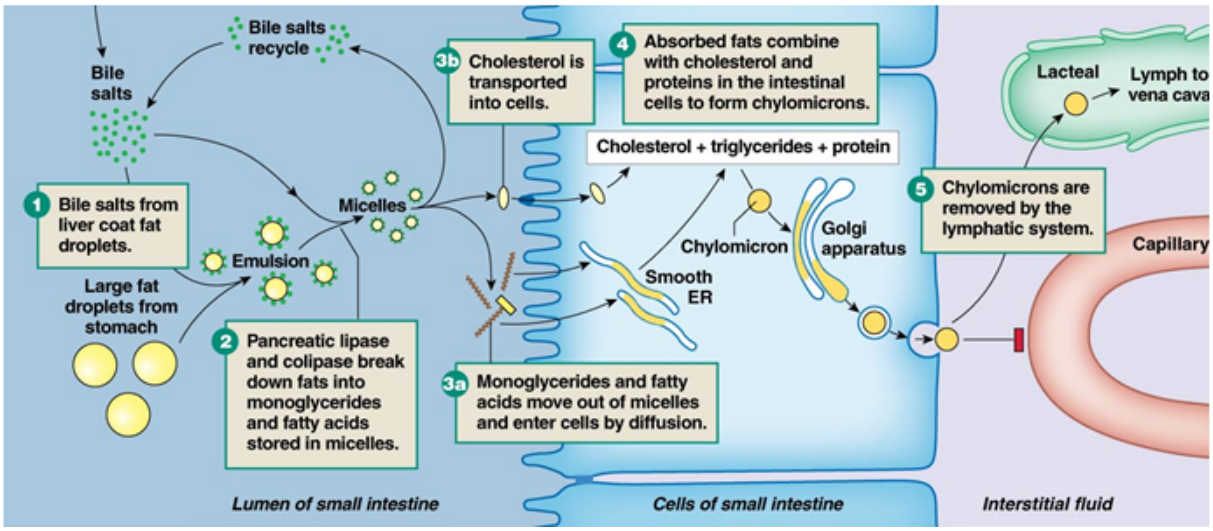


Figure 29 : Digestion et absorption des triglycérides alimentaires. Source : https://www.easynotecards.com.

Depuis mon arrivée au laboratoire CarMeN en septembre 2017, j'ai contribué, via les analyses statistiques, à plusieurs projets visant à caractériser des facteurs susceptibles de moduler le devenir des lipides alimentaires, au niveau de la lipolyse luminale, de la traversée des entérocytes ou du passage dans le système lymphatique. Ces contributions, décrites ci-dessous, m'ont permis de me familiariser avec le type de données que je pourrai utiliser pour mon projet de modélisation (décrit, lui, dans le chapitre suivant).

II.2 Influence de l'incorporation de traceurs au ^{13}C

Stage de L3 : Juliette Geoffray (2017). Stage encadré à 100%.

Stage de M1 : Damien Agopian (2018). Stage co-encadré avec Marion Létisse (CarMeN).

Dans les études cliniques visant à suivre le devenir des lipides alimentaires, la méthode de référence (gold standard) consiste à incorporer dans la matière grasse ingérée des triglycérides marqués au carbone 13, un isotope non radioactif du carbone. Un bon traceur se doit d'être neutre vis-à-vis du processus étudié : il est donc important de vérifier que les traceurs ne modifient pas la cinétique de digestion de la matière grasse dans laquelle ils sont incorporés.

Dans le cadre de la collaboration de l'équipe de Marie-Caroline Michalski (CarMeN) avec Sabine Danthine (Gembloux Agro-Bio Tech, Belgique) et Frédéric Carrière (Université d'Aix-Marseille) et du stage de M1 de Damien Agopian que j'ai co-encadré avec Marion Létisse, nous avons montré que l'incorporation de 5.7% de traceurs au ^{13}C modifie la cinétique de lipolyse de la matière grasse laitière, dans un modèle *in vitro* de l'étape de lipolyse luminale (Figure 30). Le "lag time", défini comme l'intersection de la tangente de plus forte pente avec l'axe du temps, passe de 2 à 4 minutes en présence de 5.7% de traceurs.

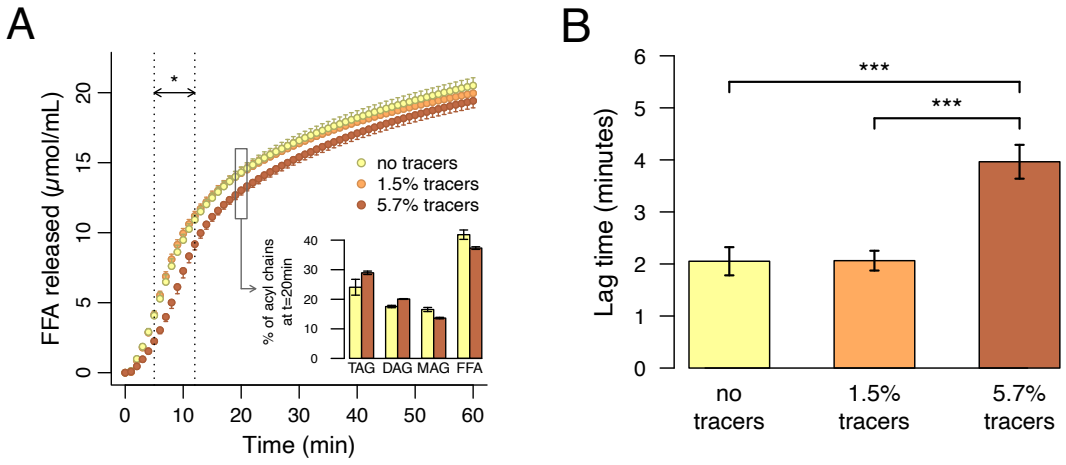


Figure 30 : Influence de l'incorporation de traceurs au ^{13}C dans de la matière grasse laitière sur la cinétique de lipolyse *in vitro*. (A) Acides gras libérés par la lipase pancréatique à 37°C et pH=6.25, en fonction du temps (n=6 à 8 assays pour chaque concentration de traceurs testée). Analyse statistique : ANOVA mixte à deux facteurs à effets fixes (temps et concentration de traceurs) et un facteur à effets aléatoires (assay, imbriqué dans le facteur concentration). Sommes des carrés de type I (séquentielles) avec l'ordre suivant : temps, concentration, interaction temps:concentration. L'effet du temps est significatif ($F_{60,\ 1080} = 2128.3$, $p < .0001$). Celui de la concentration ne l'est pas, mais l'interaction temps:concentration l'est ($F_{120,\ 1080} = 2.5$, $p < .0001$). Après correction de Bonferroni, les coefficients des interactions temps-concentration sont significativement différents de 0 pour 5.7% de traceurs entre t=5 et t=12 minutes. **(B)** "Lag time", défini comme l'intersection de la tangente de plus forte pente avec l'axe du temps, en fonction de la concentration de traceurs. Analyse statistique : ANOVA à un facteur suivi d'un test post-hoc de Tukey. D'après (Danthine et al. 2019).

Les analyses physico-chimiques réalisées par Sabine Danthine suggèrent que ce délai est dû au fait qu'à 37°C, la matière grasse avec 5.7% de traceurs reste partiellement solide, alors qu'elle est complètement fondue en l'absence de traceurs ou lorsque ceux-ci sont incorporés à plus faible concentration (1.5%) (Danthine et al. 2019). Ce travail suggère que les traceurs doivent être utilisés à faible concentration dans les études cliniques qui cherchent à détecter des différences fines de cinétique postprandiale.

Il reste cependant difficile de prédire précisément l'impact des traceurs dans une étude clinique, où les mesures sont en général effectuées dans le plasma, les selles et l'air expiré[14], et à une granularité temporelle plus grossière que celle utilisée dans l'étude *in vitro*. Les étapes intermédiaires entre la lipolyse luminale et l'arrivée dans le plasma ou l'air expiré pourraient soit tamponner soit amplifier l'impact identifié ici sur la lipolyse luminale.

[14] Le CO_2 est en effet le devenir ultime des atomes de carbone des lipides qui ont été β-oxydés, c'est-à-dire brûlés pour produire de l'énergie.

II.3 Influence de la pasteurisation

BQR INSA "SiMoLip : Impact de l'organisation spatiale des matières grasses sur leur digestion et sur leur absorption : couplage entre approches expérimentales et simulations" soutenu par l'INSA de Lyon à hauteur de 24000 euros sur deux ans. Trois laboratoires impliqués, en biologie de la nutrition, en informatique et en matériaux. CK porteuse du projet.

Stage de M2 : Julie Etienne (2019). Stage co-encadré avec Hugues Berry (LIRIS et Inria Beagle).

Au-delà de l'incorporation de traceurs — un procédé relativement exceptionnel puisqu'utilisé seulement dans les études cliniques et animales —, d'autres procédés plus courants de transformation des aliments peuvent impacter leurs cinétiques de digestion et d'absorption intestinale. La pasteurisation est l'un d'entre eux. Elle est couramment utilisée dans l'industrie agroalimentaire, mais aussi dans les lactariums des services de néonatalogie, pour pasteuriser le lait maternel de mères donneuses, destiné à l'alimentation des nourrissons prématurés lorsque le lait de leur mère n'est pas disponible.

Dans le cadre d'un partenariat avec le service de néonatalogie de l'hôpital de la Croix-Rousse et l'équipe STLO de l'INRAE Rennes, et des stages de M2 de Marine Vincent et Julie Etienne, nous avons caractérisé l'impact de la pasteurisation du lait maternel de don sur la cinétique de lipolyse et d'absorption intestinale, en utilisant des modèles *in vitro* des deux étapes. Le modèle *in vitro* pour la lipolyse luminale, développé à Rennes (Ménard et al. 2018), prenait en compte les spécificités des nourrissons (pH gastrointestinal plus élevé, concentrations plus faibles des différentes enzymes et des sels biliaires) et était calibré pour un nourrisson prématuré de 2kg. Le modèle *in vitro* pour l'absorption intestinale était des cultures de cellules Caco-2, réalisées au laboratoire CarMeN par Marine Vincent et Armelle Penhoat. Cette lignée cellulaire est au départ dérivée d'un cancer humain du côlon, mais ces cellules ont la capacité de se différencier spontanément en une monocouche de cellules partageant de nombreuses propriétés des entérocytes matures (Sambuy et al. 2005).

Comme le montre la Figure 31A, la pasteurisation réduit le degré de pré-lipolyse, c'est-à-dire de lipolyse qui se produit avant même la mise en contact avec les fluides simulant l'environnement gastrique et intestinal. En effet, le lait maternel humain contient une enzyme capable d'hydrolyser les triglycérides, la BSSL (bile-salt stimulated lipase). Une expérience complémentaire a confirmé que la pasteurisation telle que pratiquée dans les lactariums réduit de 86 % l'activité de la BSSL mesurée en conditions standard, c'est-à-dire à pH 8 et avec la tributyrine comme substrat.

Après les phases de digestion gastrique et intestinale, le lait maternel pasteurisé contenait 13% de moins d'acides gras libres que le lait maternel cru. Cependant, lorsqu'on normalise le degré de lipolyse après les phases gastro-intestinales par le degré de pré-lipolyse (Figure 31C), on ne détecte plus de différence significative, ce qui suggère que les phases gastro-intestinales de lipolyse ne sont pas très sensibles en elles-mêmes à la pasteurisation. Elles ne réduisent pas la différence initiale liée à la pré-lipolyse, et ne l'amplifient pas non plus.

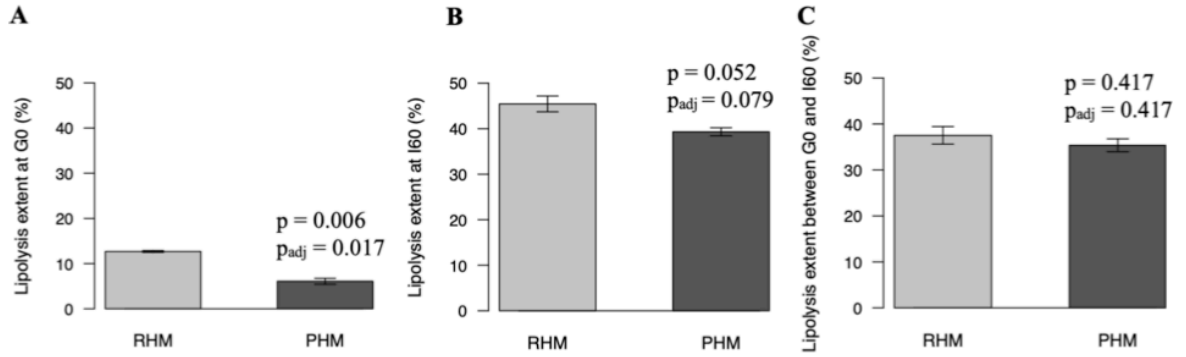


Figure 31 : Impact de la pasteurisation sur la pré-lipolyse (A) et sur la lipolyse gastrointestinale (B et C). RHM : lait maternel non pasteurisé ; PHM : lait maternel pasteurisé. Plus de lipides sont hydrolysés après les phases de lipolyse gastrique et intestinale (B), mais cette différence est principalement due à une plus grande pré-lipolyse avant le début de la phase gastrique (A), car elle disparaît si on normalise par le niveau de pré-lipolyse (C). Les valeurs de p sont les valeurs brutes des tests de Welch (Student avec variances inégales), tandis que les valeurs p_{adj} sont celles ajustées par la correction FDR (False Discovery Rate) pour les tests multiples. D'après (Vincent et al. 2020).

L'effet de la pasteurisation sur la pré-lipolyse continue de se voir sur l'étape d'absorption intestinale, simulée en incubant des cellules Caco-2 avec le même volume de lait maternel digéré, pasteurisé ou non. Après 16h, les cellules sont colorées à l'Oil-Red-O, un colorant liposoluble, ce qui permet de voir les gouttelettes lipidiques intracellulaires en microscopie. Durant son stage de M2, que j'ai co-encadré avec Hugues Berry (LIRIS et Inria Beagle), Julie Etienne a adapté le protocole d'analyse d'image de (Deutsch et al. 2014) pour identifier les gouttelettes lipidiques sur les images. Comme le montre la Figure 32, les cellules incubées avec le lait maternel pasteurisé digéré contenaient environ 13% de lipides de moins que celles incubées avec le lait cru digéré, les gouttelettes lipidiques étant aussi nombreuses dans les deux cas mais plus petites en moyenne dans le cas du lait pasteurisé. Ainsi, la différence initiale liée à l'inactivation de la BSSL n'est pas non plus tamponnée (ni amplifiée) par les processus cellulaires d'absorption des acides gras.

Au final, cette étude *in vitro* suggère que (i) l'impact principal de la pasteurisation du lait maternel se situe au niveau de l'inactivation de la BSSL, ce qui réduit le degré de pré-lipolyse, et que (ii) cette différence initiale persiste lors des étapes subséquentes de lipolyse gastro-intestinale et d'absorption intestinale, conduisant à environ 13% de lipides absorbés en moins.

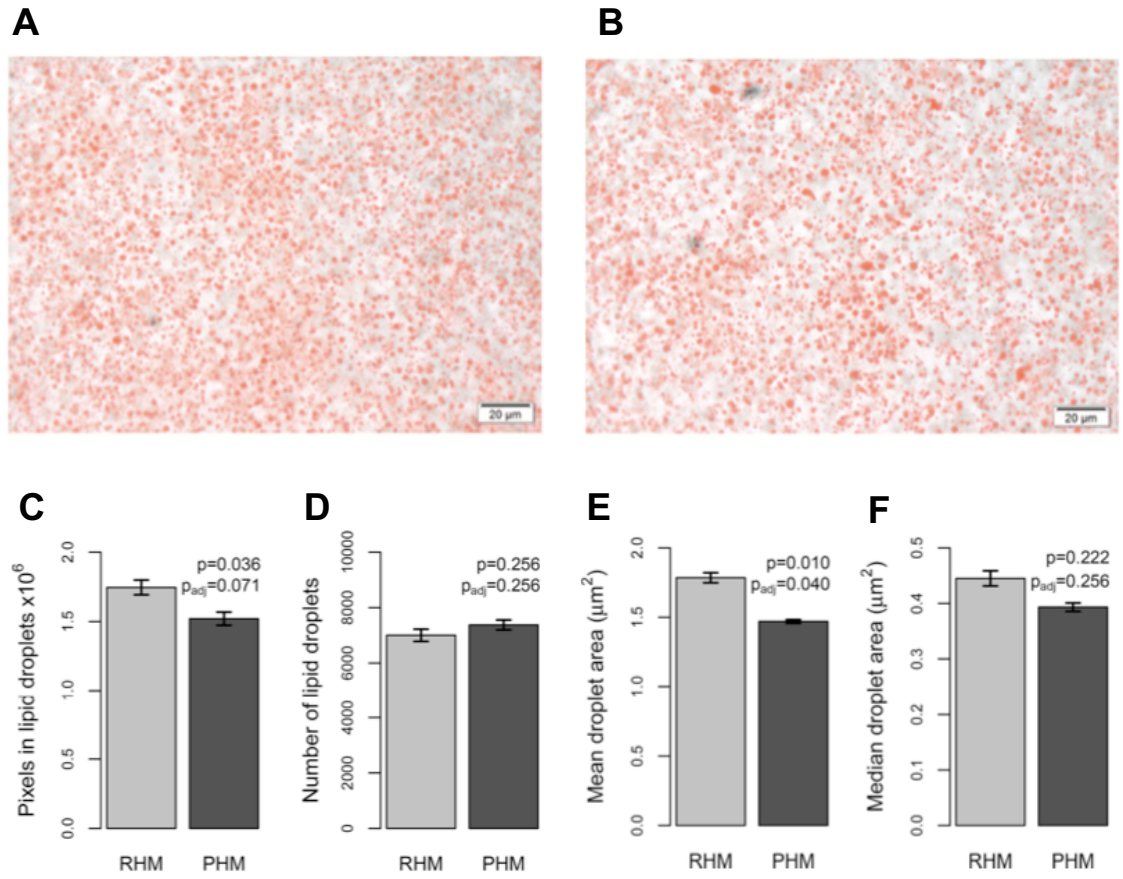


Figure 32 : Quantification par analyse d'images de l'impact de la pasteurisation du lait maternel sur l'absorption des lipides par les cellules Caco-2. (A) Cellules Caco-2 colorées à l'Oil-Red-O, un colorant liposoluble, après 16h d'incubation avec du lait maternel cru, digéré en conditions infantiles. **(B)** Cellules Caco-2 colorées à l'Oil-Red-O après 16h d'incubation avec du lait maternel pasteurisé, digéré en conditions infantiles. **(C)** Nombre de pixels identifiés comme appartenant à des gouttelettes lipidiques intracellulaires après analyse des images de microscopie. RHM : lait maternel cru ; PHM : lait maternel pasteurisé. **(D)** Nombre de gouttelettes lipidiques intracellulaires par image. **(E)** Aire moyenne des gouttelettes lipidiques intracellulaires. **(F)** Aire médiane des gouttelettes lipidiques intracellulaires. D'après (Vincent et al. 2020).

II.4 Influence de la vectorisation des acides gras

Comme nous l'avons vu, les acides gras peuvent être apportés via les triglycérides, mais ils peuvent aussi être apportés (dans une moindre quantité) par les phospholipides. A quantité totale d'acides gras égale, la question se pose de savoir quelle est l'influence du vecteur (triglycérides ou phospholipides) sur leurs cinétiques postprandiales.

Pour répondre à cette question, Chloé Robert, doctorante à CarMeN sous la direction de Marie-Caroline Michalski et de Carole Vaysse (ITERG), a mesuré pendant 5h les concentrations d'acides gras dans la lymphe de rats gavés avec des formulations lipidiques présentant les mêmes quantités totales d'acides gras, mais différant par la proportion d'acides gras apportés via des phospholipides. Pour cela, les formulations lipidiques étaient réalisées

en mélangeant, dans des proportions contrôlées, des huiles végétales et de la lécithine de colza, riche en phospholipides.

Comme le montre la Figure 33A-B, plus la proportion de phospholipides dans la formulation augmente, plus on retrouve d'acides gras dans la lymphe, ce qui suggère que les acides gras sont mieux assimilés lorsqu'ils sont apportés sous forme de phospholipides.

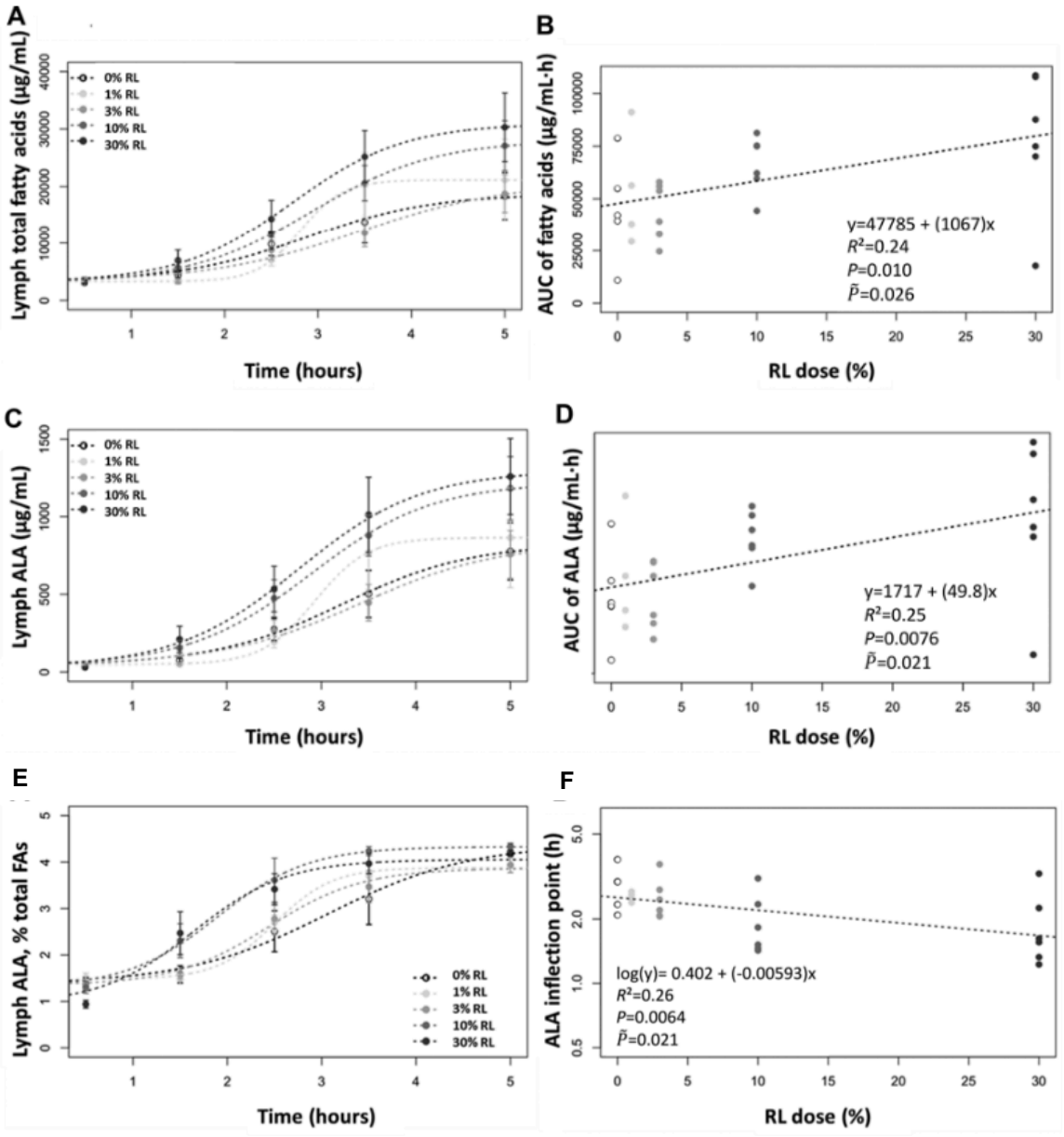


Figure 33 : Influence de la proportion de lécithine de colza dans la formulation huileuse sur les cinétiques postprandiales d'acides gras dans la lymphe de rats. Symboles évidés : cinétique de contrôle avec la formulation huileuse sans lécithine de colza. Symboles pleins : cinétiques obtenues pour les formulations huileuses avec 1%, 3%, 10% ou 30% de lécithine de colza, selon la nuance de gris. Analyses statistiques : régressions non linéaires de la forme $\boldsymbol{y(t) = y_{\min} + \frac{\delta}{1+e^{\frac{t_{infl}-t}{s}}}}$, puis régression linéaire de δ, $\log(t_{infl})$, s et de l'aire sous la courbe. P : p-valeurs brutes ; $\widetilde{\boldsymbol{P}}$: p-valeurs corrigées selon la méthode FDR (False Discovery Rate). D'après (Robert et al. 2020).

Ceci est vrai en particulier pour l'acide alpha-linolénique (ALA, Figure 33C-D), un oméga-3 essentiel que l'organisme humain ne sait pas synthétiser et qui doit donc être apporté dans l'alimentation. Précurseur de la synthèse des omégas-3 EPA et DHA, normalement apportés par le poisson, il est indispensable pour les personnes végétariennes ou végan. La proportion d'ALA apportée sous forme de phospholipides plutôt que sous forme de triglycérides influence non seulement la quantité assimilée par l'organisme (ALA, Figure 33C-D), mais aussi le timing d'apparition de l'ALA dans la lymphe (Figure 33E-F) (Robert et al. 2020). En effet, avec 30% de lécithine de colza dans le mélange huileux, le point d'inflexion de la cinétique postprandiale du pourcentage d'ALA dans les acides gras lymphatiques est plus précoce d'environ 1 heure.

Il faut toutefois souligner qu'une proportion de 30% de lécithine ne peut pas être atteinte dans une alimentation normale, seule une supplémentation permettrait de le faire. En effet, de par leurs propriétés émulsifiantes et épaississantes, les lécithines ne peuvent pas être ajoutées à haute dose dans les aliments : la plus haute proportion rapportée est de 10% (EFSA Panel on Food Additives and Nutrient Sources added to Food (ANS) et al. 2017).

II.5 Conclusion du chapitre

Dans l'ensemble, ces résultats confirment que la digestion et l'absorption des lipides sont sensibles à d'autres facteurs que la composition des aliments en macronutriments. Or le principal outil actuellement mis à disposition du grand public dans l'objectif de prévention des maladies liées à une nutrition inadéquate (diabète de type 2, obésité, maladies cardiovasculaires) est le Nutri-Score, conçu dans le cadre du Programme National Nutrition Santé. Celui-ci est calculé uniquement d'après la composition de l'aliment, et indépendamment de qui le consomme. Or, pour raffiner la relation entre un aliment et son impact sur la santé, il faut des outils permettant de quantifier et d'expliquer l'impact d'autres facteurs que la composition (organisation spatiale des nutriments, "effet matrice", voir (Thorning et al. 2017)), et d'identifier les sous-populations les plus sensibles à ces facteurs et les plus à même de bénéficier de certains types d'interventions nutritionnelles. Le projet de recherche décrit dans le chapitre suivant vise notamment à concevoir des modèles mécanistiques et quantitatifs du devenir des lipides dans l'organisme pour contribuer à cet objectif.

Chapitre III : Projet

> *Nutritional science is presently undergoing a data explosion as an increasing number of studies are incorporating methods from genomics, transcriptomics, proteomics, and metabolomics. However, it is presently unclear how these high-dimensional datasets can be related to the physiological characterization of phenotype using traditional nutritional research methods such as indirect calorimetry, nutrient balance, body composition assessment, and isotopic tracer methods. Thus, a fundamental challenge for nutrition research is to connect these data that are collected at vastly different spatial, temporal, and dimensionality scales. Although statistical analysis is still the method of choice to deal with the high dimensionality of "-omics" datasets, systems biology and computational modeling approaches begin to reveal quantitative mechanistic relationships between these various measurements.*
>
> (de Graaf et al. 2009)

Mon projet est d'animer, au sein du laboratoire CarMeN et de l'équipe Inria Beagle, un axe sur la modélisation mécanistique et quantitative du devenir des lipides dans l'organisme, à court terme (durant les heures qui suivent un repas) et à long terme (à l'échelle de plusieurs années). Le présent chapitre présente ce projet de recherche tel que je l'envisage pour les dix prochaines années. Il se décline en trois volets :

1. modélisation macroscopique, compartimentale, à l'échelle de l'organisme, du devenir des triglycérides alimentaires durant les heures qui suivent un repas ;
2. modélisation microscopique du flux de triglycérides à travers les entérocytes durant les heures qui suivent un repas ;
3. modélisation de la physiologie des lipides adipeux (stockage, mobilisation) en lien avec déséquilibre énergétique entre l'apport alimentaire et la dépense énergétique sur plusieurs mois ou années (collaboration avec Peter Arner au Karolinska University Hospital et Kirsty Spalding au Karolinska Institute de Stockhlom, et avec Samuel Bernard, Institut Camille Jordan et équipe Inria Dracula).

Avant de présenter ces volets en détail, je voudrais faire deux remarques méthodologiques générales, qui expliquent la nécessité d'ancrer ce projet à l'Inria, en plus du laboratoire CarMeN.

D'une part, bien que la plupart des modèles soient formulés mathématiquement au départ, il est en général impossible de les résoudre analytiquement. En pratique, il faudra recourir à la simulation numérique, en choisissant soigneusement les solveurs et leur paramétrage. De plus, certains paramètres des modèles ne seront pas disponibles dans la littérature, et il faudra recourir à des techniques d'estimation. Comme nous le verrons un peu plus loin, une simple approche de moindres carrés non linéaires n'est pas toujours satisfaisante, et il est parfois préférable de recourir à des méthodes bayésiennes, plus informatives au final mais au

prix d'un nombre beaucoup plus grand de jeux de paramètres testés. Cela pose des problématiques de temps de calcul, en particulier dans les régions de l'espace de paramètres où le système est raide.

D'autre part, pour caractériser l'impact d'autres facteurs que la composition des aliments, comme leur organisation spatiale (effet matrice, finesse de l'émulsion, ...), les modèles classiquement utilisés en biochimie ne sont pas toujours adaptés, car ils font typiquement une hypothèse d'homogénéité spatiale en trois dimensions. Par exemple, dans le métabolisme des lipides, qui sont hydrophobes, de nombreux processus enzymatiques se passent à l'interface huile-eau, c'est-à-dire à la surface des gouttelettes, donc en deux dimensions. Le ratio surface/volume des gouttelettes est donc important pour la cinétique de ces processus. Les formalismes et outils développés dans l'équipe Inria Beagle visent précisément à intégrer les phénomènes d'hétérogénéité spatiale dans la modélisation des processus moléculaires et cellulaires. Ils s'avèreront donc précieux pour les deux premiers volets de ce projet.

Ma double affiliation, CarMeN-Inria, constitue donc une configuration idéale pour le projet de recherche que je propose.

III.1 Modélisation du devenir des lipides à court terme

III.1.1 Pourquoi construire des modèles mécanistiques en nutrition ?

Le concept de "phénotype nutritionnel" (Zeisel et al. 2005) a été proposé comme un ensemble de mesures génétiques, métabolomiques, fonctionnelles et comportementales permettant de quantifier de façon multivariée le statut nutritionnel humain. Ce phénotype dépend à la fois de facteurs génétiques et environnementaux, dont bien sûr l'alimentation. Les cinétiques postprandiales, sensibles à l'état physiologique du sujet mais aussi aux caractéristiques des aliments et donc aux interventions nutritionnelles, font partie intégrante de ce phénotype nutritionnel. Bien qu'étant potentiellement très riches d'information, elles sont généralement résumées par un ou plusieurs indicateurs phénoménologiques, ce qui facilite ensuite l'analyse multivariée : timing du ou des pics, amplitude maximale des pics, aire sous la courbe, délai de retour au basal, etc.

Par exemple, (Bouwman et al. 2012) puis (van den Broek et al. 2017) ont proposé une méthode de la visualisation des sujets et de leur réponse à une intervention nutritionnelle. Il s'agit de projeter les phénotypes nutritionnels pré- et post-intervention dans un espace 3D appelé "health space", dont chacun des 3 axes est une combinaison prédéfinie de certaines mesures. Un algorithme de clustering hiérarchique est ensuite utilisé dans ce "health space" pour découvrir des clusters de sujets qui présentent des phénotypes nutritionnels similaires et/ou des réponses similaires à l'intervention. Cette approche permet, le cas échéant, de distinguer des clusters de sujets "répondeurs" ou "non répondeurs" à une intervention, au sein de sujets qui étaient a priori tous "sains". Cependant, elle souffre de plusieurs écueils. Un premier problème est que même si certains biomarqueurs ont été mesurés de façon cinétique, ces cinétiques sont préalablement résumées par un ou deux indicateurs phénoménologiques (typiquement la valeur initiale et l'aire sous la courbe), ce qui correspond à une perte importante d'information. Un second problème, plus gênant, est souligné par (van Ommen, Cavallieri, et al. 2008) : "This type of inventory studies does not necessarily lead to a deeper understanding of the biological processes. For this it is inevitable to combine “omics” with firm functional assays and mechanistic studies as part of a comprehensive phenotyping."

De fait, l'analyse statistique des indicateurs phénoménologiques ne permet pas d'identifier des processus physiologiques candidats ; elle ne permet pas de comprendre pourquoi telle catégorie de sujets répond mieux à telle intervention nutritionnelle. Disposer d'un modèle mathématique mécanistique, prédictif des cinétiques postprandiales, permettrait d'aller plus loin dans l'interprétation biologique des données, et de suggérer quel(s) processus physiologique(s) pourraient différer dans leurs paramètres pour expliquer des différences entre deux cinétiques.

Dans une stratégie mécanistique, le phénotype nutritionnel pourrait être redéfini comme l'ensemble des valeurs des paramètres du modèle mécanistique qui permettent de reproduire les cinétiques postprandiales dans leur intégralité. L'analyse multivariée et le clustering resteraient possibles, mais on éviterait ainsi de perdre trop d'information, et, surtout, les clusters de patients pourraient être mis en relation avec les mécanismes physiologiques pris en compte dans le modèle.

III.1.2 Modèles existants pour les cinétiques postprandiales

L'abondance de données de glycémie et d'insulinémie postprandiales (données de type OGTT) a permis l'élaboration et le paramétrage de modèles mathématiques quantitatifs et mécanistiques, prédictifs de ces réponses physiologiques pour des sujets sains et pour des patients atteints de diabète (Ajmera et al. 2013). L'un de ces modèles (Dalla Man et al. 2007) a été agréé en 2008 par la Food and Drug Administration comme substitut à l'expérimentation animale pour les tests précliniques de certains traitements du diabète de type I (Kovatchev et al. 2009), ouvrant ainsi la voie aux essais précliniques *in silico*.

Cet avancement contraste avec la situation des cinétiques de lipémie postprandiale (données de type OFTT), pour lesquelles il n'existe que très peu de modèles. L'un des modèles les plus avancés est celui de (Sips et al. 2015). Il s'agit d'une extension du modèle glucose-insuline de (Dalla Man et al. 2007), incorporant des équations supplémentaires pour prédire la concentration des acides gras non estérifiés dans le plasma. Toutefois, en phase postprandiale, la majeure partie des acides gras circulants se trouve sous forme estérifiée dans des triglycérides[15] et des phospholipides, eux-mêmes empaquetés avec du cholestérol et des protéines dans des assemblages appelés chylomicrons et lipoprotéines. Il manque donc actuellement dans la littérature un modèle mathématique prédisant la réponse physiologique de l'organisme à l'absorption de lipides alimentaires, non seulement en termes de cinétique postprandiale des acides gras non estérifiés, mais aussi en termes de cinétique postprandiale des acides gras estérifiés sous forme de triglycérides.

Au niveau du plasma, la cinétique postprandiale des triglycérides résulte (i) de l'afflux de chylomicrons provenant de l'intestin via la lymphe, (ii) de l'afflux de lipoprotéines provenant du foie, et (iii) de leurs clairances au niveau des capillaires du tissu adipeux, du cœur, des muscles squelettiques et des glandes mammaires, découlant de l'action de la lipoprotéine lipase.

Mon projet de modélisation du devenir à court terme des triglycérides alimentaires comporte deux volets. Le premier volet consiste à construire un modèle de type "modèle à compartiments", à gros grain mais pour l'ensemble de l'organisme, prenant en compte les

[15] Dans le plasma de sujets sains de poids normal en phase postprandiale, il y a environ 7 fois plus d'acides gras estérifiés sous forme de triglycérides que d'acides gras non estérifiés (d'après les données de l'étude LIPINFLOX menée par le laboratoire CarMeN et le CRNH, Vors et al. 2013).

trois phénomènes ci-dessus et prédisant la cinétique des triglycérides dans le plasma, dans les heures qui suivent un repas (données de type OFTT). Le second volet consiste à modéliser de façon détaillée l'afflux de triglycérides provenant de l'intestin, c'est-à-dire le flux (i) du modèle à gros grain, afin d'identifier les mécanismes moléculaires qui modulent, à l'intérieur de l'entérocyte, la phase montante des cinétiques OFTT. Ces deux volets sont décrits dans les deux sections suivantes.

III.1.3 Volet 1 : Modélisation macroscopique, à l'échelle de l'organisme

Stage de M2 : Ella Beaumann (2019). Stage co-encadré avec Marie-Caroline Michalski (CarMeN) et Samuel Bernard (Institut Camille Jordan et Inria Dracula).

L'objectif final de ce volet est de caractériser physiologiquement les sujets répondant à différents types de repas grâce à une modélisation quantitative et mécanistique des cinétiques postprandiales de lipémie et de glycémie, pour contribuer à définir des profils phénotypiques de flexibilité nutritionnelle. Même si nous nous focalisons sur la lipémie, la prise en compte de la glycémie est nécessaire, car celle-ci est directement liée à la production d'insuline, et l'insuline stimule la clairance de la lipémie. En effet, elle active la captation des lipides par les tissus périphériques via la stimulation de la production de la LPL et l'inhibition de la lipolyse adipocytaire.

Ce volet se décline selon les 4 sous-objectifs suivants (qui, ensemble, dépassent le cadre d'une seule thèse) :

O1. Concevoir un modèle mathématique des cinétiques postprandiales des triglycérides, des acides gras non estérifiés, du glucose et de l'insuline, en étendant et adaptant les modèles de (Dalla Man et al. 2007) et de (Sips et al. 2015);

O2. Estimer les paramètres de ce modèle pour reproduire, pour un sujet donné et un repas donné, les cinétiques postprandiales du glucose, de l'insuline, des triglycérides et des acides gras non estérifiés ;

O3. Quantifier l'effet d'une intervention nutritionnelle (portant idéalement sur l'organisation spatiale des nutriments dans l'aliment, e.g. finesse de l'émulsion des matières grasses, interaction avec les fibres alimentaires) et identifier les processus physiologiques potentiellement impactés en comparant les valeurs des paramètres avant et après l'intervention ;

O4. Regrouper les sujets (clustering) en fonction de leurs paramètres pré-intervention et de leur réponse à l'intervention, et étudier statistiquement la corrélation entre le cluster de réponse physiologique et des indicateurs plus facilement mesurables à grande échelle comme le poids, l'indice de masse corporelle, le tour de taille, la masse grasse, etc.

Les données cliniques utilisées seront celles déjà obtenues par le laboratoire CarMeN et le Centre de Recherche en Nutrition Humaine (CRNH) Rhône-Alpes. Le CRNH dispose non seulement de données de glycémie et de lipémie postprandiale, mais aussi de données de flux basées sur l'utilisation de traceurs isotopiques (glucides ou lipides), pour des différents types de sujets (hommes et femmes ; normopondérés, en surpoids ou obèses ; sains ou diabétiques de type 2, etc).

Nous avons pu utiliser une partie de ces données pour l'étude préliminaire menée pendant le stage de M2 d'Ella Beaumann, co-encadré avec Samuel Bernard (Institut Camille Jordan et équipe Inria Dracula) et Marie-Caroline Michalski (laboratoire CarMeN). Le travail d'Ella a révélé que le module "absorption intestinale" dans le modèle de (Dalla Man et al. 2007) est

très difficile à paramétrer sur d'autres données que celles de l'article original et souffre peut-être d'overfitting. Dans le cadre du sous-objectif O1, nous utiliserons donc probablement une version simplifiée de ce module.

Le sous-objectif O2 est particulièrement délicat, car le travail de M2 d'Ella nous a également montré que le système glucose-insuline pose déjà des problèmes d'identifiabilité pratique. Une simple approche de type moindres carrés non linéaires avec minimisation par l'algorithme de Levenberg-Marquardt donne des valeurs de paramètres très bruitées, très sensibles à de petites variations dans les données. Une approche bayésienne avec des algorithmes de type Metropolis-Hastings, bien que beaucoup plus coûteuse en temps de calcul, permet de reconstruire les densités postérieures des paramètres et ainsi de diagnostiquer les sources de non-identifiabilité (par exemple, la compensation entre deux paramètres). Cela permet de détecter les combinaisons de paramètres réellement identifiables. La Figure 34, issue du rapport de master d'Ella, montre par exemple que les paramètres β et m_5 du module Insuline sont difficilement identifiables séparément, alors que la distribution conjointe des paramètres k_i et k_{p1} est plus favorable : les valeurs échantillonnées sont organisées en cercle, avec au centre les valeurs qui minimisent la somme des moindres carrés.

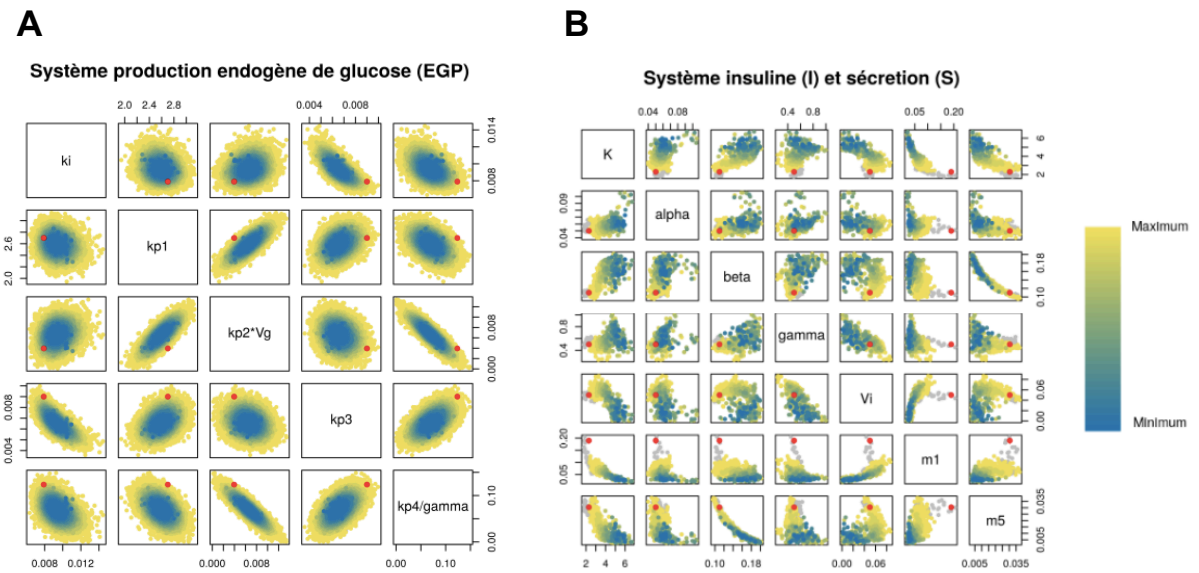


Figure 34 : Illustration des problèmes d'identifiabilité pratique des paramètres du modèle de (Dalla Man et al. 2007), notamment dans le module Insuline. Dans l'article original, les valeurs des paramètres avaient été estimées par une méthode classique de moindres carrés non linéaires, probablement de type Levenberg-Marquardt. Nous les avons ré-estimées sur les données originales avec une approche bayésienne, utilisant une méthode de Monte Carlo par chaînes de Markov (MCMC) pour échantillonner la distribution *a posteriori* des paramètres, étant données les cinétiques observées et une distribution *a priori* uniforme entre deux bornes. Spécifiquement, nous avons utilisé la méthode modMCMC du package FME de R, qui repose sur l'algorithme de Metropolis-Hastings. L'échelle de couleur correspond à la valeur des moindres carrés. Les points rouges correspondent aux valeurs de paramètres dans (Dalla Man et al. 2007). (A) Paramètres du module de la production endogène de glucose. (B) Paramètres du module de production et de sécrétion d'insuline. Pour la plupart des paramètres, la valeur publiée n'est pas optimale, au sens où elle ne minimise pas la somme des moindres carrés. D'après le rapport de M2 d'Ella Beaumann.

Deux pistes sont possibles pour obtenir le jeu réduit de paramètres réellement identifiables : (i) simplifier le modèle mathématique en imposant a priori des relations entre paramètres, ou (ii) utiliser des méthodes plus automatiques de réduction de dimension linéaires (e.g. ACP) ou non linéaires (e.g. diffusion maps, t-distributed stochastic neighbor embedding) sur le nuage des valeurs de paramètres acceptées par l'algorithme de Metropolis-Hastings. La première solution demande plus de travail mais a l'avantage de préserver la signification physiologique des paramètres, ce qui est important car c'est là que se trouve la plus-value d'un modèle mécanistique. Le clustering des sujets serait ensuite réalisé sur le jeu réduit de paramètres.

III.1.4 Volet 2 : Modélisation microscopique dans les entérocytes

Travail amorcé grâce :

- *au projet "Lipuscale: Simulation hybride de la digestion et de l'absorption des triglycérides par l'intestin" soutenu à hauteur de 5000 euros sur deux ans par l'Institut Rhône-Alpin des Systèmes Complexes (IXXI). Deux laboratoires impliqués, en biologie de la nutrition et en mathématiques. CK porteuse du projet.*
- *au BQR INSA "SiMoLip : Impact de l'organisation spatiale des matières grasses sur leur digestion et sur leur absorption : couplage entre approches expérimentales et simulations" soutenu par l'INSA de Lyon à hauteur de 24000 euros sur deux ans. Trois laboratoires impliqués, en biologie de la nutrition, en informatique et en matériaux. CK porteuse du projet.*
- *au financement de thèse Inria-Inserm obtenu pour Julie Etienne.*

Stage de M2 : Julie Etienne (2019). Stage co-encadré avec Hugues Berry (LIRIS et Inria Beagle).

Thèse : Julie Etienne (2019-2022). Co-dirigée à 70% avec Marie-Caroline Michalski (CarMeN).

Des modèles existent dans la littérature pour l'étape de lipolyse luminale (Li & McClements 2010; Marze & Choimet 2012; Giang et al. 2015; Giang et al. 2016). Ils permettent, à partir de la composition en triglycérides et de leur organisation supramoléculaire et physicochimique (e.g. finesse de l'émulsion), de prédire le flux d'acides gras et de monoglycérides arrivant dans l'intestin. Par contre, il n'existe pas à ce jour de modèle pour l'étape suivante, celle du franchissement des entérocytes. Or les entérocytes ne se comportent pas comme de simples "baignoires" qui se rempliraient et se videraient selon la différence entre flux d'entrée et de sortie, mais sont au contraire susceptibles de moduler ces flux, par exemple du fait d'une saturation des mécanismes de transport ou d'un stockage temporaire des triglycérides dans l'entérocyte.

L'objectif de ce volet est donc de développer un modèle mécanistique du franchissement des entérocytes, basé sur les réseaux moléculaires (prioritairement transport et voies métaboliques) impliqués dans le trafic des triglycérides à l'intérieur des entérocytes. Ce modèle computationnel, formé par des équations différentielles non-linéaires couplées, permettra de prédire le flux de triglycérides en sortie des entérocytes à partir d'un flux d'entrée en acides gras et monoglycérides. Il pourra s'inspirer de modèles développés pour les cellules du foie (Shorten & Upreti 2005; Wallstab et al. 2017). Dans un second temps, ce modèle pourrait être combiné avec un des modèles existants de lipolyse, formant ainsi un pipeline (voir **Figure 35**) qui permettrait de prédire la cinétique de l'afflux de triglycérides dans la circulation à partir de la composition des triglycérides ingérés et de leur organisation

supramoléculaire (e.g. finesse de l'émulsion dans le cas d'un aliment de type émulsion huile-dans-eau). Ultimement, il s'agit de contribuer à élucider les bases moléculaires et cellulaires de l' "index lipidémique" des aliments, par analogie avec l'index glycémique qui permet aux diabétiques de mieux gérer leur maladie.

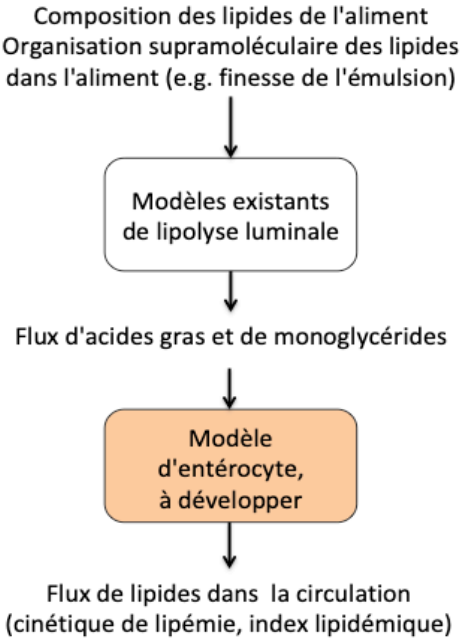


Figure 35 : Projet de modélisation mécanistique du flux de triglycérides à travers les entérocytes. Pour être mis en relation avec les caractéristiques des aliments, le modèle pourra être couplé avec les modèles existants de lipolyse luminale.

Ce travail a commencé, grâce à un BQR INSA, par le stage de M2 de Julie Etienne, co-encadré avec Hugues Berry (LIRIS et Inria Beagle). Grâce à un financement de thèse Inserm-Inria, Julie continue actuellement ce travail en thèse, sous ma direction et celle de Marie-Caroline Michalski (CarMeN). Dans un premier temps, nous nous focalisons sur les réactions de transport et de métabolisme, en considérant comme constantes les concentrations des transporteurs et des enzymes. Dans un second temps, nous permettrons à ces concentrations de varier sous l'effet de la régulation de l'expression des gènes.

Les deux principaux défis de ce travail ambitieux (qui dépasse le cadre d'une seule thèse) sont d'une part le choix des hypothèses simplificatrices, c'est-à-dire le choix d'un compromis adéquat entre réalisme et nombre de paramètres, et d'autre part la calibration des paramètres. La Figure 36, issue du rapport de master de Julie, illustre l'hétérogénéité des différentes étapes moléculaires en termes de volume de données quantitatives disponibles dans la littérature. Il se peut que pour certains paramètres, nous manquions de valeurs dans la littérature et que nous devions les estimer, avec les mêmes difficultés que pour le premier volet du projet.

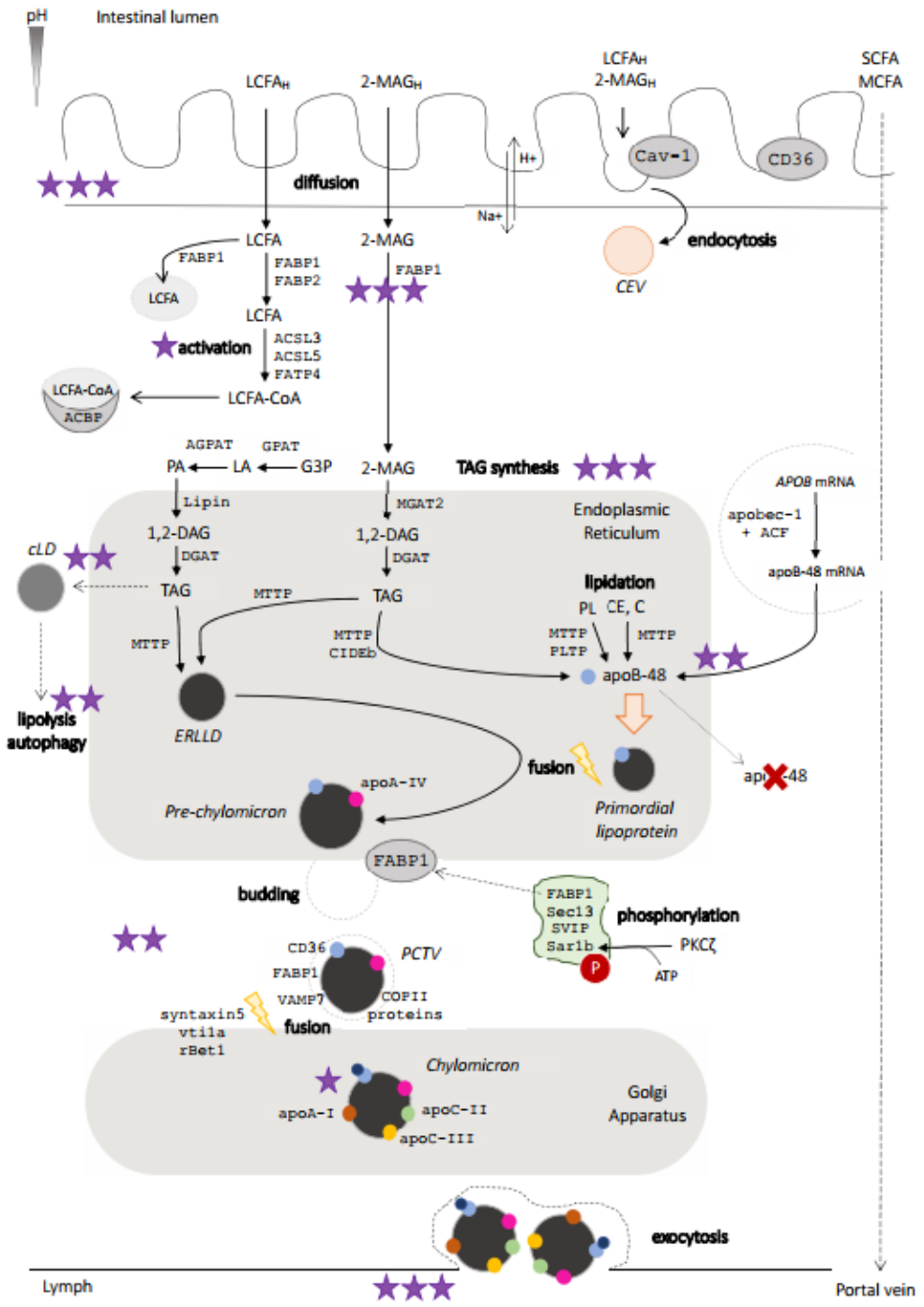


Figure 36 : Volume de données quantitatives disponibles dans la littérature pour les différentes étapes moléculaires du franchissement des entérocytes par les triglycérides. Si x est le nombre d'articles trouvés avec des données quantitatives, alors : absence d'étoile: x = 0; ⋆ : x ≤ 5; ⋆⋆ :5 < x ≤ 10; ⋆ ⋆ ⋆ : x > 10. LCFA: long-chain fatty acid; SCFA: short chain fatty acid ; MCFA: medium-chain fatty acid ; MAG: monoacylglycerol (monoglyceride) ; DAG: diacylglycerol (diglyceride) ; TAG: triacylglycerol (triglyceride) ; PL: phospholipid ; C: cholesterol ; CE: cholesterol esters ; CEV: caveolae endocytic vesicles ; cLD: cytosolic lipid droplet ; ERLLD: endoplasmic reticulum luminal lipid droplet ; PCTV: pre-chylomicron transport vesicle. D'après le rapport de M2 de Julie Etienne.

III.2 Modélisation du devenir des lipides à long terme

Travail amorcé grâce au financement du projet "Modelling lipid physiology and energy imbalance during weight loss" à hauteur de 4000 euros par Biosyl, la fédération de recherche lyonnaise en biologie des systèmes. CK co-porteuse du projet, avec Samuel Bernard (Institut Camille Jordan et Inria Dracula). Collaboration avec Peter Arner et Kirsty Spalding (Karolinska University Hospital et Karolinska Institute, Stockhlom, Suède).

Stage de M1 : Justine Antoine (2019). Stage co-encadré avec Samuel Bernard.

Le devenir des lipides à long terme, c'est-à-dire à l'échelle de plusieurs années, dépend à la fois de la dynamique éventuelle de prise ou de perte de poids, mais aussi, même en période de poids stable, du "turnover" continuel qui s'opère dans le tissu adipeux, à travers le déstockage de lipides anciens et le stockage de nouveaux lipides provenant du dernier repas. En effet, deux éléments montrent que les lipides du tissu adipeux contribuent à la dépense énergétique quotidienne et sont donc sujets à un renouvellement continu. Premièrement, le traçage au ^{13}C des lipides alimentaires montre qu'entre 40 et 57% d'entre eux sont stockés plutôt que brûlés pour fournir l'énergie quotidienne (Vors et al. 2013). Deuxièmement, la datation au ^{14}C des lipides adipeux montre que les lipides stockés ont une demi-vie d'environ 1.5 ans (Arner et al. 2011). Ainsi, pour avoir une compréhension complète des mécanismes de gain et perte de masse grasse, il est nécessaire de prendre en compte les mécanismes physiologiques de stockage et mobilisation des lipides. D'un point de vue clinique, il est important de savoir si certains groupes de patients ont des taux particuliers de stockage ou de mobilisation. Par exemple, une thérapie basée seulement sur l'exercice physique risque d'échouer si le taux de mobilisation des lipides ne peut pas être augmenté.

Les modèles mathématiques les plus utilisés en clinique pour prédire la prise ou la perte de poids sont ceux basés sur l'équilibre (ou le déséquilibre) en calories entre l'apport alimentaire et la dépense énergétique (K. D. Hall et al. 2011). Bien que ces modèles soient étonnamment performants pour prédire l'évolution de la masse grasse et de la masse maigre, ils ne font pas la distinction entre l'énergie provenant du tissu adipeux ou des lipides de la dernière prise alimentaire.

D'un autre côté, les modèles existants pour l'inférence des taux de stockage et de mobilisation des lipides à partir de leur datation au ^{14}C (Arner et al. 2011) font l'hypothèse que le poids du patient est stable. Ils ne permettent donc pas de cerner la façon dont ces taux s'adaptent dans le contexte d'une prise ou d'une perte de poids.

Dans le cadre de ma collaboration avec Samuel Bernard (Institut Camille Jordan et équipe Inria Dracula) et avec les équipes de Peter Arner et Kirsty Spalding (Karolinska University Hospital et Karolinska Institute, Stockhlom, Suède), nous avons le projet de combiner ces deux modèles pour caractériser la façon dont la physiologie des lipides (stockage, mobilisation) s'adapte au déséquilibre énergétique consécutif à une intervention thérapeutique visant une perte de poids (régime ou chirurgie bariatrique par exemple). Pour cela, nous pourrons utiliser les données longitudinales des cohortes suédoises SOWOT et DEOSH, pour lesquelles nous connaissons l'évolution de la masse grasse et de la masse maigre des patients pendant 4 à 10 ans, ainsi que l'évolution de l'âge de leurs lipides pendant cette période. Ce projet se décompose en quatre sous-objectifs :

O1. Calibrer le modèle d'équilibre énergétique de (K. D. Hall et al. 2011) pour estimer l'évolution du déséquilibre énergétique en fonction de l'évolution de la masse grasse et de la masse maigre ;

O2. Etendre le modèle de (Arner et al. 2011) pour permettre les variations de masse grasse, et utiliser les données d'âge des lipides pour inférer l'évolution des taux de stockage et de mobilisation des lipides ;

O3. Analyser la corrélation entre les résultats des deux approches pour caractériser la façon dont la physiologie des lipides s'adapte au déséquilibre énergétique lors de la perte de poids.

O4. Sur de plus longues échelles de temps (plus de 10 ans), prendre en compte le renouvellement des lipides, mais aussi le renouvellement des adipocytes, car le nombre de cellules adipeuses peut impacter la capacité de stockage à long terme. Des données de datation au ^{14}C de l'ADN des cellules adipeuses sont également disponibles (Spalding et al. 2008) et pourront être utilisées pour construire un modèle plus intégratif du tissu adipeux.

Dans le cadre du stage de M1 de Justine Antoine, co-encadré avec Samuel Bernard, nous avons pu réaliser l'objectif O1 pour la cohorte DEOSH, constituée de 39 femmes de 30 à 64 ans, suivies pendant 5 ans après une chirurgie bariatrique. Nous avons également commencé à travailler vers l'objectif O2, en testant deux scénarios simples pour l'évolution des taux de stockage et de mobilisation des lipides :

- scénario 1 : des lipides ne sont stockés que si l'apport énergétique est supérieur à la dépense énergétique, et de façon réciproque, des lipides du tissu adipeux ne sont mobilisés que si la dépense énergétique excède l'apport.
- scénario 2 : une proportion constante de l'apport énergétique est stockée sous forme de lipides dans le tissu adipeux, cette proportion étant déterminée grâce à l'âge initial des lipides (avant le début de la perte de poids).

Les résultats préliminaires montrent que le scénario 2 prédit beaucoup mieux que le 1 l'évolution de l'âge des lipides du tissu adipeux. Le scénario 1 est en fait clairement mauvais, alors qu'il est celui que nombre de cliniciens ou scientifiques ont en tête par défaut. Cela suggère que même dans une période de perte de poids rapide, un renouvellement des lipides continuerait de s'effectuer dans le tissu adipeux, impliquant donc qu'une partie des lipides de l'apport alimentaire quotidien va être stocké, même si cet apport alimentaire est loin de couvrir les dépenses énergétiques nécessaires.

III.3 Conclusion du chapitre

En résumé, je propose d'animer une équipe, ou une sous-équipe dans un premier temps, dont l'objectif serait de construire des modèles mécanistiques et quantitatifs des processus moléculaires, cellulaires et physiologiques qui gouvernent le destin des lipides alimentaires dans l'organisme, à court terme et à long terme. Je conçois la démarche de modélisation et de simulation comme complémentaire à l'approche plus traditionnelle des analyses statistiques. La modélisation et la simulation permettent de tester si le modèle conceptuel que l'on a en tête pour les mécanismes biologiques est cohérent avec les observations expérimentales, de réaliser *in silico* des expériences impossibles, d'identifier des connaissances manquantes, de guider la conception des expériences futures — en un mot d'exploiter le mieux possible les données expérimentales.

Ce projet nécessite des collaborations solides avec les expérimentateurs et les cliniciens, qui sont maintenant bien établies grâce à mon intégration dans le laboratoire CarMeN et ma

collaboration avec le Karolinska Institute, qui se sont traduites par quatre co-encadrements de stages de master et par des résultats préliminaires qui ont permis de cerner les principaux verrous. Il nécessite aussi des moyens de calcul, mais aussi et surtout un environnement scientifique de modélisateurs, tout deux accessibles grâce à mon appartenance à l'équipe Inria Beagle. Enfin, et c'est le plus important, il nécessite le recrutement de doctorants et de post-doctorants formés à l'intersection entre biologie, mathématiques, statistiques et informatique — a priori des ressources rares, mais qui existent en l'occurrence dans le département d'enseignement dont je suis actuellement responsable, à travers le parcours "Bioinformatique et Modélisation". Un enjeu important est donc l'obtention de financements de thèse et de post-doctorat. Une première allocation de thèse a été obtenue grâce à un financement Inserm-Inria, et la doctorante en question, Julie Etienne, vient de terminer sa première année de thèse. D'autres financements devront être obtenus, soit à travers des projets ANR ou européens, soit à travers des collaborations industrielles.

Conclusion générale

Les travaux menés sur la période 2007-2014 avec mes étudiantes, mes étudiants et mes collaborateurs a mis en lumière les limites de l'expérience de pensée en matière d'évolution moléculaire, et l'apport des approches de type "modèle minimal" (ou modèle "proof-of-concept" selon (Servedio et al. 2014)) pour explorer l'ensemble des possibles sous un certain jeu d'hypothèses.

Ces travaux ont aussi illustré comment on peut, partant d'une approche d'évolution expérimentale *in silico*, la combiner avantageusement avec d'autres approches pour aller plus loin : approche mathématique d'une part pour aller vers un modèle encore plus minimal des mécanismes en jeu, approche de génomique comparative d'autre part pour étudier des patterns évolutifs dans des génomes réels et pour tester les méthodes d'inférence. Cela nécessite cependant d'accepter d'être plus un "couteau suisse" méthodologique que l'experte internationalement reconnue de telle ou telle approche... or il n'y a pas de section CNU "couteau suisse méthodologique" ! Tant pis, j'assume, parce que je suis convaincue que c'est à l'intersection de toutes ces approches que se trouvent les résultats les plus éclairants.

La période 2014-2017 a été une période charnière. Le développement d'un modèle multi-échelles pour l'évolution des génomes, des réseaux métaboliques et des réseaux de régulation m'a permis d'amorcer un virage vers la modélisation des réseaux cellulaires d'une part, et vers le calibrage quantitatif des paramètres de ces réseaux d'autre part.

Les compétences développées durant cette période m'ont permis de compléter le virage vers la modélisation quantitative et prédictive. Rien n'est gratuit cependant : faire ce virage implique de se concentrer sur un système biologique particulier, en renonçant à rechercher des mécanismes universels et des invariants. C'est un choix conscient et qui correspond à un désir de voir plus concrètement les effets de ma recherche sur des enjeux biomédicaux.

Qui dit quantitatif et prédictif dit données expérimentales, et celles-ci furent au cœur de la période 2017-2020, avec les analyses statistiques réalisées à CarMeN sur les données de digestion et d'absorption de lipides, et les travaux préliminaires de modélisation menés avec mes étudiantes de master à CarMeN et à l'Inria.

En cet automne 2020, tous les éléments semblent maintenant réunis pour commencer à construire et animer une équipe dont l'objectif, ambitieux je le reconnais, sera la modélisation quantitative et mécanistique du devenir des lipides alimentaires, à court et à long terme.

Références bibliographiques

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

Ajmera, I. et al., 2013. The impact of mathematical modeling on the understanding of diabetes and related complications. *CPT: Pharmacometrics & Systems Pharmacology*, 2, p.e54.

Alexeev, N. & Alekseyev, M.A., 2017. Estimation of the true evolutionary distance under the fragile breakage model. *BMC Genomics*, 18(4), pp.1–9.

Altschul, S.F. et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17), pp.3389–3402.

Armand, M., 2008. Digestibilité des matières grasses chez l'homme. *Sciences des Aliments*, 28(1-2), pp.84–98.

Arner, P. et al., 2011. Dynamics of human adipose lipid turnover in health and metabolic disease. *Nature*, 478(7367), pp.110–113.

Bansal, S. et al., 2007. Fasting compared with nonfasting triglycerides and risk of cardiovascular events in women. *Journal of the American Medical Association*, 298(3), pp.309–316.

Bar-Even, A. et al., 2011. The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry*, 50(21), pp.4402–4410.

Batut, B., 2014. *Étude de l'évolution réductive des génomes bactériens par expériences d'évolution in silico et analyses bioinformatiques*. Thèse de doctorat, INSA de Lyon.

Batut, B. et al., 2013. In silico experimental evolution: a tool to test evolutionary scenarios. *BMC Bioinformatics*, 14(Suppl 15), p.S11.

Batut, B. et al., 2014. Reductive genome evolution at both ends of the bacterial population size spectrum. *Nature Reviews Microbiology*, 12(12), pp.841–850.

Bedau, M.A. et al., 2001. Open problems in artificial life. *Artificial life*, 6(4), pp.363–376.

Beiko, R.G. & Charlebois, R.L., 2007. A simulation test bed for hypotheses of genome evolution. *Bioinformatics*, 23(7), pp.825–831.

Beslon, G. & Knibbe, C., 2010. Petits bricolages en évolution. In *Des mondes bricolés ?* PPUR Presses polytechniques.

Beslon, G. et al., 2010. Scaling laws in bacterial genomes: a side-effect of selection of mutational robustness? *BioSystems*, 102(1), pp.32–40.

Bénichou, O. et al., 2005. Optimal search strategies for hidden targets. *Physical review letters*, 94(19), p.198101.

Biller, P., Guéguen, L. & Tannier, E., 2015. Moments of genome evolution by Double Cut-and-Join. *BMC Bioinformatics*, 16(Suppl 14), p.S7.

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

Biller, P., Knibbe, C., et al., 2016. Comparative Genomics on Artificial Life. In A. Beckmann, L. Bienvenu, & N. Jonoska, eds. *Pursuit of the Universal: Proceedings of the 12th Conference on Computability in Europe, CiE 2016, Paris, France, June 27 - July 1, 2016*. Springer, pp. 35–44.

Blount, Z.D. et al., 2012. Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature*, 489(7417), pp.513–518.

Blount, Z.D., Borland, C.Z. & Lenski, R.E., 2008. Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 105(23), pp.7899–7906.

Bourguignon, T. et al., 2020. Increased Mutation Rate Is Linked to Genome Reduction in Prokaryotes. 30(19), pp.3848–3855.e4.

Bouwman, J. et al., 2012. Visualization and identification of health space, based on personalized molecular phenotype and treatment response to relevant underlying biological processes. *BMC Medical Genomics*, 5(1), pp.1–9.

Bruen, T.C., Philippe, H. & Bryant, D., 2006. A simple and robust statistical test for detecting the presence of recombination. *Genetics*, 172(4), pp.2665–2681.

Charlesworth, J. & Eyre-Walker, A., 2006. The rate of adaptive evolution in enteric bacteria. *Molecular Biology and Evolution*, 23(7), pp.1348–1356.

Chow, S.S. et al., 2004. Adaptive radiation from resource competition in digital organisms. *Science*, 305(5680), pp.84–86.

Crombach, A. & Hogeweg, P., 2007. Chromosome Rearrangements and the Evolution of Genome Structuring and Adaptability. *Molecular Biology and Evolution*, 24(5), pp.1130–1139.

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

Crombach, A. & Hogeweg, P., 2009. Evolution of resource cycling in ecosystems and individuals. *BMC Evolutionary Biology*, 9, p.122.

Dalla Man, C., Rizza, R.A. & Cobelli, C., 2007. Meal simulation model of the glucose-insulin system. *IEEE transactions on bio-medical engineering*, 54(10), pp.1740–1749.

Dalquen, D.A. et al., 2012. ALF--A Simulation Framework for Genome Evolution. *Molecular Biology and Evolution*, 29(4), pp.1115–1123.

Danthine, S. et al., 2019. Homogeneous triacylglycerol tracers have an impact on the thermal and structural properties of dietary fat and its lipolysis rate under simulated physiological conditions. *Chemistry and Physics of Lipids*, 225, p.104815.

de Graaf, A.A. et al., 2009. Nutritional Systems Biology Modeling: From Molecular Mechanisms to Physiology. *PLoS Computational Biology*, 5(11), p.e1000554.

Deutsch, M.J. et al., 2014. Digital image analysis approach for lipid droplet size quantitation of Oil Red O-stained cultured cells. *Analytical Biochemistry*, 445, pp.87–89.

Dufresne, A., Garczarek, L. & Partensky, F., 2005. Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biology*, 6(2), pp.R14–10.

EFSA Panel on Food Additives and Nutrient Sources added to Food (ANS) et al., 2017. Re-evaluation of lecithins (E 322) as a food additive. *EFSA journal. European Food Safety Authority*, 15(4), p.e04742.

Elena, S.F. & Sanjuán, R., 2008. The effect of genetic robustness on evolvability in digital organisms. *BMC Evolutionary Biology*, 8(1), pp.284–9.

Eriksen, N. & Hultman, A., 2004. Estimating the expected reversal distance after a fixed number of reversals. *Advances in Applied Mathematics*, 32, pp.439–453.

Espinosa-Soto, C. & Wagner, A., 2010. Specialization can drive the evolution of modularity. *PLoS Computational Biology*, 6(3), p.e1000719.

Fischer, S., 2013. *Modélisation de l'évolution de la taille des génomes et de leur densité en gènes par mutations locales et grands réarrangements chromosomiques*. Thèse de doctorat, INSA de Lyon.

Fischer, S. et al., 2014. A model for genome size evolution. *Bulletin of Mathematical Biology*, 76(9), pp.2249–2291.

Flombaum, P. et al., 2013. Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus. *Proceedings of the National Academy of Sciences of the United States of America*, 110(24), pp.9824–9829.

Floreano, D. et al., 2007. Evolutionary conditions for the emergence of communication in robots. *Current Biology*, 17(6), pp.514–519.

Forbes, N., 2005. *Imitation of Life*, Cambridge, MA: MIT Press.

Giang, T.M. et al., 2015. Dynamic modeling highlights the major impact of droplet coalescence on the in vitro digestion kinetics of a whey protein stabilized submicron emulsion. *Food Hydrocolloids*, 43(C), pp.66–72.

Giang, T.M. et al., 2016. Dynamic modeling of in vitro lipid digestion: individual fatty acid release and bioaccessibility kinetics. *Food Chemistry*, 194(C), pp.1180–1188.

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

Grimm, V., 1999. Ten years of individual-based modelling in ecology: what have we learned and what could we learn in the future? *Ecological Modelling*, 115(2), pp.129–148.

Guindon, S. et al., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*, 59(3), pp.307–321.

Hall, B.G., 2008. Simulating DNA Coding Sequence Evolution with EvolveAGene 3. *Molecular Biology and Evolution*, 25(4), pp.688–695.

Hall, K.D. et al., 2011. Quantification of the effect of energy imbalance on bodyweight. *The Lancet*, 378(9793), pp.826–837.

Hanage, W.P. et al., 2006. Modelling bacterial speciation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 361(1475), pp.2039–2044.

Hannenhalli, S. & Pevzner, P.A., 1995. Transforming Men into Mice (polynomial algorithm for genomic distance problem). In Proceedings of 36th Annual Symposium on Foundations of Computer Science. pp. 1–16.

Hindré, T. et al., 2012. New insights into bacterial adaptation through in vivo and in silico experimental evolution. *Nature Reviews Microbiology*, 10(5), pp.352–365.

Hoban, S., Bertorelle, G. & Gaggiotti, O.E., 2012. Computer simulations: tools for population and evolutionary genetics. *Nature Reviews Genetics*, 13(2), pp.110–122.

Ishihama, Y. et al., 2008. Protein abundance profiling of the *Escherichia coli* cytosol. *BMC Genomics*, 9, p.102.

Jackson, K.G., Poppitt, S.D. & Minihane, A.M., 2012. Postprandial lipemia and cardiovascular disease risk: Interrelationships between dietary, physiological and genetic determinants. *Atherosclerosis*, 220(1), pp.22–33.

Jacob, F., 1977. Evolution and tinkering. *Science*, 196(4295), pp.1161–1166.

Johnson, T., 1999. The approach to mutation-selection balance in an infinite asexual population, and the evolution of mutation rates. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 266(1436), pp.2389–2397.

Kaneko, K., 2011. Proportionality between variances in gene expression induced by noise and mutation: consequence of evolutionary robustness. *BMC Evolutionary Biology*, 11, p.27.

Kashtan, N. & Alon, U., 2005. Spontaneous evolution of modularity and network motifs. *Proceedings of the National Academy of Sciences of the United States of America*, 102(39), pp.13773–13778.

Kashtan, N. et al., 2014. Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science*, 344(6182), pp.416–420.

Knibbe, C. & Parsons, D.P., 2014. What Happened to My Genes? Insights on Gene Family Dynamics from Digital Genetics Experiments. In H. Sayama et al., eds. *ALIFE 2014: Proceedings of The Fourteenth International Conference on the Synthesis and Simulation of Living Systems, July 30-August 2, 2014, New York, USA.* Cambridge, MA: MIT Press, pp. 30–40.

Knibbe, C. et al., 2007. A long-term evolutionary pressure on the amount of noncoding DNA. *Molecular Biology and Evolution*, 24(10), pp.2344–2353.

Kolovou, G.D. et al., 2011. Assessment and clinical relevance of non-fasting and postprandial triglycerides: an expert panel statement. *Current Vascular Pharmacology*, 9(3), pp.258–270.

Kovatchev, B.P. et al., 2009. In silico preclinical trials: a proof of concept in closed-loop control of type 1 diabetes. *Journal of Diabetes Science and Technology*, 3(1), pp.44–55.

Kuo, C.-H. & Ochman, H., 2009. Deletional bias across the three domains of life. *Genome Biology and Evolution*, 1, pp.145–152.

Kuo, P.D., Banzhaf, W. & Leier, A., 2006. Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence. *BioSystems*, 85(3), pp.177–200.

Langton, C.G. et al., 1991. *Artificial Life II* Santa Fe Institute Studies in the Sciences of Complexity. C. G. Langton et al., eds., Redwood, CA: Addison-Wesley.

Larget, B., Simon, D.L. & Kadane, J.B., 2002. Bayesian phylogenetic inference from animal mitochondrial genome arrangements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), pp.681–693.

Lenski, R.E. & Bennett, A.F., 1993. Evolutionary response of *Escherichia coli* to thermal stress. *The American Naturalist*, 142 Suppl 1, pp.S47–64.

Lenski, R.E. et al., 2003. The evolutionary origin of complex features. *Nature*, 423(6936), pp.139–144.

Li, Y. & McClements, D.J., 2010. New Mathematical Model for Interpreting pH-Stat Digestion Profiles: Impact of Lipid Droplet Characteristics on in Vitro Digestibility. *Journal of Agricultural and Food Chemistry*, 58(13), pp.8085–8092.

Liard, V. et al., 2020. The Complexity Ratchet: Stronger than Selection, Stronger than Evolvability, Weaker than Robustness. *Artificial Life*, 26(1), pp.38–57.

Lindell, D. et al., 2004. Transfer of photosynthesis genes to and from Prochlorococcus viruses. *Proceedings of the National Academy of Sciences of the United States of America*, 101(30), pp.11013–11018.

Lipinski, K.J. et al., 2011. High spontaneous rate of gene duplication in *Caenorhabditis elegans*. *Current Biology*, 21(4), pp.306–310.

Löytynoja, A. & Goldman, N., 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences of the United States of America*, 102(30), pp.10557–10562.

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

Lynch, M. et al., 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 105(27), pp.9272–9277.

Marais, G.A.B., Batut, B. & Daubin, V., 2020. Genome Evolution: Mutation Is the Main Driver of Genome Size in Prokaryotes., 30(19), pp.R1083–R1085.

Marze, S. & Choimet, M., 2012. In vitro digestion of emulsions: mechanistic and experimental models. *Soft Matter*, 8(42), p.10982.

Mattiussi, C. & Floreano, D., 2007. Analog genetic encoding for the evolution of circuits and networks. *IEEE Transactions on Evolutionary Computation*, 11(5), pp.596–607.

Maurizi, M.R., 1992. Proteases and protein degradation in *Escherichia coli*. *Experientia*, 48(2), pp.178–201.

Ménard, O. et al., 2018. A first step towards a consensus static in vitro model for simulating full-term infant digestion. *Food Chemistry*, 240, pp.338–345.

Miller, G.F., 1995. Artificial Life as Theoretical Biology: How to do real science with computer simulation. *Cognitive and Computing Sciences Research Reports*, 378, 33 pages.

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

Misevic, D., Ofria, C. & Lenski, R.E., 2006. Sexual reproduction reshapes the genetic architecture of digital organisms. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 273(1585), pp.457–464.

Molina, N. & Van Nimwegen, E., 2008. The evolution of domain-content in bacterial genomes. *Biology Direct*, 3(1), p.51.

Nilsson, A.I. et al., 2005. Bacterial genome size reduction by experimental evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 102(34), pp.12112–12116.

Nordestgaard, B.G. et al., 2016. Fasting Is Not Routinely Required for Determination of a Lipid Profile: Clinical and Laboratory Implications Including Flagging at Desirable Concentration Cutpoints-A Joint Consensus Statement from the European Atherosclerosis Society and European Federation of Clinical Chemistry and Laboratory Medicine. *Clinical Chemistry*, 62(7), pp.930–946.

Palmer, M.E. & Feldman, M.W., 2011. Spatial environmental variation can select for evolvability. *Evolution*, 65(8), pp.2345–2356.

Parsons, D.P., 2011. *Sélection Indirecte en Évolution Darwinienne : Mécanismes et Implications*. Thèse de doctorat, INSA de Lyon.

Parsons, D.P., Knibbe, C. & Beslon, G., 2011. Homologous and Nonhomologous Rearrangements: Interactions and Effects on Evolvability. In T. Lenaerts, ed. *Advances*

in Artificial Life, ECAL 2011: Proceedings of the Eleventh European Conference on the Synthesis and Simulation of Living Systems, Paris, France August 8-12, 2011. Cambrige, MA: European Conference on Artificial Life (ECAL), pp. 622–629.

Parsons, D.P., Knibbe, C. & Beslon, G., 2010. Importance of the Rearrangement Rates on the Organization of Genome Transcription. In H. Fellermann et al., eds. *Artificial Life XII: Proceedings of the Twelfth International Conference on the Synthesis and Simulation of Living Systems, Odense, Denmark, August 19-23, 2010*. Cambridge, MA, pp. 479–486.

Partensky, F. & Garczarek, L., 2010. *Prochlorococcus*: advantages and limits of minimalism. *Annual Review of Marine Science*, 2, pp.305–331.

Peignier, S., Rigotti, C. & Beslon, G., 2015. Subspace Clustering Using Evolvable Genome Structure. In S. Silva & A. I. Esparcia-Alcázar, eds. *GECCO'15: Proceedings of the 2015 Genetic and Evolutionary Computation Conference, Madrid, Spain, July 11-15, 2015*. New York, NY, United States: ACM Press, pp. 575–582.

Penel, S. et al., 2009. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*, 10 Suppl 6, p.S3.

Ray, T.S., 1991. An approach to the synthesis of life. In C. G. Langton et al., eds. *Artificial Life II*. Redwood, CA: Addison-Wesley, pp. 317–408.

Rennard, J.P., 2002. *Vie Artificielle : Où La Biologie Rencontre L'informatique*, Paris: Vuibert.

Robert, C. et al., 2020. Rapeseed Lecithin Increases Lymphatic Lipid Output and α-Linolenic Acid Bioavailability in Rats. *The Journal of Nutrition*, in press.

Rocabert, C., 2017. *Étude de l'évolution des micro-organismes bactériens par des approches de modélisation et de simulation informatique*. Thèse de doctorat, INSA de Lyon.

Rocabert, C. et al., 2017. Beware batch culture: Seasonality and niche construction predicted to favor bacterial adaptive diversification. *PLoS Computational Biology*, 13(3), p.e1005459.

Sambuy, Y. et al., 2005. The Caco-2 cell line as a model of the intestinal barrier: influence of cell and culture-related factors on Caco-2 cell functional characteristics. *Cell Biology and Toxicology*, 21(1), pp.1–26.

Schrider, D.R. et al., 2013. Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. *Genetics*, 194(4), pp.937–954.

Servedio, M.R. et al., 2014. Not just a theory--the utility of mathematical models in evolutionary biology. *PLoS Biology*, 12(12), p.e1002017.

Shorten, P.R. & Upreti, G.C., 2005. A mathematical model of fatty acid metabolism and VLDL assembly in human liver. *Biochimica et Biophysica Acta*, 1736(2), pp.94–108.

Sips, F.L.P. et al., 2015. Model-Based Quantification of the Systemic Interplay between Glucose and Fatty Acids in the Postprandial State. *PLoS ONE*, 10(9), pp.e0135665–21.

Smith, J.M., 1992. Evolutionary biology. Byte-sized evolution. *Nature*, 355(6363), pp.772–773.

Sniegowski, P.D. et al., 2000. The evolution of mutation rates: separating causes from consequences. *BioEssays*, 22(12), pp.1057–1066.

Spalding, K.L. et al., 2008. Dynamics of fat cell turnover in humans. *Nature*, 453(7196), pp.783–787.

Stroock, D.W., 2006. *An Introduction to Markov Processes*, Berlin/Heidelberg: Springer Science & Business Media.

Strope, C.L. et al., 2009. Biological Sequence Simulation for Testing Complex Evolutionary Hypotheses: indel-Seq-Gen Version 2.0. *Molecular Biology and Evolution*, 26(11), pp.2581–2593.

Sun, S. et al., 2012. Genome-Wide Detection of Spontaneous Chromosomal Rearrangements in Bacteria. *PLoS ONE*, 7(8), p.e42639.

Taddei, F. et al., 1997. Role of mutator alleles in adaptive evolution. *Nature*, 387(6634), pp.700–702.

Tenaillon, O. et al., 1999. Mutators, population size, adaptive landscape and the adaptation of asexual populations of bacteria. *Genetics*, 152(2), pp.485–493.

Tenaillon, O. et al., 2001. Second-order selection in bacterial evolution: selection acting on mutation and recombination rates in the course of adaptation. *Research in Microbiology*, 152(1), pp.11–16.

Thorning, T.K. et al., 2017. Whole dairy matrix or single nutrients in assessment of health effects: current evidence and knowledge gaps. *The American Journal of Clinical Nutrition*, 105(5), pp.1033–1045.

Tusscher, ten, K.H.W.J. & 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, p.159.

van den Broek, T.J. et al., 2017. Ranges of phenotypic flexibility in healthy subjects. *Genes & Nutrition*, 12, p.32.

van Ommen, B., Cavallieri, D., et al., 2008. The challenges for molecular nutrition research 4: the "nutritional systems biology level". *Genes & Nutrition*, 3(3-4), pp.107–113.

van Ommen, B., Keijer, J., et al., 2008. The challenges for molecular nutrition research 2: quantification of the nutritional phenotype. *Genes & Nutrition*, 3(2), pp.51–59.

Vincent, M. et al., 2020. Human milk pasteurisation reduces pre-lipolysis but not digestive lipolysis and moderately decreases intestinal lipid uptake in a combination of preterm infant in vitro models. *Food Chemistry*, 329, p.126927.

Volkmer, B. & Heinemann, M., 2011. Condition-dependent cell volume and concentration of Escherichia coli to facilitate data conversion for systems biology modeling. *PLoS ONE*, 6(7), p.e23126.

Vors, C. et al., 2014. Intérêt de la phase postprandiale pour la santé de l'Homme. *Obésité*, 9(1), pp.31–41.

Vors, C. et al., 2013. Modulating absorption and postprandial handling of dietary fatty acids by structuring fat in the meal: a randomized crossover clinical trial. *The American Journal of Clinical Nutrition*, 97(1), pp.23–36.

Vors, C., Lecomte, M. & Michalski, M.-C., 2016. Impact de la structure émulsionnée des lipides sur le devenir métabolique des acides gras alimentaires. *Cahiers de Nutrition et de Diététique*, 51(5), pp.238–247.

Waibel, M., Floreano, D. & Keller, L., 2011. A quantitative test of Hamilton's rule for the evolution of altruism. *PLoS Biology*, 9(5), p.e1000615.

Wallstab, C. et al., 2017. A unifying mathematical model of lipid droplet metabolism reveals key molecular players in the development of hepatic steatosis. *The FEBS Journal*, 284(19), pp.3245–3261.

Weiße, A.Y. et al., 2015. Mechanistic links between cellular trade-offs, gene expression, and growth. *Proceedings of the National Academy of Sciences of the United States of America*, 112(9), pp.e1038–47.

Wilke, C.O. et al., 2001. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844), pp.331–333.

Woods, R.J. et al., 2011. Second-order selection for evolvability in a large Escherichia coli population. *Science*, 331(6023), pp.1433–1436.

Xu, Z., McClure, S.T. & Appel, L.J., 2018. Dietary Cholesterol Intake and Sources among U.S Adults: Results from National Health and Nutrition Examination Surveys (NHANES), 2001-2014. *Nutrients*, 10(6), p.771.

Zeisel, S.H. et al., 2005. The nutritional phenotype in the age of metabolomics. *The Journal of Nutrition*, 135(7), pp.1613–1616.