Étude bioinformatique de l'évolution de l'usage du code génétique

Fanny Pouyet

▶ **To cite this version:**

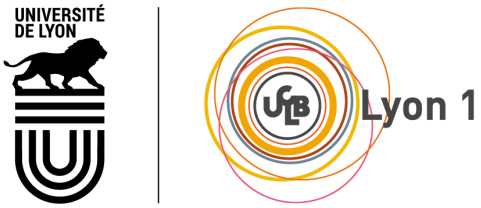
Fanny Pouyet. Étude bioinformatique de l'évolution de l'usage du code génétique. Génétique humaine. Université de Lyon, 2016. Français. NNT : 2016LYSE1140 . tel-01410446

HAL Id: tel-01410446
https://theses.hal.science/tel-01410446

Submitted on 6 Dec 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



N° d'ordre NNT : 2016LYSE1140

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON

opérée au sein de

l'Université Claude Bernard Lyon 1

École Doctorale ED341

E2M2 - Évolution Écosystèmes Microbiologie Modélisation

Soutenue publiquement le 13/09/2016, par :

Fanny Pouyet

Étude Bio-informatique de l'Évolution de l'Usage du Code Génétique

---

Devant le jury composé de :

Achaz Guillaume, Maître de Conférences Universitaire, Collège de France	Rapporteur
Galtier Nicolas, Directeur de Recherche CNRS, ISEM Montpellier II	Rapporteur
Salamin Nicolas, Professeur, Université de Lausanne	Rapporteur
Gascuel Olivier, Directeur de Recherche, Pasteur	Examinateur
Guéguen Laurent, Maître de Conférences, Lyon 1	Directeur de thèse
Mouchiroud Dominique, Professeure des Universités, Lyon 1	Directrice de thèse
Bailly-Bechet Marc, Maître de Conférences, Lyon 1	Encadrant de thèse

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. le Professeur Frédéric FLEURY**
Prédisent du Conseil Académique	M. le Professeur Hamda BEN HADID
Vice-président du Conseil d'Administration	M. le Professeur Didier REVEL
Vice-président du Conseil Formation et Vie Universitaire	M. le Professeur Philippe CHEVALIER
Vice-président de la Commission Recherche	M. Fabrice VALLÉE
Directeur Général des Services	M. Alain HELLEU

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur J. ETIENNE
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. le Professeur Y. MATILLON
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : Mme. la Professeure A-M. SCHOTT

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. le Professeur F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. THEVENARD
Département Chimie Biochimie	Directeur : Mme C. FELIX
Département GEP	Directeur : M. H. HAMMOURI
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur G. TOMANOV
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : M. le Professeur J-C PLENET
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. Y.VANPOULLE
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. le Professeur E. PERRIN
École Supérieure de Chimie Physique Électronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. le Professeur C. VITON
Institut Universitaire de Formation des Maîtres	Directeur : M. le Professeur A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Directeur : M. N. LEBOISNE

À Agathe Benvenuti et Vincent Hennion,
avec qui j'ai implémenté mon premier modèle d'évolution de séquences en T.I.P.E.
Cela remonte à 2009 alors, merci !

Résumé

Le code génétique est la table de correspondance entre codons (unité structurelle d'un gène) et acides aminés (brique élémentaire des protéines). Le code génétique est (1) universel, tous les êtres vivants ou presque partagent le même code ; (2) univoque, chaque codon spécifie un seul acide aminé et (3) dégénéré, les acides aminés peuvent être codés par plusieurs codons. Ce code dégénéré est donc utilisé par l'ensemble du vivant mais pas de la même manière, certains codons synonymes étant utilisés préférentiellement chez des espèces et pas d'autres. Pour comprendre l'émergence des biais d'usage du code (BUC) génétique entre espèces, je me place dans un contexte évolutif.

Dans ce manuscrit, je présente mes travaux de recherche en quatre parties. La première partie introductive décrit la mise en évidence et les propriétés du code génétique, son biais d'usage et les diverses caractéristiques de précédents modèles de codons. La deuxième partie présente le modèle d'évolution de codons SENCA pour Sites Evolution at the Nucleotides, Codons and Amino-acids layers que j'ai développé durant ma thèse. SENCA prend en compte la structure du code génétique. Je valide sa paramétrisation par des simulations numériques et une étude sur des espèces bactériennes ou archées. La partie suivante décrit deux extensions de SENCA qui modélisent plusieurs hypothèses d'origines évolutives du BUC et une application de SENCA sur les conséquences génomiques d'adaptations environnementales. La dernière partie étudie les origines de variations de BUC le long du génome humain par une approche de génomique comparative.

Abstract

The genetic code is the correspondence table between codons (structural units of proteic genes) and amino acids (molecular units of proteins). This genetic code is (1) universal as almost all living being use it; (2) univocal, each codon specify one amino acid and (3) redundant, an amino acid can be coded by several synonymous codons. Degeneracy of the code does not however lead to a uniform usage of those synonymous codons: at species or at genomic level, a particular codon is preferrentially used over other synonymous ones. Codon usage bias (CUB) vary between species and between genes. To fully understand how this bias arose and is maintained in a set of genes, I study codon usage in an evolutionary approach.

In this manuscript, I introduce my doctoral research in four parts. The first introductive part highlights the properties of the genetic code and its usage bias but also the caracteristics of previous published codons models. The second part presents an evolutionary codons models named SENCA for Sites Evolution at the Nucleotides, Codons and Amino-acids layers that I developped. SENCA takes into account the genetic code structure. I perform simulations and study prokaryotes species to confirm its parametrization. The following part provides two extensions of SENCA to test the hypotheses concerning the evolutive origins of CUB and an application of SENCA to study the genomic consequences of an environmental adaptation. The last part studies the origins of CUB variation within the human genome using a comparative genomic strategy.

Remerciements

Je ne pensais pas que la pire page blanche soit celle-ci. Je voudrais écrire des pages entières pour chaque personne qui a été là pour moi mais ça risque d'être long et pénible. Titre.
Conforme. Je tiens à remercier Guillaume Achaz, Nicolas Galtier et Nicolas Salamin d'avoir accepté de lire et d'évaluer ce manuscrit.Merci également à Olivier Gascuel d'avoir accepté de juger mon travail. Merci aussi aux membres de mon comité de pilotage qui m'ont effectué des retours essentiels sur l'avancement de ma thèse : Christophe Douady, Manolo Gouy, Bénédicte Lafay et Eduardo Rocha.
Évident. Merci à Marc Bailly-Bechet, Laurent Guéguen et Dominique Mouchiroud de m'avoir encadrée durant ces 3 dernières années. Je les remercie sincèrement de m'avoir fait confiance, parfois un peu trop, de m'avoir poussée à toujours présenter mon travail, à réfléchir à haute voix et permis de constater qu'une bonne recherche se fait toujours à plusieurs. Bien évidemment, je les remercie de m'avoir tout appris en recherche ! Je souhaite plutôt faire des remerciements personnels alors, merci à Marc de m'avoir aidée, surtout au début, à apprendre R et LateX, pour les enseignements et d'avoir si bien su gérer les timings pour la rédaction de l'article et de la thèse. Merci à Laurent, tout particulièrement d'avoir été présent, patient puis il faut le dire, de m'avoir appris à déstresser quand le programme ne suit pas mes attentes ! J'ai vraiment aimé faire ma thèse avec toi, je pense que tu m'as bien plus appris que ce que je pourrais écrire en quelques mots. Enfin merci à Dominique, déjà pour avoir été ma référente de L3 à l'ENS, pour m'avoir présenté le LBBE et pour m'avoir suivie durant cette thèse tout en étant directrice du laboratoire. Il y a ces personnes qu'on admire, tu en fais partie.
Indispensable. Merci à Laurent Duret et Marie Sémon pour notre projet de recherche ensemble. Merci en fait d'avoir voulu travailler avec moi car notre projet est vraiment très excitant et je n'aurai jamais osé venir vous le proposer. Le cheminement scientifique de Laurent est remarquable. J'ai adoré travaillé avec Laurent car, en plus des résultats c'est la façon de le présenter et d'extraire ce qui est utile ou non qui m'ont impressionné. Merci à Marie, pour sa spontanéité et aussi, pour les idées qui fusent et se révèlent toujours excellentes ! J'ai vraiment eu de la chance de travailler avec toi car cela m'a permis de mieux comprendre les enjeux et comment définir les questions auxquelles il faut répondre.
Essentiel. C'est peut-être la partie la plus intime. Titre. Je voudrais remercier mes co-bureaux car finalement, c'est avec eux que j'ai passé la majeure partie de ma thèse. Ça a pris du temps pour me sentir à l'aise dans un bureau aussi impressionnant. Vincent Daubin et Éric Tannier

qui refont le monde politique ou scientifique plus d'une fois par semaine, et Bastien Boussau qu'on nous présente en cours à l'ENS comme le meilleur chercheur de sa génération. Au début, t'écoutes, tu apprends puis enfin tu participes et tu te rends compte qu'ils sont sympas et que vraiment t'espères garder contact. Scientifiquement bien entendu. Mais pas que. Merci Vincent pour avoir été un voisin et pour avoir organisé des conférences comme celle de Roscoff ou les meetings d'Ancestrome. Je me suis bien marrée. Merci Éric pour m'avoir écouté quand ça n'allait pas fort, pour les questions scientifique et pour ton coté critique. Tu prends ton temps pour répondre, alors au début ces silences mettent mal à l'aise, merde, j'ai dit quelque chose de nul ou d'inintéressant. En fait, j'ai vraiment l'impression qu'on s'entend bien puis j'aime bien écouter ta façon d'appréhender une question, c'est toujours enrichissant. Merci à Bastien aussi, j'ai déjà lancé trop de fleurs avant, alors simplement merci pour tes conseils avant de commencer ma thèse mais aussi, pour m'avoir motivée pour la recherche de post-doc et pour nos discussions scientifiques.

Fun. Il y a vraiment énormément de collègues que je souhaite remercier. Merci tout particulièrement à Magali pour avoir si bien partagé mes sentiments sur la thèse et nos directeurs. Merci bien évidemment à Thomas (il se vexerait si ces lignes n'existaient pas) pour m'avoir aidée en informatique et pour nos discussions sur les relations humaines (c'est bien dit ;)) ! Merci à Héloïse pour avoir partagé ces 3 années dans le labo, m'avoir aidée quand je ne comprenais rien à l'informatique et m'avoir accompagnée dans les mouvements d'idées comme Sciences en Marche. Merci à Marie C. pour m'avoir si bien acceuillie au labo, pour les soirées chez toi, les sorties cinéma (sauf une) et les discussions sur le féminisme. Merci aux filles pour les hammams (Aline, Cécile et les 3 ci-dessus. Titre.). Merci à Magali, Wandrille, Simon, Rémi, Michel, Thomas et Murray pour avoir accepté de se mettre en slip dans le bureau pour une photo ! Merci à tout le deuxième étage pour la journée de travail-ski, pour le samedi bricolage chez Damien mais aussi pour les BBQ chez Laurent D. Merci à G3K pour les conseils d'enseignements et nos discussions. Merci à Clément G. pour les séminaires des doctorants, nos affiches étaient top ! Merci aussi à Clément G. et Christophe pour les Happy Hour, Sciences en Marche et tout ça. Merci à Laurent J. pour les restaurants, les discussions politiques même si après quelques cocktails on n'était pas toujours d'accord. Merci à ceux que j'ai appris à connaître un peu plus tard, Simon, Philippe, Guillaume G., Adrian, Frédéric, Damien et Daniel pour les pauses et les discussions de couloir. Merci en particulier à Wandrille que j'ai appris à connaître, qui m'a appris beaucoup sur le moyen-âge et l'Histoire (d'ailleurs, on a une mission en cours, que je ne peux pas mettre dans ces remerciements, mais qu'on devrait réaliser avant que je ne parte !). Je ne dis

pas merci à Ghislain, parce qu'il veut que je le paye pour qu'il me remercie dans son propre manuscrit. Abusé. Merci aux informaticiens pour avoir sauvé mon ordi plus d'une fois. Merci au secrétariat pour avoir si bien géré mes missions. Merci aussi au PRABI pour leur prestations de services et à Dominique pour cette belle formulation. Merci aussi aux anciens Yann, Sophie, Mathieu et Florent et à ceux que j'oublie. Enfin, merci Nicolas, Damien, Vincent, Magali R. pour les #NuitDebout. Thanks also to Will, Peter and Murray for your useful comments on my english. Enfin, merci aux autres personnes avec qui j'ai eu l'occasion de travailler comme Clémentine François et Tristan Lefébure.

Historique. Merci à mes anciens responsables de stage qui m'ont donné envie de faire de la recherche, Vincent Daubin, Phil Donoghue, Gaël Yvert et Laurent Guéguen.

Généalogique. Merci à toute ma famille pour m'avoir toujours soutenue et donnée envie de faire ce qui me plaît. Bien évidemment, merci maman, ma première supportrice, et merci papa, vous m'avez donné le goût d'apprendre, de faire et d'essayer. Grâce à vous, j'ai confiance et ça n'a pas de prix. Merci à mes frères : Pierre m'a appris que chacun est différent, chacun est enrichissant et chacun se vaut ; Julien et sa femme Céline m'ont montré comment être efficace et savoir débattre. Des bisous à mes neveux Maximilien, Célestin et Théophile. Un merci particulier à mes grands-parents. Merci Michel, pour m'avoir donnée la joie de vivre et l'espièglerie. Merci aussi à mes autres grand-parents qui ne sont plus là, Fernande qui m'a appris à bien me tenir, Ginette et Jean qui m'ont appris la curiosité et m'ont fait connaître le monde de la recherche. C'est qu'une question de timing et je prends un peu d'avance mais merci à ma belle-famille. Merci Anne pour ta douceur et ton écoute. Merci Julien et Géraldine pour les bons moments passés à Lipari ou à Paris. Des bisous aussi à mon neveu et ma nièce Abel et Olga.

Affectif. Bien sûr merci aux copains ! Le plus grand de tous les mercis aux Relous, cette bande de potes pour qui les meilleures blagues sont les plus ... longues ! Titre ? Merci donc à la meilleure des colocs avec Augustin, Avelyne, ClémentS, Tomàs et Simon. C'était si bien que j'y ai rencontré mon Clément. Merci à Augustin et Simon pour nos randonnus à la coloc ou en vrai ! Merci à Avelyne pour son enthousiasme et son pep's, à Clément D. pour sa musique et son envie de nature. Merci à Tomàs pour nos mardis soir gargantuesques, je regrette que tu ne lises pas ces lignes. C'est le moment le plus douloureux là car je n'en suis toujours pas remise. Merci à la pompom coloc, Laura, Carole, Sara et Lise. Merci en particulier à Sara, ma témoin qui a su rester lyonnaise jusqu'au bout, à Laura pour nos soirées jeux et à Carole pour nos cluboufs et en particulier pour notre dessert de mariage (c'est les mêmes quantités). Même si "ce[tte

thèse] n'est pas un[e thèse] sur le cyclimse", merci au sous-groupe vélo, j'ai nommé Clément Debin, Marlène Sasoeur (ou Blanchet, j'ai jamais su), Avelyne, Solenn, Lauren, Tomàs et Clément Tauber. Merci au couple bidochon Théo et Léonie. On ne partira plus jamais en vacances en Irlande mais la Sicile ou la Macédoine c'est quand vous voulez ! Merci à Camille, mon amie de domaine de thèse, dommage que tu deviennes prof, j'aurai bien collaboré avec toi. Merci aux autres copains de Lyon, Leslie, Maxime, Magali R., Florent, Arthur, Richard et Gopi. Enfin, merci à mes éternels copains de lycée, Minh-tu, Michael et Céline H. J'apprécie toujours de voir mes copains parisiens pour parler du bon vieux temps et de nos projets, si différents maintenant ! Et aussi, merci à tous ceux que j'ai oublié bien évidemment.
Sentimental. Merci Clément Tauber, tu pourrais faire partie de plusieurs de ces paragraphes, du coup tu en mérites bien un à toi tout seul. Tu m'apaises, sans toi je ne suis ni sereine ni forte. Merci pour notre mariage, pour nos 6 années déjà passées ensemble et pour l'envie de continuer notre vie ensemble. Merci pour nos mignonneries, pour ton humour si précis et efficace, pour partager mon amour des animaux (et d'Émile). Titre. En ce qui concerne cette thèse, merci pour ton soutien dans les moments difficiles, pour avoir répondu à mes questions de comment faire, pour m'avoir encouragée à essayer, à faire.

Bonne lecture à toi, personne lectrice de la suite !

Abréviations

A	Adénine
AA	Acides aminés
ADN	Acide Désoxyribonucléique
ARN	Acide Ribonucléique
ARNm	ARN messager
ARNt	ARN de transfert
BUC	Biais d'Usage du Code
C	Cytosine
CDS	Coding DNA sequence ; correspond à la partie du gène codant la protéine
CUB	Codon Usage Bias
ENC	Effective Number of Codons (Wright, 1990) ; mesure du biais d'usage du code qui varie entre vaut 61 (si l'usage est aléatoire) et 20 (si un seul codon est utilisé par acide aminé)
G	Guanine
gBGC	Biais de Conversion Génique biaisé vers GC ; intervient lors de la réparation des mésappariements de la recombinaison.
GC	Taux de G+C dans une région (gène, région flanquante etc.)
GC3	Taux de G+C en 3ème position des codons (la position la plus redondante)
GO	Gene Ontology
T	Thymine

Table des matières

I Introduction **21**

1 Le code génétique **23**
1.1 Mise en évidence du code génétique 24
1.2 Les propriétés physiques et chimiques du code 26

2 Le biais d'usage du code génétique **31**
2.1 Définition et mise en évidence 31
2.2 Origines du biais d'usage du code 33
2.3 Comment mesurer le biais d'usage ? 37

3 Les modèles évolutifs de séquences **45**
3.1 Les modèles nucléotidiques 49
3.2 Les modèles de codons 52

II SENCA : étude de l'origine et de l'évolution du biais d'usage du code à l'aide d'un modèle de codons multi-couches 59

4 Avant-propos 61

III Applications et extensions de SENCA 79

5 Le Biais d'Usage du Code chez *Homo sapiens* : extension avec le gBGC et l'hypermutabilité des CpG 83

5.1 Organisation génomique et mécanismes évolutifs chez *Homo sapiens* 84
5.2 Paramétrisation du gBGC et des CpG 90
5.3 Données analysées 94
5.4 Résultats 97
5.5 Discussion et perspectives 106

6 Extension : Expression de gènes 111

6.1 Extension : le paramètre `expr` 112
6.2 Simulation 113
6.3 Perspectives 117

7 Application : Les proaselles **119**
7.1 Le modèle biologique 120
7.2 Matériels et méthodes 122
7.3 Résultats 122
7.4 Discussion et perspectives 129

8 Perspectives d'utilisation de SENCA **131**

IV Pourquoi l'usage des codons synonymes varie entre différentes catégories fonctionnelles chez l'Homme ? **133**

9 Avant-propos **135**

V Conclusion Générale **167**

A Supplementary Material of: SENCA : a multi-layered codon model to study the origins and dynamics of codon usage paper **187**

B Supplementary Material of: Why does synonymous codon usage vary among different functional categories of humans genes? paper **207**

PREMIÈRE PARTIE

Introduction

CHAPITRE 1

Le code génétique

Pour parler du code génétique, il est nécessaire d'introduire le dogme central de la biologie moléculaire. Ce dogme proposé en 1958 par Francis Crick et publié en 1970 (Crick, 1970), affirme que le matériel génétique est le seul à contenir de l'information pour coder les protéines (les protéines ne peuvent pas s'auto-générer). L'information génétique et plus précisément les gènes se situent sur un brin d'acide désoxyribonucléique (ADN), composé de quatre éléments : l'adénine (A), la cytosine (C), la guanine (G) et la tyrosine (T); alors que les protéines sont composées de 20 types d'acides aminés (AA) différents. Un gène est transcrit en acide ribonucléique messager (ARNm) qui est alors traduit en protéine par des ARN de transfert (ARNt). Les ARNt sont des ARN non codant assurant la correspondance entre l'information génétique contenue dans l'ARNm et les protéines à l'aide d'une séquence de 3 nucléotides, appelée anticodon, qui s'apparie au codon de l'ARNm conformément au code génétique. Le code génétique est donc l'alphabet qui permet la traduction d'un brin d'ADN en une séquence protéique (figure 1.1).

Chapitre écrit d'après les livres : Biochimie de Voet et Voet (2005) et L'essentiel de la Biologie cellulaire de Alberts, Bray, Hopkin, Johnson, Lewis, Raff, Roberts et Walter (2004)

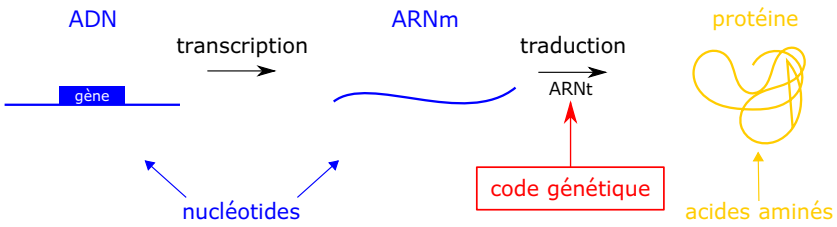


FIGURE 1.1 – Le dogme central de la biologie moléculaire, proposé par Crick (1970). En bleu, les éléments contenant des nucléotides, en jaune des acides aminés (AA). Le code génétique est représenté en rouge. Il est utilisé par les ARNt qui traduisent l'ARNm en protéines. Ces couleurs seront utilisées tout le long de ce manuscrit.

1.1 Mise en évidence du code génétique

Le code génétique peut théoriquement être spécifié de différentes manières. Avec 4 bases pour 20 AA, un groupe de plusieurs bases appelé codon est nécessaire pour spécifier un acide aminé. Un codon est au minimum un triplet puisque cela permet d'avoir $4^3 = 64$ codons différents alors qu'un doublet n'autorise que $4^2 = 16$ possibilités. Avec les triplets, soit il y a 44 codons qui ne codent pas pour un AA soit, plusieurs codons codent un même AA. Ce dernier code redondant, est dit dégénéré selon un terme emprunté aux mathématiques. Par ailleurs, il existe plusieurs manières de déchiffrer une séquence d'ADN. Par exemple, la séquence :

$$ABCDEFGHI...\tag{1.1}$$

peut être lue codon après codon, ABC puis DEF puis GHI ou bien, le code est chevauchant et lu, ABC puis BCD puis CDE etc. Dernière possibilité, un motif peut séparer chaque codon qui doit être décodé, si D et H sont des motifs particuliers, les codons lus sont ABC puis EFG. Enfin, on sait que les gènes sont contenus sur un brin d'ADN, il existe donc un motif qui définit le début et la fin de chacun d'eux. Je présente ici les découvertes historiques qui ont permis de déterminer que le code génétique est : composé de triplets, dégénéré, non chevauchant et sans motifs séparant les codons.

Un code à triplets

Crick et al. (1961) ont déterminé que le code est à triplets grâce aux études de mutations sur le cistron rIIB du phage T4. Les mutations sont des changements de séquences. Elles se divisent en

deux grandes catégories : (1) les mutations ponctuelles de type substitution que sont les transitions dans lesquelles une purine, A ou G (respectivement une pyrimidine T ou C) est remplacée par une autre purine (resp. pyrimidine) et les transversions où la nature chimique du nucléotide est modifiée, une purine étant remplacée par une pyrimidine et vice-versa. Il existe aussi (2) les mutations par insertion ou délétion d'un ou plusieurs nucléotides. Crick et al. (1961) ont étudié des mutants du phage T4 d'*Escherichia coli* qui a pour phénotype sauvage de lyser son hôte. Lorsque ce phage est soumis à une simple mutation du cistron rIIB, il ne peut plus détruire son hôte, le phage ne peut recouvrir sa capacité à infecter *E. coli* que si le nombre de mutations est paire ou divisible par 3. Ces résultats impliquent que les mutations étudiées sont des insertions ou des délétions de nucléotides (qui modifient le cadre de lecture du gène). Lorsque le nombre de mutations est paire, il y a autant d'insertions que de délétions. Le code génétique est lu de manière séquentielle car les mutations décalent le cadre de lecture. Par ailleurs, ils ont montré que si le nombre de mutations est divisible par 3, les mutants ont toujours un phénotype sauvage. Le code est donc à triplet. Enfin, les triplets codent tous ou presque tous pour un acide aminé, le code est dégénéré. A noter qu'en évolution moléculaire, une substitution désigne une mutation de type substitution qui est fixée dans une population/une espèce. Par la suite, j'emploierai le terme de substitution en me référant à la définition d'évolution moléculaire (sauf indication contraire).

Décryptage du code

Le code a ensuite été décrypté entre 1961 et 1966 (figure 1.2). Aujourd'hui pour décrypter

FIGURE 1.2 – Le code génétique "standard". En orange, les acides aminés non polaires ; en rouge, les acides ; en violet, les polaires non chargés et en bleu, les basiques. Le codon AUG code pour la méthionine (Met) est le codon d'initiation de la traduction (en vert). La structure chimique commune des 20 acides aminés est représentée au-dessus, seule la chaîne latérale notée *R* est variable. Les chaînes latérales spécifiques des acides aminés est présentée à coté de chacun d'eux, tiré de Biochimie de Voet et Voet (2005)

le code génétique la comparaison d'une séquence nucléotidique à la séquence protéique correspondante suffirait mais à l'époque cette technique n'existait pas encore. Nirenberg & Matthaei (1961) ont montré que le codon UUU spécifie la phénylalanine, Phe (sur les ARN l'uracile, U, remplace la thymine T de la séquence d'ADN correspondante) : pour cela, ils ont étudié un système de synthèse de protéine, une séquence d'ARN poly(U) et un mélange de

20 AA dont un seul radioactif. C'est seulement en présence de Phe radioactive, que la protéine synthétisée est radioactive. Les expériences semblables ont pu être conduites pour les codons AAA et CCC (les séquences poly(G) ne sont pas stables, se compactent et précipitent). Pour décrypter les codons restants, Nirenberg & Leder (1964); Ochoa & Weinstein (1964) ont étudié des di-nucléotides tels que poly(UG) avec des fractions différentes de U et de G. Par exemple, avec 76% de U et 24% de G, on s'attend à avoir $0.76^3 = 44\%$ de UUU donc de Phe, $3 * 0.76^2 * 0.24 = 42\%$ de codons avec 2 U et 1G, $3 * 0.76 * 0.24^2 = 12\%$ avec 1U et 2G et enfin, $0.24^3 = 2\%$ de codons avec 3G. Ils ont aussi utilisé des polymères contenant 2, 3 et 4 bases pour finir de déchiffrer ce code. Ces expériences ont par ailleurs prouvé que le code est dégénéré car les expériences avec poly(UA), poly(UC) et poly(UG) permettent tous la synthèse de protéine avec de la leucine (Leu). Nirenberg & Leder (1964) ont étudié les liaisons d'ARNt aux différents codons :l'$ARNt^{Phe}$ se lie à UUU. Les poly(UG) autorisent la liaison d'ARNt spécifiant la leucine, la cystéine et la valine. Cette expérience a permis d'identifier à peu près 50 codons. Nishimura et al. (1964) décryptent la fin du code génétique grâce à la synthèse d'ARN de séquences connues. La séquence $UCUCUCUCUCUC...$ permet de synthétiser $Ser-Leu-Ser-Leu....$ Avec les résultats précédents, il a montré que la serine est codée par UCU et la leucine par CUC. Ces résultats confirment aussi que les codons sont formés par un nombre impair de nucléotides (il n'y a que 2 AA) donc que les codons sont de taille 3. Enfin, les tri-nucléotides tels que poly(UAC) permettent de synthétiser poly(Tyrosine), poly(Thréonine) et poly(Leucine) à cause des 3 cadres de lecture. Les codons STOP qui spécifient l'arrêt de traduction ont été découvert car les polypeptides synthétisés étaient trop courts. Ces codons STOP sont UAA, UAG et UGA (figure 1.2). A noter que pour déchiffrer le code, le codon d'initiation de la traduction AUG n'est pas indispensable car la traduction peut être initiée dans un milieu avec une forte concentration de Mg^{2+} (ce qui était le cas dans les expériences ci-dessus). Enfin, dernière précision, les expériences précédentes permettent aussi de faire correspondre l'extrémité 5' de l'ARN à l'extrémité N-terminale du polypeptide spécifiant le sens de lecture (Nirenberg & Leder, 1964).

Évidemment, le prix Nobel de médecine a été attribué a Nirenberg, Khorona et Holley en 1968 pour leur contribution à l'interprétation du code génétique et à la synthèse de protéines.

1.2 Les propriétés physiques et chimiques du code

La structure du code

Universel Tous les êtres vivants possèdent le même code à de rares exceptions près. Par exemple, les mitochondries de mammifères ont AUG et AUA comme codons d'initiation, et UGA spécifie le tryptophane plutôt que STOP. Cette universalité prouve que tous les êtres vivants partagent une histoire, un ancêtre (LUCA - Last Common Unique Ancestor) qui utilisait déjà le code génétique pour produire des protéines à partir de l'information génétique. Cette universalité permet aujourd'hui de synthétiser des protéines humaines ou de vaccins chez d'autres organismes. Par exemple, l'insuline qui n'est pas synthétisée chez les diabétiques, l'a été dans *E. coli* puis extraite et injectée chez les sujets diabétiques à des fins thérapeutiques.

Univoque Si chaque acide aminé peut être codé par plusieurs codons l'inverse n'est pas vrai, les séquences d'ADN ne sont pas ambiguës.

Dégénéré La redondance du code génétique n'est pas aléatoire et l'on voit figure 1.2 que la 3ème lettre d'un codon est la plus redondante. Les codons qui spécifient un même acide aminé sont des codons dits synonymes. Par exemple, la position 3 des codons synonymes des acides aminés 4-fois dégénérés (valine Val, alanine Ala, glycine Gly, proline Pro, thréonine Thr) peut être l'une des 4 bases. Dans la suite de ce manuscrit, je note d_A la dégénérescence de l'acide aminé A.

Structuré Certains AA sont codés de manière unique comme la méthionine (AUG) ou le tryptophane (UGG) alors que la leucine, la serine ou l'arginine sont 6-fois dégénérées. Cela n'est pas aléatoire dans le sens où les acides aminés 6-fois dégénérés sont plus fréquemment utilisés dans les protéines. De plus, le tryptophane qui est le seul acide-aminé avec un cycle (donc un fort encombrement stérique) est proche en terme de séquence des codons STOP. Par exemple, Freeland & Hurst (1998) ont montré que le code génétique n'est pas aléatoire et que sa structure est optimale pour réduire les effets des mutations.

Distances entre acides aminés

D'après la figure 1.2 on observe que les acides aminés ayant des propriétés physico-chimiques proches sont codés par des codons similaires. L'exemple le plus frappant est celui de l'asparagine (Asp) et de la glutamine (Glu), les deux seuls acides aminés acides du vivant. Ils sont tous deux 2-fois dégénérés et codés par GAN (N étant n'importe lequel des 4 nucléotides).

Ces acides aminés sont donc proches chimiquement mais aussi séquentiellement. De manière générale, ces observations intriguent, et certains chercheurs tentent de comprendre comment le code génétique s'est mis en place au cours de l'évolution (Higgs, 2009; Francis, 2013; Massey, 2015; Carlevaro et al., 2016). Ces dégénérescences concordent avec l'effet wobble (flottement) proposé par Crick : les ARNt qui décodent une séquence nucléique respectent les règles d'appariement de Watson-Crick (A :T et C :G) pour les deux premières positions alors que l'appariement en 3ème position est imparfait (Crick, 1966).
Les ARNt sont en effet moins nombreux (par exemple, 48 chez l'Homme, voir partie IV) et reconnaissent pourtant les 61 codons sens (64 - 3 codons STOP) : un ARNt peut reconnaître plusieurs codons synonymes. Par exemple, *Escherichia coli* possède 3 ARNt différents pour traduire les 6 codons synonymes codant pour l'Arginine. La table 1.1 représente les codons synonymes dans les ARNm mais aussi les ARNt qui les reconnaissent chez *E. coli*.

codon (ARNm)	anti-codon (ARNt)
CGU CGC CGA	GCG
CGG	GCC
AGA AGG	UCU

TABLEAU 1.1 – Reconnaissance de plusieurs codons synonymes par les ARNt chez *E. coli* via l'effet wobble. A gauche, les codons codant pour l'Arginine des ARNm. A droite, les anticodons des ARNt qui reconnaissent les codons de gauche.

Dans cette thèse je m'intéresse aux séquences de codons et à la distance entre deux acides aminés. Il existe plusieurs manières de définir cette distance : (1) la distance génétique est le nombre minimal de mutations ponctuelles nécessaires pour passer d'un acide aminé à un autre AA, elle est utilisée dans la table d'AA CPM (Codons Probability Mutations, (Thorvaldsen, 2015)) ; (2) la distance physico-chimique entre deux AA, ou (3) la distance phylogénétique. Cette dernière est définie comme le taux de changement d'un AA vers un autre AA observé dans un jeu de séquences prédéfini ; elle est utilisée dans les tables d'AA telles que les matrices PAM (point accepted mutation, Dayhoff et al. (1978)) ou BLOSUM (block substitution matrix, Henikoff & Henikoff (1992)). Nous venons de voir que le code génétique est universel, univoque, dégénéré et structuré, et il est intéressant d'observer que l'usage de ce code génétique

chez les êtres vivants est biaisé et non-uniforme.

CHAPITRE 2

Le biais d'usage du code génétique

L'utilisation des codons synonymes au sein d'un organisme n'est pas aléatoire. Il existe des préférences de certains codons par rapport à d'autres, c'est le biais d'usage du code (BUC). Dans une première partie, je présente l'historique de mise en évidence du BUC et les observations qui en ont été faites. Dans un second temps je décris les différentes hypothèses avancées quant à l'origine de ce BUC au sein d'un génome. Puis, je présente les mesures ou statistiques qui permettent de calculer ce BUC et enfin les limites de ces mesures pour étudier l'évolution du BUC entre les organismes.

2.1 Définition et mise en évidence

L'usage du code génétique est la fréquence d'utilisation des codons synonymes. Ces derniers ne sont pas employés à la même fréquence soit entre espèces soit au sein d'une espèce, c'est le biais d'usage du code génétique (BUC). Fitch (1976) étudie la fréquence d'utilisation de codons synonymes chez le phage MS2 et son lien avec la composition cellulaire en ARNt chez l'hôte du

Chapitre écrit d'après le chapitre "Codon Usage Bias" du livre Codon evolution, edité par Cannarozzi et Schneider, 2008

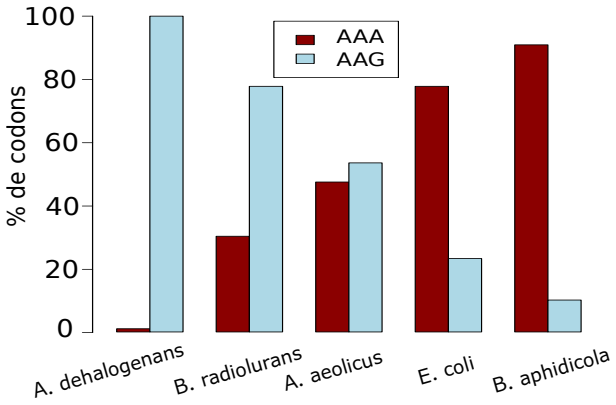
phage (i.e. *E. coli*). D'une part, chez *E. coli* les ARNt capables de traduire les codons synonymes terminant par des C ou T, ont en première position de l'anticodon un G. Ces ARNt peuvent reconnaître ainsi plusieurs codons synonymes. D'autre part, chez le phage, parmi les codons synonymes terminant par C ou T, ceux terminant par C sont plus fréquents (voir la table 2.1). Cette observation introduit la notion d'un biais d'usage du code génétique.

Codons synonymes Position 1 et 2	3ème position –C	3ème position –U	Acide aminé correspondant
AA-	28	17	Asn
AG-	16	8	Ser
CA-	9	6	His
GA-	22	28	Asp
TA-	32	9	Tyr
TG-	6	6	Cys
TT-	29	19	Phe
Total	142	93	

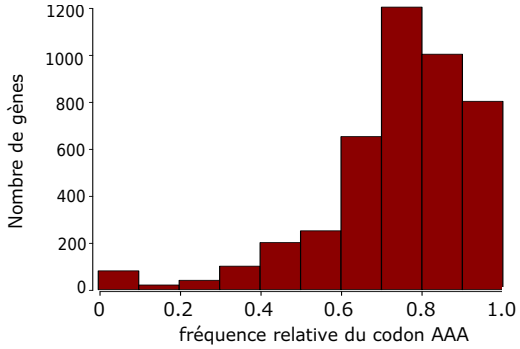
TABLEAU 2.1 – Distribution des codons synonymes terminant par C ou T dans le phage MS2. La première colonne correspond aux 2 premières positions des codons. La deuxième colonne correspond au nombre de fois que sont utilisés les différents codons synonymes qui se terminent par C ou T. La dernière colonne correspond à l'acide aminé encodé par ces codons. Table reproduite à partir de l'article de Fitch (1976)

Par la suite, de nombreuses études ont montré que ce biais n'est pas universel (à l'inverse du code génétique ou de la relative conservation de la machinerie de traduction dans le vivant) et qu'il dépend entre autre de l'espèce considérée. La figure 2.1a) montre la différence d'utilisation des codons synonymes de la lysine parmi plusieurs bactéries. Cette différence d'usage peut être extrêmement importante, allant pour le codon AAA de 1% chez *Anaeromyxobacter dehalogenans* à 91% chez *Buchnera aphidicola*. Il est intéressant de noter que ces différences de biais s'observent aussi au sein d'un génome. La figure 2.1b) montre le biais d'utilisation du codon AAA de la lysine au sein des gènes d'*Escherichia coli*.

La figure 2.1 montre un biais d'usage chez les bactéries mais il existe dans l'ensemble du domaine du vivant. Les différentes hypothèses sur l'origine du BUC sont détaillées ci-après.



(a) Chez différentes espèces bactériennes.



(b) Chez *Escherichia coli*.

FIGURE 2.1 – Variation du biais d'usage des codons de la lysine chez différentes bactéries ou au sein d'une bactérie, d'après la thèse Bailly-Bechet (2007)

2.2 Origines du biais d'usage du code

Qualitativement, les mécanismes précis expliquant le biais d'usage du code sont connus. Il existe différentes explications qui sont classées en deux grandes catégories concernant les processus mutationnels à l'origine de ce biais : (1) les processus non sélectifs et (2) ceux résultant de la sélection. Ce n'est pas tant l'existence de ces deux catégories qui est débattue mais c'est la contribution quantitative de chacun de ces deux mécanismes. La vision adaptative privilégie l'importance de la sélection : un codon synonyme est préféré à un autre car cela permet une augmentation de la fitness (capacité de transmettre les gènes à la génération suivante) au niveau traductionnel. La vision non-adaptative estime que les explications neutres sont prédominantes : le BUC résulte d'un biais mutationnel ou d'un biais de conversion génique (BGC, voir la partie III chapitre 5). Ces deux forces agissent sur le BUC (Bulmer, 1991; Sharp et al.,

1993; Akashi, 1994; Akashi & Eyre-Walker, 1998; Rocha, 2004) et la question est de quantifier le rôle de ces deux processus sur le BUC. En effet, ces processus ne s'opposent pas (Daubin & Perrière, 2003) et la variation du BUC peut se faire à deux échelles : en intra-spécifique, le BUC serait adaptatif alors qu'à l'échelle inter-spécifique, il serait dirigé par des processus non-adaptatifs.

La sélection traductionnelle

Parler de sélection traductionnelle impose de définir la notion d'expression de gènes. Sachant qu'un gène est transcrit en plusieurs ARNm et que chaque ARNm est traduit en plusieurs copies de protéines, l'expression d'un gène s'étudie soit en estimant la quantité d'ARNm cellulaire soit en regardant le nombre de protéines synthétisées. Par la suite, sauf mention contraire, l'expression d'un gène renvoie au nombre de protéines synthétisées dans la cellule.

Historiquement, certains auteurs ont montré qu'il existe un ensemble de codons pour lesquels le taux d'erreur lors de la traduction ou les erreurs de repliement sont minimisés, ce sont les codons synonymes optimaux (Sharp et al., 1986; Duret & Mouchiroud, 1999; Behura & Severson, 2011) qui sont différents entre les espèces. L'hypothèse génomique ("the genome hypothesis" en anglais) défini par Grantham et al. (1980) suppose que la sélection s'effectue au sein d'un génome et non pas au niveau du gène.

Trois prédictions résultent de la sélection traductionnelle : (1) les gènes fortement exprimés (en nombre de protéines) montrent un BUC plus important que les autres gènes ; (2) le contenu en ARNt de la cellule est corrélé au BUC de ces gènes (Duret & Mouchiroud, 1999; Behura & Severson, 2011), voir la figure 2.2 et (3) le taux de changement silencieux en position 3 des codons synonymes est plus faible pour les gènes fortement exprimés. Ces prédictions ont d'abord été observées chez les bactéries (Ikemura, 1985; Sharp & Li, 1987) où les gènes fortement exprimés ont un usage du code positivement corrélé avec le contenu en ARNt codant pour ces codons (Ikemura, 1981a; Akashi, 1994). Les différences d'usage des codons synonymes n'ont pas de conséquence sur la séquence protéique mais elles peuvent impacter l'efficacité ou la précision de la traduction d'un ARNm en protéine. La sélection sur l'usage des codons synonymes est donc spécifique des gènes voire des codons où l'optimisation traductionnelle est la plus importante (Novoa & Ribas de Pouplana, 2012).

Ikemura (1985) a montré que la disponibilité des ARNt dans la cellule est une contrainte très importante pour traduire les ARNm. Les codons optimaux sont définis comme étant (1) les

FIGURE 2.2 – Lien entre la concentration en ARNt d'une cellule et la fréquence d'usage des codons associés à ces ARNt (Dong et al., 1996).

plus fréquents dans les gènes fortement exprimés (Sharp & Li, 1987) ou (2) les plus fréquents dans les gènes de protéines ribosomales (protéines dont on sait qu'elles sont fortement exprimées). Les codons les plus fréquents dans l'ensemble des gènes connus d'une espèce (Carbone et al., 2003) sont appelés codons préférés.
La figure 2.3 extraite de la revue de Novoa & Ribas de Pouplana (2012) montre comment la sélection peut agir sur l'usage du code. La distribution des codons optimaux le long d'un gène n'est pas aléatoire, ce qui affecte la vitesse de progression du ribosome et donc la traduction (Marquez et al., 2005). Par exemple, dans un gène les codons synonymes les plus utilisés ont tendance à être reconnus par les mêmes molécules d'ARNt, ce qui limiterait le coût en ARNt (un ARNt pouvant être recyclé) (Cannarrozzi et al., 2010). De plus, les codons synonymes "non-préférés" sont plus fréquents au début des gènes, limitant la vitesse de traduction, assurant ainsi une diminution du taux d'erreur et un bon repliement (Stoletzki & Eyre-Walker, 2007). Enfin, les gènes d'une même catégorie fonctionnelle présentent un usage du code plus homogène entre eux qu'avec les autres gènes (Gingold et al., 2014). Le BUC peut être spécifique des stades cellulaires voire des cycles circadiens chez les cyanobactéries (Xu et al., 2013) et il serait lié au contenu en ARNt de ces stades (voir la partie IV pour plus de détails sur cette hypothèse chez l'Homme).

FIGURE 2.3 – Les hypothèses principales des mécanismes responsables d'un usage du code biaisé et leurs effets sur l'efficacité de la traduction, extrait de Novoa & Ribas de Pouplana (2012). (a) Effet sur l'efficacité des ribosomes (b) Le BUC serait corrélé au contenu en ARNt et à leurs spécificités de reconnaissance des codons synonymes liées elles-mêmes aux enzymes de modifications post-transcriptionnelles (uridine méthyl-transférases UMs chez les bactéries et adénosines déaminases ADAT chez les eucaryotes). (c) Le BUC est lié au moment où les gènes sont exprimés.

D'autres hypothèses (notamment avancées chez les bactéries) suggèrent que le BUC est associé à l'environnement et aux traits d'histoire de vie (Botzman & Margalit, 2011). Les explications sélectives ne sont pas suffisantes pour expliquer l'existence chez certaines espèces d'un biais d'usage des codons synonymes lié ni au contenu en ARNt ni au niveau d'expression des gènes.

Les processus non-adaptatifs

Les processus non-adaptatifs peuvent également expliquer le biais d'utilisation des codons synonymes. En effet, chez les vertébrés et certaines plantes le BUC d'un gène est corrélé à la composition en GC de la région où se trouve le gène. Chez ces espèces, Bernardi et al. (1985) a montré que les génomes sont des mosaïques relativement homogènes en GC : ce sont les isochores (voir la partie III, chapitre 5 pour connaître l'organisation des génomes eucaryotes). Les gènes contenus dans les isochores sont affectés par le fond nucléotidique et ainsi leur usage du code l'est aussi. Ainsi, un gène situé dans une isochore GC-riche présentera un biais d'usage du code vers les codons synonymes se terminant par C ou G (voir la partie IV sur les gènes humains). Les explications non-adaptatives suggèrent que le BUC est soit (1) une résultante d'un biais mutationnel globalement neutre, fixé par dérive génétique (Marais & Duret, 2001) soit (2) une résultante d'un biais de fixation des mutations. Les mutations sont neutres lorsqu'elles n'impactent pas la capacité d'un individu à transmettre cette mutation aux générations suivantes. La fixation de ces mutations neutre est donc aléatoire, c'est la dérive génétique. Cette dérive est d'autant plus visible que la population étudiée est de petite taille. En effet, dans le cas d'une large population, la fixation d'une mutation est plus longue que dans une petite population, d'autant plus si la mutation ne confère pas d'avantage. La dérive génétique dans les petites populations permet aussi à certaines mutations faiblement délétères de s'y fixer. Par ailleurs, certaines mutations sont favorisées par rapport à d'autres : les transitions ($A \leftrightarrow G$ ou $C \leftrightarrow T$) sont plus fréquentes que les transversions car il n'y a pas de changement de nature chimique des nucléotides (A et G sont des purines et C et T des pyrimidines). De plus, Hershberg & Petrov (2010) ont aussi montré qu'il existe un biais universel vers AT chez les procaryotes qui impacte leur BUC : les mutations de G vers A ou de C vers T sont plus fréquentes que les autres mutations et notamment plus fréquentes que A vers G ou T vers C (figure 2.4).

FIGURE 2.4 – Le biais mutationnel chez quelques bactéries aux positions synonymes, d'après Hershberg & Petrov (2010)

Le biais de fixation des mutations ou biais de conversion génique (BGC) est un biais de fixation des allèles lors de la recombinaison. On parle de gBGC lorsque ce biais se fait vers les allèles GC-riches. Chez les Primates, Galtier et al. (2009) montrent que les mésappariements au niveau des cassures double brin de la recombinaison méiotique sont réparés de manière biaisée vers G ou C. Ce mécanisme sera détaillé ultérieurement dans la partie III, chapitre 5. Plus

généralement, Clay & Bernardi (2011) montrent que les codons 4-fois dégénérés présentent un choix des bases en position 3 similaire à la composition en bases des régions non-codantes. Ainsi, la composition en GC génomique est le facteur principal de la variation du BUC et plus précisément du biais observé en troisième position des codons (voir Duret (2002) pour une revue). Il semblerait que cela soit le cas chez les eucaryotes.

Dans la partie suivante, je présente les différentes méthodes qui permettent de mesurer le BUC. Nous verrons que les méthodes dépendent des hypothèses explicatives testées.

2.3 Comment mesurer le biais d'usage ?

Le biais d'usage du code est un terme unique qui définit des réalités disparates : le code génétique étant composé de 61 codons pour 20 acides aminés, certaines catégories de codons peuvent être biaisées et pas d'autres. La mesure du BUC s'emploie à résumer, à l'aide d'une statistique simple, un ensemble de processus complexes. Nous venons de voir qu'il existe deux grandes catégories d'hypothèses qui expliquent l'existence d'un tel biais (biais de composition ou sélection traductionnelle). De même, il existe deux principales catégories d'étude du BUC. Un biais, au sens général, peut se mesurer comme une déviance à une utilisation aléatoire ou bien, à une utilisation optimale. Si on se place dans une hypothèse de sélection traductionnelle, on s'intéresse à l'utilisation "optimale" et à l'inverse, dans le cas d'un biais de composition, on se focalisera sur l'utilisation aléatoire. Cette section permet de classer et comparer les principales méthodes qui permettent de mesurer le biais d'usage du code (BUC). Quelque soit la mesure du BUC choisie il faut dans un premier temps répertorier le nombre de codons observés c'est-à-dire calculer la table de contingence des codons présents. À noter que ces mesures sont sensibles à la taille des gènes étudiés : en effet, si ces derniers sont trop courts (nous caractériserons ce terme ultérieurement), on ne s'attend pas à observer chaque type de codon. On préférera étudier des gènes de longueur au moins égale à 80-100 codons ou bien des concaténats de gènes. Généralement, la table de contingence est divisée par la longueur du gène pour n'avoir plus que des fréquences d'utilisation des codons, notée par la suite f, et s'affranchir de l'effet de longueur. On a :

$$f_I = \frac{C_I}{L} \tag{2.1}$$

où C_I est le nombre de codons I observés et L la longueur totale de la séquence en codons. Cette normalisation a néanmoins le défaut de favoriser les acides aminés fréquents. On peut normaliser autrement, en calculant la fréquence relative des codons (notée p) au sein de chaque acide aminé. Dans le cas d'un codon I codant pour l'acide aminé A, on a :

$$p_I = \frac{C_I}{\sum_{k \in cod(A)} C_k} \tag{2.2}$$

avec $cod(A)$ les différents codons synonymes de A et C_k le nombre de codons k observés.

Les mesures sans référence

Certaines mesures du BUC n'ont pas besoin d'un ensemble de codons de référence, ce qui les rend facilement utilisables lorsqu'on n'a pas d'*a priori* sur les séquences étudiées. Dans ce cas, on mesure la déviance d'usage des codons à un usage uniforme, où il n'y a pas de préférence de codons. Autrement dit, l'hypothèse nulle considère que les codons synonymes pour un AA donné s'utilisent à la même fréquence. Je vais présenter en détail quatre mesures, largement employées dans la communauté scientifique et que j'utilise par la suite : le GC3, le RSCU (utilisé dans partie IV), l'ENC et l'ENC' (voir partie II).

GC3 Le GC3 mesure le taux de $G+C$ au niveau de la troisième position des codons. En effet, j'ai présenté dans le paragraphe 1.2 le fait que la 3ème position étant la plus redondante d'après le code génétique standard (figure 1.2), cette position est un bon indicateur du BUC. Le GC3 s'exprime en % de GC. Il a été montré que le GC3 est un bon indicateur du GC régional, car la 3ème position étant plus redondante, elle est moins soumise à la sélection (Mouchiroud et al., 1991; Eyre-Walker & Hurst, 2001). Nous verrons ceci plus en détails dans la partie IV chez l'Homme.

RSCU Le Relative Synonymous Codon Usage a été proposé par Sharp et al. (1986) et calcule l'usage relatif des codons par rapport à un usage uniforme. Pour un codon I, il est défini comme :

$$r_I = \frac{C_I}{\frac{1}{d_A}\sum_{k \in cod(A)} C_k} \tag{2.3}$$

$$r_I = \frac{p_I}{p_{I_attendu}} \tag{2.4}$$

avec d_A la dégénérescence de l'acide aminé A et $p_{I_attendue}$ le nombre de codons I attendu si l'usage des codons synonymes est aléatoire. Le RSCU qu'on note r se calcule par codon. Si $r = 1$ alors le codon est utilisé comme attendu. Si $r < 1$ (respectivement, $r > 1$), il est moins (resp. plus) utilisé. Au final, la somme des r des codons synonymes traduit la dégénérescence de l'acide aminé codé. Autrement dit, prenons l'exemple de la lysine codée par AAA et AAG. Sans biais, on s'attend à une fréquence de 0,5 pour ces deux codons. Si on observe une fréquence de 0,25 pour AAA et 0,75 de AAG alors $r_{AAA} = \frac{0,25}{0,5} = 0,5 < 1$ et $r_{AAG} = \frac{0,75}{0,5} = 1,5 > 1$ donc AAG est plus utilisé qu'attendu et AAA moins.

ENC L'Effective Number of Codons (Wright, 1990) est une mesure qui calcule l'usage effectif des codons théoriquement disponibles. La formule de l'ENC est détaillée ici selon deux formules synonymes, l'équation 2.5 est très générique alors que l'équation 2.6 détaille les catégories de codons en fonction de leur dégénérescence :

$$ENC = \sum_d \frac{k_d^2}{\sum_{A \in d} \frac{1}{n_A - 1}\left(\left(n_A \sum_{I \in cod(A)} p_I^2\right) - 1\right)} \tag{2.5}$$

$$ENC = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6} \tag{2.6}$$

avec $d \in [1;2;3;4;6]$ la classe de dégénérescence des AA, k_d le nombre d'AA d'une telle classe d, n_A le nombre de codons observés pour un acide aminé A, $cod(A)$ les différents codons synonymes de A, p_I la fréquence relative du codon I parmi les codons synonymes. Les F_d représentent le nombre effectif des codons des AA de dégénérescence d. Ces formules se rapprochent du calcul d'un χ^2 et tiennent compte de la structure du

code génétique. L'ENC varie de 20 (biais maximal) à 61 (pas de biais). Lorsqu'il vaut 20, seul un codon synonyme est utilisé par acide aminé et lorsqu'il est égal à 61, cela veut dire que les 61 codons sens sont uniformément employés. En général, l'ENC calculé est intermédiaire, ce qui rend compte d'un usage préférentiel de certains codons synonymes par rapport aux autres. Sans biais, on a $F_2 = 0.5$, $F_3 = 0.33$, $F_4 = 0.25$ et $F_6 = 0.1666$. On en déduit $ENC = 2 + \frac{9}{0.5} + \frac{1}{0.33} + \frac{5}{0.25} + \frac{3}{0.1666} = 2 + 18 + 3 + 20 + 18 = 61$.

ENC' est une mesure de l'ENC adaptée par Novembre (2002). L'ENC' est un peu différente des deux précédentes mesures car elle considère comme hypothèse nulle un usage uniforme des codons uniquement expliqué par le contenu nucléotidique du gène ou de la région :

$$ENC' = \sum_{R \in ARC} \frac{n_R^2}{\sum_{A \in R} \frac{1}{d_A(n_A - 1)} \left(\left(n_A \sum_{I \in cod(A)} \frac{\left(p_I^{obs} - p_I^{att}\right)^2}{p_I^{att}} \right) + n_A - d_A \right)} \tag{2.7}$$

avec p_I^{att} la fréquence attendue du codon I connaissant la composition nucléotidique biaisée du génome et p_I^{obs} la fréquence du codon I dans la séquence. L'ENC' varie aussi entre 20 (un seul codon par AA) et 61 (utilisation des codons due à la composition en nucléotides). Ainsi, si on étudie le BUC d'un génome riche en AT, le gène aura tendance à utiliser préférentiellement les codons terminant par A ou T sans avoir besoin d'invoquer la sélection.

Les acides aminés codés par un seul codon (Trp ou Met pour le code standard) ne sont pas informatifs dans ces mesures. Les codons stop (3 dans le code standard) ne sont pas pris en compte dans l'étude du BUC car ils sont de fait rares et souvent biaisés vers un codon particulier.

Les mesures relatives à une référence

De nombreuses mesures comparent les codons utilisés dans une séquence à ceux d'un ensemble de séquences particulières. En effet, Ikemura (1985) a montré que les gènes fortement exprimés sont soumis à une forte sélection traductionnelle et ont un usage du code très biaisé. Une hypothèse très largement répandue estime que ce biais est optimisé pour améliorer l'efficacité de traduction en augmentant la vitesse ou en diminuant les erreurs, voir la revue Duret (2002). Les mesures qui dépendent d'une référence sont dépendantes d'un jeu de codons particulier : les codons optimaux (définis dans le paragraphe 2.2. Je présente ici les statistiques Fop et CAI :

Fop La Fréquence des codons OPtimaux (Ikemura, 1981a) est le rapport des codons optimaux utilisés dans la séquence sur le nombre total de codons synonymes utilisés.

$$Fop = \sum_{opt \in cod_{opt}} \frac{C_{opt}}{L} \tag{2.8}$$

où cod_{opt} est la liste des codons optimaux et L la longueur de la séquence en codons. Le Fop vaut 1 au maximum, lorsque la séquence n'utilise que les codons optimaux et 0 lorsqu'aucun des codon optimaux n'est utilisé.

CAI Le Codon Adaptation Index est la méthode la plus utilisée pour calculer le BUC (Sharp & Li, 1987). Le CAI s'intéresse à l'adaptivité relative des codons. Dans le paragraphe précédent, nous avons parlé de deux façons de normaliser les occurrences de codons, l'adaptivité relative en est une troisième. Notée w_I, c'est le nombre d'observations d'un codon I par rapport au nombre d'observations du codon synonyme le plus fréquent/optimal :

$$w_I = \frac{C_I}{C_{opt}} \tag{2.9}$$

Ici, les codons optimaux sont les codons les plus fréquents dans une séquence fortement exprimée. L'adaptivité relative vaut donc 1 pour le codon le plus fréquent/l'optimal. Le CAI est alors une moyenne des w des codons d'une séquence de L codons :

$$CAI = \Big(\prod_{i \in [1:L]} w_i \Big)^{\frac{1}{L}} \tag{2.10}$$

Ces deux mesures sont reliées entre elles, du fait de leur définition. Pour certaines espèces, telles que *Escherichia coli*, il est intéressant de noter que le CAI ou Fop est corrélé avec le niveau d'expression de protéines (von der Haar, 2008). Nous verrons ce point plus en détail dans le paragraphe 2.3.

Autres méthodes

Il existe une quantité importante de méthodes pour mesurer le BUC. J'ai choisi de vous présenter en détail les méthodes les plus utilisées et courantes. Néanmoins, il est intéressant de voir que la mesure du biais est un champ disciplinaire en tant que tel et qu'il existe un nombre

important de statistiques. Je liste quelques unes d'entre elles dans le tableau 2.2.

Abbréviation	Nom et courte définition	Publication
Ew	Weighted sum of relative Entropy : considère le biais de composition nucléotidique mais aussi en acides aminés (mesure basée sur la théorie de l'information)	Suzuki et al. (2004)
MCB	Maximum likelihood Codon Bias : permet de tester différentes hypothèses nulles, comme un biais de composition nucléotidique ou di-nucléotidique	Urrutia & Hurst (2001)
P	codon Preference : vraisemblance d'un ensemble de codons selon un ensemble de référence prédéterminé	Gribskov et al. (1984)
P1	mesure l'influence de la disponibilité en ARNt sur le taux d'erreur de traduction d'un gène	Gouy & Gautier (1982)
RCB	Relative Codon usage Bias : mesure la contribution relative des codons, connaissant la composition nucléotidique aux 3 positions d'un codon	Roymondal & Sahoo (2009)
SEMPPR	Stochastic Evolutionary Model of Protein Production Rate : relie le coût de production d'un gène à son usage du code.	Gilchrist (2007)
tAI	tRNA Adaptation Index : se base sur l'hypothèse que la disponibilité en ARNt est une force motrice de la sélection traductionnelle	dos Reis et al. (2003, 2004)

TABLEAU 2.2 – Table non-exhaustive de différentes méthodes de mesures du BUC. Ew, MCB, RCB mesurent le BUC selon une hypothèse nulle prédéfinie ; P se base sur un ensemble de codons prédéfini ; P1 et tAI, prennent en compte le besoin d'ARNt pour traduire les séquences et SEMPPR considère le BUC comme résultant des propriétés intrinsèques du code génétique et du coût de production car les séquences observées sont dépendantes des effets de la mutation, de la dérive et de la sélection.

Utilisation de ces mesures

Ces statistiques sont largement répandues dans l'analyse de nouvelles séquences. En effet, on associe souvent un BUC moyen par organisme, avec un BUC plus accentué pour les gènes les plus exprimés. Historiquement, ces mesures ont été développées pour trouver le cadre de

lecture des ORFs (Open Reading Frame) et la détection des décalages de cadre de lecture (Gribskov et al., 1984; Ghaemmaghami et al., 2003). Par ailleurs avec la mise en évidence de transferts horizontaux (TH) de gènes entre espèces, les mesures de BUC ont permis d'identifier ces gènes soumis au TH. Si un gène a récemment été transféré d'une espèce avec un certain usage du code vers une autre, il a un BUC très différent du reste du génome de l'espèce receveuse (Carbone et al., 2003; Sugaya et al., 2004; Bodilis & Barray, 2006). Enfin, chez les procaryotes, l'étude du BUC permet de prédire l'expression de gènes (Comeron & Aguadé, 1998; Supek & Vlahovicek, 2005; Suzuki et al., 2008). Les corrélations entre le BUC et le taux d'expression du gène (en ARNm ou en nombre de protéines) ne sont pas parfaites puisqu'un gène est transcrit en plusieurs ARNm et parce que chaque ARNm est lui-même traduit en plusieurs protéines. L'étude de Belle et al. (2006) est une des rares études des demi-vies de protéines et de l'efficacité de traduction.

Limites de ces mesures

Les mesures du BUC sont dépendantes de plusieurs caractéristiques du génome que je liste ici. Les mesures du BUC prennent en compte l'une ou l'autre des limites. Elles dépendent de :

1. la composition nucléotidique. Si une espèce est très riche en AT alors il est difficile, par hasard, d'avoir un codon riche en C ou G. L'ENC' pallie cette limite.

2. la longueur de la séquence étudiée. Si le gène étudié est court, par exemple 50 codons, alors de nombreux codons ne seront pas présents simplement en raison d'un tirage trop faible. Les statistiques présentées précédemment sont consistantes si le nombre attendu de chacun des codons est au moins de 1. Une méthode généralement utilisée est d'étudier le BUC d'un ensemble de gènes concaténés ou d'ajouter un pseudo-compte (pc) ($C_{c_{pc}} = C_c + 1$).

3. la dégénérescence du codon. On a déjà abordé ce point lors des différentes méthodes de normalisation des comptes de codons. En effet, la leucine par exemple a 6 codons synonymes alors que le tryptophane n'a qu'un seul codon. Si les codons étaient utilisés de manière entièrement uniforme (quelque soit l'AA produit), la leucine serait 6 fois plus fréquente que le tryptophane. Le RSCU et l'ENC permettent de comparer l'utilisation des codons selon leur dégénérescence.

4. la composition en AA. A l'inverse du point précédent, si par exemple le tryptophane est un AA très abondant dans un gène par rapport à la leucine, son codon sera enrichi par rapport aux 6 codons de la leucine. L'ENC sépare les codons selon la dégénérescence du code.

Le BUC dont nous venons de parler est observable et facilement mesurable dans les séquences actuelles. Nous avons vu aussi qu'il est spécifique aux espèces et qu'ainsi, l'étude du BUC peut permettre d'identifier des gènes soumis aux transferts horizontaux (Carbone et al., 2003; Sugaya et al., 2004; Bodilis & Barray, 2006). Or, c'est à partir de cette simple observation, que nous nous sommes intéressés à son évolution. Le BUC est une composante espèce spécifique et les espèces ont évolué à partir d'ancêtres communs le long d'un arbre phylogénétique (ou d'un buisson (Rokas & Carroll, 2006; Doolittle, 1999), mais ce n'est pas le propos). Ainsi, le BUC évolue lui aussi le long de l'arbre des espèces et la comparaison simple des BUC observés entre différentes espèces ne prend pas en compte les relations phylogénétiques qui peuvent exister entre espèces. De fait, c'est tout l'enjeu de ce manuscrit : proposer une méthode d'étude non pas du BUC mais de son évolution. Dans le chapitre suivant, je vais tout d'abord vous présenter les différents modèles d'évolution de codons qui existent. Généralement, les modèles de codons s'appliquent à distinguer les substitutions synonymes (changement d'un codon synonyme vers un autre) des substitutions non-synonymes (changement d'acide aminé) mais ils ne considèrent pas la structure du code génétique.

CHAPITRE 3

Les modèles évolutifs de séquences

L'évolution moléculaire est l'étude du changement des molécules biologiques au cours de l'Histoire du vivant. Un tel changement est un processus stochastique (évolution d'une variable aléatoire de façon discrète ou continue en temps) qui a lieu dans une espèce ou une population. Les évolutionnistes cherchent à répondre à la question : comment en est-on arrivé là ? Dans cette thèse, je me place dans un contexte phylogénétique selon lequel les espèces évoluent le long d'un arbre phylogénétique au sein duquel les feuilles sont les espèces actuelles et les noeuds sont les ancêtres communs hypothétiques. Au cours de ma thèse je m'intéresse à l'évolution de séquences de codons le long d'un arbre phylogénétique connu. J'étudie en particulier les mutations ponctuelles et ne considère ni les délétions ni les insertions.

Modéliser l'évolution

Je présente ici une partie seulement des modèles d'évolution, les modèles de substitutions. Le nombre de différences observées entre deux séquences est le nombre minimum de substitutions

Chapitre écrit d'après la partie "Modelling codon evolution" du livre Codon evolution, edité par Cannarozzi et Schneider, 2008.

qu'il y a eu entre ces séquences mais il n'est pas forcément égal au nombre de substitutions qui ont effectivement eu lieu. Par exemple, un site inchangé (S_1) peut avoir eu deux substitutions : $S_1 \rightarrow S_2 \rightarrow S_1$ (avec $S_2 \neq S_1$). De même, un changement ($S_1 \rightarrow S_2$) observé peut en cacher plusieurs : $S_1 \rightarrow S_3 \rightarrow S_2$ (avec $S_3 \neq S_2 \neq S_1$).
Les modèles de substitutions cherchent à déterminer la nature et le nombre de substitutions qui ont effectivement eu lieu. De façon générale, les modèles de substitutions sont des modèles de Markov en temps continu. Un tel modèle est défini par :

1. un ensemble d'états S. S est fini ou dénombrable :
2. une matrice de taux de transition Q (de dimension $|S|^2$). Pour un état $I \neq J$, le taux de transition $q_{I,J}$ est un réel positif et la somme de chaque ligne vaut 0 :

$$q_{I,I} = -\sum_{J \neq I} q_{I,J}$$

3. une distribution initiale des états.

En phylogénie, un modèle a toujours une distribution d'équilibre (la notation $\star$ est utilisée par la suite pour décrire l'équilibre) et, sous ce modèle les distributions convergent vers cet équilibre. Dans le cas de modèles de substitutions de codons les états sont les 61 codons sens (les codons STOP n'étant généralement pas considérés) et dans le cas des modèles nucléotidiques les états sont les 4 bases.
Par ailleurs, dans tout modèle d'évolution, on étudie des séquences de différentes souches ou espèces qui sont alignées puis comparées. Je tiens ici à rappeler l'importance d'étudier des jeux de données de qualité. Fletcher & Yang (2010) étudient l'effet d'insertion, de délétion et plus généralement de mauvais alignement de séquences qui jouent un rôle dans l'estimation des paramètres d'équilibre. Ils montrent que, de manière générale, l'ensemble des modèles de substitutions (en fait, de codons) sont sensibles aux mauvais alignements de séquences.

Comment paramétrer ?

Les modèles d'évolution de séquences sont la plupart du temps réversibles ce qui induit que :

- le taux de substitution instantané (processus continu dans le temps) d'un état I vers un état J peut se définir comme :

$$q_{I,J} = \mu_{I,J} \times \pi^{\star}_{J} \tag{3.1}$$

  où, $\mu_{I,J}$ est le taux de mutation ou l'échangeabilité de I vers J et $\pi^{\star}_{J}$ la fréquence d'équilibre de J.

- l'égalité $\pi^{\star}_{I} q_{I,J} = \pi^{\star}_{J} q_{J,I}$ est vérifiée.

Ces matrices s'estiment le long d'un arbre phylogénétique. Un modèle peut être non-homogène en temps, c'est-à-dire que la probabilité de passer de I à J est différente entre les branches (figure 3.1).

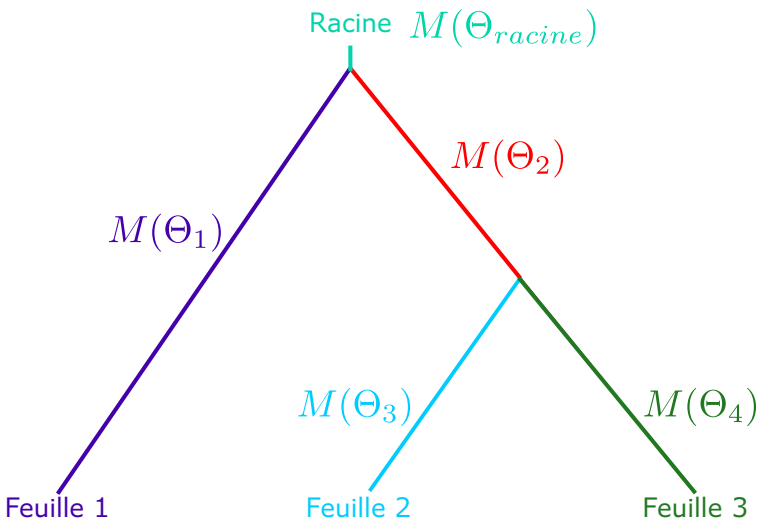


FIGURE 3.1 – Modèle d'évolution M non-homogène en temps. La couleur d'une branche correspond à une paramétrisation Θ du modèle M.

Comment estimer les paramètres d'un modèle et comparer les modèles ?

L'étude de l'évolution n'a accès qu'aux séquences actuelles (sauf quelques exemples très rares de paléo-génomique dans lesquels des échantillons de dents de fossiles ou de cadavres permettent d'avoir accès à des séquences anciennes (Damgaard et al., 2015; Duchemin et al., 2015) ou les expériences d'évolution expérimentale, Kawecki et al. (2012) pour une revue). La vraisemblance du modèle correspond à la probabilité conditionnelle d'obtenir les données actuelles sachant le modèle. Le meilleur modèle est celui qui a la plus grande probabilité d'expliquer les données, donc la plus grande vraisemblance. La vraisemblance dépend du nombre de paramètres et il peut y avoir des effets de sur-paramétrisation. C'est le cas lorsqu'un modèle avec 150 paramètres est employé pour expliquer l'évolution de séquences de 10 codons. Certains paramètres ne sont pas informatifs pour ces données. Le meilleur modèle est donc celui qui explique le mieux les données avec un minimum de paramètres. Je considère que modéliser c'est définir le paramètre essentiel qui explique nos données, puis d'ajouter étape par étape d'autres paramètres. La question est de les ajouter judicieusement pour extraire le signal informatif qui existe dans les séquences. Lorsque plusieurs modèles peuvent être appliqués aux données il est donc nécessaire de pouvoir comparer les résultats de ces modèles. Il existe trois méthodes principales permettant cette comparaison. Ce sont : l'AIC (Aikake Information Criterium, Akaike (1973)), le BIC (Bayesian Information Criterium, (Burnham & Anderson, 2002)) et le LRT (Likelihood Ratio Test, Huelsenbeck (1997)), voir le tableau 3.1.
Enfin, une dernière étape pour justifier un choix de modèle est de le tester avec :

1. des tests de bootstrap (un même jeu de données est échantillonné plusieurs fois) pour vérifier que le modèle est fiable,

2. des tests de simulation pour tester la performance,

3. des tests sur des jeux de données connus ou dont on connaît certaines particularités.

Chacune des différentes étapes de justification d'un modèle est importante voire nécessaire avant de le valider. Ce travail revient bien évidemment aux concepteurs de modèles qui le mettent à disposition de la communauté scientifique. L'utilisation d'un modèle publié implique de connaître les limites d'application de celui-ci.

Tests	Description	Formules	Limites
LRT	Pénalise le fait d'ajouter des paramètres à un modèle déjà existant. Le LRT suit une loi du χ^2 : on peut déterminer statistiquement si l'ajout d'un paramètre est informatif ou non.	$2 \times (ln(M_{\Theta_1}\|S) - ln(M_{\Theta_0}\|S))$	Modèles emboîtés
AIC	Pénalise les grands nombres de paramètres	$2\times k - 2\times ln(L(M_{\Theta_1}\|S))$	Non-statistique
BIC	Pénalise le nombre de paramètres plus fortement que l'AIC en fonction du nombre d'observations dans S	$ln(n) \times k - 2 \times ln(L(M_{\Theta_1}\|S))$	Non-statistique

TABLEAU 3.1 – Résumé des méthodes de comparaisons de modèles. On note M le modèle utilisé et Θ les paramètres à leur valeur estimée ; S les données observées ; k le nombre de paramètres et n le nombre d'observations de S (par exemple, le nombre de sites et le nombre de branches de l'arbre phylogénétique).

3.1 Les modèles nucléotidiques

La distribution à l'équilibre d'un modèle d'évolution permet de décrire le processus évolutif du modèle. Dans ce paragraphe, les modèles de nucléotides sont illustrés par un schéma des taux de mutations entre les 4 nucléotides et de leurs fréquences d'équilibre. Les flèches d'une même couleur ont les mêmes valeurs. Plus il y a de couleurs, plus il y a de paramètres. Lorsque les fréquences d'équilibre des nucléotides sont colorées, ce sont des paramètres. Le premier modèle de substitutions de Markov, JC69, a été publié par Jukes & Cantor (1969).

JC69 est un modèle de substitutions nucléotidiques où les nucléotides ont la même fréquence d'équilibre ($\pi^\star_A = \pi^\star_C = \pi^\star_G = \pi^\star_T = 1/4$). **JC69** est le modèle le plus simple, avec un seul paramètre : μ.

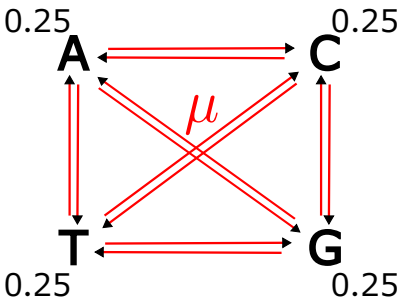


FIGURE 3.2 – *JC69*

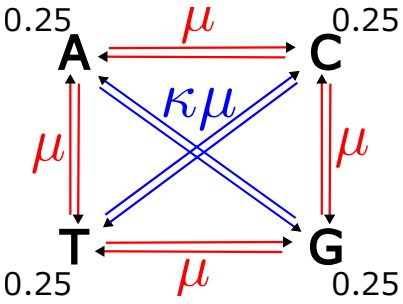


FIGURE 3.3 – *K80*

Les transitions sont plus fréquemment observées que les transversions sur les données. Le modèle **K80**, développé par Kimura (1980) considère en plus du taux de mutations, le paramètre κ qui est le rapport du taux de transitions sur le taux de transversions (fig. 3.3). A l'équilibre : $\pi^\star_A = \pi^\star_T = \pi^\star_C = \pi^\star_G = 1/4$.

D'autres modèles tels que celui de Felsenstein (1981) suppriment l'hypothèse d'égalité des fréquences des nucléotides de **JC69**. Dans **F81**, les fréquences à l'équilibre des nucléotides (la somme des fréquences vaut 1) sont des paramètres, voir fig. 3.4

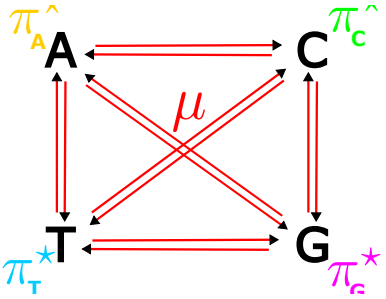


FIGURE 3.4 – *F81*

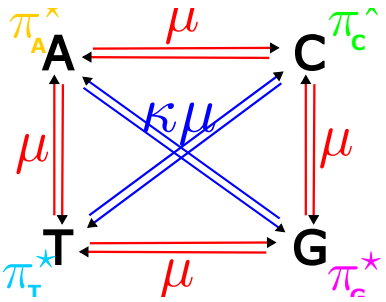


FIGURE 3.5 – *HKY85*

Le modèle d'Hasegawa et al. (1985) différencie les fréquences d'équilibre des 4 nucléotides mais aussi les transitions des transversions, **HKY85**. Il y a 6 paramètres : μ, les fréquences de nucléotides et κ, voir fig. 3.5.

T92 (Tamura, 1992) considère la symétrie des brins pour diminuer le nombre de paramètres par rapport à HKY85 donc : $\pi^\star_A = \pi^\star_T$ et $\pi^\star_C = \pi^\star_G$. **T92** a 4 paramètres et 3 degrés de liberté : μ, κ et le taux de GC et celui de AT (avec $\pi^\star_{GC} + \pi^\star_{AT} = 1$), voir fig. 3.6

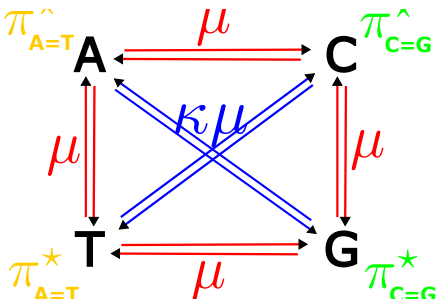


FIGURE 3.6 – *T92*

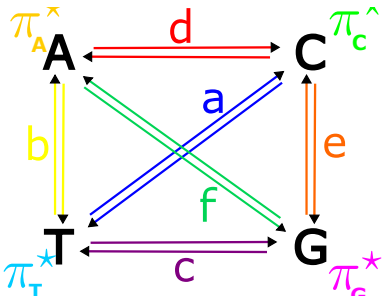


FIGURE 3.7 – *GTR*

GTR est le modèle réversible le plus générique dans lequel les taux de mutations et les fréquences d'équilibres sont des paramètres (General Time-Reversible, (Tavaré, 1986)). Il y a 10 paramètres (fig. 3.7).

Il existe d'autres modèles qui prennent en compte certaines hypothèses biologiques. Par exemple, l'appariement des bases impose qu'au niveau génomique $\pi^\star_A = \pi^\star_T$ et $\pi^\star_C = \pi^\star_G$. **SSR** est un modèle réversible et brin-symétrique (Yap & Speed, 2004) (fig. 3.8). **SSR** possède 5 paramètres.

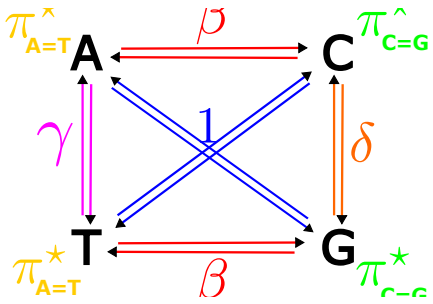


FIGURE 3.8 – *SSR*

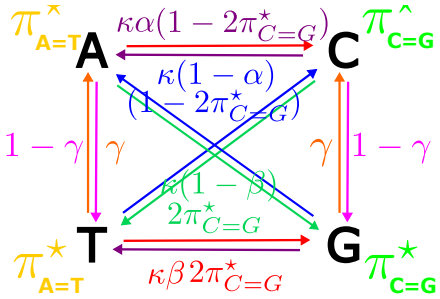


FIGURE 3.9 – *L95*

Le dernier modèle que je souhaite présenter est **L95**. C'est le modèle brin-symétrique le plus général, mais celui-ci n'est plus réversible (Lobry, 1995) (fig. 3.9). Il possède 6 paramètres. Ce modèle postule que le taux de substitution d'un nucléotide N1 vers N2 est le même que le taux du complémentaire de N1 vers le nucléotide complémentaire N2.

J'ai présenté tous ces modèles de substitutions nucléotidiques car j'en emploierai certains dans le modèle d'évolution de codons que j'ai développé lors de ma thèse, voir partie II.

3.2 Les modèles de codons

Les modèles de codons sont extrêmement puissants car ils combinent à la fois l'information nucléotidique et protéique mais il en résulte une matrice de taille importante (61x61, on ne considère que les codons sens car les substitutions vers un codon STOP sont généralement délétères). Ces modèles ont pour objectif de comprendre quels sont les mécanismes principaux qui régissent l'évolution de séquences codantes. Ils résument l'évolution grâce à un nombre limité de paramètres mathématiques, qu'on peut ensuite analyser de manière biologique. La plupart des modèles de codons posent un certain nombre d'hypothèses simplificatrices : (1) seuls les codon-sens sont considérés (61) ; (2) le taux de substitution d'un codon à un autre est 0 si il y a plus de deux changements de nucléotides ; (3) les modèles de codons mesurent la sélection sur les substitutions non-synonymes avec le paramètre ω qui est le rapport du taux de substitutions non-synonymes sur le taux de substitutions synonymes. En d'autres termes, ω rend compte du lien qui existe entre génotype (nucléotides) et phénotype (via les protéines).

Estimation de la sélection : ω

Le paramètre ω est une mesure de la force de sélection portant sur des substitutions non-synonymes par rapport aux substitutions synonymes. Ceci suppose que les substitutions synonymes sont neutres (Yang, 2006). Ainsi dans le cas d'une évolution totalement neutre, le taux de substitutions non-synonymes doit être égal aux taux de substitutions synonymes. Le paramètre ω permet de qualifier la nature de la sélection au niveau d'un gène :

- si $\omega > 1$, le taux de substitutions non-synonymes est supérieur à celui des substitutions synonymes. Ceci traduit une sélection positive de type diversificatrice.
- si $\omega = 1$, l'évolution est sous-modèle neutre. Dans le cas d'une évolution neutre, le taux de substitutions non-synonymes doit être égal à celui des substitutions synonymes ;
- si $\omega < 1$, le taux de substitutions non-synonymes est inférieur au taux de substitutions synonymes. Les mutations non-synonymes sont éliminées par la sélection naturelle : ceci traduit une sélection négative de type purificatrice.

Les modèles de codons ont principalement été développés dans l'objectif de mesurer ω précisément et d'obtenir des informations concernant la nature de la sélection sur les gènes. Par exemple, les gènes codant pour les protéines de virus reconnues par le système immunitaire de

leur hôte sont un cas connu de sélection diversificatrice. Le changement en acides aminés au sein de la protéine codée permet de limiter la reconnaissance de l'élément étranger par l'hôte.

Les modèles de codons classiques ou récents

Je présente ici les principaux modèles de codons utilisés dans la communauté scientifique ainsi que certains modèles originaux. Parmi ceux-ci, de nombreux modèles ont été proposés par Ziheng Yang et Rasmus Nielsen. Ces modèles sont implémentés dans le logiciel PAML (Yang, 2007) mais aussi dans Bio++ (Guéguen et al., 2013). PAML (Phylogenetic Analysis Using Maximum Likelihood) est un logiciel développé à Londres par Ziheng Yang, qui permet la comparaison de différents modèles de codons dans un contexte phylogénétique connu : PAML permet de tester des hypothèses de processus évolutifs qui peuvent être sites-spécifiques et/ou branches-spécifiques. Il n'est pas conçu pour reconstruire des arbres phylogénétiques. Bio++ est un logiciel développé par Julien Dutheil et Laurent Guéguen qui permet les mêmes fonctionnalités que PAML ainsi que d'autres fonctions telles que la génération de séquences simulées. La plupart des modèles présentés ci-dessous sont implémentés dans Bio++. D'autres logiciels existent comme RevBayes (Höhna et al., 2015). Ce dernier permet de tester des processus évolutifs sur des arbres connus mais dans un contexte d'estimation par méthodes bayésiennes. C'est une méthode différente de celle d'estimation par maximum de vraisemblance que j'ai évoqué dans le paragraphe précédent.

YN98 *YN*98 (Yang & Nielsen, 1998) est le premier modèle qui prend en compte ω. Il est incorporé dans l'équation 3.1, page 47 comme suit, avec I et J deux codons :

$$q_{IJ} \propto \begin{cases} 0 & \text{si } I \text{ et } J \text{ diffèrent à 2 ou 3 positions,} \\ \mu_{IJ} \quad \pi_J^\star & \text{si } I \to J \text{ est une substitution synonyme,} \\ \mu_{IJ}\omega\pi_J^\star & \text{si } I \to J \text{ est une substitution non-synonyme,} \end{cases} \tag{3.2}$$

Pour l'anecdote, les deux premiers modèles de codons ont été publiés en 1994 dans le journal "Molecular Biology Evolution" : GY (Goldman & Yang, 1994) et MG (Muse & Gaut, 1994). La principale différence entre ces modèles est que GY suppose que le taux de substitutions instantanées entre codons dépend de la fréquence d'équilibre du codon cible alors que MG suppose que ce taux dépend de la fréquence d'équilibre du nucléotide qui a changé entre le codon de départ et le codon cible. Ainsi, dans l'équation 3.2, $\pi_J^\star$ est la fréquence d'équilibre de J dans

le cas de modèles de type GY et, $\pi^{\star}_{J_k}$ est la fréquence d'équilibre du nucléotide J_k qui change entre I et J dans le cas de modèles de type MG. $YN98$ est de type GY et aujourd'hui il existe trois paramétrisations différentes de ces fréquences d'équilibre des codons :

F1x4 chaque nucléotide a sa fréquence d'équilibre ($\{\pi^{\star}_A, \pi^{\star}_C, \pi^{\star}_G, \pi^{\star}_T\}$) donc $\pi^{\star}_I = \pi^{\star}_{I_1}\pi^{\star}_{I_2}\pi^{\star}_{I_2}$;

F3x4 il existe une fréquence d'équilibre pour chaque nucléotide et pour chaque position d'un codon ($\{\pi^{\star}_A, \pi^{\star}_C, \pi^{\star}_G, \pi^{\star}_T\}_{1,3}$) donc aussi $\pi^{\star}_I = \pi^{\star}_{I_1}\pi^{\star}_{I_2}\pi^{\star}_{I_2}$;

F61 chacun des 61 codons a une fréquence d'équilibre, $\pi^{\star}_I$.

Le modèle KCM proposé par Zaheri et al. (2014) considère pour chaque position dans un codon, une matrice d'échangeabilité. L'échangeabilité entre deux codons est donc la multiplication de trois matrices de taille 4x4. Ils considèrent aussi la fréquence des 61 codons sens. Cette paramétrisation offre de nombreux avantages, notamment le fait que les mutations simultanées à deux positions d'un codon sont possibles et indépendantes. En d'autres termes, ce modèle permet des substitutions multiples sans pour autant accroître de manière importante le nombre de paramètres. En outre, ce modèle est très générique et si les hypothèses biologiques le permettent, il est tout à fait possible de contraindre les mutations multiples à une probabilité de 0. Ici, les fréquences des 61 codons sont égales aux fréquences observées.

Mugal et al. (2014) étudient l'effet de l'échelle de temps sur l'estimation de ω. En effet, dans le paragraphe précédent, j'ai présenté ω dans un contexte phylogénétique, mais il est parfois utile d'étudier les valeurs de sélection à une plus courte échelle de temps, celle des populations. La génétique des populations est l'étude de populations dans lesquelles des allèles (des copies de gènes) ne sont pas encore fixés alors que la phylogénie étudie les espèces dans lesquelles les allèles sont fixés (et peuvent être différents d'autres espèces). Mugal et al. (2014) offrent donc un contexte de génétique des populations dans l'étude du paramètre de sélection ω. Ainsi, ils décomposent la matrice de taux de mutations (M) en trois sous-matrices, chacune définie par un générateur : celle des mutations synonymes (M^{syn}), des non-synonymes(M^{nonsyn}) et des autres (M^{STOP}, mutations vers un codon STOP).

$$M = M^{syn} + M^{nonsyn} + M^{STOP}$$

Cette décomposition permet d'estimer pour chaque codon les probabilités conditionnelles d'avoir une mutation synonyme et une mutation non-synonyme sachant qu'une mutation a eu lieu dans le codon. Ainsi, ils estiment les probabilités globales d'avoir des mutations synonymes (π^{syn}), non-synonymes (π^{nonsyn}) ou autres (π^{STOP}) et peuvent alors estimer précisément ω.

Les modèles de type MutSel

Il existe une classe de modèles de codons, notés MutSel, pour lesquels la sélection se mesure avec le paramètre S de sélection sur le taux de fixation, en plus de ω. Ces modèles empruntent des concepts à la génétique des populations et notamment $S = 2N_e s$ avec s le coefficient de sélection mesuré et N_e la taille efficace de la population étudiée. Ces modèles estiment si la sélection favorise ou non des changements de codons ou d'acides aminés particuliers.

YN08 (Yang & Nielsen, 2008) est un modèle de codons qui calcule la sélection sur l'usage du code. Je détaille ce modèle car c'est à partir de celui-ci que j'ai développé un nouveau modèle de codons. Ce modèle est de type MG. Pour une substitution de I vers J le changement de codons dépend d'un facteur $h(S_{IJ})$ qui est le rapport des probabilités de fixation de $I \rightarrow J$ par rapport à une fixation neutre. Autrement dit, $h(S_{IJ})$ est une fonction de la force de sélection. On a :

$$q_{IJ} \propto \begin{cases} 0 & \text{si } I \text{ et } J \text{ diffèrent à 2 ou 3 positions,} \\ \mu_{I_k J_k} \quad h(S_{IJ}) & \text{si } I_k \rightarrow J_k \text{ est une substitution synonyme,} \\ \mu_{I_k J_k} \omega h(S_{IJ}) & \text{si } I_k \rightarrow J_k \text{ est une substitution non-synonyme,} \end{cases} \tag{3.3}$$

Pour définir mathématiquement $h(S_{IJ})$, Yang et Nielsen s'inspirent de concepts de génétique des populations (McVean & Vieira, 2001). Il existe un paramètre de fitness f_I pour chaque codons et un coefficient de sélection d'une mutation de $I \rightarrow J$ noté s_{IJ} tel que $s_{IJ} = f_J - f_I$. La probabilité de fixation de cette mutation est $\frac{2s_{IJ}}{1-\exp^{2N_e s_{IJ}}}$. Le coefficient de sélection normalisé est noté $S_{IJ} = 2N_e s_{IJ}$. On note $h(S_{IJ}) = \frac{S_{IJ}}{1-\exp^{S_{IJ}}}$. Yang & Nielsen (2008) discutent de la difficulté d'interprétation de leur paramètre de fitness de codons. En effet cette fitness est relative aux 61 codons sens ($\sum_{i \in cod} f_i = 1$) alors que l'usage du code génétique est une comparaison de l'usage des codons synonymes. Une grande avancée de ce modèle est de s'intéresser à la sélection sur les sites synonymes mais ils ne considèrent pas de manière explicite le BUC. Ainsi, dans le

cadre de ma thèse qui porte sur l'évolution du BUC, j'adapte ce modèle : je considère en fait que la somme des préférences des codons synonymes vaut 1 pour chaque acide aminé et que la somme des préférences des acides aminés vaut elle aussi 1. En d'autres termes, je propose de prendre en considération la structure du code génétique (voir partie II).

Halpern & Bruno (1998); Rodrigue et al. (2010); Tamuri et al. (2012) proposent des modèles de mutation-sélection site-spécifiques. Dans une protéine, la sélection peut avoir lieu spécifiquement au niveau de certains sites (au niveau du site-actif par exemple). Cette sélection n'est pas identifiable avec des modèles qui sont homogènes en sites. Les auteurs définissent alors un processus de mutation neutre sur les nucléotides et un paramètre de sélection site-spécifique qui distingue les sites sélectionnés des autres. Ce modèle permet d'étendre les modèles de codons à de la sélection qui a lieu au niveau protéique.

Récemment, Spielman & Wilke (2015) ont démontré l'existence de certaines limites dans l'estimation de ω ou de S. Leur point de départ est le fait que les modèles MutSel possèdent deux paramètres distincts pour regarder la "sélection". Ils regardent donc les relations qui existent entre ω et S. Premièrement, ils estiment à partir de plusieurs modèles la sélection ω sur des données simulées selon un modèle MutSel. Ils trouvent que la sélection est mieux estimée (i.e, plus proche de la valeur simulée) dans certains modèles qui n'ont pourtant pas la meilleure vraisemblance (les modèles sont comparés par AIC et BIC). Deuxièmement, Spielman & Wilke (2015) prouvent que si les mutations synonymes sont toutes neutres, alors ω est inférieur à 1, ce qui est en contradiction avec l'hypothèse d'évolution neutre. Par ailleurs, ω peut prendre des valeurs arbitrairement élevées dans le cas de sélection purificatrice. En conclusion, les auteurs soulèvent les limites qui peuvent exister parmi ces modèles.

Les modèles précédents ne considèrent pas explicitement le BUC et supposent que les substitutions synonymes sont neutres. Il existe une troisième classe de modèles (pour le moment un seul modèle est publié et le notre) qui proposent de combiner les modèles standard de codons (YN98) et les modèles de sélection qui incluent de la sélection sur les codons synonymes. Le modèle MutNSE (Kubatko et al., 2015) est site-spécifique. A chaque position k d'une séquence de codons, un codon I à une probabilité q_{ij}^k de muté vers J qui vaut :

$$q_{ij}^k = q_{ij} \times N_e \times \pi^k(I \rightarrow J)$$

avec q_{ij} le taux de substitution du codon I à J, Ne la taille de la population efficace et $\pi^k(I \to J)$ la probabilité de fixation de I à J en position k. La fonction π est une fonction exponentielle qui rend compte de la sélection due au rapport coût-bénéfice de la mutation sur la production de la protéine. Cette sélection est une fonction du taux d'erreur de sens (changement d'AA dans une séquence) site-spécifique. Ces erreurs permettent de séparer la sélection qui opère sur les sites synonymes des autres types de sélection de manière explicite. La fonction π de MutNSE est homogène en temps. Les auteurs relient leurs paramètres de sélection synonyme au nombre de copies de gènes d'ARNt et aux niveaux d'expression des gènes. Ils retrouvent que le coût des erreurs de traduction est d'autant plus fort dans les gènes fortement exprimés et que la sélection sur l'usage des codons synonymes est un déterminant important de leur taux d'évolution. Ce modèle cherche à introduire des paramètres qui ont un sens biologique et donc sont facilement analysables à la lumière des caractéristiques cellulaires connues.

Dans cette partie introductive, j'ai présenté mon objet d'étude. Ce n'est pas un organisme mais le code génétique, code qui permet de lier les informations de séquences nucléotidiques au phénotype d'organismes. Ce code redondant est utilisé par l'ensemble du vivant (il existe quelques variantes) mais pas de la même manière, certains codons synonymes étant utilisés préférentiellement chez des organismes et pas d'autres. Pour comprendre l'émergence des biais d'usage du code génétique entre espèces, je me place dans un contexte évolutif : les espèces actuelles avec des BUC différents ont évolué à partir d'ancêtres communs qui avaient eux-mêmes un certain BUC. J'étudie donc l'évolution du BUC. En effet, j'ai présenté plusieurs modèles d'évolution de séquences de codons mais ils ne considèrent pas explicitement les variations de BUC : par exemple, ils considèrent généralement que les substitutions synonymes sont neutres. Dans le cas des modèles de type MutSel, ils ne distinguent pas la sélection qui opère au niveau de l'usage du code génétique de la sélection sur l'usage des acides aminés. Lors de ma thèse, je cherche à comprendre si les substitutions synonymes sont non-adaptatives (dans le sens, où elles peuvent être biaisées par un biais mutationnel ou biaisé par de la conversion génique) ou bien sélectionnées (il existe effectivement une préférence sélective due aux forces de sélection de l'étape de traduction). Je vais donc dans une seconde partie vous présenter un modèle d'évolution de codons qui prend en compte la structure du code génétique. Ce modèle se nomme SENCA pour Sites Evolution at the Nucleotides, Codons and Amino-acids layers.

Deuxième partie

SENCA : étude de l'origine et de l'évolution du biais d'usage du code à l'aide d'un modèle de codons multi-couches

CHAPITRE 4

Avant-propos

L'article présenté dans cette partie introduit le modèle SENCA (Sites Evolution of Nucleotides, Codons and Amino-acids). SENCA est un modèle de codons qui utilise le point de vue des modèles de type MutSel présentés en Introduction (chapitre 3). En plus de la sélection sur les changements non-synonymes implémentés via le paramètre ω, SENCA considère la sélection entre codons synonymes. SENCA est un des premiers modèles à prendre en compte l'évolution de l'usage des codons synonymes et donc à considérer l'évolution du biais d'usage du code génétique le long de l'arbre phylogénétique (Kubatko et al., 2015). SENCA sépare l'évolution de codons en trois couches distinctes :

1. la couche N (pour Nucléotidique) qui définit un taux de mutation global le long de la séquence nucléotidique ;

2. la couche C (pour Codons) qui définit les préférences intra-acides aminés des codons synonymes. Cette couche prend en compte la structure du code génétique et englobe la sélection traductionnelle et les processus non-adaptatifs qui impactent la 3ème position des codons (*i.e.* par exemple le biais de conversion génique qui bien qu'impactant toutes les positions n'est pas contre-sélectionné aux positions synonymes, comme nous

le verrons en partie III) ;

3. la couche A (pour Acides aminés) qui définit l'usage des AA au niveau génomique et qui prend en compte le paramètre ω.

Ce modèle est largement inspiré de $YN08$ (Yang & Nielsen, 2008), la différence principale étant que $YN08$ réunit dans le paramètre de sélection "S", la sélection sur les codons synonymes et la sélection sur les AA alors que nous les distinguons en 2 composantes. Une autre avancée importante de SENCA est qu'il peut être employé dans un contexte phylogénétique hétérogène et/ou non-stationnaire. Nous verrons dans cette partie mais aussi la suivante que nous pouvons fixer la stationnarité de chaque couche indépendamment. Cette caractéristique est un atout pour les utilisateurs de SENCA qui peuvent prendre en compte leur connaissance biologique sur le jeu de données étudié.
Au cours de ma thèse, un important effort a été de proposer des statistiques simples permettant l'analyse des résultats de SENCA. Ces statistiques sont le $dGC^\star$, le $dGC3^\star$ et l'$ENC^\star$. Elles permettent de quantifier la part relative des 3 couches dans la composition génomique ($dGC^\star$ ou $dGC3^\star$) ou dans le biais d'usage du code génétique ($ENC^\star$) :

$dGC^\star$ estime le biais de GC par rapport à une utilisation aléatoire des 4 bases. L'article montre que le dGC est la somme des effets dus à la couche N, C et A.

$dGC3^\star$ idem pour le biais de GC en 3ème position des codons.

$ENC^\star$ a déjà été présenté en introduction. L'ENC estime le BUC par rapport à un usage uniforme des 61 codons sens. SENCA, à l'inverse d'autres modèles classiques tels que $YN98F61$ permet de retrouver le BUC observé chez les espèces actuelles. Nous proposons de quantifier la part du BUC qui résulte de la couche N et celle qui résulte de la couche C.

SENCA est implémenté dans Bio++ (Guéguen et al., 2013). Dans cet article, nous présentons en détail notre modèle, nous testons sa performance sur différents jeux de données simulées, nous montrons qu'il est mathématiquement identifiable puis nous appliquons SENCA sur 21 espèces bactériennes dont on connaît les caractéristiques génomiques. Il est publié depuis juillet 2016 dans le journal "Genome Biology Evolution" ; http://doi.org/10.1093/gbe/evw165.

GBE

Genome Biology and Evolution

SMBE

SENCA: A Multilayered Codon Model to Study the Origins and Dynamics of Codon Usage

Fanny Pouyet, Marc Bailly-Bechet, Dominique Mouchiroud, and Laurent Guéguen*

Laboratoire de Biologie et Biométrie Evolutive, University Claude Bernard Lyon 1—University of Lyon, Villeurbanne, France

*Corresponding author: E-mail: laurent.gueguen@univ-lyon1.fr.

Accepted: July 1, 2016

Abstract

Gene sequences are the target of evolution operating at different levels, including the nucleotide, codon, and amino acid levels. Disentangling the impact of those different levels on gene sequences requires developing a probabilistic model with three layers. Here we present SENCA (site evolution of nucleotides, codons, and amino acids), a codon substitution model that separately describes 1) nucleotide processes which apply on all sites of a sequence such as the mutational bias, 2) preferences between synonymous codons, and 3) preferences among amino acids. We argue that most synonymous substitutions are not neutral and that SENCA provides more accurate estimates of selection compared with more classical codon sequence models. We study the forces that drive the genomic content evolution, intraspecifically in the core genome of 21 prokaryotes and interspecifically for five Enterobacteria. We retrieve the existence of a universal mutational bias toward AT, and that taking into account selection on synonymous codon usage has consequences on the measurement of selection on nonsynonymous substitutions. We also confirm that codon usage bias is mostly driven by selection on preferred codons. We propose new summary statistics to measure the relative importance of the different evolutionary processes acting on sequences.

Key words: codon usage bias, nonstationary homogeneous model, evolutionary codon model.

Introduction

Nucleotide substitutions that act on coding DNA sequences can be classified as either: 1) Synonymous substitutions, which cause no change in the encoded protein; or 2) nonsynonymous substitutions, which change the encoded protein sequence. Evolutionary studies thus aim to distinguish between these two kinds of substitutions (Miyata and Yasunaga 1980; Nei and Gojobori 1986). As the substitution type depends on its position within a codon, this led to the emergence of codon substitution models (Goldman and Yang 1994; Muse and Gaut 1994; Yang and Nielsen 1998; Pond and Muse 2005; Kosiol et al. 2007; Mayrose et al. 2007), taking the codon as the unit of evolution. Such models are currently used to estimate the strength of selection acting on coding sequences, usually assuming that synonymous substitutions are neutral. In addition, they can be used to model nonuniform frequencies of synonymous codons in real coding sequences.

Indeed, the usage of synonymous codons in genes and genomes is not random and shows for every organism a specific set of preferences (Grantham et al. 1980), called codon usage bias (CUB). In prokaryotes, codon preferences are stable enough within a genome to be a useful tool to detect, for example, recent horizontal transfer between genomes, based on differences in CUB (Karlin 2001). Furthermore, CUB intensity is variable within a genome, which helps to predict gene expression levels (Gouy and Gautier 1982; Sharp et al. 1986; Thomas et al. 1988; Agashe et al. 2013; Wallace et al. 2013; Gilchrist et al. 2015). Two explanations for the existence of CUB are usually proposed: Mutational bias (neutral or nonadaptative) or selective pressures to optimize translational efficiency or accuracy (Akashi and Eyre-Walker 1998; Hershberg and Petrov 2008; Sharp et al. 2010). Mutational biases can be due to either mutational processes (Sueoka 1988; Rocha et al. 2006; Hershberg and Petrov 2010; Hildebrand et al. 2010; Palidwor et al. 2010) or biased gene conversion (Duret 2002), whereas selective pressures act for coadapting codon usage and tRNA content in the cell (Gouy and Grantham 1980; Sharp and Li 1986; Bulmer 1987; Kanaya et al. 1999; Rocha 2004). These hypotheses explain the existence of CUB through evolutionary processes. However, CUB is usually

© The Author 2016. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from http://gbe.oxfordjournals.org/ by Fanny Pouyet on August 26, 2016

Genome Biology and Evolution

SMBE

studied on extant sequences, using statistics that do not consider its evolution, such as Codon adaptive index (CAI) (Sharp and Li 1986, 1987) or ENC (effective number of codons) (Wright 1990), which respectively, measure the similarity of a gene's CUB relative to a reference gene set or to a uniform distribution. Some studies propose CUB measures that account for mutational bias (e.g., Knight et al. 2001; Supek et al. 2010; O'Neill et al. 2013), and the more widespread *ENC′* by Novembre (2002). These descriptive statistics are insufficient to quantify the level of selection acting on CUB through time as they only take extant genomic nucleotide composition into account. Hence evolutionary models are needed to infer and quantify the processes acting on sequences, by gathering information from the phylogenetic signal.

By construction, usual codon models assume that CUB arises by neutral mechanisms. However, the existence of selection on synonymous codons may have strong consequences on these models (Spielman and Wilke 2015). McVean and Vieira (2001) built a codon model restricted to synonymous mutations that jointly use neutral rates of mutation, and a model of relative fitness between synonymous codons to derive codon substitution rates that involve codon-specific selection coefficients. This idea of relative fitness of codons has been first adapted in a broader phylogenetic context in Nielsen et al. (2007), to add selection on CUB to synonymous and nonsynonymous substitutions. Their modeling is more simple, because codons are either preferred or not-preferred, and all codons of the same category share the same fitness. A more realistic model was after proposed by Yang and Nielsen (model FMutSel in Yang and Nielsen [2008]), where each codon has its own fitness. However, in both models, the relative fitness between two codons is computed in the same way whether they encode for the same amino acid or not. But amino acids themselves have their own specific fitness, as their distribution is not uniform in proteins. In these models, the fitness of codons does not only consider CUB, but also amino acid preferences, which may blur the specific analysis of CUB. In Halpern and Bruno (1998), Rodrigue et al. (2010), and Tamuri et al. (2012), amino acid fitness is explicitly modeled in site-specific context as the main feature of selection in addition to a neutral mutation process on nucleotides.

Here, we extend the work of McVean and Vieira to synonymous and nonsynonymous substitutions in a model that disentangles the selective processes acting on synonymous codons and on amino acids, and considers explicitly the fitness of amino acids, as in the work of Halpern and Bruno. Specifically, we add a process of substitution between amino acids to the nucleotide and synonymous codon substitution process. We organize then the substitution processes of coding sequences in three layers: The nucleotide layer describes the neutral mutation process that every site undergoes, the amino acid layer describes how the nonsynonymous substitutions change the coded amino acids, and the codon layer describes how each codon is preferred among its synonymous codons. Our model name is SENCA for site evolution of nucleotides, codons, and amino acids and is implemented in Bio++ (Guéguen et al. 2013).

SENCA allows us to explicitly estimate mutational processes, preferences on codon usage and on the usage of amino acids. Because of this organization, we can propose summary statistics, based on GC content and ENC, to measure the relative importance of mutational processes and selection on the evolution of codon usage. In this article, we show how SENCA disentangles qualitatively and quantitatively the effect of mutational processes and selection upon CUB and GC content. For this, we use SENCA in an homogeneous and nonstationary way, first on 21 groups of prokaryotes (Lassalle et al. 2015) that span a wide diversity of genomic GC content (between 27% and 65%), and then, at a deeper evolutionary scale, on five species of the Enterobacteria clade.

New Approaches

Theoretical Model

We modeled the evolution of codon sequences by specifying the substitution rate from sense codon $I = I_1I_2I_3$ to $J = J_1J_2J_3$, where I_k changed to J_k ($k \in [1; 3]$). The instantaneous substitution rate from I to J is

$$q_{IJ} \propto \begin{cases} 0 \text{ if } I \text{ and } J \text{ differ at two or three different positions,} \\ m_{I_kJ_k} g(x_I, x_J) \text{ if } I_k \rightarrow J_k \text{ is a synonymous mutation,} \\ m_{I_kJ_k} \omega g(x_I, x_J) \text{ if } I_k \rightarrow J_k \text{ is a nonsynonymous mutation,} \end{cases} \quad (1)$$

where $m_{I_kJ_k}$ is the mutation parameter from nucleotide I_k to J_k; x_I (respectively x_J) is the overall preference of codon I (respectively J) and g is the part of the substitution rates due to fixation bias from the formula introduced in McVean and Vieira (2001) and Yang and Nielsen (2008), in a similar way as in Halpern and Bruno (1998):

$$g(x_I, x_J) = \begin{cases} \dfrac{-\log\left(\dfrac{x_I}{x_J}\right)}{1 - \dfrac{x_I}{x_J}} \text{ if } x_I \neq x_J, \\ 1 \text{ if } x_I = x_J. \end{cases} \quad (2)$$

We considered that g depends on the product of synonymous codon preference with the respective amino acid preference (if they code for different amino acids). Thus, we defined the overall preference of codon I as the product of the relative preference of the amino acid encoded by codon I, AA_I, over the other amino acids $\psi(AA_I)$; the relative preference of codon I over synonymous codons, $\varphi_{AA_I}(I)$; and d_{AA_I} the degeneracy of amino acid AA_I. Thus

Downloaded from http://gbe.oxfordjournals.org/ by Fanny Pouyet on August 26, 2016

Genome Biology and Evolution SMBE

$$x_I = \psi(\mathrm{AA}_I) \times d_{\mathrm{AA}_I} \times \varphi_{\mathrm{AA}_I}(I). \quad (3)$$

g ranges between 0 and $+\infty$. Interestingly, $\frac{g(x_I,x_J)}{g(x_I,x_J)} = \frac{x_I}{x_J}$ (see supplementary equation 1, Supplementary Material online), which means that, considering only preferences between codons, the ratio of substitution rates between two codons equals the ratio of their preferences.

SENCA is based on three substitution layers: Nucleotide (*N*), codon (*C*), and amino acid (*A*) layers. These layers act simultaneously as represented on figure 1.

- The nucleotide layer *N* accounts for a neutral process of nucleotidic mutations, and is modeled through a classic nucleotide model (see http://biopp.univ-montp2.fr/manual/html/bppsuite/2.2.0/Nucleotide.html#Nucleotide for a list of available models). We can compute equilibrium frequencies of A,...,T nucleotides: $\pi^\star_A,\ldots,\pi^\star_T$ from the mutation parameter from nucleotide I_k to J_k, $m_{I_kJ_k}$. The number of free parameters depends on the chosen model.
- The codon layer *C* accounts for the relative preferences between synonymous codons; let us denote cod(AA_I) the set of synonymous codons translated into AA_I. The relative preference of codon *I* over synonymous codons is $\varphi_{\mathrm{AA}_I}(I) \in [0, 1]$, and for each amino acid these preferences are normalized such that $\sum_{I\in\mathrm{cod}(\mathrm{AA}_I)} \varphi_{\mathrm{AA}_I}(I) = 1$. This layer has 61 parameters and only $61 - 20 = 41$ free ones due to our intra amino acid normalization process.
- The amino acid layer *A* accounts for the preferences between amino acids in the case of nonsynonymous substitutions; in our case, we modeled it with a unique overall selection parameter on nonsynonymous substitutions (as is usually done in codon models), called ω, and a preference profile on amino acids. We then have 20 free parameters: ω represents the ratio of the nonsynonymous over synonymous substitution rates, and for any amino acid AA the relative preference of AA over the other amino acids is $\psi(\mathrm{AA})$, and they are normalized such that $\sum_{\mathrm{AA}\in \text{ amino acids}} \psi(\mathrm{AA}) = 1$.

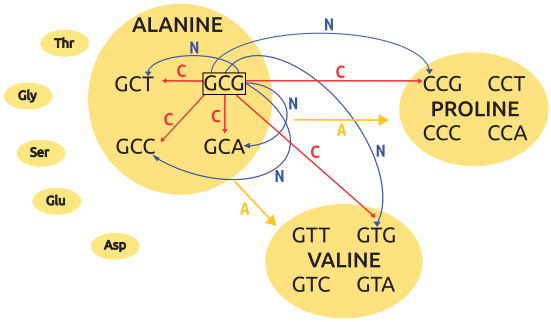


Fig. 1.—SENCA representation. Here is an example of construction of some—not all—instantaneous substitutions from sense codon GCG, that codes for Alanine, to other codons, for example, coding for Alanine, Valine, or Proline. In blue is the nucleotide (*N*) layer, in red the codon (*C*) layer, and in yellow the amino acids (*A*) layer. Arrows indicate which layer affects each substitution.

After this parameterization, the generator q is normalized as usual, with one substitution per site per unit of time on the stationary distribution.

Hereafter we use the notation SENCA[layers] to indicate the "layers" that are considered under a particular set of assumptions. In the case of uniform codon usage (i.e., no CUB), the *C* layer follows a null hypothesis—we denote that assumption as SENCA[NA]—and $\varphi_{AA_I}(I) = \frac{1}{d_{AA_I}}$. The preference of codon *I* is then the preference of its amino acid ψ_{AA_I}. In the case of no preference on the amino acids, the *A* layer follows a null hypothesis—denoted as SENCA[NC]—and $\psi(AA) = \frac{1}{20}$ for each amino acid AA. There the overall preference of codon *I* is proportional to $d_{AA_I} \times \varphi_{AA_I}(I)$. One can notice that in the joint case of no preference of amino acids nor on codons—that is, null model, denoted SENCA[N]—the preferences of the 61 sense codons are equal (stop codons are not considered in the model).

Equilibrium Frequencies

From equation (1), when the nucleotidic model is reversible we can compute the equilibrium frequency of codon *I*, $f^\star(I)$:

$$f^\star(I) \propto \left(\underbrace{\prod_{k=1}^{3} \pi^\star_{I_k}}_{N \text{ layer}}\right) \times \left(\underbrace{d_{AA_I} \times \varphi_{AA_I}(I)}_{C \text{ layer}}\right) \times \left(\underbrace{\psi(AA_I)}_{A \text{ layer}}\right). \quad (4)$$

This illustrates how processes induced by SENCA are separated into three layers *N*, *C*, and *A*. We computed partial equilibrium frequencies of codons that result from either *N* (i.e., model SENCA[N]), *C* (i.e., SENCA[C]), or *A* (i.e., SENCA[A]) layer only, by setting the other layers' parameters to their null hypothesis value in equation (4). Under SENCA[N], amino acids and codons preferences are ignored and for each codon *I* equation (4) becomes $f^\star_N(I) \propto \Pi_{k=1}^3 \pi^\star_{I_k}$. Under SENCA[C] equation (4) becomes $f^\star_C(I) \propto \varphi_{AA_I}(I) \times d_{AA_I}$, and under SENCA[A] it becomes $f^\star_A(I) \propto \psi(AA_I)$. These partial equilibrium frequencies are useful for comparing evolutionary layers, but as they are deduced from extant sequences which have been applied simultaneously to all (*N*, *C*, and *A*) layers, they should always be interpreted together and not separately.

Summary Statistics

As SENCA has many free parameters, we developed three summary statistics to estimate the overall role played by each layers based on classical codon usage statistics: GC and GC3 composition, and ENC (Wright 1990). First, we deduced from equation (4) the GC equilibrium frequency for each layer, respectively, noted $GC^\star_N$, $GC^\star_C$, and $GC^\star_A$ in order to estimate the influence of each layer on the equilibrium genome

Downloaded from http://gbe.oxfordjournals.org/ by Fanny Pouyet on August 26, 2016

Genome Biology and Evolution

SMBE

composition. Similarly, we estimated the equilibrium frequency at the third position within codons, denoted $GC3^\star$ for each layer. Because the redundancy in the genetic code is greater at this position compared with others, GC3 is often used as a proxy for underlying mutational bias in prokaryotic genomes (Muto and Osawa 1987). As genomes with very different CUB can be similar in terms of $GC^\star$ and $GC3^\star$, we used the same approach by computing $ENC^\star$ statistics for each layer.

We computed distance of genomic $GC^\star$ content to a uniform usage of the 61 sense codons (i.e., 51.4% of GC, after removing the stop codons, which are AT-rich) as

$$dGC^\star = GC^\star - 0.514. \quad (5)$$

We also defined $dGC^\star_N$, $dGC^\star_C$, and $dGC^\star_A$ as the distances to unbiased content for each layer.

We defined similar statistics for genomic $GC3^\star$ content. A uniform usage of the 61 sense codons leads to $GC3 = 0.508$. Then

$$dGC3^\star = GC3^\star - 0.508. \quad (6)$$

Similarly, we defined $dGC3^\star_N$, $dGC3^\star_C$, and $dGC3^\star_A$.

To study more specifically CUB, we computed ENC (Wright 1990) for observed sequences using codons frequencies. ENC is a measure of CUB as it goes from 20 (maximum bias) to 61 (no bias):

$$ENC = \sum_{R\in \mathrm{ARC}} \frac{k_R^2}{\sum_{\mathrm{AA}\in R} \frac{1}{n_{\mathrm{AA}}-1}\left(\left(n_{\mathrm{AA}} \sum_{l\in \mathrm{cod(AA)}} p^2(l)\right) - 1\right)} \quad (7)$$

with ARC the set of all degeneracy classes of amino acids, k_R the number of amino acids of such a class, n_{AA} the observed number of codons coding for AA, cod(AA) the set of codons of amino acid AA, $p(l)$ the relative frequency of codon l among its synonymous.

We computed $ENC^\star$, the effective number of codons on sequences at equilibrium of the model, by replacing in equation (7): $n_{\mathrm{AA}} = L \times f^\star(\mathrm{AA})$ with L the length (in codons) of the data and $f^\star(AA)$ the equilibrium frequency of amino acid AA and replacing $p(l) = \frac{f^\star(l)}{\sum_{J\in \mathrm{cod(AA)}} f^\star(J)}$ the relative frequency of codon l at equilibrium. We also defined $ENC^\star_{\mathrm{layer}}$ induced by each layer, by computing ENC using partial codon equilibrium frequencies described previously: $n_{\mathrm{AA}}{}^\star_{\mathrm{layer}} = L \times \sum_{l\in \mathrm{cod(AA)}} f^\star_{\mathrm{layer}}(l)$ and $p^\star_{\mathrm{layer}}(l) = \frac{f^\star_{\mathrm{layer}}(l)}{\sum_{J\in \mathrm{cod(AA)}} f^\star_{\mathrm{layer}}(J)}$ the relative equilibrium frequency of the codon l of this layer. For both N and C layers, $f^\star_N(l)$ and $f^\star_C(l)$ were computed as described in the section "Equilibrium Frequencies."

From the $ENC^\star$ estimates, we computed the distance from uniform usage (61) to the effective codon usage. We denoted $dENC^\star = 61 - ENC^\star$, for any layer: $dENC^\star_{\mathrm{layer}} = 61 - ENC^\star_{\mathrm{layer}}$.

Materials and Methods

Data and Model Implementation

Intraspecies Data Set

Our data set came from Lassalle et al. (2015), see table 1. We used coding DNA sequences from the core genomes of 20 bacterial pathogens and of one archeal group. These species were chosen because they encompass the diversity of genome composition among prokaryotes and that we could select nonrecombinant genes. We obtained between 6 and 35 strains per species. For each species, we built codonwise nucleotide alignments using seqinR package in R (Charif and Lobry 2007) to translate nucleotide sequences, ClustalW (Larkin et al. 2007) to align protein sequences, and PAL2NAL (Suyama et al. 2006) to retrieve nucleic alignments. Within a species, we sorted genes by increasing ENC values and concatenated them by groups of around 50 genes, to ensure we had enough data for precise parameter estimation. In total, we obtained 166 concatenates (from 1 to 16 per species, see table 1). Then, we computed a phylogenetic tree for each concatenate using CodonPhyML (Gil et al. 2013), with an Nearest Neighbor Interchange (NNI) tree topology search and the GY + W+K + F, F3x4 model. We selected one tree per species, as trees topologies were consistent within each species (see supplementary fig. S1 for topology, Supplementary Material online). We rooted our trees using TPMS (Bigot et al. 2013) and a reference species tree built with BIONJ (Gascuel 1997) from a distance matrix of the complete genomes of HOGENOM V6 database (Penel et al. 2009).

Interspecies Data Set

We considered five enterobacteria that present an average GC content: *Klebsiella pneumoniae* 342 (KLEP3), *Escherichia coli* E24377A (ECO24), *Citrobacter koseri* ATCC BAA-895 (CITK8), *Salmonella enterica* subsp. *enterica* serovar *Typhimurium* str (SAENT1), and *Escherichia fergusonii* ATCC 35469 (ESCF3). These species present a similar GC content of 55% and the phylogenetic depth of the tree is such that we can perform our SENCA analysis in a homogenous context. Moreover, we chose this data set as it contains *S. enterica* and *E. coli*, two species present in the intraspecies data set. Indeed, we will compare the results of both data sets. From HOGENOM, we selected the 1,797 gene families containing these five species, and only kept gene families for which the topology correspond to the reference HOGENOM species tree (see supplementary fig. S2, Supplementary Material online) and for which there were neither duplications nor deletions. We obtained 222 HOGENOM families that were then aligned

Downloaded from http://gbe.oxfordjournals.org/ by Fanny Pouyet on August 26, 2016

Genome Biology and Evolution SMBE

Table 1
Summary of the Data Set Characteristics

Data Set	Taxon Name	No. of Strains	No. of Concatenates	Mean GC %	Median ENC
Clostridium	*Clostridium botulinum*	8	11	29.6	35.3
Campylo	*Campylobacter jejunii*	6	7	31.6	39.8
Francis	*Francisella tularensis*	8	7	33.8	41.6
Staph	*Staphylococcus aureus*	15	11	34.2	40.5
Sulfo[a]	*Sulfolobus* spp.	8	9	35.4	45.0
B_anthracis	*Bacillus antharcis/aureus* group	17	6	37.0	42.5
Listeria	*Listeria* spp.	8	6	38.8	47.6
Strep_pyo	*Streptococcus pyogenes*	12	7	39.6	48.5
Helico	*Helicobacter pylori*	14	2	40.4	46.6
Acineto	*Acinetobacter* spp.	6	10	40.8	43.7
Clamy_trach	*Clamydia trachomatis*	13	7	41.8	50.7
Strep_pneu	*Streptococcus pneumoniae*	13	7	42.0	48.8
Yersinia	*Yersinia pestis*	11	13	49.3	51.8
Escherichia	*Escherichia coli*	35	3	53.3	45.5
Salmo	*Salmonella enterica*	14	12	54.6	45.3
Neisseiria	*Neisseiria meningitidis*	8	4	55.3	43.8
Brucella	*Brucella* spp.	9	8	58.8	41.6
Bifido_longum	*Bifidobacterium longum*	6	7	61.9	38.2
Mycobacterium	*Mycobacterium tuberculosis* complex	7	1	66.1	41.5
Burk_ceno	*Burkholderia cenocepacia* complex	8	16	68.2	31.0
Burk_mal	*Burkholderia mallei* group	9	12	68.7	31.0

Note.—Data comes from Lassalle et al. (2015). On each line is indicated the species and the corresponding number of strains in the alignments, the number of concatenates, the mean observed GC content and the median observed ENC, each concatenate being approximately 50 genes long. Genes are from the core genome, at least 900 nt long and classified as nonrecombinant in Lassalle et al. (2015).

[a]Archeal species of the data set.

codonwise as previously described. We concatenated genes sorted by increasing ENC values into four concatenates of around 50 genes each.

Implementation

SENCA was implemented in Bio++ (Guéguen et al. 2013) and likelihood optimized with bppml (Dutheil and Boussau 2008). For the *N* layer, under the hypothesis that the mutation process is strand symmetric and reversible, as, in our study, we are interested in broad tendencies in GC content at equilibrium, we used the T92 model (Tamura 1992) which depends on two free parameters, the equilibrium frequencies of the GC pairs $\pi^{\star}_{CG}$ and κ which is the transition/transversion ratio. Additionally, to reduce computational complexity in the intraspecies analysis, we supposed that the *A* layer is stable within a species, that is, ψ_{AA_I} stationary (which is more realistic than assuming stationary amino acids frequencies). This assumption is reasonable as we studied intraspecies evolution, with short tree depths. We relaxed this assumption in the interspecies analysis. We tested the informativeness of SENCA layers *N*, *C*, and *A* with likelihood ratio tests (LRT, see supplementary table S1, Supplementary Material online). In order to demonstrate the usefulness of our approach, we compared SENCA with the more classical YN98 + F61 codon model (Yang and Nielsen 1998), noted YN98 hereafter, in which synonymous substitutions are neutral, but where any CUB can be modeled, as each codon has its own equilibrium frequency. We performed nonstationary analyses using a homogeneous modeling for all models (numbers of parameters in supplementary table S1, Supplementary Material online). We compared SENCA and YN98 using Akaike information criterion (AIC) and Bayesian information criterion (BIC) (see supplementary table S1, Supplementary Material online). Please note that, if we use HKY85 (Hasegawa et al. 1985) model for the *N* layer and assume stationarity, then the fitness of codon *I*, noted F_I, presented in Yang and Nielsen (2008) is equal to $F_I = d_{AA_I} \times \varphi_{AA_I}(I) \times \psi(AA_I)$.

Simulations

We performed simulation studies using bppseqgen sequences generator with SENCA model (Dutheil and Boussau 2008). We used a species trees with 13 leaves and median branch length ≈ 0.10 (see supplementary fig. S3, Supplementary Material online) and simulated an alignment of 20,000 sites. Root was set equal to the global null hypothesis, that is, uniform codon usage, and we simulated with combinations of GC^{*}_{N} at 0.3, 0.5 and 0.7, and $GC3^{*}_{C}$ at 0.3, 0.5 and 0.7. We tested different classical nucleotidic models for the *N* layer of SENCA: T92 (Tamura 1992), HKY85 (Hasegawa et al. 1985, as in FMutSel of Yang and Nielsen [2008]), GTR (Tavar 1986),

Downloaded from http://gbe.oxfordjournals.org/ by Fanny Pouyet on August 26, 2016

Genome Biology and Evolution

SMBE

SSR (Yap and Speed 2004), and L95 (Lobry 1995, the most general strand symmetric model). As there are many ways to set the parameters of the codon layer for a given $GC3^*_C$, we used the scenario that may be the most difficult to discriminate, where the amino acids of a same redundancy class share the same codon preferences and where for each amino acid all GC ending synonymous codons share uniformly this $GC3^*_C$ preference (and symmetrically for AT). The AA preferences of the *A* layer were chosen randomly, to be different from the root preferences which equals to $\frac{1}{20}$ for each amino acid. Hence, the *A* layer is not stationary (see supplementary material, Supplementary Material online, for the values). For each parametrization, we ran five replicates.

We also performed parametric bootstrap tests on *Burkholderia cenocepacia* complex, *Campylobacter jejunii* species (GC-rich and AT-rich, respectively) to check the variance of real estimates that considers particular codon bias. We performed 30 replicates for each concatenate.

Results

We studied 21 groups of prokaryotes that are diverse in terms of genomic content (GC content ranges from 29% to 68%). We showed two main results. First, SENCA better predicts genomic content and CUB than YN98 + F61. Second, SENCA parameterization is relevant to distinguish mutational effects from selection on codons, and to compare them. Finally, we studied a deeper Enterobacteria tree of five species to see how the different layer effects scale with the depth of the tree.

Model Identifiability and Validation

Simulations

In theory, SENCA is identifiable (see supplementary material, Supplementary Material online, for demonstration), but we wanted to check its practical identifiability on our data. For this, we performed a simulation study and parametric bootstraps. Results are shown in figure 2, for controlled parameters (red dots) and for parametric bootstraps on *C. jejunii* (AT-rich species, blue dots) and *B. cenocepacia* complex (GC rich species, green dots). In both cases maximum-likelihood estimates from SENCA retrieved with good precision the values used for simulations, confirming the model identifiability. In particular, one concern may be that opposite effects from the nucleotidic and codon layers may be hard to grasp by SENCA. Here we see that SENCA retrieves the input parameters correctly, even in those difficult cases.

We also tested a simulation study with *N* layer modeled by HKY85, SSR, GTR, or L95. We saw that using complex nucleotidic models, such as SSR, GTR or L95, reduced the practical identifiability of the model, and that HKY85 and T92 gave similar results justifying our usage of the T92 nucleotidic model. Results are shown in supplementary figure S4, Supplementary Material online.

Model Validation

We compared likelihoods of SENCA and YN98 models using AIC and BIC criteria (see supplementary table S1, Supplementary Material online). Using AIC, SENCA is better-fit than YN98 for 152 concatenates out of 166. Using BIC (ΔBIC > 2), SENCA is better-fit than YN98 in 121 concatenates. SENCA has fewer parameters to estimate than YN98: Both models approximately share the same number of total parameters, but in SENCA we can hypothesize the stationarity layer by layer, and doing it for the AA layer reduces the number of free parameters by 19. This possibility of tuning each layer in the model according to the biological signal under study is one of the most relevant features of SENCA. To check the importance of each layer, we also performed estimations by fixing one layer to its null hypothesis at a time: $SENCA_{[NC]}$, $SENCA_{[NA]}$, $SENCA_{[CA]}$. We computed LRT to validate the significance of our parametrization. Layers *N*, *C*, and *A* are always informative (*P* value < 0.05 after Bonferroni correction) except for the layer *N* of the enterobacteria study.

Comparison to YN98 + F61 Model

GC Content at Equilibrium

In figure 3, we compare the equilibrium *GC** content of YN98 and SENCA with T92 model of the *N* layer (analyses using HKY85 are similar, results are not shown). For most of the species, global *GC** estimates of SENCA are below GC_{obs}, indicating a global tendency toward AT enrichment at equilibrium. In particular, for all AT-rich species, *GC** is close to 0.3, a value observed in some recent studies (Hershberg and Petrov 2010; Hildebrand et al. 2010) as the equilibrium of mutation forces. This overall tendency is not identified by YN98, whose estimates are often closer to a uniform GC content relative to SENCA estimates.

As already observed in many species, in figure 3b, we found GC3 content more biased than GC content. Comparing equilibrium *GC3** of both models, we see that SENCA estimates are often closer to the observed values than YN98, especially for AT-rich and GC-rich species, even though models are theoretically both able to retrieve such extreme GC3 biases. It suggests that explicitly taking into account the structure of the genetic code in the substitution process is an important modeling feature.

It is interesting to understand how these results depend on the evolutionary scale. In particular, intraspecific results for *Escherichia* and *Salmonella* can be compared with those of the interspecific study (fig. 3) which includes these species. For global GC content, results are quite similar between and inside species, with YN98 still closer to a uniform GC content relative to SENCA. For GC3, the equilibrium estimates both by SENCA and YN98 are higher than the intraspecific estimates, which reveals the difference between studies at inter- versus intraspecific scales, where synonymous mutations may still be

Downloaded from http://gbe.oxfordjournals.org/ by Fanny Pouyet on August 26, 2016

Genome Biology and Evolution SMBE

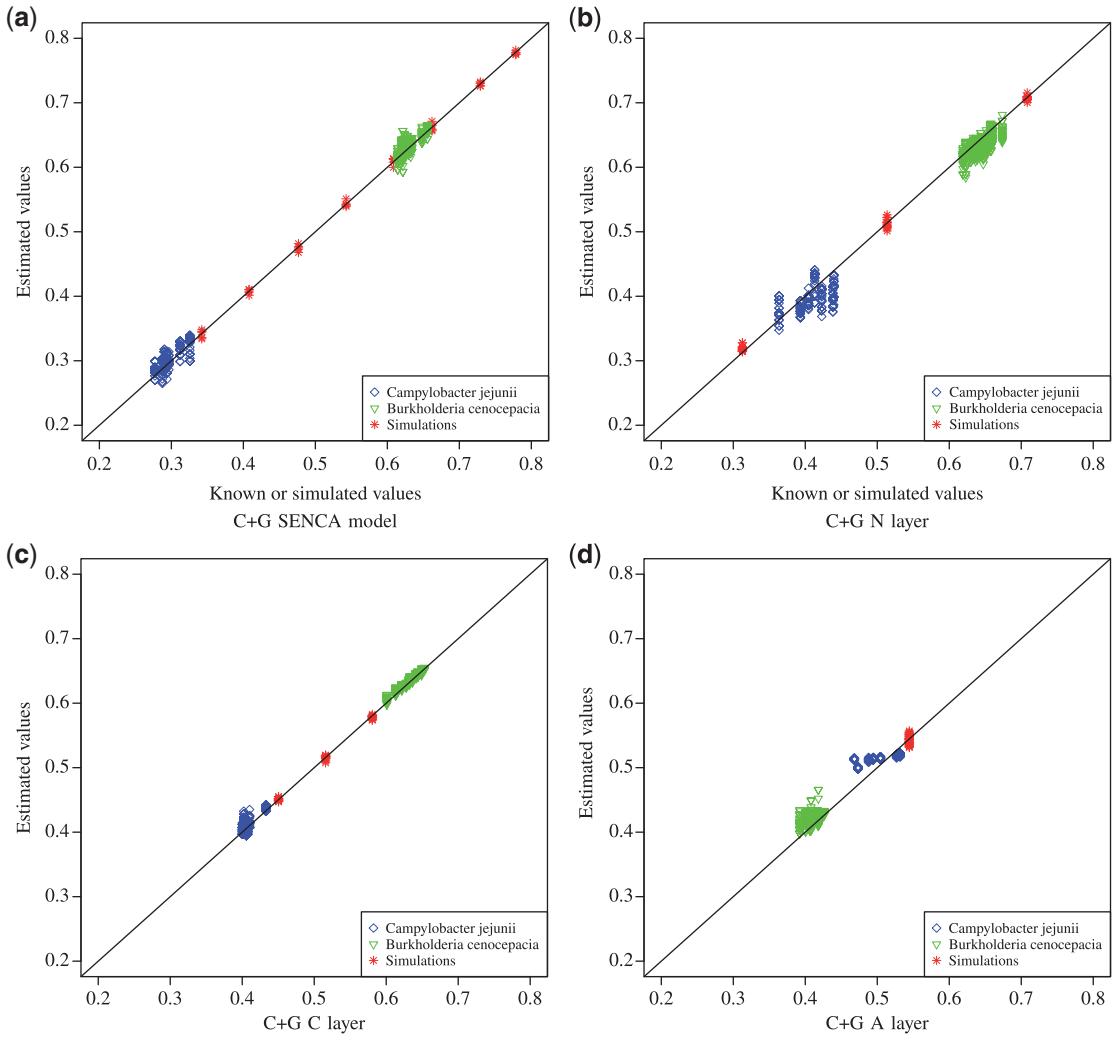


Fig. 2.—Simulation results for each layer. *x* axis corresponds to the chosen values (red dots) or known values (green and blue dots) used to simulate data, *y* axis to the values estimated by maximum likelihood. Red dots correspond to simulations with GC_N and GC_C ranging from 0.3 to 0.7. Green and blue dots correspond to parametric bootstrap where parameter values taken from previous estimations are used to first simulate, then infer, the evolutionary processes, respectively, of *Burkholderia cenocepacia* complex and of *Campylobacter jejunii*.

polymorphic. We can see that the difference is more marked for SENCA, which we attribute to a better capacity to grasp the evolutionary signal at the third position in the intraspecific study.

Codon Usage Bias

We then explored CUB using ENC (Wright 1990). We computed equilibrium *ENC** estimated by YN98 and by SENCA (see fig. 3c). As expected, ENC_{obs} is lower in AT-rich or GC-rich species: The higher the bias in genomic content, the higher the bias in codon usage. By comparison to ENC_{obs}, $ENC^{\star}_{YN98}$ almost invariably shows lower CUB than the observed values, whereas $ENC^{\star}_{SENCA}$ and ENC_{obs} are closer for all species. At the interspecific level, SENCA also predicts an equilibrium ENC closer to the observed one, whereas YN98 has higher values, matching what is seen on the intraspecific analysis of *E. coli* and *S. enterica*. Moreover, we can notice that *ENC** of *Enterobacteria* is lower than of *E. coli* and *S. enterica*. This is explained because at the intraspecific scale, slightly deleterious mutations are expected to be still present whereas they must have been deleted at the interspecific scale. Those deleterious mutations increase the frequency of unpreferred codons and lead to an increase of *ENC** values.

Effects on Selection Measure

ω is used as an index for the strength of selection—the lower the value of ω, the stronger the purifying selection. ω is considered as the ratio between nonsynonymous substitutions

Downloaded from http://gbe.oxfordjournals.org/ by Fanny Pouyet on August 26, 2016

Genome Biology and Evolution

SMBE

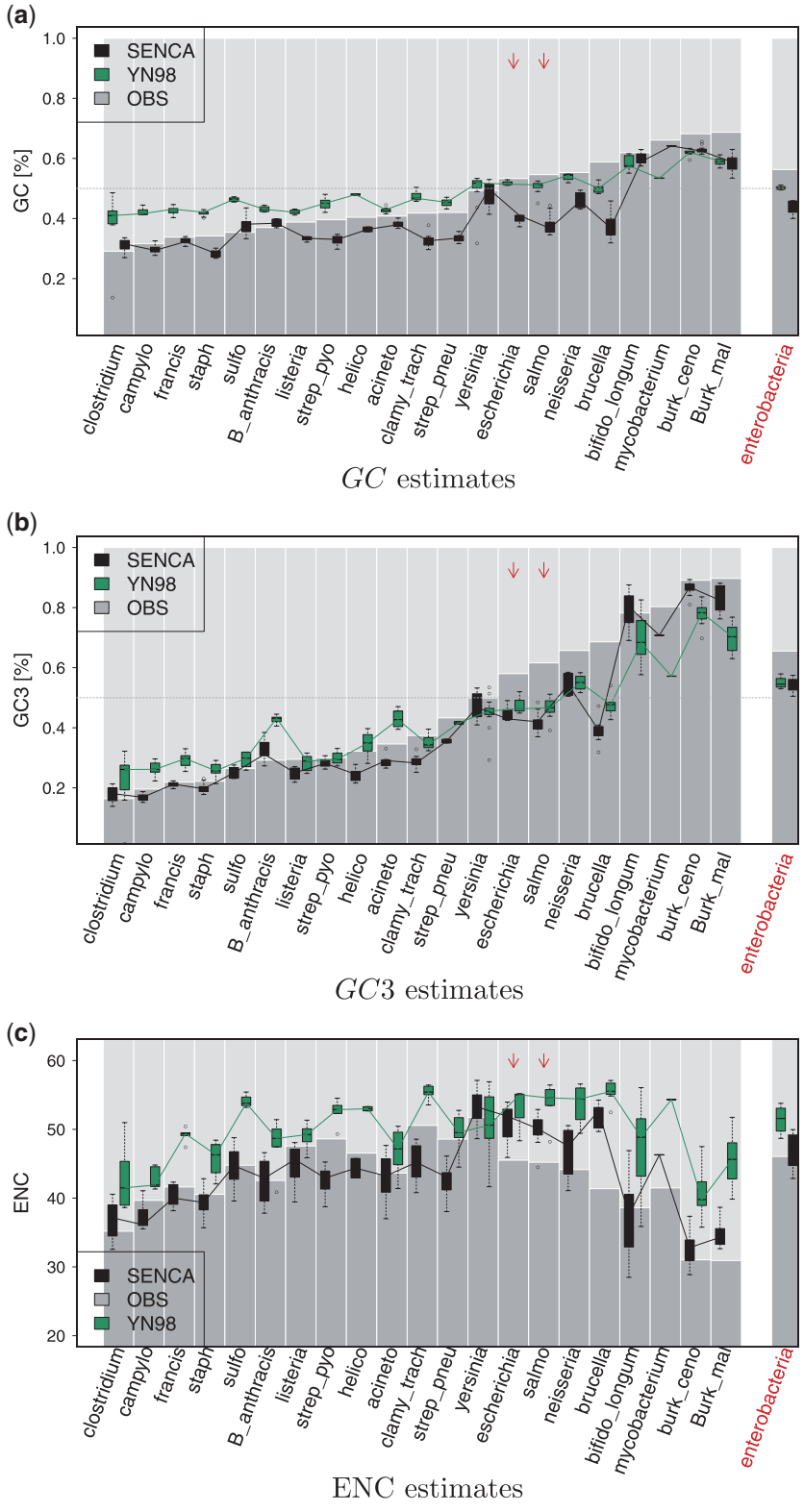


Fig. 3.—GC, GC3 contents and ENC estimates at equilibrium from SENCA and YN98. Species are ordered by increasing GC content in (*a*) and (*c*), and by increasing GC3 in (*b*). Interspecific results are shown on the right. Gray bars represent observed GC in (*a*), observed GC3 in (*b*), and observed ENC in (*c*). Boxplots span the different concatenates within a species. Black stands for SENCA estimates, green for YN98 estimates. Arrows indicate *Escherichia coli* and *Salmonella enterica*.

Downloaded from http://gbe.oxfordjournals.org/ by Fanny Pouyet on August 26, 2016

Genome Biology and Evolution SMBE

and synonymous substitutions. In a context where some synonymous substitutions are slightly deleterious, they are less frequent than if considered as neutral substitutions, and SENCA will estimate higher ω than YN98. As shown in figure 4a, we indeed observe that ω values inferred with SENCA are significantly higher than with YN98 ($P < 210^{-16}$, unilateral paired Wilcoxon test). Moreover, these differences are even greater in the enterobacteria estimates of ω than in *E. coli* or *S. enterica* estimates (except for two concatenates of *Salmonella*), see figure 4b. Indeed, for enterobacteria, the median difference between ω estimates is 0.0075 (variance 5.8×10^{-7}) whereas for *E. coli* or *S. enterica*, the median is 0.0035 (variance 1.1×10^{-4}). In fact, in intraspecific studies, slightly deleterious mutations may not yet have been suppressed, and less difference is expected between neutral and synonymous substitutions than in interspecific studies.

This demonstrates that taking CUB into account for evolutionary studies is important as it can change the classical estimates of selection acting on genomic sequences.

SENCA: A Multilayered Model

GC Content at Equilibrium

As SENCA is a multilayered model, it is possible to examine the different layers separately. At first, we blocked one layer at a time, which leads to the loss of useful information (see LRT results; supplementary table S1, Supplementary Material online). Even though, in this case, the global *GC** content is mostly reliable (see supplementary fig. S5, Supplementary Material online), the dynamics between layers are different and the CUB (through *ENC**; see supplementary fig. S5*c*, Supplementary Material online) is highly impacted if *C* (for average GC species) or *A* is fixed (for every species). This is explained as each layer refines the model, and it confirms the importance to examine the joint contribution of *N*, *C*, and *A* on *GC** and *GC3** estimates.

We looked at $dGC^\star$, the distance between $GC^\star$ and uniform composition (see eq. 5). We checked whether the effects of the different layers may be summed to explain the equilibrium GC content. Indeed the correlation between $dGC^\star$ and the sum $dGC^\star_A + dGC^\star_C + dGC^\star_N$ is highly significant ($R^2 = 0.996$, $P < 10^{-16}$, see supplementary fig. S6*a*, Supplementary Material online), and the slope of the regression is 0.95 (intercept was fixed to 0). Therefore, $dGC^\star$ estimates can be seen as different forces acting separately on the global $GC^\star$ content. Thus, we looked at the contribution of *N*, *C*, and *A* layers on equilibrium GC content (fig. 5a). We observed that in most of the cases *C* and *N* layers influence GC in the same direction. This leads to a more biased $dGC^\star$—that is, further from 0.514—than any layer taken independently. For AT-rich species, the *N* layer has negative $dGC^\star$ values, whereas for GC-rich species (> 60%), $dGC^\star_N$ is positive. The *C* layer follows the same pattern in a smoother way.

Furthermore, we saw that similar $dGC^\star$ can be due to very different $dGC^\star_N$, $dGC^\star_C$, and $dGC^\star_A$. As an example, the species *Clostridium botulinum*, *Staphylococcus aureus* and *Streptococcus pyogenes* present similar $dGC^\star$ values, approximately −0.2, but different layers effect, with the *N* layer dominating in *St. aureus*, or *C* and *A* layers having opposite effects in *Cl. botulinum*. This illustrates the ability of our multilayered model to explain the nucleotide composition of sequences.

Overall, we observed two large categories of intraspecific results. For all the species but the GC-rich ones and *Yersinia pestis*, the *N* layer has a negative effect on $dGC*$ and $dGC3*$.

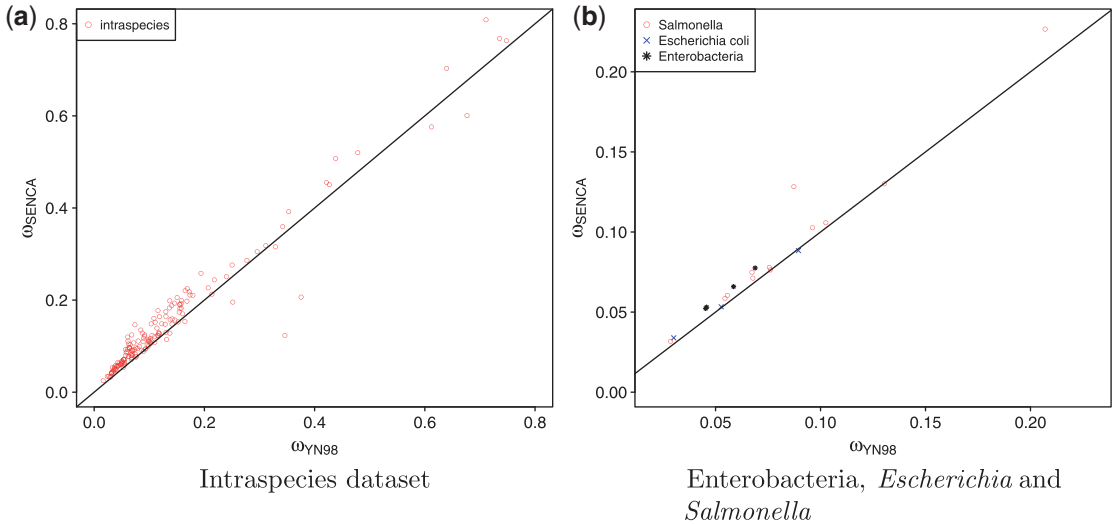


Fig. 4.—Estimates of ω from SENCA and YN98. The line represents $\omega_{SENCA} = \omega_{yn98}$. Estimates of SENCA are significantly higher than those of YN98 (see main text). (*a*) represents the intraspecies data set and (*b*) represents in blue *Escherichia coli*, in red *Salmonella enterica*, and in black the concatenates from the interspecific data set. Each point is a concatenate.

Downloaded from http://gbe.oxfordjournals.org/ by Fanny Pouyet on August 26, 2016

Genome Biology and Evolution

SMBE

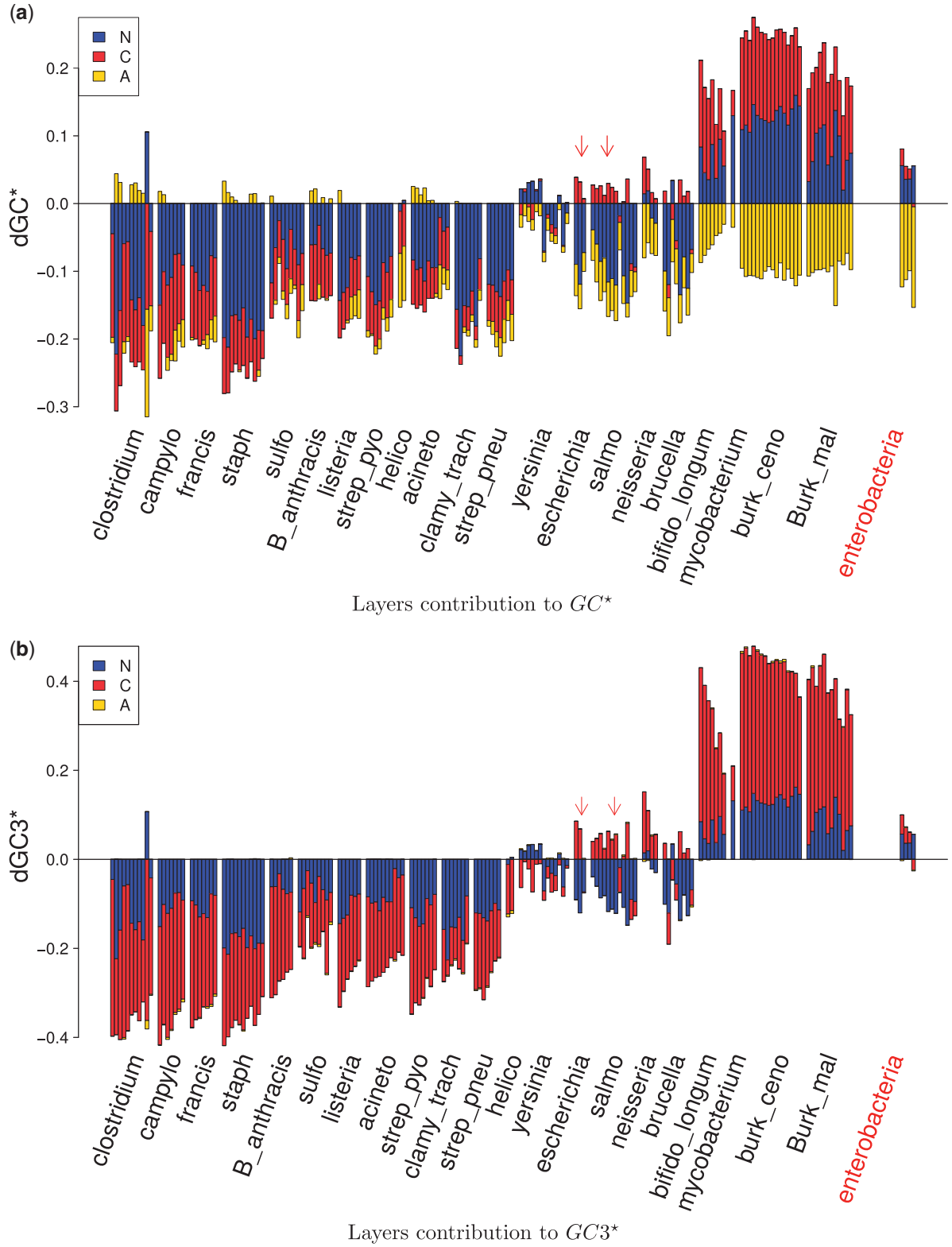


Fig. 5.—Layers contribution to GC and GC3 contents at equilibrium from SENCA. Blue stands for $GC^{\star}_{N}$, red for $GC^{\star}_{C}$, and yellow for $GC^{\star}_{A}$. (*a*) represents the distances of *N*, *C*, and *A* to a uniform GC content ($dGC^{\star}_{layer} = GC^{\star}_{layer} - 0.514$) and (*b*) the distances of *N*, *C*, and *A* to a uniform GC3 content ($dGC3^{\star}_{layer} = GC3^{\star}_{layer} - 0.508$). Each bar represents one concatenate. Species are ordered by increasing observed GC in (*a*) in and observed GC3 in (*b*). Interspecific results are shown on the right. Arrows indicate *Escherichia coli* and *Salmonella enterica*.

Downloaded from http://gbe.oxfordjournals.org/ by Fanny Pouyet on August 26, 2016

Genome Biology and Evolution SMBE

This makes sense in relation to the theory that mutations are universally biased toward AT (Hershberg and Petrov 2010; Hildebrand et al. 2010). For average GC species, selection may compensate for such a bias by being GC-driven. However, the behavior on GC-rich species is very different. This difference is not due to the model, as it is symmetric with GC, and the causes should be looked in the evolution process. The *C* layer contributes far more to GC in these species than in other ones. Both *N* and *C* layers are toward high GC, and the *A* layer is strongly in the opposite direction, all with equal strength, suggesting a complex process on content equilibrium.

Comparing the interspecific analysis with results from *Escherichia* and *Salmonella* is interesting as the decomposition in the interspecific analysis is different than the one of *Escherichia* and *Salmonella* species. In the interspecific data, mutations are fixed, which means that the *N* layer is concerned by substitutions, that is, mutations plus selection. These substitutions are GC-driven, and as in the intraspecific studies the mutations are toward AT, we can hypothesize that selection biased toward GC. At this evolution scale, preferences on amino acids have a strong impact toward AT, much stronger than in the intraspecific studies, with the exception of GC-rich species. This unexpected result is connected to previous hypotheses published as Lobry (1997). Indeed, the preference toward AT in the *A* layer is probably related to the chemical constraints of the bacterial proteome.

Finally, we studied $GC3^\star$ (fig. 5b). The correlation between $dGC3^\star$ and the sum $dGC3^\star_A + dGC3^\star_C + dGC3^\star_N$ is highly significant ($R^2 = 0.997$, $P < 10^{-16}$, see supplementary fig. S6*b*, Supplementary Material online) with a slope of 0.87 and an intercept fixed to 0. Globally, $dGC3^\star$ is clearly driven by the *C* layer. Red barplots are predominant for every species but *S. enterica*. This is different from the behavior of $GC^\star$ estimates but it is expected as most—but not all—of the *C* layer action should be seen in the third codon position. This is consistent with the classical observation that GC3 is more biased than GC12 (GC at the first and second codon positions) in prokaryote genomes (Muto and Osawa 1987). The *N* layer effect is weak, but not null at this position. By definition, it is equal to the global *N* effect which acts identically on all positions. Note that for GC^*3 there is nearly no impact from the *A* layer because of the degeneracy of the genetic code at this position.

Codon Usage Bias

We computed partial $ENC^\star$, that is, ENC computed on codon partial equilibrium frequencies due to each layer separately, and compared this with $ENC^\star$ and ENC_{obs} (see supplementary fig. S7, Supplementary Material online). Our main result is that $ENC^\star$ is quite close to $ENC^\star_C$ and that $ENC^\star_N$ was very high (mean value is 58.4). This suggests that the *C* layer dominates the establishment of CUB at equilibrium, with a relatively small effect of the *N* layer. There are a few interesting exceptions: *St. aureus*, *Chlamydia trachomatis*, or, among GC-rich species, *Burkholderia cenocepaia* show a lower value of $ENC^\star_N$, indicating a marked effect of the *N* layer on CUB. These effects need to be studied in context, that is, to be compared with $ENC^\star$.

To quantify the effects of *C* and *N* layers on CUB at equilibrium, we defined dENC as the distance of ENC to 61 (no bias). In figure 6, we see that the *C* layer is predominant in the

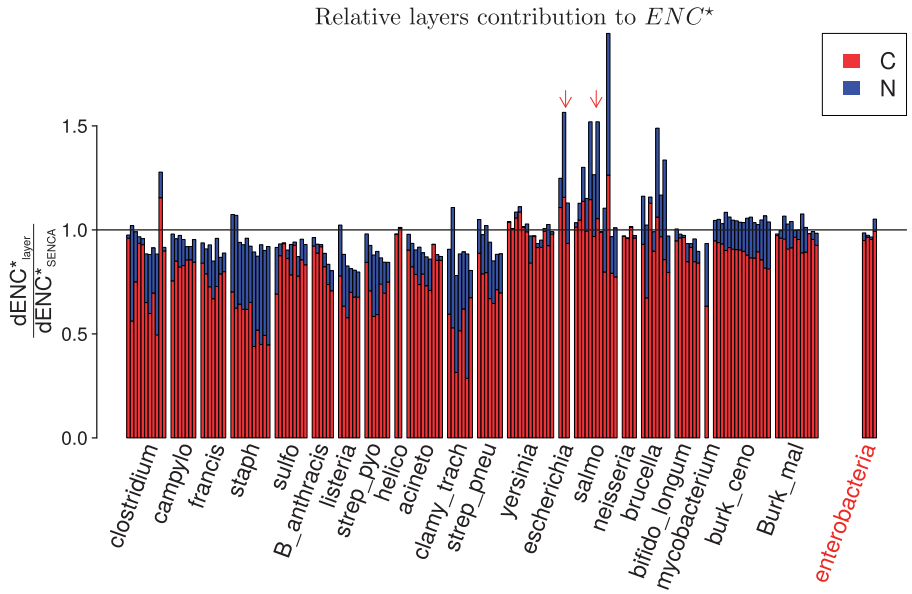


Fig. 6.—Quantification of *N* and *C* layers' effect on CUB. Blue represents $\frac{dENC^\star_N}{dENC^\star}$ and red $\frac{dENC^\star_C}{dENC^\star}$. Species are ordered by increasing observed GC content. Interspecific results are shown on the right. Arrows indicate *Escherichia coli* and *Salmonella enterica*.

Downloaded from http://gbe.oxfordjournals.org/ by Fanny Pouyet on August 26, 2016

Genome Biology and Evolution

SMBE

estimation of CUB for the 21 species: $dENC^{\star}_{C} > dENC^{\star}_{N}$. Similarly to our procedure for $dGC^{\star}$, we checked whether $dENC^{\star}$ could be predicted from the layer estimates, by fitting $dENC^{\star}$ to $dENC^{\star}_{N} + dENC^{\star}_{C}$ with a linear regression model. The fit indicates that the intuitive idea of adding separate layer effects to estimates CUB at equilibrium works quite well ($R^2 = 0.964$ with $P < 10^{-16}$, slope: 0.98 with an intercept fixed to 0, see supplementary fig. S8, Supplementary Material online). The slope of 0.9806 of the linear model indicates that the direct sum of $dENC^{\star}_{N}$ and $dENC^{\star}_{C}$ slightly overestimates $dENC^{\star}$ in these data. Moreover, we can see that this tendency varies with GC content, as the ratio $\frac{dENC^{\star}_{C}+dENC^{\star}_{N}}{dENC^{\star}}$ is mostly below 1 for GC-poor species, and above 1 for GC-rich species (see fig. 6).

As ENC is a statistic computed on multidimensional data, the sum of $dENC^{\star}$ for the *C* and *N* layers neglects the overlapping effects of these layers. By direct comparison, our model allows us to measure how *N* and *C* layers interact to influence the overall CUB, either positively or negatively. One clear example of negative interaction is *Salmonella*: In figure 5b, we see that $dGC3^{\star}_{N}$ and $dGC3^{\star}_{C}$ do not have the same sign for this species. Correspondingly, in figure 6, the sum $dENC^{\star}_{C} + dENC^{\star}_{N}$ overestimates $dENC^{\star}$, for all *Salmonella* concatenates but the three where the *N* and *C* layers agree on $GC3^{\star}$, which means that *N* and *C* layers interact negatively, as expected. One can also see a pattern of "descending staircase" for the red bars in figure 6, for many species (in particular *St. aureus* and *B. cenocepaia*, respectively, for AT- or GC-rich examples). This is related to the data structuration, as genes were concatenated according to their observed ENC values, higher ENC (and then lower CUB) last. This pattern then indicates that for genes having a low level of observed CUB, SENCA finds the *C* layer effect to be less important than in genes with higher CUB.

Discussion

In sequence evolution, several biological processes act together at nucleotides, codons, and amino acids scales. In order to quantify the effects of mutation and selection at each of these scales, we developed an evolutionary model, SENCA, divided into three layers: nucleotide (*N*), codon (*C*), and amino acid (*A*). SENCA, by construction, is very flexible, and can be employed to tackle a variety of biological questions. As an example, we can set each layer to be stationary or not in function of the data. The decomposition of evolutionary signals in different layers allows for treating each layer separately; for example, by using specific amino acid substitution models for the *A* layer, or specific nucleotide substitution models for the *N* layer. Moreover, because the genetic code is explicit in this model, selection on CUB and on nonsynonymous substitutions can be studied simultaneously. This different modeling makes the most prominent difference with model FMutSel, where layers *A* and *C* are not distinguishable. Moreover, FMutSel is all stationary, which is a strong hypothesis (actually not supported by our data). Considering the model described in Nielsen et al. (2007), the authors assume that 1) CUB is only defined through an optimal codon per amino acid, 2) selection on CUB shows the same intensity for all amino acids, and 3) the set of optimal codons is known a priori. In this model, this unique fitness on all preferred codons neutralizes all preferences on amino acids (which is not supported by our data). Moreover, SENCA does not require the optimal codons to be known—which is particularly useful when using a nonhomogeneous codon layer where preferences may change over time. One additional feature of SENCA is that we can easily study the overall equilibrium of the model in a mixture of equilibrium from each layer, through summary statistics, such as $dGC^{\star}_{\text{layer}}$, $dGC3^{\star}_{\text{layer}}$, and $dENC^{\star}_{\text{layer}}$. We have shown that these statistics can be manipulated intuitively, as the effects of all layers can be summed up almost linearly to give the global equilibrium. Moreover, these statistics all account for the phylogenetic signal, which was not considered in previous studies such as Novembre (2002), Supek et al. (2010), and O'Neill et al. (2013).

We performed a nonstationary analysis of the core genomes of 21 bacterial and archaeal species from Lassalle et al. (2015), and of five Enterobacteria. We estimated equilibrium frequencies using SENCA in comparison with similar estimates using classical codon model YN98 + F61. The main mechanistic difference between the two models is that SENCA considers explicitly the genetic code, and synonymous substitutions are a priori not neutral. Indeed, $ENC^{\star}$ of YN98 is higher than $ENC^{\star}$ of SENCA (fig. 3c), which challenges the assumption that synonymous substitutions are neutral. As expected, and in accordance with simulations in Spielman and Wilke (2015), we show that this assumption leads to a systematic bias in the estimation of the strength of selection acting on nonsynonymous substitutions. When synonymous substitutions have a selective cost, they are less frequent, leading to higher estimates of ω. These estimates are in most cases more accurate than those of YN98, as shown by maximum-likelihood comparisons with the AIC and BIC. On the other hand, it is possible that codon preferences change, in which case synonymous substitutions may be advantageous, and lead to lower estimates of ω. SENCA is then useful for detecting selective pressure on nonsynonymous substitutions, as it better estimates the cost of synonymous substitutions by distinguishing them from the background mutational bias (Lawrie et al. 2011).

Moreover, taking into account selection on CUB allows our model to better predict genome composition. This is unexpected, as in comparison with YN98 + F61 there is no additional composition specific feature in our modeling. First, our estimates of the evolutionary processes acting on genome composition in all these species are in agreement with the recent findings of Hershberg and Petrov (2008) and

Downloaded from http://gbe.oxfordjournals.org/ by Fanny Pouyet on August 26, 2016

Genome Biology and Evolution SMBE

Hildebrand et al. (2010), as $GC_N^\star$ is low, indicating a bias toward AT in the mutational process. Second, our model describes more accurately how GC3 is more biased than GC. Interestingly, although this higher variability of the third codon position is often hypothesized to come from mutational processes unrestricted by selection (as is the case in first and second positions of codons, e.g., Muto and Osawa 1987), SENCA explains most of this variability through selection on CUB. On the other hand, the influence of nucleotide processes is stronger when considering the global genome composition, as CUB has a much weaker impact on the first and second positions.

The SENCA approach allows us to draw conclusions with respect to the relative influence of selection and mutation on codon usage. In our analysis, multiple AT-rich pathogens have very similar *GC** values, which are decomposed in different effects of each layer. We also show that the *A* layer effects is prominent in GC-rich species, with an amino acid composition depleting the genome in GC, whereas the *A* layer is quantitatively less important in AT-rich species. Finally, our results clearly indicate that CUB is driven by the *C* layer (fig. 6). These differences may arise from differences in host, population size or species evolutionary history (Losada et al. 2010).

Globally, our results on intraspecific data can be interpreted in the context of the current thinking that mutations are universally biased toward AT. For middle and low GC species, we observe a quite constant effect of the *N* layer with a partial equilibrium GC of 30%, in agreement with Hershberg and Petrov (2010) and Hildebrand et al. (2010). The *C* layer effect on *GC**, on the contrary, goes smoothly upwards with increasing observed GC content. Then, it appears that non-GC-rich species all share the same nucleotide processes, and their actual GC content depends on the level of selection on CUB.

A surprising result is the inversion of the *N* pattern for GC-rich species. One explanation could be the selection on CUB: In those species, there would be such a strong selection deleting AT-driven mutations, that the *N* layer would stand for substitutions, and not mutations, even though the data are intraspecific. Indeed, comparing *Brucella* and *Bifidobacterium*, two *GC*-rich species with close observed GC, we can see that their *N* layers are very different, and *ENC** is much lower in *Bifidobacterium*, indicating a stronger selection on CUB. Another hypothesis is that nucleotidic processes in those species are more complex; in particular, one may think of GC-biased gene conversion, which may push the GC content of those genomes higher. A third hypothesis would be selection on GC content itself by the environment, an hypothesis hotly debated at the turn of the century (Galtier and Lobry 1997; Naya et al. 2002; Musto et al. 2006; Palmeira et al. 2006) and still driving research nowadays (Reichenberger et al. 2015)

Our interspecific analysis shows that, if the average results on genome composition are quite similar with those of the corresponding species studied intraspecifically (*Salmonella*) and *Escherichia*, the internal evolutionary dynamics can be quite different. This may be related to the evolutionary scale or the rate of fixation of mutations in the intraspecific data. These results emphasize the interest of decomposing the evolutionary signals in layers, as done by SENCA, to better test hypotheses on the evolution of those species.

One future SENCA development is to distinguish gBGC from other genomic signals. This would be particularly relevant for applications to metazoan, where gBGC acts as a spurious mode of positive selection, promoting the fixation of deleterious mutations (Ratnakumar et al. 2010) whereas selection on CUB might also be effective (Gingold et al. 2014). Concerning bacteria, which have been shown to also be subject to gBGC in recombining genes (Lassalle et al. 2015), application of SENCA is also in theory possible but much more difficult because the method would require the knowledge of several site-specific phylogenetic trees, which is hard to infer when between species recombination is strong. Eventually, likelihood inference will have to consider all these trees simultaneously.

Finally, flexibility of our model allows for an investigation of biological questions focused on each particular layer. With SENCA, rates of substitutions between amino acids are only based on a profile of 20 preferences. To be more realistic, an ongoing project is to use empirical matrices of preferences between amino acids, as done in models of protein evolution. But specific matrices will be needed, as in our case overall nucleotide biases are handled by the nucleotide layer and classical protein models already include them. Several methods to model site-specific amino acid fitness have been proposed previously (Halpern and Bruno 1998; Rodrigue et al. 2010; Tamuri et al. 2012), with similar formula for the selection, and it may be straightforward to adapt them on our modeling. However, the additional complexity may prevent the direct estimation of the whole process, and perhaps it will be necessary to estimate this site-specificity in a second step.

The codon layer accounts for the relative preferences between synonymous codons. We could compare these preferences to biological correlates such as tRNA content and gene expression. For example, are the most frequent tRNA in cells linked to codon preferences estimates? There is a known correlation between tRNA content and codon usage (e.g., Kanaya et al. 1999; Rocha 2004). Using SENCA, we could quantify if this correlation is only due to the *C* layer, or if CUB originating from *N* layer has an impact on this correlation. Moreover, nonhomogeneous modeling will permit us to analyze how and when CUB has evolved. This could be applied to cases of genome reduction caused by ecological changes, such as the marine cyanobacteria *Prochlorococcus* (Batut et al. 2014) or *Mycobacterium leprae* (Gómez-Valero et al. 2007).

Last but not least, the evolutionary estimation of CUB by SENCA could be used as a predictive factor instead of observed CUB in multiple applications. One potential application

Downloaded from http://gbe.oxfordjournals.org/ by Fanny Pouyet on August 26, 2016

Genome Biology and Evolution SMBE

is the correlation between CUB and gene expression, and we hope that SENCA will provide a relevant estimator along these lines. Using techniques such as stochastic mapping (Minin and Suchard 2008; Romiguier et al. 2012), it is possible to infer heterogeneous ancestral patterns of evolution from an homogeneous model, and then to infer ancestral gene expression. As an extension of SENCA, we plan to parametrize site-specific selection on codon usage, and use mixtures of these site models to obtain site-specific and gene-specific estimates of the effect of selection on codon usage.

Supplementary Material

Supplementary equation, material, figures S1–S8, and table S1 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by French National Research Agency (ANR) grant Ancestrome (ANR-10-BINF-01-01). F.P. received a doctoral scholarship from Ecole Normale Suprieure de Lyon (http://www.ens-lyon.eu/). This work was performed using the computing facilities of the CC LBBE/PRABI. The authors thank Will Pett for expert English corrections, Bastien Boussau for comments, and Vincent Miele for providing the species tree of HOGENOM complete genomes to process the rooting of our trees. They also thank the anonymous reviewers for their relevant comments that helped us to improve this work greatly.

Literature Cited

Agashe D, Martinez-Gomez NC, Drummond DA, Marx CJ. 2013. Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. Mol Biol Evol. 30(3):549–560.

Akashi H, Eyre-Walker A. 1998. Translational selection and molecular evolution. Curr Opin Genet Dev. 8(6):688–693.

Batut B, Knibbe C, Marais G, Daubin V. 2014. Reductive genome evolution at both ends of the bacterial population size spectrum. Nat Rev Microbiol. 12(12):841–50

Bigot T, Daubin V, Lassalle F, Perriere G. 2013. TPMS: a set of utilities for querying collections of gene trees. BMC Bioinformatics 14:109.

Bulmer M. 1987. Coevolution of codon usage and transfer RNA abundance. Nature 325(6106):728–730.

Charif D, Lobry JR. 2007. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. Biological and Medical Physics, Biomedical Engineering. New York: Springer.

Duret L. 2002. Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev. 12(6):640–649.

Dutheil JY, Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. BMC Evol Biol. 8(1):255.

Galtier N, Lobry JR. 1997. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J Mol Evol. 44(6):632–636.

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol. 14(7):685–695.

Gil M, Zanetti MS, Zoller S, Anisimova M. 2013. CodonPhyML: fast maximum likelihood phylogeny estimation under codon substitution models. Mol Biol Evol. 30(6):1270–1280.

Gilchrist M, Chen W, Shah P, Landerer C, Zaretzki R. 2015. Estimating gene expression and codon-specific translational efficiencies, mutation biases, and selection coefficients from genomic data alone. Genome Biol Evol. 7(6):1559–1579.

Gingold H, et al. 2014. A dual program for translation regulation in cellular proliferation and differentiation. Cell 158(6):1281–1292.

Gómez-Valero L, Rocha EPC, Latorre A, Silva FJ. 2007. Reconstructing the ancestor of *Mycobacterium leprae*: the dynamics of gene loss and genome reduction. Genome Res. 17(8):1178–85.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 11(5):725–736.

Gouy M, Gautier C. 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 10(22):7055–7074.

Gouy M, Grantham R. 1980. Polypeptide elongation and tRNA cycling in *Escherichia coli*: a dynamic approach. FEBS Lett. 115(2):151–155.

Grantham R, Gautier C, Gouy M, Mercier R, Pavé A. 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8(1):r49–r62.

Guéguen L, et al. 2013. Bio++: efficient extensible libraries and tools for computational molecular evolution. Mol Biol Evol. 30(8):1745–1750.

Halpern AL, Bruno WJ. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol. 15(7):910–917.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 22(2): 160–174.

Hershberg R, Petrov DA. 2008. Selection on codon bias. Annu Rev Genet. 42:287–299.

Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet. 6(9):e1001115.

Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. PLoS Genet. 6(9):e1001107.

Kanaya S, Yamada Y, Kudo Y, Ikemura T. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238(1): 143–155.

Karlin S. 2001. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. Trends Microbiol. 9(7): 335–343.

Knight RD, Freeland SJ, Landweber LF. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol. 2(4):RESEARCH0010.

Kosiol C, Holmes I, Goldman N. 2007. An empirical codon model for protein sequence evolution. Mol Biol Evol. 24(7):1464–1479.

Larkin M, et al. 2007. Clustal W and Clustal X Version 2.0. Bioinformatics 23(21):2947–2948.

Lassalle F, et al. 2015. GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands. PLoS Genet. 11(2):e1004941.

Lawrie DS, Petrov DA, Messer PW. 2011. Faster than neutral evolution of constrained sequences: the complex interplay of mutational biases and weak selection. Genome Biol Evol. 3:383–395.

Lobry JR. 1995. Properties of a general model of DNA evolution under no-strand-bias conditions. J Mol Evol. 40:326–330.

Lobry JR. 1997. Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. Gene 205 (1-2):309–316

Losada L, et al. 2010. Continuing evolution of *Burkholderia mallei* through genome reduction and large-scale rearrangements. Genome Biol Evol. 2:10216.

Downloaded from http://gbe.oxfordjournals.org/ by Fanny Pouyet on August 26, 2016

Mayrose I, Doron-Faigenboim A, Bacharach E, Pupko T. 2007. Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates. Bioinformatics 23(13):319–327.
McVean GA, Vieira J. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. Genetics 157(1):245–257.
Minin V, Suchard M. 2008. Fast, accurate and simulation-free stochastic mapping. Phil Trans R Soc B. 363:3985–3995.
Miyata T, Yasunaga T. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J Mol Evol. 16(1):23–36.
Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol. 11(5):715–724.
Musto H, et al. 2006. Genomic GC level, optimal growth temperature, and genome size in prokaryotes. Biochem Biophys Res Commun. 347(1):1–3.
Muto A, Osawa S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci U S A. 84(1):166–169.
Naya H, Romero H, Zavala A, Alvarez B, Musto H. 2002. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. J Mol Evol. 55(3):260–264.
Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3(5):418–426.
Nielsen R, DuMont VLB, Hubisz MJ, Aquadro CF. 2007. Maximum likelihood estimation of ancestral codon usage bias parameters in Drosophila. Mol Biol Evol. 24(1):228–235.
Novembre JA. 2002. Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol. 19(8):1390–1394.
O'Neill PK, Or M, Erill I. 2013. scnRCA: a novel method to detect consistent patterns of translational selection in mutationally-biased genomes. PLoS One 8(10):e76177.
Palidwor GA, Perkins TJ, Xia X. 2010. A general model of codon bias due to GC mutational bias. PLoS One 5(10):e13431.
Palmeira L, Guéguen L, Lobry JR. 2006. UV-targeted dinucleotides are not depleted in light-exposed prokaryotic genomes. Mol Biol Evol. 23(11):2214–2219.
Penel S, et al. 2009. Databases of homologous gene families for comparative genomics. BMC Bioinformatics 10 (6 Suppl):S3.
Pond SK, Muse SV. 2005. Site-to-site variation of synonymous substitution rates. Mol Biol Evol. 22(12):2375–2385.
Ratnakumar A, et al. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. Phil Trans R Soc Lond B Biol Sci. 365(1552):2571–2580.
Reichenberger ER, Rosen G, Hershberg U, Hershberg R. 2015. Prokaryotic nucleotide composition is shaped by both phylogeny and the environment. Genome Biol Evol. 7(5):1380–1389.
Rocha EPC. 2004. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res. 14(11):2279–2286.
Rocha EPC, Feil EJ. 2010. Mutational patterns cannot explain genome composition: are there any neutral sites in the genomes of bacteria? PLoS Genet. 6(9):e1001104.
Rocha EPC, Touchon M, Feil EJ. 2006. Similar compositional biases are caused by very different mutational effects. Genome Res. 16(12):1537–1547.
Rodrigue N, Philippe H, Lartillot N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. Proc Natl Acad Sci U S A. 107(10):4629–4634.
Romiguier J, et al. 2012. Fast and robust characterization of time-heterogeneous sequence evolutionary processes using substitution mapping. PLoS One 7:1–10.
Sharp PM, Emery LR, Zeng K. 2010. Forces that influence the evolution of codon bias. Phil Trans R Soc Lond B Biol Sci. 365(1544):1203–1212.
Sharp PM, Li WH. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol. 24(1-2):28–38.
Sharp PM, Li WH. 1987. The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15(3):1281–1295.
Sharp PM, Tuohy TM, Mosurski KR. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 14(13):5125–5143.
Spielman SJ, Wilke CO. 2015. The relationship between dN/dS and scaled selection coefficients. Mol Biol Evol. 32(4):1097–1108.
Sueoka N. 1988. Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci U S A. 85(8):2653–2657.
Supek F, Skunca N, Repar J, Vlahovicek K, Smuc T. 2010. Translational selection is ubiquitous in prokaryotes. PLoS Genet. 6(6):e1001004.
Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34(2 Suppl):W609–W612.
Tamura K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G + C-content biases. Mol Biol Evol. 9(4):678–687.
Tamuri AU, dos Reis M, Goldstein RA. 2012. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. Genetics 190(3):1101–1115.
Tavaré S. 1986. In: Miura R.M, editor. Location Providence Some probabilistic and statistical problems in the analysis of DNA sequences. Vol. 17. Lectures on Mathematics in the Life Sciences. American Mathematical Society. p. 57–86.
Thomas LK, Dix DB, Thompson RC. 1988. Codon choice and gene expression: synonymous codons differ in their ability to direct aminoacylated-transfer RNA binding to ribosomes in vitro. Proc Natl Acad Sci U S A. 85(12):4242–4246.
Wallace EWJ, Airoldi EM, Drummond AD. 2013. Estimating selection on synonymous codon usage from noisy experimental data. Mol Biol Evol. 30(6):1438–1453.
Wright F. 1990. The 'effective number of codons' used in a gene. Gene 87(1):23–29.
Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol. 46:409–418.
Yang Z, Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol. 25(3):568–579.
Yap V, Speed T. 2004. Modeling DNA base substitution in large genomic regions from two organisms. J Mol Evol. 58(1):12–18.

Associate editor: Ruth Hershberg

Downloaded from http://gbe.oxfordjournals.org/ by Fanny Pouyet on August 26, 2016

TROISIÈME PARTIE

Applications et extensions de SENCA

Dans cette partie, je présente un ensemble de projets qui utilisent SENCA. Un grand avantage de SENCA, implémenté dans Bio++, est qu'il est flexible. J'ai donc testé certaines extensions de SENCA. En particulier, comme vu dans la partie I, il existe plusieurs hypothèses d'origines évolutives du biais d'usage du code. L'origine du BUC peut être sélective (sélection traductionnelle) ou non-adaptative (le BUC résulte de biais de patrons mutationnels ou de biais de fixation lors de la recombinaison). Dans le chapitre 5, je présente une extension de SENCA qui permet de tester l'hypothèse d'origine non-adaptative du BUC (i.e. l'origine due au gBGC). J'ai testé ce modèle SENCA étendu sur un jeu de données Primates pour lequel le BUC résulte du biais de conversion génique mais aussi de l'hypermutabilité des di-nucléotides CpG. Dans le chapitre 6, j'étends SENCA avec un paramètre sélectif site-spécifique. J'ai déjà montré dans la partie II que l'origine du BUC chez les bactéries est essentiellement sélective. Au sein d'un organisme, cette sélection traductionnelle peut être encore plus forte chez les gènes fortement exprimés : l'intensité de sélection traductionnelle varie au sein du génome. J'ai ajouté un paramètre d'intensité du BUC qui varie selon les sites. Enfin, dans le dernier chapitre (chapitre 7), nous verrons que SENCA s'utilise chez de nombreux organismes. SENCA permet de répondre à différentes questions biologiques sur l'origine évolutive du BUC. Dans ce chapitre, nous étudions l'évolution des proaselles qui sont des crustacés dont certaines espèces se sont adaptées au milieu souterrain. Cette adaptation est rapide et nous testons si elle s'accompagne d'un changement de patron mutationnel ou bien par un changement de préférences de codons ou d'acides aminés.

CHAPITRE 5

Le Biais d'Usage du Code chez *Homo sapiens* : extension avec le gBGC et l'hypermutabilité des CpG

Il existe quatre mécanismes principaux pouvant affecter la composition en GC et le biais d'usage du code des gènes chez les Vertébrés dont *Homo sapiens* (Eyre-Walker & Hurst, 2001; Mugal et al., 2015) :

1. le biais mutationnel, le patron de mutation global étant biaisé vers l'un ou l'autre des nucléotides ;

2. la sélection sur l'usage des codons synonymes, qui existe dans le cadre de l'hypothèse de sélection traductionnelle détaillée en introduction ;

3. le mécanisme du biais de conversion génique (gBGC) et la recombinaison de l'ADN. Le gBGC est un biais de réparation des cassures double brin lors de la recombinaison méiotique en faveur des nucléotides G ou C ;

4. l'hypermutabilité des di-nucléotides CpG méthylés et la méthylation de l'ADN. Ce biais est un cas particulier de (1) car il rend compte d'un patron mutationnel qui dépend du contexte.

SENCA tel que présenté précédemment considère déjà le facteur (1) de biais mutationnel (via la couche N) et (2) de sélection sur l'usage des codons synonymes (via la couche C). Dans ce chapitre je propose d'étudier le BUC de l'Homme d'un point de vue évolutif et donc d'étendre SENCA en incorporant un facteur (3) d'intensité du gBGC et un facteur (4) d'hypermutabilité des CpG. Après avoir présenté quelques spécificités génomiques chez l'Homme comparé aux bactéries ou archées, je détaillerai les mécanismes biologiques (3) et (4) qui interviennent sur le BUC. Je présenterai la paramétrisation proposée et le jeu de données test de Primates. Je montrerai en partie l'identifiabilité des paramètres et les résultats de mon modèle étendu.

5.1 Organisation génomique et mécanismes évolutifs chez Homo sapiens

L'espèce Homo sapiens, de l'ordre des Primates est un Mammifère et un Vertébré du domaine des eucaryotes. Les eucaryotes sont des êtres vivants dont les cellules contiennent un noyau, dans lequel est localisé leur matériel génétique. Je vais présenter les caractéristiques principales des génomes d'*Homo sapiens* avant de détailler les mécanismes de conversion génique biaisée (gBGC) et d'hypermutabilité des di-nucléotides CpG.

Organisation génomique eucaryote

Les génomes de mammifères et d'oiseaux montrent des variations de contenu en GC. Ces génomes sont des mosaïques de régions plus ou moins homogènes en GC, ce sont les isochores. Ces dernières ont été découvertes par l'étude biochimique de fragments d'ADN de grande taille ($\approx$ 100kb) (Bernardi et al., 1985; Clay & Bernardi, 2011). Ces variations de GC affectent l'ensemble du génome (Clay & Bernardi, 2011). Cette organisation en isochores le long du génome et leur variabilité à l'échelle des vertébrés posent la question de leur origine évolutive. Historiquement, 3 grandes hypothèses ont été proposées pour expliquer cette organisation en isochores : (1) les isochores se forment via un patron de mutation différent entres les régions, (2) il y a une sélection qui favorise un enrichissement en GC ou AT de certaines régions, (3) les

régions riches en GC sont celles soumises au gBGC sur le long terme, les isochores n'étant que la résultante mécanique du gBGC. La première hypothèse a été développée par Wolfe et al. (1989) et la seconde par Bernardi et al. (1985). Selon Bernardi et al. (1985), les isochores seraient la résultante d'une adaptation à la température chez les homéothermes : les paires de GC (3 liaisons covalentes) étant plus stables thermiquement que les paires AT (2 liaisons covalentes), les gènes des régions GC-riches seraient ainsi protégés de pressions environnementales. La troisième hypothèse est la plus récente, proposée par Galtier et al. (2001). Le mécanisme sera détaillé dans le paragraphe suivant 5.1.
Enfin, à une échelle plus fine, les gènes eucaryotes sont eux-mêmes organisés différemment de ceux des procaryotes. La figure 5.1 détaille les particularités eucaryotes :

- Le CDS (Coding DNA Sequence) est la partie traduite en protéine. Le CDS est une sous-partie du gène. Ce dernier comporte en plus du CDS, les introns et certaines régions régulatrices.

- L'organisation exonique des gènes permet un épissage alternatif : les gènes peuvent être formés d'exons (régions transcrites qui sont retenues dans l'ARNm mature) et d'introns (non-traduites). Les exons ne sont pas forcément tous traduits. Prenons l'exemple d'un gène à trois exons, en théorie, il peut produire jusqu'à 7 protéines différentes car les exons sont traduits séquentiellement.

  $$Exon1-Exon2-Exon3$$

  Ceci donne comme protéines possibles : Exon1, 2, 3 mais aussi 12, 23, 13 ou 123. Cet épissage est spécifique des eucaryotes et permet de diversifier le nombre de protéines traduites.

- Les régions 3'UTR et 5'UTR sont les régions transcrites et non-traduites (untranslated en anglais). Elles sont régulatrices et permettent le passage de l'ARNm (m pour messager) du noyau au cytoplasme.

L'organisation du génome humain dépendant du gBGC je propose ici de détailler ce mécanisme qui joue sur le BUC de l'Homme.

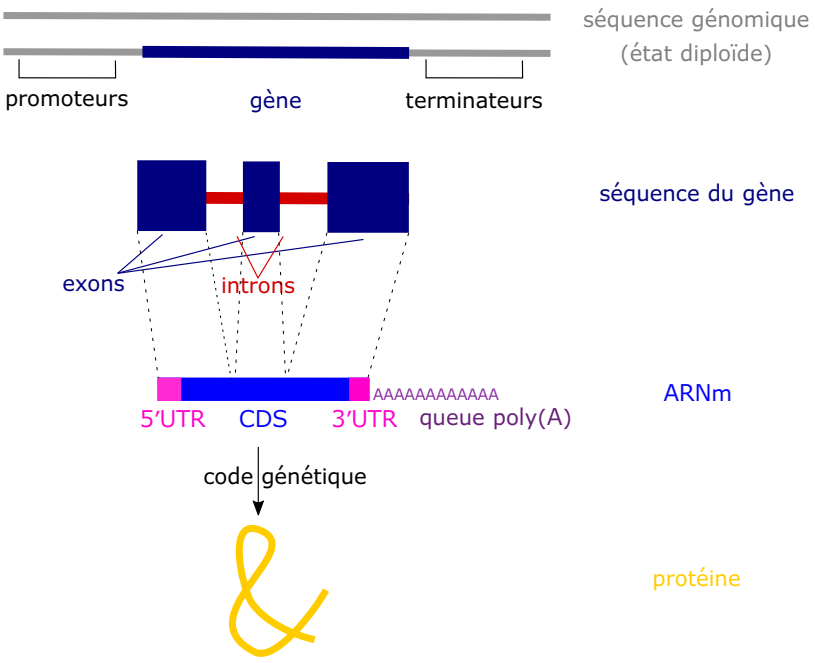


FIGURE 5.1 – Organisation d'un gène eucaryote : transcription et traduction. (1) Le gène se situe sur le génome, il existe des régions régulatrices hors du gène qui permettent de réguler sa transcription (dans les régions promotrices ou terminatrices). (2) Le gène comporte des exons et des introns. L'ensemble du gène est transcrit en ARN messager (ARNm) primaire (3) La maturation de l'ARNm primaire (excision des introns et ajout d'une queue poly(A)) aboutit à la production d'un ARNm mature qui est exporté du noyau au cytoplasme (4) Le CDS porté par l'ARNm est ensuite traduit dans le cytoplasme en protéine.

Le gBGC, déterminant du GC3

Le biais de conversion génique vers GC ou gBGC, est un processus non-adaptatif qui biaise la fixation des mutations lors de la recombinaison.

Le gBGC, conséquence de la recombinaison

Chez les eucaryotes sexués, un nouvel individu est généré à partir d'une gamète femelle et d'un mâle. Les gamètes à la différence de la plupart des autres cellules de l'organisme sont haploïdes (possèdent N chromosomes) et non pas diploïdes (2N chromosomes). Ces cellules sont

"Le gBGC, déterminant du GC3" écrit à partir du manuscrit de thèse de Yann Lesecque (Lesecque, 2014)

transmises à la génération suivante puisque c'est la fusion d'un gamète femelle et d'un mâle qui produit la cellule-oeuf et la restauration de l'état diploïde. Pour passer de l'état diploïde à haploïde, lors de la formation des gamètes, il survient un événement méiotique qui fait suite à une réplication de l'ADN. La méiose est un type de division cellulaire des eucaryotes dans lequel le génome est divisé en quatre (passage d'une cellule diploïde à 4 cellules filles haploïdes). Lors de la méiose, il y a des cassures double-brins le long du génome (DSB, double-strand break), voir figure 5.2. La recombinaison résulte au niveau moléculaire de la réparation de ces DSB qui a lieu en trois principales étapes (Szostak et al., 1983). Premièrement, les brins sont coupés et restent attachés de manière covalente au génome. Deuxièmement, les brins coupés sont excisés de 5' vers 3' ce qui produit deux simples brins du chromosome "receveur" (le chromosome intact est le "donneur"). Les parties simples brins proches physiquement des parties intactes se lient ensuite à leurs chromosomes homologues, c'est l'invasion de brin (SEI, single-end invasion en anglais). Troisièmement, le brin excisé est réparé à partir du brin homologue intact. C'est la recombinaison. Si les brins parentaux sont identiques (homozygotes), la recombinaison n'a pas de conséquence mais si les brins parentaux possèdent deux allèles différents (hétérozygotes) alors la recombinaison induit un événement de mésappariement de nucléotides et donc une réparation ultérieure, voir figure 5.2.

Le gBGC est un biais de réparation des mésappariements qui favorise spécifiquement les allèles GC-riches par rapport aux allèles AT-riches. En effet, les mésappariements sont généralement réparés de manière biaisée vers G ou C. Le gBGC est une conséquence mécanique de la recombinaison. L'intensité du gBGC est donc corrélée entre autre au taux de recombinaison à long terme le long des génomes. Sans plus de détails, il faut noter que le taux de recombinaison n'est pas constant le long du génome : il existe des points chauds de recombinaison où cet événement est plus fréquent (Lesecque et al., 2014). Chez les Mammifères, le gBGC enrichit localement les génomes en GC et détermine un biais d'usage des codons synonymes en faveur des codons avec G ou C en 3ème position (la troisième position étant la position la plus redondante).

Mise en évidence expérimentale du gBGC

Je détaille succinctement comment le gBGC a été mis en évidence chez la levure et le faisceau concordant de preuves du gBGC chez d'autres eucaryotes et notamment chez l'Homme.

Chez la levure, les cellules issues d'une méiose sont groupées en tétrade. Mancera et al. (2008)

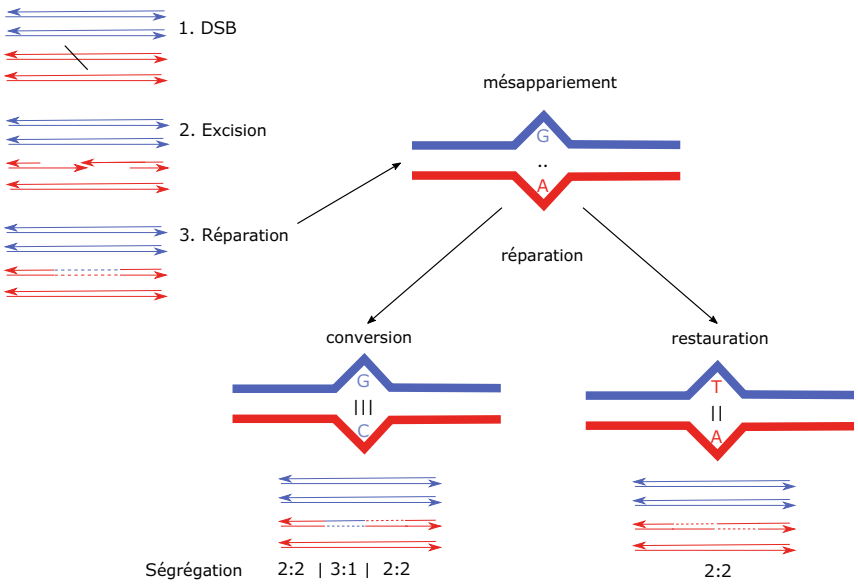


FIGURE 5.2 – La conversion génique d'un mésappariement lors de la recombinaison (gBGC). (1) Une cassure double brins (DSB) d'une chromatide se produit en méiose. Les chromatides soeurs sont de même couleur (bleu ou rouge). Elle est représentée par un trait noir. (2) Les nucléotides sont excisés en 5' et en 3' de ce DSB (3) La réparation de la recombinaison peut se faire à partir du chromatide ayant le même allèle ou non. Les profils de ségrégation du locus sont donnés, soit il y a conversion et la ségrégation est déséquilibrée en faveur du "donneur" soit il y a restauration et aucun biais n'est observé. En rouge, le chromatide "receveur" en bleu, le "donneur", adapté de Lesecque (2014)

a produit une carte à haute résolution des événements de recombinaison suite à l'hybridation de 2 souches polymorphiques de la levure et ainsi a montré qu'il existe un biais de conversion des bases A ou T vers G ou C. La réanalyse de la carte de recombinaison par Lesecque et al. (2013) a permis de mettre en évidence que le gBGC est spécifiquement associé aux événements de crossing-over lors de la recombinaison (l'allèle cassé lors de la recombinaison peut être échangé avec l'allèle du chromosome homologue, ce qui permet le brassage génétique entre générations). Sans gBGC la ségrégation est équilibrée de type "2 :2" alors que le gBGC déséquilibre cette ségrégation en faveur d'un haplotype. Cette ségrégation est de type "3 :1" : l'allèle en excès a converti l'autre. La réparation d'un DSB se fait (1) soit vers le génotype "receveur", il y a restauration, (2) soit vers le génotype "donneur", il y a conversion génique. Parmi les événements de recombinaison au niveau d'une population, certains allèles ont plus de chances que d'autres de convertir leur homologue. Williams et al. (2015) est la première étude qui prouve directement l'existence du gBGC chez l'Homme. Cet article fait suite à un ensemble de preuves

indirectes de l'existence du gBGC pour expliquer les variations intra-génomiques de GC et de BUC chez l'Homme (Galtier & Duret, 2007; Duret & Galtier, 2009; Galtier et al., 2009).

L'hypermutabilité des CpG et les conséquences évolutives de la méthylation de l'ADN

L'hypermutabilité des di-nucléotides CpG est un facteur déterminant des préférences de codons synonymes (Jukes, 1978; Hobolth et al., 2006; Scaiewicz et al., 2006). La question d'intégrer l'hypermutabilité des CpG (le p correspondant à la liaison phosphate entre les bases d'un brin d'ADN) dans des modèles d'évolution de codons ou de protéines a notamment été soulevée par Misawa et al. (2008), Misawa & Kikuno (2009) mais aussi Mugal et al. (2015). Après avoir présenté les mécanismes de mutations dus aux désaminations des 5-méthylcytosines des CpG, je montre en quoi cela peut avoir un impact sur le biais d'usage du code chez les Primates.

La méthylation de l'ADN et l'hypermutabilité de la 5-méthylcytosine

La méthylation de l'ADN est une modification épigénétique courante chez les eucaryotes et les bactéries qui intervient après la réplication de l'ADN. Cette méthylation permet de contrôler la réplication du chromosome ou bien le niveau d'expression du gène, elle est par ailleurs influencée ensuite par des facteurs environnementaux chez les Mammifères. La méthylation de la cytosine du di-nucléotide CpG en 5-méthylcytosine est la forme de méthylation de l'ADN la plus fréquente chez les Vertébrés. Or, la 5-méthylcytosine est 10 fois moins stable que la cytosine non méthylée (Holliday & Grigg, 1993). Elle se désamine spontanément en thymine alors que la désamination de la cytosine forme de l'uracile. Ce dernier est reconnu et éliminé par les enzymes de réparation de l'ADN tandis que la thymine est normalement présente dans l'ADN, de sorte que la conversion de la 5-méthylcytosine en thymine conduit à un mésappariement TG dans l'ADN. Ce mésappariement est ensuite réparé de façon symétrique vers CG ou vers TA, induisant alors une mutation. Bird & Taggart (1980) et Cooper & Krawczak (1989) ont ainsi montré que les CpG ont une fréquence plus faible qu'attendue aléatoirement dans les génomes de Vertébrés.

Le BUC et les contraintes de l'hypermutabilité

L'hypermutabilité des CpG induit des mutations au niveau des gènes donc des changements de patrons de substitutions de codons (Misawa et al., 2008; Misawa & Kikuno, 2009; Mugal et al., 2015). Plus précisément, Misawa et al. (2008) ont montré que l'hypermutabilité des CpG est

à l'origine d'un biais universel de gain et perte d'acides aminés. Ce biais a été proposé par Jordan et al. (2005) (il a été critiqué par Goldstein & Pollock (2006); Hurst et al. (2006) à propos de biais méthodologiques). Mugal et al. (2015) ont récemment mis en évidence que la méthylation de l'ADN via la désamination de la cytosine méthylée a un impact non-négligeable sur la composition en GC des Vertébrés. Misawa & Kikuno (2009) ont évalué l'importance de l'hypermutabilité des CpG sur le patron de substitution de codons chez l'Homme. Cet article propose un modèle qui incorpore l'hypermutabilité des CpG, le rapport κ de transitions sur les transversions et les différentes propriétés chimiques des acides aminés. Ils estiment par maximum de parcimonie les séquences ancestrales de 7 645 gènes le long d'un arbre contenant l'Homme, le Chimpanzé et la Souris. Ils montrent ainsi qu'à peu près 18% des substitutions synonymes résultent de l'hypermutabilité des CpG. Ils concluent ainsi qu'il est nécessaire d'incorporer les mutations des CpG dans les modèles de codons. À ma connaissance, seul le modèle d'évolution de codons Ying & Huttley (2011) intègre cette composante dans l'estimation des fréquences de codons à l'équilibre. Je le détaille dans le paragraphe suivant lorsque je présente la paramétrisation de l'hypermutabilité des CpG.

Dans cette section, j'ai présenté deux importantes composantes de l'évolution du BUC chez les Mammifères et en particulier chez les Vertébrés. Bien que l'objectif de ma thèse n'est pas de proposer un modèle données-spécifiques, l'hypermutabilité des CpG et le gBGC ne peuvent être ignorées dans l'estimation de l'évolution du BUC chez *Homo sapiens*. Je propose donc, dans le but d'étudier l'évolution du BUC chez l'Homme, d'incorporer dans SENCA ces deux facteurs.

5.2 Paramétrisation du gBGC et des CpG

Modèles de gBGC

Nagylaki (1983) est le premier qui propose de quantifier l'impact du gBGC sur une population. Soit une séquence dans laquelle un site est hétérozygote AT/ GC. Sans a priori, l'allèle GC est avantageux dans la moitié des cas (c'est l'hypothèse la plus simple et parcimonieuse qui puisse être faite). La fréquence moyenne des allèles GC des gamètes d'un hétérozygote sans biais vaut $x = \frac{1}{2}$. Cette fréquence est augmentée par le gBGC selon l'intensité du biais de conversion,

notée α et du taux de conversion au locus étudié, noté r :

$$x = \frac{1}{2}(1 + r \times \alpha) \tag{5.1}$$

Par la suite $r \times \alpha$ est noté b, le coefficient de gBGC ($b \in \mathbb{R}^+$). Cette équation est reprise par Galtier et al. (2009). Dans cet article, ils montrent que le gBGC promeut la fixation de changements d'acides aminés délétères chez les Primates. Ils observent en effet un taux de transition de $AT \rightarrow GC$ au moins 5 fois plus important que de $GC \rightarrow AT$. Ce changement étant dirigé préférentiellement vers GC, il peut s'expliquer avec le gBGC. Le gBGC accroît la valeur du paramètre de sélection ω en permettant l'accumulation de mutations non-avantageuses. Mathématiquement, ils se placent dans le cas d'une faible sélection (s) aux sites non-synonymes, c'est-à-dire que $s_{deletere} = -s_{avantageuse}$. Le gBGC affecte aussi bien les sites synonymes que non-synonymes : le coefficient associé à la substitution $AT \rightarrow GC$ vaut $+b$ et de $GC \rightarrow AT$ vaut $-b$. Le taux de substitution d'un codon I vers J s'écrit :

$$q_{IJ} = \mu_{IJ} \times F_{IJ}$$

avec F_{IJ} le taux de fixation :

$$F_{IJ} = \frac{2(\tilde{b} + \tilde{s})}{1 - \exp^{-4N_e(\tilde{b}+\tilde{s})}} \tag{5.2}$$

où N_e la taille de la population efficace, $\tilde{b}$ vaut $+b$ si le changement a lieu de $AT \rightarrow GC$ et $-b$ sinon et $\tilde{s}$ vaut 0 dans le cas d'un changement synonyme et s sinon. En notant, $S = 4N_e s$ et $B = 4N_e b$ et en gardant la même notation du $\tilde{}$, cela s'écrit :

$$F_{IJ} = \frac{(\tilde{B} + \tilde{S})}{2N_e(1 - \exp^{-(\tilde{B}+\tilde{S})})} \tag{5.3}$$

Je reprends la notation de SENCA (p.66) pour la mettre en relation avec celle de Galtier et al. (2009) afin de montrer comment le gBGC peut être inclus dans SENCA. Le taux de substitution entre deux codons s'exprime comme :

$$q_{IJ} = \begin{cases} \mu_{IJ} \times g_{IJ} & \text{, si les codons sont synonymes} \\ \mu_{IJ} \times g_{IJ} \times \omega & \text{, sinon.} \end{cases}$$

avec

$$g_{IJ} = \frac{(\Psi(J) - \Psi(I)) + (\Phi(J) - \Phi(I))}{1 - \exp^{-(\Psi(J) - \Psi(I)) - (\Phi(J) - \Phi(I))}}$$

où $\Phi(I) = \log(d_{AA_I}\phi(I))$ est le logarithme de la préférence du codon I parmi ses codons synonymes et de d_{AA_I} la dégénérescence de l'AA et où $\Psi(I) = \log(\psi(I))$ est le logarithme de la préférence de l'AA codant I. Le parallèle entre SENCA et Galtier et al. (2009) donne :

- Sans gBGC (i.e. $B = 0$), les deux modèles sont semblables en posant $S_{IJ} = (\Psi(J) - \Psi(I)) + (\Phi(J) - \Phi(I))$. La sélection de la mutation de I à J est la différence entre les logarithmes des préférences de codons et d'AA.

- gBGC ($B \neq 0$) intervient lors d'une substitution entre deux codons (à la fois entre codons synonymes et codons non-synonymes). Le taux de substitution est identique à celui du modèle de Galtier et al. (2009) en remplaçant $\Phi(J) - \Phi(I)$ par :

  $$(\Phi(J) - \Phi(I))' = \Phi(J) - \Phi(I) + B \times (N_J(GC) - N_I(GC))$$

  où $N_I(GC)$ est le nombre de bases G et C dans le codon I.

Nous avons implémenté ce modèle dans Bio++ (Guéguen et al., 2013), le paramètre B s'appelle `bgc`. Pour étudier comment se comporte ce nouveau modèle, que nous nommerons par la suite SENCA+bgc, nous effectuerons des tests d'identifiabilité du modèle à l'aide de simulations.

Paramétrer l'hypermutabilité des CpG

Il existe peu de modèles phylogénétiques d'évolution de codons qui paramètrent l'hypermutabilité des CpG. Certes l'article de Misawa & Kikuno (2009) évalue et estime l'effet de l'hypermutabilité des CpG sur les substitutions de codons mais il ne développe pas un modèle d'évolution de codons. L'article de Ying & Huttley (2011) propose un modèle phylogénétique de codons qui distingue les transitions CpG synonymes ou non. En effet, l'hypermutabilité induit toujours une transition : le di-nucléotide cible d'une désamination de CpG est CpA ou TpG.

Le modèle de Ying & Huttley (2011) est un modèle de codons pour lequel le taux de substitution q du codon I au codon J dépend de la fréquence d'équilibre de J (π_J) et du produit des

paramètres affectant les échanges entre codons ($r(I,J)$). Donc, :

$$q(i,j) = \begin{cases} 0 \text{ , s'il y a 2 ou 3 substitutions entre I et J} \\ \pi_J.r(I,J) \text{ , sinon} \end{cases}$$

où $r(I,J)$ est une extension du modèle GTR (Tavaré, 1986) et distingue les échanges incluant des di-nucléotides CpG des autres :

$$r_{GTR+G}(I,J) = \begin{cases} r_{GTR}(I,J) \quad \text{, substitution sans CpG} \\ r_{GTR}(I,J).G \text{ , transition avec CpG} \end{cases}$$

avec r_{GTR} le taux de substitution de I vers J de GTR et avec G le paramètre d'hypermutabilité des CpG. Ils proposent aussi de distinguer les transitions impliquant des CpG ou non mais aussi de distinguer les CpG dans le cas de substitutions non-synonymes. Soit :

$$r_{GTR+G+GK+\alpha}(I,J) = \begin{cases} r_{GTR}(I,J) & \text{, substitution synonyme sans CpG} \\ r_{GTR}(I,J).G\kappa & \text{, transition synonyme avec CpG} \\ r_{GTR}(I,J) \quad .\omega & \text{, non-synonyme sans CpG} \\ r_{GTR}(I,J).G\kappa.\omega & \text{, transition non-synonyme avec CpG} \end{cases}$$

avec κ le rapport des transitions sur les transversions et ω le rapport de substitutions non-synonymes sur synonymes. Ils présentent aussi d'autres paramétrisations de leur modèle, par exemple en distinguant les acides aminés comportant des CpG des autres. Tous ces modèles sont emboîtés et cela leur permet de tester hiérarchiquement la significativité des paramètres. Sur un jeu de données comprenant l'Homme et d'autres Primates, ils mettent en évidence l'existence d'une forte sélection purificatrice agissant spécifiquement sur les codons contenant des CpG.

Leur paramétrisation distingue l'hypermutabilité des CpG lors d'une transition synonyme ($G\kappa$) ou lors d'une transition non-synonyme ($G\kappa\omega$). Autrement dit, l'hypermutabilité est définie selon le contexte comme : $G\kappa > 1$ ou bien $G\kappa\omega > 1$. Dans SENCA, nous proposons de définir l'hypermutabilité avec un autre facteur, ρ ($\rho = G\kappa$). La substitution de I vers J est multipliée par ρ lorsque I contient un CpG et J contient un di-nucléotide cible d'une désamination de CpG c'est-à-dire CpA ou TpG. Cette paramétrisation est différente de celle de Ying & Hutley (2011) car le paramètre ρ est directement l'hypermutabilité d'un CpG vers TpG ou CpA. Par ailleurs, nous ne considérons pas les CpG entre un codon terminant par C et un commen-

çant par G.

SENCA avec gBGC et CpG

Ainsi le gBGC et l'hypermutabilité des Cpg sont inclus dans la fonction g_{IJ} de SENCA comme suit :

$$q_{IJ} = \begin{cases} \mu_{IJ} \frac{(\Psi(J)-\Psi(I))+(\Phi(J)-\Phi(I)+B\times(N_J(GC)-N_I(GC)))}{1-\exp^{-(\Psi(J)-\Psi(I))-(\Phi(J)-\Phi(I)+B\times(N_J(GC)-N_I(GC)))}} & \text{, substitution synonyme sans CpG} \\ \mu_{IJ}\omega \frac{(\Psi(J)-\Psi(I))+(\Phi(J)-\Phi(I)+B\times(N_J(GC)-N_I(GC)))}{1-\exp^{-(\Psi(J)-\Psi(I))-(\Phi(J)-\Phi(I)+B\times(N_J(GC)-N_I(GC)))}} & \text{, non-synonyme sans CpG} \\ \rho\mu_{IJ} \frac{(\Psi(J)-\Psi(I))+(\Phi(J)-\Phi(I)+B\times(N_J(GC)-N_I(GC)))}{1-\exp^{-(\Psi(J)-\Psi(I))-(\Phi(J)-\Phi(I)+B\times(N_J(GC)-N_I(GC)))}} & \text{, synonyme avec CpG} \\ \rho\mu_{IJ}\omega \frac{(\Psi(J)-\Psi(I))+(\Phi(J)-\Phi(I)+B\times(N_J(GC)-N_I(GC)))}{1-\exp^{-(\Psi(J)-\Psi(I))-(\Phi(J)-\Phi(I)+B\times(N_J(GC)-N_I(GC)))}} & \text{, sinon} \end{cases}$$

NOus avons implémenté ce modèle dans Bio++ (Guéguen et al., 2013), que nous nommerons par la suite SENCA+bgc+CpG. Pour étudier comment se comporte ce nouveau modèle, nous allons le tester sur un jeu de données de Primates. Dans un premier temps, je présente le jeu de données puis les résultats de SENCA+bgc+CpG. Enfin, je montre les simulations qui ont été effectuées sur SENCA+bgc et qui prouvent son identifiabilité (les simulations de SENCA+bgc+CpG seront effectuées prochainement). Je tiens dès a présent à préciser que de nombreuses analyses restent encore à effectuer, de nombreuses hypothèses à tester et je détaillerai cela dans le paragraphe de discussion.

5.3 Données analysées

Le jeu de données des Primates

Pour étudier l'évolution du BUC chez l'Homme, j'utilise un jeu de données publié par Kosiol et al. (2008). Il contient les données chez 6 Mammifères dont 3 Primates. Il est de très bonne qualité car les 6 espèces sont entièrement séquencées : l'Homme (*Homo sapiens*), le chimpanzé (*Pan troglodytes*), le macaque (*Macaque rhésus*), la souris (*Mus musculus*), le rat (*Rattus norvegicus*) et le chien (*Canis familiaris*). De plus, la phylogénie des espèces est connue voir la figure 5.3. Lartillot (2013) a montré que l'intensité du gBGC évolue très vite et n'est pas stable à l'échelle des mammifères. Cette intensité vaut en moyenne 0,1 chez les primates et peut aller jusque 10 chez certaines branches telles que les musaraignes (figure 5.4). Il est important ici de préciser

FIGURE 5.3 – Phylogénie des six espèces étudiées : l'Homme, le macaque, le chimpanzé, le rat, la souris et le chien, adapté de Kosiol et al. (2008). Les branches bleues représentent notre jeu de données d'intérêt sur lequel est appliqué notre modèle étendu et dont les résultats sont présentés par la suite.

FIGURE 5.4 – Reconstruction de l'évolution de l'intensité de bgc (paramètre B) chez les mammifères. Les nombres au niveau de chaque branche correspondent au pourcentage d'exons avec $B > 1$ et $B > 10$, tiré de (Lesecque, 2014) et adapté de Lartillot (2013)

que le gBGC joue un rôle sur des millions d'années et donc une valeur faible (par exemple, 0,1) impacte tout de même sur la composition génomique d'équilibre. L'objectif de ce chapitre est de mesurer l'évolution du BUC chez l'Homme et puisque le taux de gBGC varie rapidement en temps, nous restreignons notre analyse de SENCA au sous-ensemble des primates. Néanmoins, ayant accès aux données de 6 espèces et à leur phylogénie, la souris, le rat et le chien sont ici utilisés comme groupe externe de l'analyse. Kosiol et al. (2008) a sélectionné les gènes n'ayant connu aucun événement récent de duplication ce qui nous assure qu'ils suivent la même topologie que l'arbre d'espèces. Au total, l'analyse porte sur 15 236 gènes codant pour des protéines et pour lesquels les données d'expression sont disponibles. Ces gènes sont concaténés par groupe de 100 afin d'avoir suffisamment de signal phylogénétique pour estimer les paramètres de SENCA. Puisque j'étudie le BUC chez l'Homme, j'ai choisi de concaténer les gènes par (1) GC3 croissant qui rend compte de l'usage des codons synonymes puis par (2) niveau d'expression croissant qui rend compte de la sélection sur l'usage des codons synonymes sous l'hypothèse de sélection traductionnelle. Plus précisément, les gènes sont concaténés selon :

- Le GC3 moyen des gènes de l'Homme calculé directement sur les gènes humains issus de Kosiol et al. (2008). J'ai quatre quartiles de GC3.

- L'expression médiane chez l'Homme. Les données d'expression sont les médianes des expressions estimées sur plusieurs tissus en FPKM (Fragments Per Kilobase of transcript per Million mapped reads), voir la table 5.1. Les données d'expression sont obtenues à partir de l'article (Brawand et al., 2011).

- Pour chaque sous-groupe, les gènes sont concaténés par groupe de $\approx$ 100 de manière aléatoire.

Expression (log(FPKM)) \ GC3%	30.9	52.5	66.9	86.9
-0.187	721 (7)	753 (7)	716(7)	788 (7)
0.241	1361 (13)	1009 (10)	842 (8)	873 (8)
0.517	1212 (12)	1052 (10)	907 (9)	915 (9)
1.473	573 (5)	902 (9)	1300 (13)	1312 (13)

TABLEAU 5.1 – Répartition des 15 236 gènes selon leur GC3 médian et le taux médian d'expression (en logarithme base 10). La médiane d'expression est calculée en FPKM (Fragments Per Kilobase of transcript per Million mapped reads) à partir de l'expression chez les mâles et les femelles dans le cerveau, le coeur, le foie et les reins.

La table 5.1 donne le nombre de gènes et entre parenthèses le nombre de concaténats par catégorie.

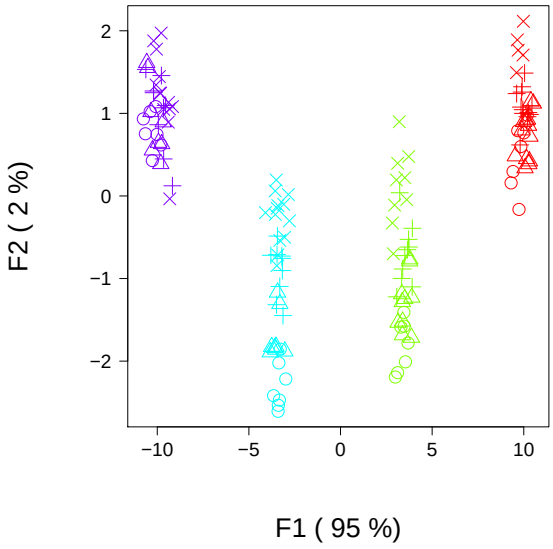
Analyse du jeu de données observé

Pour chaque concaténat, je calcule le RSCU observé des 59 codons sens ayant des synonymes chez *Homo sapiens*. La figure 5.5 représente une analyse en composantes principales des RSCU observés. Le premier axe (F1) explique 95% de la variance et le second axe (F2) 2%. F1 explique donc presque toute la variance, néanmoins F2 n'est pas négligeable car cet axe explique plus de $1/59 = 1.6\%$ de la variance. Comme attendu, F1 sépare les concaténats par classe de GC3 (figure 5.5(a)) et F2 distingue les concaténats selon un gradient de niveau d'expression. La figure 5.5(b) représente les codons sur cette ACP. F1 distingue les codons terminant par G ou C des codons terminant par A ou T (codons GC3 en rouge versus AT3 en noir). Étonnamment, les codons TTG (au milieu à droite en rouge) et AGG (en bas à droite en rouge) sont regroupés selon F1 avec les codons AT3. Ils sont particuliers dans le code génétique (voir figure page 25). En effet, ils codent respectivement pour la leucine et l'arginine, deux codons 6 fois dégénérés (respectivement codés par CTN, TTA, TTG et CGN, AGA, AGG). Les codons TTG et AGG sont dans la sous-catégorie avec un seul autre codon synonyme ce qui peut peut-être expliquer leur coordonnée selon F1. Il est étonnant de ne pas observer cela pour le codons AGC de la sérine, elle même 6 fois dégénérée. F2 sépare principalement le codon AGG (en bas à droite) des autres codons mais aussi les codons NCG (points en étoile en haut à gauche) des autres qui contiennent un CpG. Ce résultat est concordant avec les précédentes analyses de Scaiewicz et al. (2006) qui ont aussi effectué une ACP sur les RSCU des gènes humains et qui ont montré que le second facteur de leur analyse rend compte de l'effet des îlots CpG sur le BUC humain. Les îlots sont des régions dans lesquelles plus de CpG sont retrouvés que dans le reste

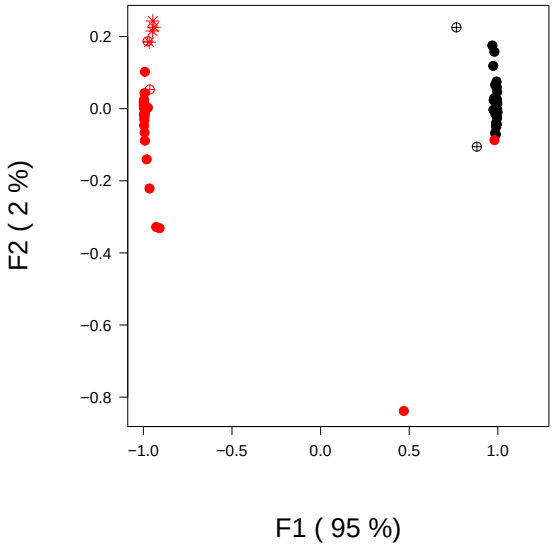
du génome : elles sont généralement peu méthylées et les gènes de ces régions sont exprimés. Dans notre analyse, les gènes les plus exprimés dans chaque classe de GC3 ont un RSCU des codons avec CpG plus fort (en haut) alors que les gènes peu exprimés ont un RSCU de AGG favorisé (en bas). AGG est probablement favorisé car c'est un des 6 codons d'Arginine et qu'il ne contient pas de CpG. F2 rend compte d'un effet CpG sur l'usage des codons.
Sur ce jeu de données, l'hypothèse est qu'il n'y a pas de sélection sur le BUC mais principalement du gBGC qui détermine le GC3 et un effet des CpG. La question est de voir comment réagit SENCA sur ce jeu de données dont l'attendu est connu.

5.4 Résultats

J'estime les paramètres de SENCA+bgc+CpG le long des branches primates, le long du sous-arbre des groupes externes et à la racine. Je fixe les contraintes et paramétrisations suivantes : la couche N est non stationnaire et hétérogène entre les 2 groupes (primates/ groupes externes) ; le modèle nucléotidique appliqué est HKY85 ; les couches C et A, bgc et ρ sont non stationnaires et hétérogènes entre les 2 groupes (primates/ groupes externes). Premièrement, j'effectue des tests hiérarchiques de significativité de préférences de codons, de bgc et de ρ. (1) Sachant que je n'attends pas de préférence de codons, j'ai fixé la couche C à l'hypothèse nulle (i.e. aucune préférence entre codons synonymes) ou bien optimisé ces paramètres pour les Primates. La couche C est toujours optimisée pour le groupe externe (2) Par ailleurs, je m'attends à ce que bgc et ρ rendent comptent des variations de préférences de codons, j'ai donc comparé les estimations de SENCA à SENCA+bgc ou SENCA+bgc+CpG. La table 5.2 montre les comparaisons de vraisemblance entre ces modèles. L'estimation des paramètres de la couche C est significativement informative ; il en est de même pour l'estimation de bgc mais ρ ne l'est que dans la moitié des cas (cas qui sont distribués dans toutes les classes de GC3 ou de niveau d'expression).
Deuxièmement, je regarde les résultats de SENCA+bgc+CpG en effectuant une ACP sur les RSCU d'équilibre (figure 5.6). Le premier axe explique 76% de la variance des résultats, il sépare effectivement les concaténats par classe de GC3 bien que cette variance soit moins forte que pour les données observées. F2 explique 3% de la variance des résultats. F2 ne distingue pas les concaténats par niveau d'expression croissant. La figure 5.6(b) montre les codons sur l'ACP, F1 sépare les codons GC3 des AT3 (respectivement points rouges et noirs) à l'exception de TTG comme pour les données observées (voir paragraphe précédent). F2 sépare les



(a) ACP selon les concaténats



(b) ACP selon les codons

FIGURE 5.5 – Analyse en Composantes Principales (ACP) de l'usage du code chez l'Homme. Les données correspondent au RSCU des 59 codons sens ayant des synonymes chez *Homo sapiens*. F1 correspond au premier facteur de l'ACP qui explique 95% de la variance et, F2 au second facteur qui explique 2%. En (a) chaque point correspond à un concaténat. Les couleurs correspondent aux 4 classes de GC3 : en rouge le GC3 à 30,9%, en vert à 52,5% en bleu à 66,9% et en violet à 86,9%. Les formes des points correspondent aux classes d'expression, les croix à l'expression forte (1,473), les plus à 0,517, les triangles à 0,241 et les ronds à -0,187 (l'expression est en logarithme de FPKM). En (b) la même ACP où chaque point correspond aux coordonnées des codons. Les points rouges correspondent aux codons terminant par C ou G et les points noirs par A ou T. Les points en étoile correspondent aux codons NCG, les ronds avec une croix aux CGN et les ronds pleins aux autres codons.

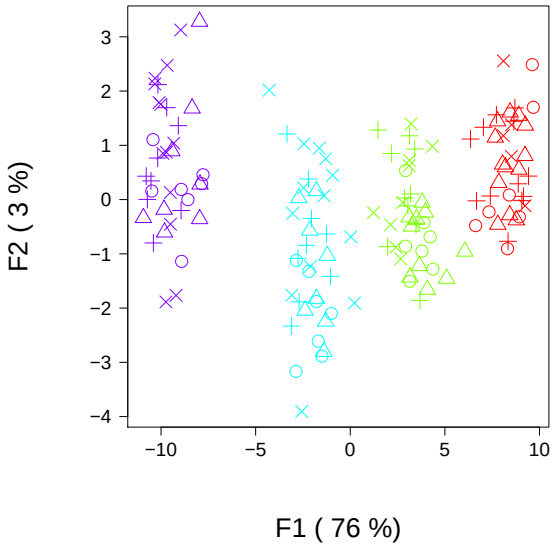
Hyp. nulle	Hyp.alternative	df	Nb de concat LRT<0.05
Couche C neutre +bgc+ ρ	Couche C estimée +bgc+ ρ	41	147/147
Couche C estimée	Couche C estimée +bgc	1	147/147
Couche C estimée +bgc	Couche C estimée +bgc+ ρ	1	75/147

TABLEAU 5.2 – Table de Likelihood Ratio Test (LRT) entre des modèles SENCA avec ou sans bgc, ρ et estimation de préférences de codons. La colonne hypothèse nulle correspond au modèle comportant le moins de paramètres et la colonne hypothèse alternative au modèle avec le plus de paramètres. La différence de degrés de liberté est indiquée dans la colonne df et le nombre de concaténats pour lesquels la p-value du test LRT est inférieure à 0.05 est indiquée dans "Nb de concat LRT <0.05". La p-value de LRT est corrigée par la correction de Bonferroni.

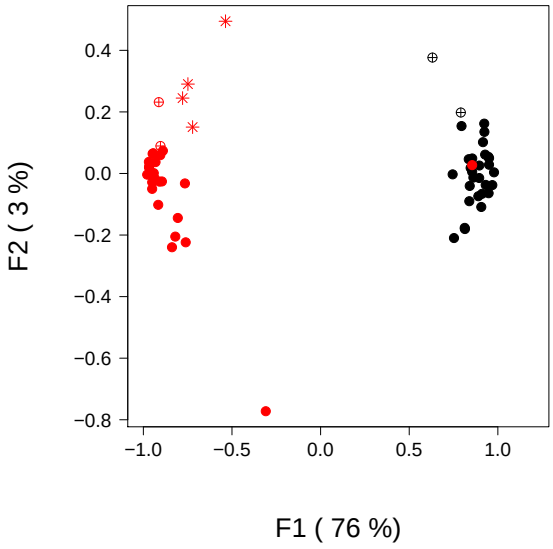
codons contenant des CpG (en haut) des autres. Les distinctions sont globalement aussi nettes que celles sur le jeu de données observées, ce qui est rassurant.

Troisièmement, je m'intéresse à l'étude des biais d'équilibre de GC et de GC3. Dans un premier temps, je vérifie que j'obtiens une corrélation entre $dGC^\star$ (resp. $dGC3^\star$) et $dGC^\star_N + dGC^\star_C + dGC^\star_A$ (resp. $dGC3^\star_N + dGC3^\star_C + dGC3^\star_A$). C'est le cas pour $dGC3^\star$ ($R^2 = 0.94$, p-value $< 2e^{-16}$), figure 5.7(a) mais pas pour $dGC^\star$ ($R^2 = 0.5$, p-value $< 2e^{-16}$, résultat non montré). La corrélation étant forte, je peux étudier la part des couches N, C (qui est estimé uniquement à partir des préférences codons donc sans bgc ni ρ) et A dans la composition globale en GC3 à l'équilibre (SENCA* prend en compte bgc et ρ, voir la discussion), figure 5.7(b). Les concaténats sont ordonnés par classe de GC3 croissant et, au sein de chaque classe, par expression croissante. La couche N est relativement stable entre les différentes classes de GC3 : le patron de mutation est commun le long du génome et tend vers un équilibre GC riche ($dGC3^\star_N > 0$). L'effet de la couche C est importante dans l'estimation du $dGC3^\star$ car c'est la couche C qui varie le plus entre classes de GC3 : les préférences de codons augmentent en intensité dans les classes GC3 riches. Dans tous les cas, au sein de chaque classe de GC3, les différentes classes d'expression sont plus homogènes entre elles que comparées aux autres classes. Cette tendance est retrouvée dans l'ACP où le premier axe sépare les différentes classes de GC3, voir la figure 5.6.

Quatrièmement, l'estimation des deux nouveaux paramètres bgc et ρ, est présentée figure 5.8. (1) L'intensité de la gBGC (bgc) est plus forte dans les classes GC3 riches que GC3 pauvres et de même, lors de la comparaison entre classe de niveau d'expression forte ou faible. (2) Il en est de même pour l'hypermutabilité des CpG, bien que les estimations soient très variables (ce

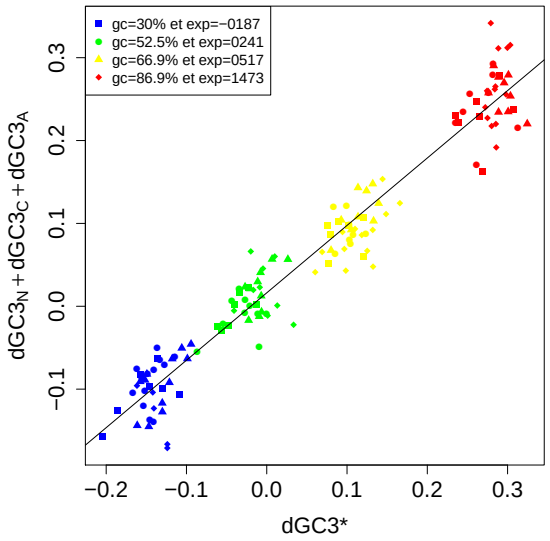


(a) ACP selon les concaténats

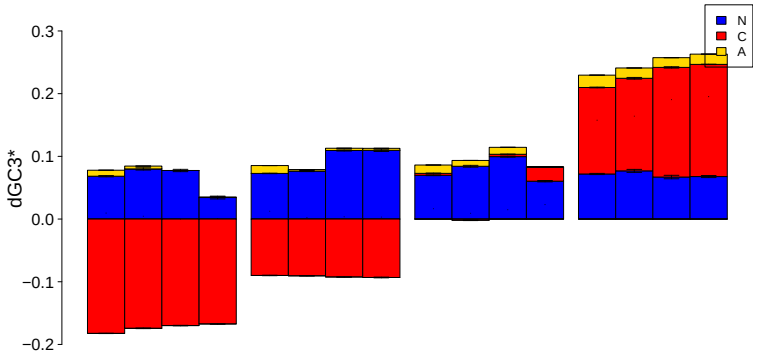


(b) ACP selon les codons

FIGURE 5.6 – Analyse en Composantes Principales (ACP) chez les Primates. Les valeurs utilisées sont les RSCU à l'équilibre de SENCA des 59 codons sens ayant des synonymes des branches primates. F1 correspond au premier facteur de l'ACP qui explique 76% de la variance et, F2 au second facteur de l'ACP qui explique 3%. Les formes et couleurs en (a) (resp. en (b)) sont les mêmes que pour la figure 5.5(a) (resp. (b))



(a) $dGC3^{\star}$ en fonction de $dGC3^{\star}_N + dGC3^{\star}_C + dGC3^{\star}_A$



(b) $dGC3^{\star}$

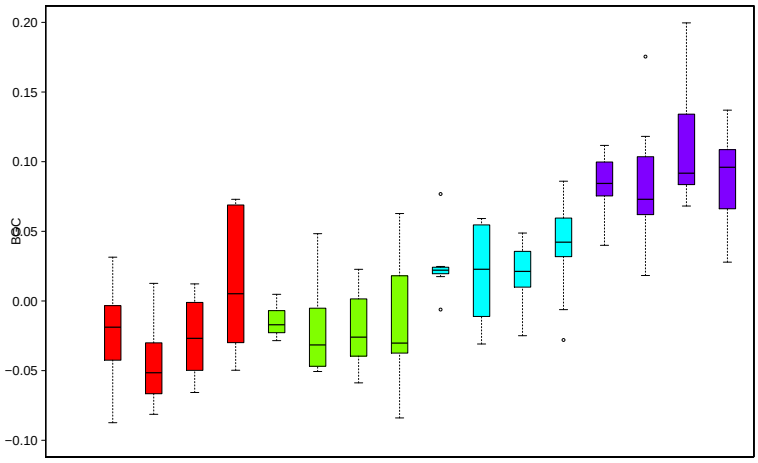
FIGURE 5.7 – Estimation du biais de GC3 à l'équilibre en considérant la couche N, le couche C et la couche A comme dans la partie II. En (a), régression linéaire : $dGC3^{\star} = \alpha(dGC3^{\star}_N + dGC3^{\star}_C + dGC3^{\star}_A)$, avec $\alpha = 0.84$. La corrélation a un $R^2 = 0.94$, p-value $< 2e^{-16}$. En (b) les barplots représentent les contributions relatives de la couche N (HKY85) en bleu, des préférences de codons en rouge et des préférences d'AA en jaune dans l'estimation du GC3 à l'équilibre. Les barplots sont ordonnés en 4 groupes qui représentent les classes de GC3 croissant et dans chaque groupe, un barplot représente une classe d'expression (elles mêmes ordonnées par niveau croissant).

qui s'explique par le fait que ρ n'est pas significativement différent de 0 dans la moitié des cas). (3) L'ordre de grandeur de ρ ne correspond pas à l'attendu et `bgc` n'a pas le signe attendu. Le gBGC doit normalement enrichir le génome humain en GC et donc être positif. Ici, `bgc` est négatif (de l'ordre de -0,1) pour les classes GC3 pauvres et vaut au maximum, 0,2. Néanmoins, l'ordre de grandeur de ce résultat est en accord avec Lartillot (2013) qui estime B autour de 0,1 dans les exons des branches primates. Concernant l'estimation de ρ, (Holliday & Grigg, 1993) a montré que les C méthylées sont 10 fois moins stables que les C non-méthylées, ρ devrait donc être de l'ordre de grandeur de 10. Or, nous l'estimons compris entre 0,8 et 1,5 ce qui n'est pas attendu. Néanmoins, ce paramètre croit avec les classes de GC3 croissant comme attendu. Nous pouvons donc conclure que le comportement qualitatif de ces paramètres correspond à l'attendu (ils croissent par classes de GC3 croissant) mais leur ordre de grandeur ou leur signe, non. Nous détaillerons les stratégies à adopter pour analyser ce résultat en discussion.
Pour résumer, l'extension de SENCA présentée ici estime une forte sélection sur le biais d'usage du code chez l'Homme ce qui est en contradiction avec Galtier & Duret (2007); Duret & Galtier (2009); Galtier et al. (2009) et avec l'étude de génomique comparative de la partie suivante. Néanmoins, SENCA+bgc+CpG met en évidence qu'il existe un patron d'évolution complexe qui joue un rôle sur l'établissement du BUC humain. Avant de discuter ces résultats, nous devons nous assurer de l'identifiabilité de notre modèle. Pour ce faire, le paragraphe suivant présente une analyse par simulation de ce modèle.

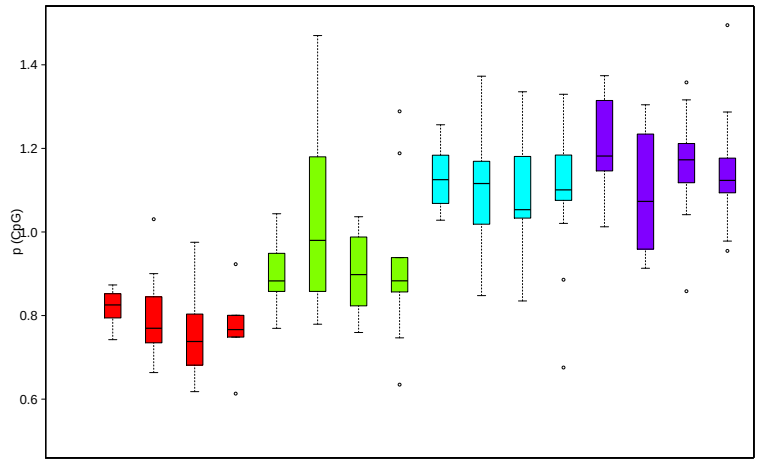
Analyse du comportement de SENCA+bgc

Je montre dans ce paragraphe les résultats de simulations du modèle de SENCA+bgc. Je présente les données simulées puis les résultats des simulations. Suite à des contraintes temporelles, je n'ai pas eu le temps de tester l'identifiabilité du modèle de SENCA+bgc+CpG (la partie CpG a été ajoutée dans un second temps de mon analyse sur le BUC chez *Homo sapiens*).

Le jeu de données simulées Pour valider le modèle SENCA+bgc, je simule des séquences selon l'arbre de la figure 5.3. Les séquences simulées le long de cet arbre ont 20 000 codons. Les simulations ont deux distributions de paramètres de SENCA aux branches, une pour les primates et une pour les autres espèces et une distribution à la racine. Au moment où j'ai réalisé ces premières simulations, j'ai effectué plusieurs hypothèses afin de réduire le nombre de paramètres à estimer, qui ont du sens biologiquement.



(a) Estimation de bgc



(b) Estimation de ρ

FIGURE 5.8 – Estimation des paramètres `bgc` et ρ. Les boxplots représentent les résultats des différents concaténats d'une même catégorie. En (a) l'estimation de `bgc` et en (b) de ρ. Les boxplots rouges sont ordonnés en classes d'expression croissante et ont un GC3 moyen de 30.9%, les boxplots verts ont un GC3 de 52,5%, les bleus de 66,9% et les violets de 86,9%

- La couche nucléotidique est hétérogène. Il y a un ensemble de paramètres pour le sous-arbre des primates ($GC^\star$ est fixé à 0.3, 0.5 ou 0.7), un pour le groupe externe ($GC^\star$ vaut 0.4) et un pour la racine ($GC^\star$ vaut 0.4).

- Le gBGC est faible chez le groupe externe et à la racine (bgc$=0.1$) et variable (bgc $\in [0,2;1;2;3;4;5]$) selon les simulations sur la branche des primates. Les simulations étant antérieures aux résultats présentés ci-dessus (dans lequel bgc ≈ 0.1), le choix des valeurs correspondent aux mêmes valeurs que celles des tests effectués dans Duret & Galtier (2009) ;

- Les préférences de codons sont neutres (i.e pour les AA deux fois dégénérés les préférences valent 0.5 ; pour les AA 4 fois dégénérés elles valent 0.25 etc.). Cette hypothèse se justifie car le BUC chez les Mammifères et en particulier chez l'Homme serait le fruit de processus non-adaptatifs (Duret & Galtier, 2009). Je reviens sur cette notion dans la partie IV ;

- Les préférences d'AA sont stationnaires le long de l'arbre. L'arbre des mammifères est court, la racine est estimée à ≈ 100 millions d'années donc l'hypothèse de stationnarité des préférences d'AA se justifie. Pour les simulations, je prends des préférences d'AA identiques à celles de l'article partie II.

La figure 5.9 présente la paramétrisation des simulations. Pour chaque ensemble de paramètres, je simule 5 séquences différentes avec bppseqgen et j'estime les paramètres de SENCA sur les données simulées avec bppml.

Résultats Les simulations permettent de vérifier que SENCA+bgc est identifiable en pratique en absence de préférences codons. Pour cela, j'ai fait varier le GC de la couche N entre 0.3 et 0.7 mais aussi `bgc` entre 0.5 et 5. Les simulations montrent que quelles que soient les valeurs simulées des paramètres, elles sont retrouvées par estimation. Autrement dit, le modèle est identifiable en ce qui concerne la BGC mais aussi les différentes couches. Je ne me suis intéressée qu'au sous-arbre des Primates.

gBGC : La valeur de bgc estimée le long des branches du sous-arbre des Primates est égale à celle simulée. La figure 5.10 montre l'identifiabilité au niveau des branches bleues de la figure 5.9.

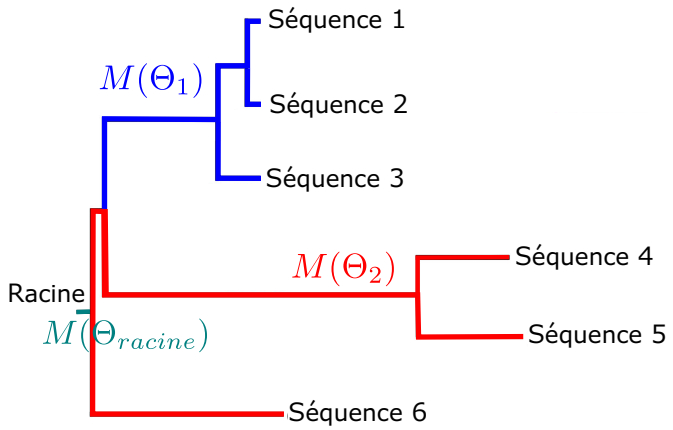


FIGURE 5.9 – Arbre des simulations. Les simulations sont non-stationnaires. (1) la distribution Θ_1 du modèle M, $M(\Theta_1)$, est simulée le long des branches bleues. $M(\Theta_1)$ suit SENCA+bgc avec : bgc $\in [0,2;1;2;3;4;5]$, la couche N suit un modèle T92 avec $\kappa = 2$ et GC $\in [0,3;0,5;0,7]$, la couche C suit un modèle neutre et la couche A a des préférences AA identiques à celles de la partie II, et $\omega = 0.12$ (2) $M(\Theta_2)$ est simulé le long des branches rouges et vaut SENCA+bgc avec bgc=0.1, $\kappa = 2.7$, GC=0.4, la couche C neutre et les préférences AA identiques aux préférences des branches bleues et $\omega = 0.05$. (3) $M(\Theta_{racine})$ vaut SENCA+bgc et suit : bgc=0.1, la couche C neutre, la couche A égale aux préférences AA des branches bleues et rouge. Les branches bleues représentent notre jeu de données d'intérêt sur lequel est appliqué SENCA+bgc.

SENCA est testé dans le cas d'un paramètre bgc faible, bgc= 0.5. Le $GC^\star$ du modèle simulé est retrouvé par SENCA+bgc, notamment lorsque la couche N simulée est GC-riche et enrichit, comme bgc, le GC global des séquences. Il est intéressant de noter dans la figure 5.11, que le GC global de SENCA+bgc est très élevé, entre 0.7 et 0.9 ! Ce résultat explique pourquoi, dans notre analyse sur les données observées, nous obtenons des valeurs de bgc très faibles, j'y reviendrai en discussion. Pour chaque valeur simulée de la couche N : (a) le $GC^\star$ de la couche N est proche de sa valeur simulée, (b) le $GC^\star$ de la couche C et la couche A sont légèrement sous-estimées.

Avec ces résultats de simulations de SENCA+bgc, je montre l'identifiabilité du modèle bgc même si la partie précédente ne me permet pas d'être confiante quant à la bonne paramétrisation de SENCA+bgc+CpG.

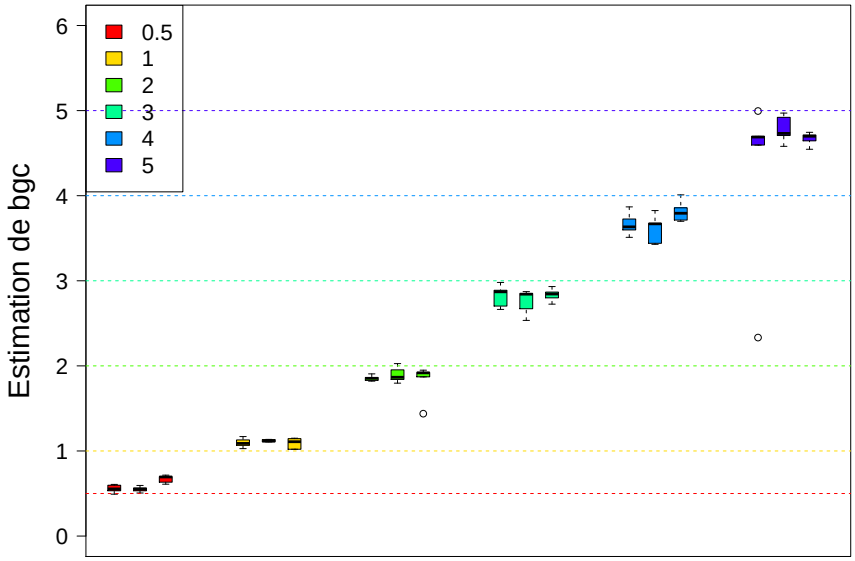
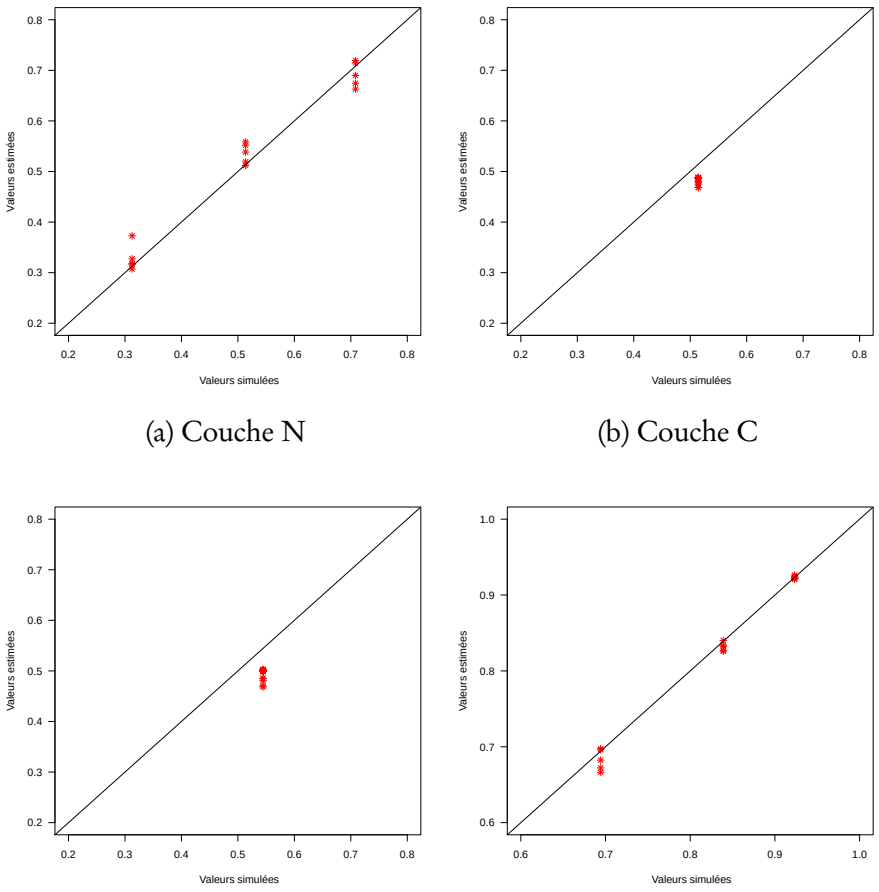


FIGURE 5.10 – Estimation du paramètre bgc dans les simulations. Chaque boxplot représente 5 réplicats indépendants de la même simulation. Pour chaque valeur de bgc, les 3 boxplots représentent la bgc estimé avec le GC de la couche N qui vaut 0.3, 0.5 ou 0.7. La ligne horizontale correspondant à la couleur d'un boxplot représente la valeur simulée du bgc. L'axe y représente les valeurs de bgc estimées. L'axe x ordonne les différents jeux de simulation.

5.5 Discussion et perspectives

Ce chapitre est une étude exploratoire sur les origines évolutives du BUC chez l'Homme. Cette question a longtemps été débattue, les uns estimant que le BUC humain résulte de phénomènes adaptatifs (Bernardi et al., 1985; Gingold et al., 2014), les autres que le BUC est une résultante d'effets non-adaptatifs tels que le gBGC (Galtier & Duret, 2007; Duret & Galtier, 2009; Galtier et al., 2009).

L'analyse des modèles emboîtés prouve que le paramètre bgc est informatif dans notre estimation du BUC humain et ρ dans la moitié des cas. bgc suit le comportement attendu : il est plus fort chez les classes GC3 riches et son ordre de grandeur est compris entre -0,2 et 0,2 ce qui est cohérent avec Lartillot (2013) où il vaut en moyenne 0,1. Néanmoins, le signe de bgc ne correspond pas à l'attendu et l'incorporation de bgc au modèle ne supprime pas l'effet de la couche C (SENCA trouve une forte sélection sur les préférences de codons). Les résultats que j'ai présentés suivent une première analyse dans laquelle j'avais testé SENCA+bgc dans



(a) Couche N (b) Couche C

(c) Couche A (d) SENCA

FIGURE 5.11 – Estimation du GC entre les différentes couches. Estimation (a) du GC de la couche N sans prendre en compte bgc, (b) du GC dû à la couche C, (c) du GC de la couche A et (d) du GC du modèle SENCA+bgc dans le cas où bgc=0.5. Les points représentent différentes simulations, il y en a 5 par valeur de paramètre.

laquelle les résultats étaient similaires, bgc étant négatif chez les concaténats AT3 riches (de l'ordre de -0,5) et positif chez les GC3 riches ($\approx$ 0,5). Pour mieux comprendre cette différence de signe avec l'attendu, nous avons décidé d'ajouter l'effet ρ. L'estimation du bgc avec ρ est comprise entre $[-0,1;0,1]$, alors que sans ρ il est compris entre $[-0,5;0,5]$. Les valeurs de bgc sont très faibles mais comme nous l'avons vu dans les simulations, bgc compris entre 0,5 et 5 mène à un $GC^\star$ compris entre 0,7 et 0,9, ce qui est bien au dessus du GC moyen chez l'Homme (autour de 0,55). Par ailleurs, notre paramétrisation de l'hypermutabilité des CpG

peut être améliorée, par exemple en incorporant un effet contextuel (i.e, cette hypermutabilité existe lorsqu'un codon terminant par C suit un codon commençant par G). Cette amélioration néanmoins n'est pas simple à mettre en oeuvre car les codons sont considérés comme des sites indépendants. Ces estimations différentes de l'attendu expliquent certainement pourquoi nous observons une forte sélection sur l'usage des codons.

Dans le cas où la couche C est fixée sans préférence de codons, bgc varie entre -0,4 et 1,4 et ρ entre 0.8 et 1,2. Ces résultats sont similaires aux résultats présentés ici.

Il est important de noter qu'avec bgc et ρ, la relation linéaire qui existe $dGC^\star \propto (dGC^\star_N + dGC^\star_C + dGC^\star_A)$ est moins importante que pour $dGC3^\star$ ($R^2 = 0.54$ et $R^2 = 0.94$ respectivement, p-values$< 2e^{-16}$). Ce résultat est étonnant car nous aurions pu nous attendre à ce que bgc et ρ aient des impacts aussi important sur le $dGC3^\star$ que le $dGC^\star$. Plus de travail sera nécessaire pour comprendre ce résultat.

L'ensemble de ces résultats indiquent que (1) les effets non-adaptatifs sont informatifs pour définir l'évolution du BUC chez l'Homme mais que (2) il reste à prendre en compte un phénomène adaptatif ou non, qui mime la sélection sur l'usage du code. Il est indispensable d'essayer d'autres paramétrisations de bgc et de ρ. La première difficulté il me semble, est que lors des simulations le modèle est identifiable mais sur des données les résultats ne sont pas cohérents avec la littérature. Les conclusions biologiques de ces résultats doivent donc être effectuées avec précaution. En ce qui concerne les données observées, une première analyse pourrait être de restreindre notre analyse au sous-arbre des primates et supprimer l'homogénéité entre Homme, Chimpanzé et Macaque. Nous pourrions par ailleurs fixer la couche A à la stationnarité ce qui nous permettrait de réduire le nombre de paramètres à estimer. Enfin, sachant que l'hypermutabilité des CpG a été quantifiée, nous pourrions tout à fait fixer ρ à 10, qui est la valeur attendue. Concernant les simulations, une analyse indispensable est de refaire des simulations en incorporant ρ et en simulant des valeurs de bgc et ρ du même ordre de grandeur que dans notre jeu de données observées. En d'autres termes, effectuer des simulations avec bgc variant de -0,5 à 0,5 et ρ de 0,8 à 10.

Les paramètres bgc et ρ sont pour le moment étudiés indépendamment de N, C et A. La question est alors de définir ce que représentent N, C et A une fois que bgc et ρ sont pris en compte. En effet, bgc est une modification du patron mutationnel (couche N) mais mime un effet de la sélection (couche C).

Ce travail est prometteur car nos paramètres sont identifiables en pratique par des tests de simulation (au moins concernant `bgc`) et facilement interprétables. De nombreux tests restent à effectuer puisque nous sommes dans une démarche très exploratoire, notamment pour préciser l'origine de ce fort effet de la couche C dans l'estimation du GC3 d'équilibre. Je vais maintenant présenter une autre extension qui ne s'appuie pas sur le savoir scientifique du BUC chez l'Homme mais sur la connaissance de l'existence de sélection traductionnelle sur le BUC chez les bactéries. Je présente un autre travail d'extension de SENCA avec un paramètre `expr` de force de sélection traductionnelle.

CHAPITRE 6

Extension : Expression de gènes

Dans ce chapitre, je m'intéresse à l'effet de l'expression des gènes sur l'évolution de leur biais d'usage du code au sein d'un organisme. Chez certaines espèces dont les bactéries, le BUC est plus fort pour les gènes fortement exprimés que pour les autres gènes (Stoletzki & Eyre-Walker, 2007; Hildebrand et al., 2010; Hershberg & Petrov, 2008). Il existe une corrélation entre l'expression d'un gène et son BUC, selon le modèle de sélection traductionnelle du BUC. Je propose ici d'introduire un paramètre d'expression `expr` dans SENCA qui permet de modéliser un BUC commun mais d'intensité variable en fonction des gènes et des sites, l'hypothèse sous-jacente étant que les gènes fortement exprimés ont un BUC plus fort que les autres gènes. Je nomme ce modèle SENCA^{e}. Pour ce faire j'étudie un jeu de données que nous avons déjà vu, celui d'*Escherichia coli* ; en effet la sélection traductionnelle sur le BUC a déjà été mise en évidence chez de nombreuses espèces bactériennes (Ikemura, 1985; Sharp & Li, 1987). Je propose dans un premier temps de définir mathématiquement l'ajout de ce paramètre puis de tester l'identifiabilité pratique de SENCA^{e}.

6.1 Extension : le paramètre `expr`

Modélisation de la sélection traductionnelle

La sélection traductionnelle est un modèle dans lequel l'origine et l'intensité du biais d'usage du code génétique sont expliquées par un processus sélectif. Ce type de sélection a été mis en évidence chez certaines espèces dont les bactéries (partie I, paragraphe 2.2). Aucun modèle d'évolution de l'usage du code génétique prenant en compte le niveau d'expression n'a encore été proposé (à l'inverse du chapitre précédent sur la BGC où des modèles existent déjà, voir chapitre 5). Ce travail repose sur des hypothèses simplificatrices, les mêmes que lors du calcul de la statistique CAI présentée en Introduction (p. 41) qui nous permettent de modéliser l'effet de la sélection traductionnelle sur le BUC :

1. il existe un BUC moyen le long des génomes. Celui-ci est identique pour tous les gènes,
2. seule l'intensité du BUC change entre les gènes,
3. l'intensité du BUC est positivement corrélée au taux d'expression des gènes.

Néanmoins, ces hypothèses soulèvent certaines questions. Par exemple, l'hypothèse (1) (un seul BUC le long du génome) par exemple impose que les codons favorisés soient les mêmes pour les gènes fortement exprimés et pour les autres. Cela est une hypothèse forte : est-ce que le BUC moyen génomique correspond à l'utilisation des codons optimaux (voir la définition en partie I)? Les hypothèses (2) et (3) quant à elles nous permettent de choisir la fonction mathématique qui définit comment l'intensité du BUC varie entre les gènes. Je justifie le choix de la fonction mathématique utilisée dans le paragraphe suivant après avoir présenté la paramétrisation de l'expression.

Modélisation selon une fonction de puissance

Sous l'hypothèse de sélection traductionnelle, un gène qui n'est pas exprimé a le même biais d'usage des codons synonymes que le biais mutationnel alors qu'un gène fortement exprimé aura un usage du code extrêmement biaisé. Ainsi, nous cherchons une fonction de l'expression pour laquelle, si l'expression est nulle alors l'usage des codons est uniforme, et si l'expression est très grande alors un seul codon par acide aminé est utilisé. Autrement dit, la force du BUC augmente quand l'expression augmente. La paramétrisation proposée dans ce chapitre suit une

loi puissance et l'augmentation n'est pas linéaire mais exponentielle (à une normalisation près). Le paramètre proposé se nomme `expr` et est un facteur d'intensité du BUC site spécifique. La préférence globale du codon I était précédemment définie comme le produit de la préférence du codon intra-acide aminé et la préférence de l'acide aminé aa, en fonction de la redondance de l'AA : $x_I = \psi(aa_I) \times d_{aa_I} \times \phi_{aa_I}(I)$, voir la partie II. Avec l'intensité du BUC, le terme x_I devient :

$$x_I = d_{aa_I} \frac{\phi_{aa}(I)^{expr}}{\sum_k \phi_{aa}(k)^{expr}} \psi_{aa_I}$$

Cette formule présente plusieurs avantages, notamment (1) la relation entre le taux d'expression (`expr`) et le BUC est positive et (2) l'implémentation est facilement réalisable dans Bio++. Nous avons utilisé une loi de mélange pour modéliser `expr` de façon site-spécifique. La loi de mélange est de moyenne 1 ce qui évite un problème de sur-paramétrisation (le BUC génomique est le BUC moyen entre les gènes). La variation de l'intensité du BUC (i.e. `expr`) suit une loi $\Gamma(\alpha, {}^{\alpha}/_{\beta} = 1)$ de moyenne de 1. La loi Γ est divisée en 4 classes ce qui permet de réduire le nombre de paramètres (il n'y a pas un `expr` estimé par site mais simplement la probabilité d'appartenir à chaque classe). Cette méthode est souvent employée dans les modèles sites-spécifiques car elle n'utilise qu'un paramètre. Ici, les 4 classes de la loi Γ représentent un groupe sans BUC (`expr` très faible), deux groupes avec un BUC moyen et un groupe avec un BUC très fortement sélectionné (`expr` élevé). Je présente ci-dessous les tests de simulation réalisés afin de justifier notre paramétrisation.

6.2 Simulation

Le jeu de données simulé

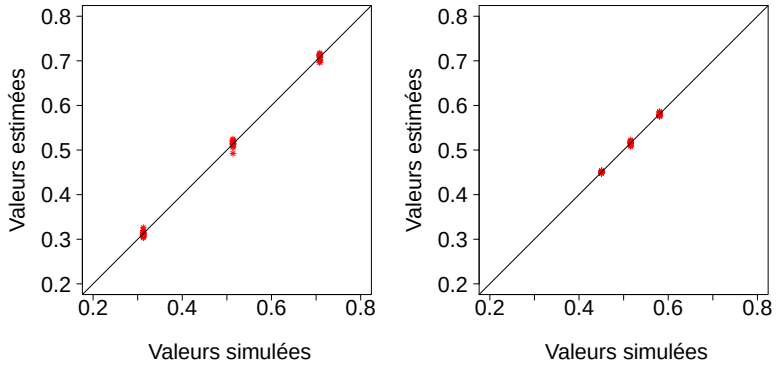
Les simulations ont été effectuées sur un arbre à 23 noeuds (le même que dans la partie II) avec le programme bppseqgen. La racine est neutre, sans préférence ni entre acides aminés ni entre codons et avec $GC_N = 0,4$. Le modèle a $\omega = 0,05$, $\kappa = 2$ et la couche N suit une loi T92 avec $GC_N \in [0,3; 0,5; 0,7]$. GC_C varie dans les mêmes gammes ($GC_C \in [0,3; 0,5; 0,7]$). Dans le cas où $GC_C = 0.5$ alors les codons synonymes ont la même préférence qui vaut l'inverse de la dégénérescence de l'acide aminé qu'ils codent. Les préférences d'AA sont identiques à celles de la partie II. En ce qui concerne `expr`, une difficulté dans cette simulation est le fait que les modèles de mélange ne sont pas disponibles dans bppseqgen. Pour pallier ce pro-

blème, pour chaque jeu de paramètres, j'ai généré 4 séquences de longueur 5 000 pour lesquelles $expr$ vaut respectivement 0,29 ; 0,65 ; 1,07 et 1,98. Ces valeurs représentant les valeurs moyennes des 4 classes divisant en 4 parts égales l'aire de répartition de la loi $\Gamma(\alpha = 2, \alpha/\beta = 1)$. Les 4 séquences sont ensuite concaténées par `expr` croissant et la question est de savoir si SENCA^e retrouve bien chaque site dans sa catégorie correspondante.

Résultats et discussion

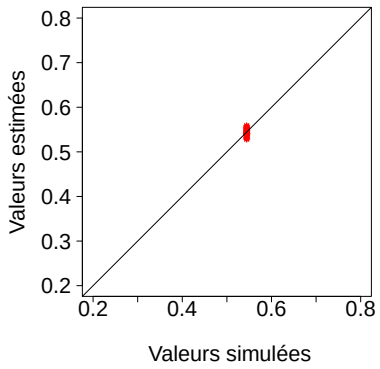
Comme précédemment, les simulations permettent de vérifier que SENCA^e est utilisable sur un vrai jeu de données. Je retrouve effectivement les valeurs simulées pour les couches N, C, A et globalement pour SENCA^e. La figure 6.1 montre l'identifiabilité pratique des GC de chaque couche : les valeurs simulées et estimées sont extrêmement proches. Il faut noter que les valeurs des estimations sont moins bruitées que pour les autres modèles (SENCA ou SENCA+bgc+CpG). Cette netteté des réplicats s'explique car le bruit qui existe entre les différents sites d'un concaténat est absorbé par `expr`.

Nous nous intéressons ensuite aux estimations de `expr` le long des séquences simulées. J'ai divisé chaque séquence simulée de 20000 codons en 200 fenêtres de 100 codons adjacents représentant des gènes artificiels. Pour chaque fenêtre, j'ai calculé la vraisemblance d'appartenir à chacune des 4 classes estimées de `expr` et j'ai calculé la moyenne de `expr` *a posteriori*. Je nomme cette valeur par fenêtre $\texttt{expr}^{\texttt{m}}$. La figure 6.2 représente $\texttt{expr}^{\texttt{m}}$ estimé selon GC_C (panel) et selon GC_N (couleur). Premièrement, GC_N n'a pas d'influence sur l'estimation d'$\texttt{expr}^{\texttt{m}}$ (les lignes se confondent dans chaque sous-figure). Deuxièmement, sans préférence de codons ($GC_C = 0,5$), il est impossible de correctement estimer $\texttt{expr}^{\texttt{m}}$ (figure 6.2(b)). Ce résultat est attendu car l'usage des codons est toujours uniforme quelque soit la valeur de `expr`. Troisièmement, les quatre paliers d'`expr` sont effectivement retrouvés lorsque $GC_C = 0,3$ ou ($GC_C = 0,7$ (figure 6.2(a) ou (c)). Au sein de chaque palier, $\texttt{expr}^{\texttt{m}}$ est bruité autour de la valeur simulée. Le palier `expr` fort est une exception car il n'y a plus de bruit et $\texttt{expr}^{\texttt{m}}$ est proche de 1,98 (la valeur simulée). Ce palier correspond en effet à un BUC d'intensité bien plus forte que pour les autres paliers. Biologiquement, ce palier devrait correspondre aux gènes pour lesquels le BUC est fortement contraint comme les gènes exprimés codant pour les protéines ribosomales par exemple. SENCA^e retrouve la variabilité du BUC en plus du BUC moyen donc ce résultat est très satisfaisant et permet de valider notre approche sur des jeux de données qui possèdent un BUC non uniforme.

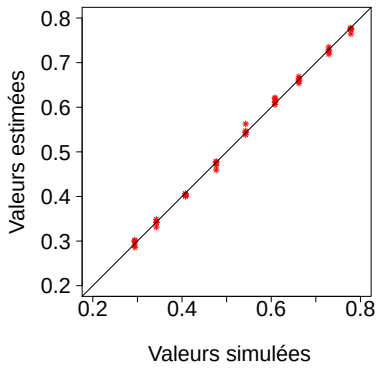


(a) Couche N

(b) Couche C

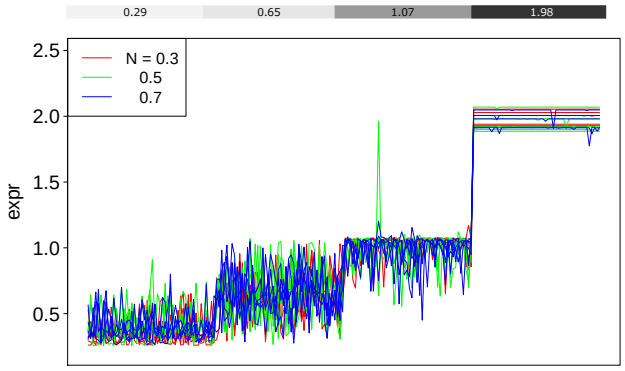


(c) Couche A

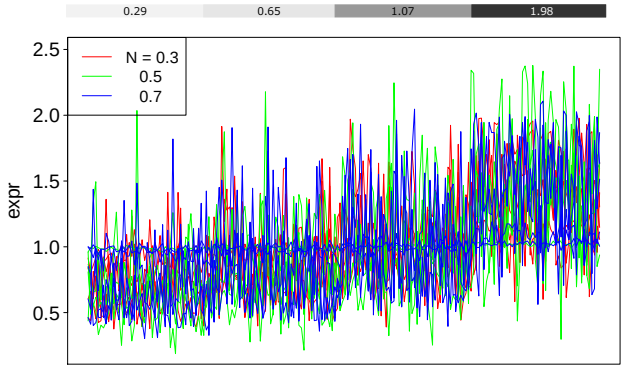


(d) SENCA

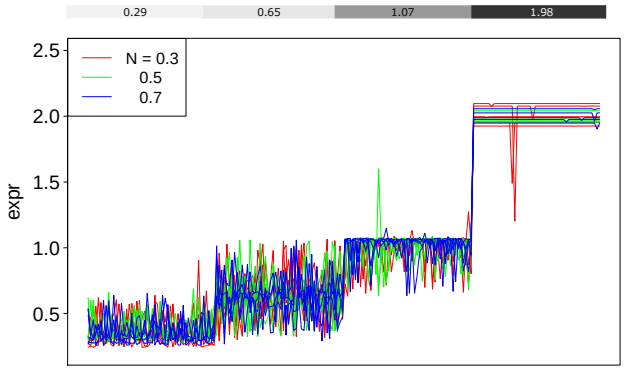
FIGURE 6.1 – Estimation du GC entre les différentes couches. Estimation (a) du GC de la couche N, (b) du GC dû à la couche C, (c) du GC de la couche A et (d) du GC du modèle SENCA Les points représentent différentes simulations, il y en a 5 par couples $\{GC_N, GC_C\} \in [0,3;0,5;0,7]$. Les GC des couches sont définis sur les 61 codons sens ce qui explique qu'un $GC = 0.3$ n'implique pas exactement un GC à 30%. Chaque séquence simulée regroupe 4 sous-séquences de 5000 codons chacune. Ces sous séquences partagent un même GC_N, GC_C, préférences AA mais une expression $\in [0,29;0,65;1,07;1,98]$. Au total, une séquence fait 20000 codons.



(a) $GC_C = 0,3$



(b) $GC_C = 0,5$



(c) $GC_C = 0,7$

FIGURE 6.2 – Estimations d'**expr**^{m} selon les différents jeux de simulations. Chaque sous-figure représente **expr**^{m} *a posteriori* de 100 sites adjacents estimé dans une classe de GC_C. (a) $GC_C = 0,3$, (b)$GC_C = 0,5$ et (c) $GC_C = 0,7$. Chaque ligne de couleur représente une séquence simulée. Les lignes rouges représentent les séquences simulées avec $GC_N = 0,3$ (il y en a 5), les vertes avec $GC_N = 0,5$ et les bleues avec $GC_N = 0,7$. L'axe x représente les index d'**expr**^{m} calculé le le long de la séquence. Une représentation schématique de la séquence simulée est proposée au dessus de chaque figure. Le gradient de gris correspond à la valeur de **expr** : les index 1-50 ont un **expr** $= 0,29$, 51-100 $= 0,65$, 101-150 $= 1,07$ et 151-200 $= 1,98$.

6.3 Perspectives

Le travail présenté ici est très encourageant mais n'est pas encore abouti, il reste de nombreux tests à effectuer avant de le présenter à la communauté scientifique. Notamment, je vais par la suite effectuer des estimations de `expr` sur les gènes d'*Escherichia coli* dont le niveau d'expression est connu. Par exemple Nuñez et al. (2013) fournit des données d'expression que je peux utiliser sur l'arbre des *Escherichia coli*, voir la partie II. Je peux aussi vérifier si les résultats sont similaires entre l'arbre d'*E.coli* et celui des *enterobactéries*; alors l'expression ne change pas rapidement au cours du temps. SENCA^e est un projet attractif car il permettra de :

1. tester si `expr` est plus corrélé au taux d'expression des ARNm ou bien au nombre de protéines produites;
2. estimer le niveau d'expression de gènes putatifs;
3. étudier l'évolution du niveau d'expression voire de reconstruire le niveau d'expression ancestral.

Toutes ces raisons sont bien évidemment motivantes pour continuer l'étude de SENCA^e, la première étant il me semble la plus simple à mettre en oeuvre.

CHAPITRE 7

Application : Les proaselles

SENCA peut-être utilisé dans de nombreux cas comme nous l'avons vu dans les perspectives de l'article de la partie II. Dans ce chapitre, je propose d'appliquer SENCA sur un jeu de données de crustacés, les proaselles. Les proaselles sont un exemple de convergences évolutives multiples et d'adaptation d'une espèce de surface à la vie souterraine (au moins 13 cas ont été référencés). Dans le laboratoire d'Écologie des Hydrosystèmes Naturels et Anthropisés, Clémentine François et Tristan Lefebure s'intéressent aux adaptations génomiques qui ont lieu chez les espèces souterraines. Avec eux, j'ai travaillé sur le jeu de données de proaselles. Après avoir présenté le modèle biologique, je montre que SENCA permet de répondre à certaines questions pour lesquelles d'autres modèles évolutifs ne sont pas informatifs : existe-t-il une adaptation à la vie souterraine sur l'usage du code génétique ? Ou bien, une adaptation sur l'usage des acides aminés ou sur la composition génomique ?

7.1 Le modèle biologique

Les proaselles sont un genre de métazoaires du clade des Arthropodes (animaux invertébrés ayant un corps segmenté et une carapace rigide), du taxon des Pancrustacés (espèces aquatiques), de l'ordre des Isopoda, de la famille des Asellidae (isopodes vivant dans les eaux douces à faible courant). La famille des Asellidae est présente dans les eaux pauvres en pesticides et pour cela, est un bio-indicateur de la qualité des eaux douces. D'un point de vue de génomique évolutive, ce genre présente l'intérêt d'avoir colonisé indépendamment et plusieurs fois deux types très différents de niches écologiques (*i.e.* de milieux environnementaux). Les proaselles vivent dans des milieux aquatiques de surface (organismes épigés) ou bien des milieux souterrains (organismes hypogés). Ces deux environnements présentent des disponibilités trophiques très contrastées : en surface, l'azote et le carbone (deux atomes indispensables) sont présents en quantité illimitée, ce qui n'est pas le cas dans les milieux souterrains. Cette observation est identique quel que soit l'indicateur énergétique ou les nutriments étudiés. De plus, le rapport de disponibilité de l'azote sur le carbone est lui aussi différent entre les deux niches trophiques (Gibert & Deharveng, 2002; Venarsky et al., 2014).
Il existe des différences morphologiques entre les organismes hypogés et épigés. Les organismes hypogés présentent (1) une perte ou au moins une forte régression des yeux (ana-ophtalmie), (2) une dépigmentation, (3) un ralentissement du métabolisme, (4) une résistance au jeûne accrue, (5) une augmentation de la longévité et (6) une facilité de détection de la nourriture en absence de lumière (Poulson & White, 1969; Hüppop, 1987; Gibert & Deharveng, 2002; Hervant & Renault, 2002; Jeffery, 2009; Hüppop, 2012; Soares & Niemiller, 2013). La figure 7.1 montre les différences phénotypiques entre une espèce épigée (Proasellus coxalis) et une hypogée (Proasellus parvulus).

FIGURE 7.1 – Comparaison morphologique de deux proaselles : (a) Proasellus coxalis, qui vit dans un milieu aquatique de surface (photographie extraite de `http://www.naturamediterraneo.com/`) et (b) proasellus parvulus qui vit en milieu souterrain (photographie de `http://america.pink/proasellus_3591081.html`).

Le cladogramme 7.2 représente l'arbre évolutif de 26 Asellidae. Ces 26 espèces vivent depuis plusieurs millions d'années dans ces environnements aquatiques soit de surface soit souterrains. La phylogénie simplifiée de la famille des Asellidae met en évidence l'existence d'au moins 13

"Le modèle biologique" écrit à partir du manuscrit de thèse de Clémentine François (Francois, 2015)

FIGURE 7.2 – Cladogramme de 26 Asellidae. L'arbre raciné contient 24 espèces qui appartiennent au genre des Proaselles (P.) et 2 qui appartiennent à celui des Bragasellus (B.). Le temps est indiqué en centaine de millions d'années. Le rond noir indique les espèces hypogées et le blanc les espèces épigées. Les noeuds internes correspondent à des ancêtres communs de surface. Les barres verticales indiquent les couples d'organismes où une transition d'un milieu de surface vers un milieu souterrain a eu lieu. Figure extraite de la thèse de Francois (2015), issue de Morvan et al. (2013)

transitions environnementales. La colonisation se fait d'un milieu de surface vers un milieu souterrain (et non pas de surface à souterrain) car elle est suivie de pertes morphologiques irréversibles (comme la perte des yeux). Ces pertes ont eu lieu pour toutes les transitions environnementales, ce qui semble indiquer une adaptation aux caractéristiques du nouvel habitat (absence de lumière, faible niveau énergétique du milieu, ...).

CES espèces d'isopodes forment un objet d'étude de choix de génomique évolutive puisqu'elles présentent, d'un point de vue morphologique, de nombreux événements d'adaptation convergente au milieu souterrain. Chacun de ces événements permet de définir un couple épigé/hypogé qui a divergé il y a plusieurs millions d'années. Ce sont des réplicats indépendants d'un changement environnemental similaire vers un milieu faible en ressources trophiques. Ces réplicats montrent une adaptation au niveau morphologique et la question est de savoir s'il existe aussi des patrons d'adaptation au niveau génomique. La phylogénie de ces couples permet d'étudier les pressions génomiques qui ont eu lieu lors du changement de milieu. Existe-t-il un changement commun de patron mutationnel ? Un changement d'usage des codons synonymes ? Ou bien, un changement de fréquence des acides aminés ? Le modèle SENCA permet de tester simultanément ces trois possibilités. Il est important ici de noter que la colonisation du milieu souterrain peut avoir eu lieu après la spéciation des espèces du couple (autrement dit, la transition environnementale a lieu le long de la branche de l'espèce hypogée).

7.2 Matériels et méthodes

J'applique SENCA sur les gènes présents en simple copie chez les 26 espèces de la figure 7.2. Sans hypothèse quant aux pressions évolutions génomiques, je concatène aléatoirement ces gènes. Au total, j'ai 386 gènes soit 4 concaténats de 96-97 gènes chacun pour lesquels les résultats doivent être identiques et peuvent donc être considérés comme des réplicats. Le genre Bragasellus ayant divergé il y a 500 millions d'années, je le considère comme groupe externe. En effet, je cherche à déterminer quels sont les pressions de sélection et les changements de motifs de mutations spécifiques aux espèces hypogées de proaselles. J'applique SENCA de manière non-stationnaire et hétérogène. SENCA est utilisé avec une couche N de type T92. Afin de limiter le nombre de paramètres, j'estime 3 jeux de paramètres pour SENCA (chacun ayant 65 degrés de liberté) :

1. Un jeu de paramètres pour toutes les feuilles qui conduisent aux espèces hypogées ;
2. un jeu pour le groupe externe Bragasellus ;
3. et un jeu pour les feuilles correspondantes aux espèces épigées et aux branches internes,
4. une distribution à la racine.

Cette stratégie permet d'identifier les patrons de substitutions globaux qui ont lieu dans les génomes des espèces hypogées. Il est important ici de préciser que je ne cherche pas à identifier un gène en particulier qui se serait adapté suite à la transition environnementale. Cela a déjà été étudié par mes collaborateurs qui ont notamment mis en évidence que l'opsine, gène impliqué dans la vision, est perdu ou fortement dégénéré chez les espèces hypogées (résultat non encore publié). Les effets attendus sont une moyenne au niveau des gènes qui sont à la fois présents chez les espèces hypogées et épigées.

7.3 Résultats

Dans cette partie, je présente les résultats des espèces épigées et ceux des hypogées des proaselles. N'ayant que 4 concaténats au total, il est difficile d'évaluer la part significative de mes résultats (voir la partie Discussion pour la stratégie à adopter par ailleurs).

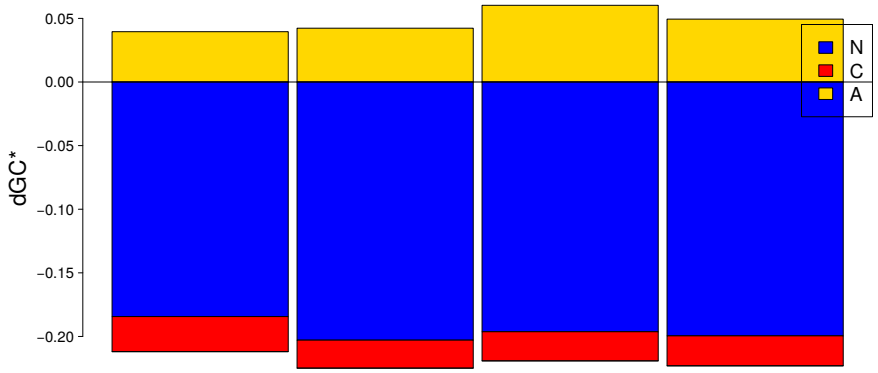
Espèces	ω	κ	$GC^{\star}[\%]$	$GC3^{\star}[\%]$
Epigées	0.070 - 0.077	1.97-2.07	34.7-35.8	30.9-27.9
Hypogées	0.104-0.109	2.23-2.35	30.9-33.2	28.1-29.6

TABLEAU 7.1 – Comparaison de certains paramètres et de certaines estimations à l'équilibre des proaselles épigées et des proaselles hypogées. Les valeurs extrêmes des 4 concaténats sont présentés ici. ω est la fraction des substitutions synonymes sur non-synonymes, le κ est la fraction des transitions sur les transversions, $GC^{\star}$ est la composition génomique globale à l'équilibre exprimée en pourcentage et $GC3^{\star}$ est la composition génomique en troisième position des codons.

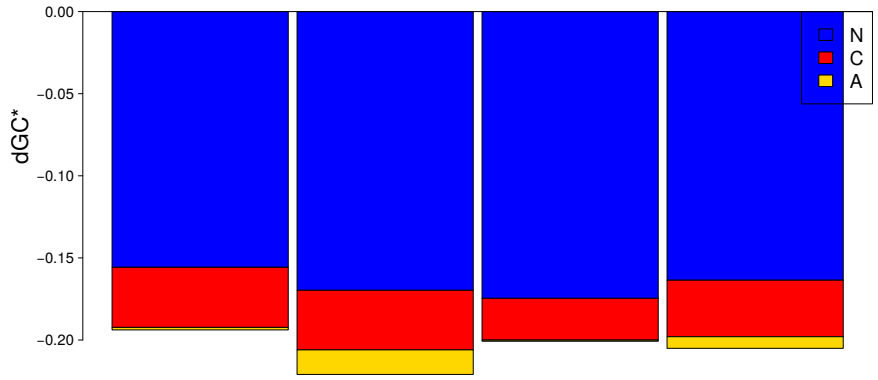
Composition génomique

À l'équilibre, les espèces épigées ont un $GC\star$ compris entre 34,7 et 35,8% alors qu'il est compris entre 30,9 et 33,2% chez les espèces hypogées, voir la table 7.1. Les deux intervalles ne se recouvrent pas, bien qu'assez proches. Qualitativement, les espèces hypogées tendent à un équilibre plus AT-riche. Ces différences se retrouvent dans la quantification de l'importance de chaque couche dans ce GC*. La figure 7.3 montre que la couche N tend vers un biais de composition génomique AT-riche avec $dGC^{\star}_{N} \approx -0,18$ pour les espèces épigées et $dGC^{\star}_{N} \approx -0,15$ pour les hypogées. La couche N a la part la plus importante dans le biais compositionnel total. Le deuxième résultat frappant est la différence de comportement de la couche A entre les espèces hypogées et les épigées. Chez les espèces épigées, (1) la couche A s'oppose à la couche N et (2) la couche A a un effet plus important que la couche C ($dGC^{\star}_{A} \approx 0.05$ est plus grand en valeur absolue que $dGC^{\star}_{C} \approx -0.02$). Chez les espèces hypogées, (3) la couche A est sans effet ($dGC^{\star}_{A} \in [-0.015; 0]$). Troisième résultat, la couche C est plus AT-riche pour les espèces hypogées que pour les espèces épigées ($dGC^{\star}_{C} \approx -0.04$ pour les hypogés alors que $dGC^{\star}_{C} \approx -0.02$ pour les épigées).

Au niveau de la composition génomique en troisième position des codons (i.e. $GC3^{\star}$), les espèces épigées ont un $GC3^{\star}$ qui varie entre 27,9 et 30,9% alors que pour les espèces hypogées $GC3^{\star}$ est compris entre 28,1 et 29,6%. Les deux intervalles se recouvrent, ce qui suggère qu'il n'existe pas de différence significative de $GC3^{\star}$. L'analyse des figures 7.3 et 7.4 montre que (1) $dGC3^{\star}$ résulte principalement de la couche N pour toutes les espèces (les valeurs $dGC^{\star}_{N}$ et $dGC3^{\star}_{N}$ sont égales puisque la couche N est identique au niveau des différentes positions de codons – aux codons STOP près –). (2) $dGC^{\star}_{C}$ est négatif pour toutes les espèces et vaut ≈ -0.05 pour les espèces épigées et ≈ -0.08 pour les hypogées. De plus, (3) les espèces épigées ont un faible effet de la couche A sur l'estimation du $dGC3^{\star}$ en faveur des acides aminés GC-



(a) Décomposition par couche du $GC^\star$ chez les espèces épigées

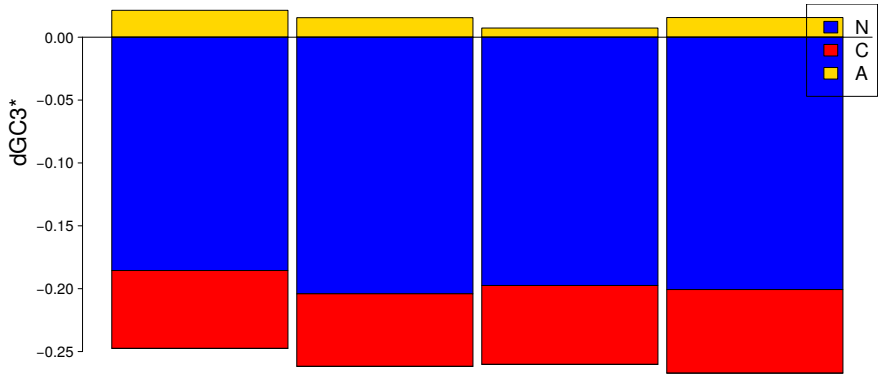


(b) Décomposition par couche du $GC^\star$ chez les espèces hypogées

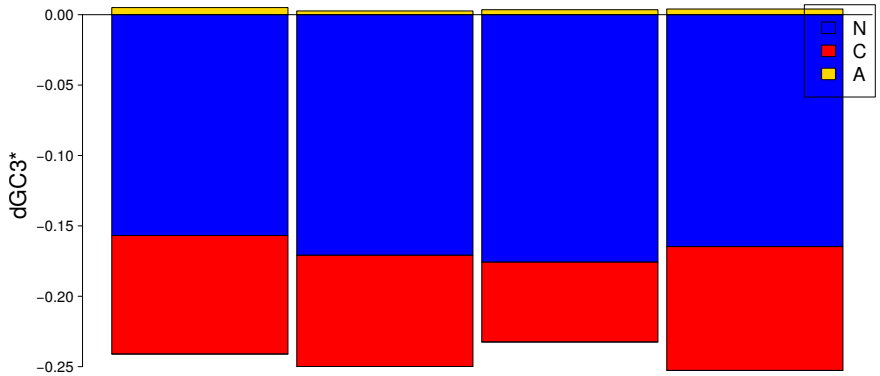
FIGURE 7.3 – Décomposition par couche du $GC^\star$ chez les proaselles (a) épigées et (b) hypogées. Les 4 barplot représentent les 4 concaténats testés, la part de la couche N est représentée en bleu, de la couche C en rouge et de la couche A en jaune. L'axe y représente le $dGC^\star$, soit le biais de composition à l'équilibre de la composition génomique (voir le chapitre II pour la définition mathématique). Si $dGC^\star < 0$ alors l'équilibre est riche en AT alors que si $dGC^\star > 0$ l'équilibre est GC-riche.

riches. Ce résultat suggère que, chez les espèces épigées, l'Isoleucine (ATT, ATA ou ATC) qui a 2 sur 3 codons synonymes terminant par A ou T est défavorisée en faveur du Tryptohane (TGG) et de la Méthionine (ATG), 2 acides aminés pour lesquels le codon termine par G. Au final, (4) le $dGC3^\star$ est similaire entre espèces hypogées et épigées mais la part relative des 3 couches est légèrement différente, en faveur d'une part plus importante de C pour les espèces hypogées.

Enfin, la table 7.1 suggère que des différences de contraintes génomiques existent entre espèces épigées et hypogées. κ qui est le rapport des transitions sur les transversions, diminue de 2,3 à



(a) Décomposition par couche du $GC3^{\star}$ chez les espèces épigées



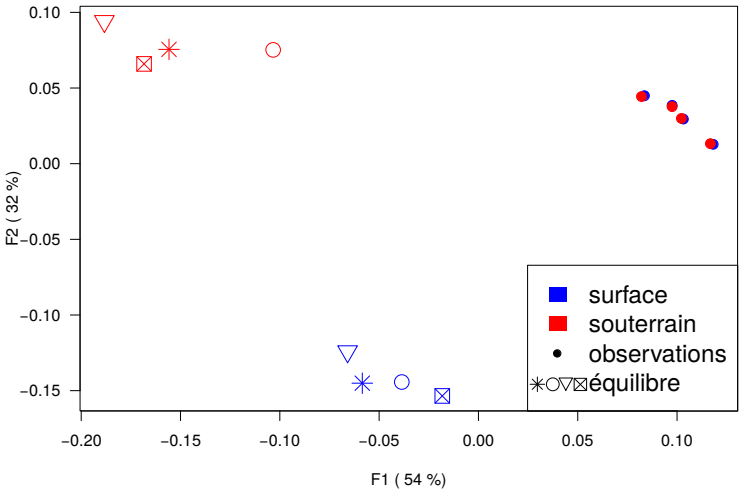
(b) Décomposition par couche du $GC3^{\star}$ chez les espèces hypogées

FIGURE 7.4 – Décomposition par couche du $GC3^{\star}$ chez les proaselles (a) épigées et (b) hypogées. Les 4 barplot représentent les 4 concaténats testés, la part de la couche N est représentée en bleu, de la couche C en rouge et de la couche A en jaune. L'axe y représente le $dGC3^{\star}$, soit le biais de composition à l'équilibre de la composition génomique en troisième position des codons (voir le chapitre II pour la définition mathématique). Si $dGC3^{\star} < 0$ alors l'équilibre est riche en AT alors que si $dGC3^{\star} > 0$ l'équilibre est GC-riche.

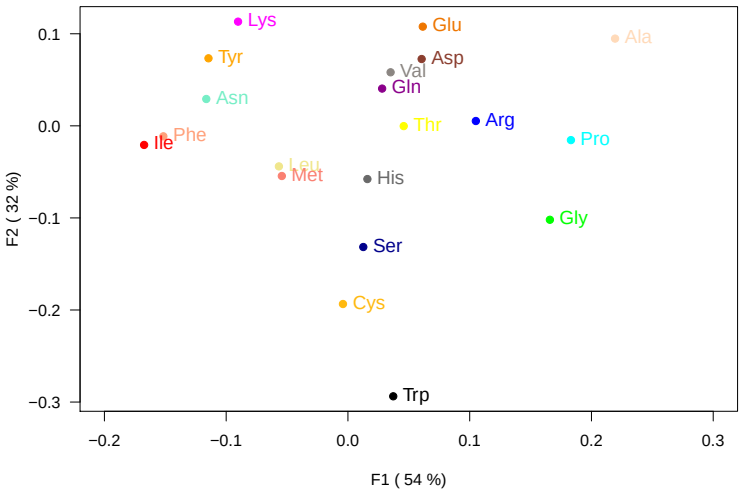
≈ 2 des espèces de surface aux souterraines. Cette diminution peut résulter (1) d'une augmentation des transitions ou (2) d'une diminution des transversions. Ce changement est accompagné d'une augmentation de ω puisqu'il passe de 0,07 à 0,10 entre surface et souterrain. Une augmentation de ω peut résulter (1) d'une augmentation des substitutions non-synonymes et (2) d'une diminution des substitutions synonymes.

SENCA permet d'estimer les codons pour lesquels le comportement (i.e. fréquence à l'équilibre) est différent entre les deux types d'espèces. La figure 7.5 représente une analyse factorielle des correspondances (AFC) effectuée sur les nombres d'acides aminés observés et estimés à l'équilibre des espèces épigées et hypogées des 4 concaténats. Les effectifs d'AA à l'équilibre rendent compte de l'effet de la couche N et A mais pas des variations de préférences entre codons synonymes. Chaque concaténat a un ensemble de paramètres observés et un d'équilibre pour les espèces épigées mais aussi un ensemble observé et un d'équilibre pour les espèces hypogées. Au total, il y a donc 4x4 jeux de paramètres qui forment 4 couples (épigées et hypogées) de données observées et 4 couples de résultats à l'équilibre. D'une part, la figure 7.5(a) montre que les effectifs d'AA observés entre espèces épigées et hypogées sont très similaires entre les différents concaténats. Le premier axe F1 explique 54% de la variance et l'axe F2 32%. Les deux axes permettent de séparer trois sous-groupes : les données observées, les estimations à l'équilibre des espèces de surface et les estimations des espèces souterraines. SENCA propose une distinction entre les deux types d'espèces, ce qui n'est pas possible d'observer lorsque seule les données observées sont disponibles. D'autre part, l'analyse de la figure 7.5(b) indique que les espèces souterraines tendent à être enrichies en acides aminés AT riches (Lysine Lys, Tyrosine Tyr, Asparagine Asn, Phénylalanine Phe et Isoleucine Ile) alors que les espèces de surfaces le sont en Tryptophane Trp et Cystéine Cys. Ce résultat est en accord avec l'analyse des figures 7.3 et 7.4.

La figure 7.6 montre une ACP effectuée sur les RSCU observés et à l'équilibre. Il faut noter ici que les données de RSCU ne prennent pas en compte l'usage des AA ; à l'équilibre elles sont le reflet de l'effet de la couche C et N. L'ACP explique 30% de la variabilité des données selon le 1er axe et 18% selon le deuxième axe. La figure 7.6(a) montre que le premier axe différencie les observations des estimations des espèces épigées (en bleu) et celles des espèces hypogées (en rouge). Cette séparation est moins distincte que dans la figure 7.5 mais reste néanmoins observable. La figure 7.6(b) indique les codons qui différencient les espèces. Le résultat frappant est l'absence de patron de changement de préférences de codons entre espèces. Par exemple, les espèces de surface préfèrent utiliser le codon de la Lysine AAG plutôt que AAA mais ils préfèrent CGA plutôt que CGG pour la Proline. Ce résultat est cohérent avec les précédentes observations effectuées sur la figure 7.4. La couche C n'a que peu d'impact sur la différentiation entre les espèces de proaselles.

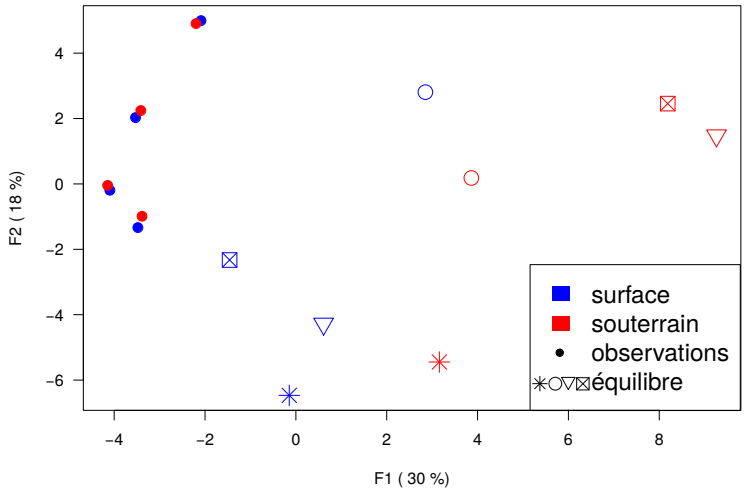


(a) AFC selon les branches et concaténats

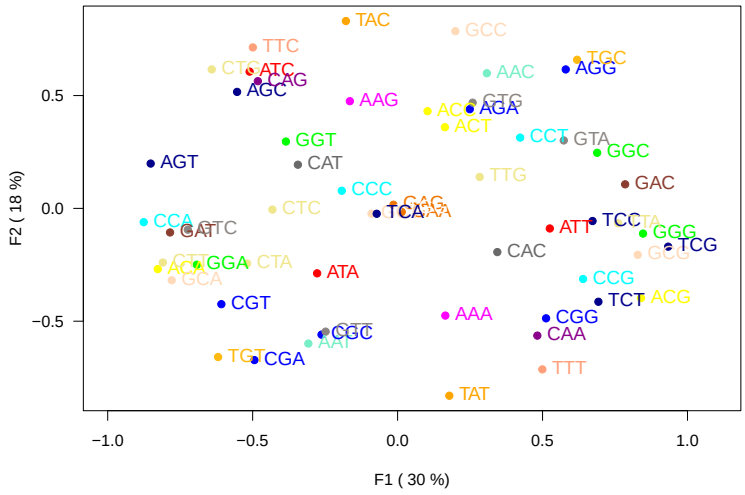


(b) AFC selon les acides aminés

FIGURE 7.5 – Analyse factorielle des correspondances (AFC) sur les proaselles selon (a) les données observées et les estimations à l'équilibre de SENCA sur les branches et les concaténats ou selon (b) les acides aminés. Les données utilisées sont les effectifs d'acides aminés sur les espèces épigées ou hypogées à partir de 4 concaténats et les effectifs d'AA estimés à partir de SENCA sur les espèces souterraines ou de surface à partir des 4 concaténats. L'effectif à l'équilibre est la fréquence de l'acide aminé à l'équilibre multiplié par la longueur du concaténat. En (a) chaque point représente un type d'espèce sur un concaténat : en rouge, les espèces souterraines et en bleu celles de surface. Les points pleins représentent les données observées, les points creux de même forme correspondent à un même concaténat. En (b) chaque point correspond aux cordonnées d'un AA.



(a) ACP selon les branches et concaténats



(b) ACP selon les codons

FIGURE 7.6 – Analyse en composantes principales (ACP) sur les proaselles. Les données utilisées sont les RSCU des 59 codons sens ayant des synonymes estimées à partir de SENCA sur les espèces souterraines ou de surface à partir de 4 concaténats. En (a) chaque point représente un type d'espèce sur un concaténat : en rouge, les espèces souterraines et en bleu celles de surface. Les points de même forme correspondent à un même concaténat. En (b) la même ACP où chaque point correspond aux coordonnées des 59 codons sens ayant des synonymes. Les codons ont la même couleur que leurs AA de la figure 7.5.

7.4 Discussion et perspectives

Discussion

Cette étude a été effectuée sur 4 concaténats de gènes présents en simple copie chez 26 espèces d'Asellidae. Les 4 réplicats fournissent des estimations de paramètres extrêmement proches, ce qui permet l'analyse de nos résultats, si ce n'est statistiquement au moins qualitativement.

SENCA permet de montrer que l'adaptation à un milieu souterrain a un faible impact sur les compositions génomiques chez les proaselles. Néanmoins, cette adaptation environnementale modifie le patron de sélection des substitutions synonymes et non-synonymes. En effet, nous avons vu qu'il y a une augmentation non négligeable du paramètre ω entre les espèces de surface et les souterraines. Cette modification, avec le contexte biologique connu, semble plutôt résulter d'une augmentation du nombre de substitutions non-synonymes que d'une diminution du taux de substitutions synonymes et indique un possible relâchement de pression de sélection purificatrice due au changement de milieu. Premièrement, il n'y a pas de changement de préférences des codons, le taux de substitutions synonymes doit être constant et l'augmentation du ω résulte donc probablement d'une augmentation du taux de substitutions non-synonymes. Nous pourrions tester cela avec des méthodes de mapping de ces taux le long des branches de l'arbre. Deuxièmement, la couche A induit un biais de composition vers GC pour les espèces épigées et pas pour les espèces hypogées. Ce changement de biais de composition induit un changement de préférences d'acides aminés (voir figure 7.5). Les différences de fréquence d'équilibre des AA sont déterminantes pour différencier les types d'espèces : certains acides aminés sont ainsi enrichis ou évités chez les espèces hypogées. La modification de ces préférences est en accord avec une augmentation du taux de substitutions non-synonymes (ces dernières seraient favorisées par les changements de préférences d'AA).

Par ailleurs, nous avons observé une augmentation de κ avec soit une augmentation des transitions soit une diminution des transversions pour les espèces hypogées comparées aux épigées. L'analyse de ce paramètre n'est pas évidente et son lien avec des facteurs environnementaux non plus. Néanmoins, une augmentation globale du κ indique qu'il y a plus de changements qui ne modifient pas la nature du cycle chimique des nucléotides (les purines restent purines et les pyrimidines restent pyrimidines).

Perspectives

Les proaselles et notamment le sous-groupe étudié ici constituent un ensemble de choix pour analyser l'effet d'un changement drastique d'environnement sur les compositions génomiques. Le travail présenté ici est préliminaire dans le sens où de nombreuses modifications peuvent facilement être réalisées mais aussi où de nombreuses questions restent encore en suspens.
Dans un premier temps, j'ai choisi de n'avoir que 4 concaténats pour être sûre de pouvoir estimer autant de paramètres (185 paramètres au total en plus des longueurs de branche). Il me semble indispensable d'étudier un plus grand nombre de concaténats. Une possibilité serait d'effectuer des concaténats deux fois plus petits (*i.e.* de taille 50 gènes) afin d'augmenter la robustesse de nos résultats.
La couche C étant similaire entre les espèces hypogées et épigées, il pourrait être intéressant d'en tester l'homogénéité. Cela permettrait de réduire le nombre de paramètres de 41 (un seul ensemble de préférence de codons pour les espèces souterraines et de surface). Il serait aussi intéressant d'avoir un jeu de paramètres pour chacune des espèces hypogées. Sachant que ce sont ces espèces qui ont changé de milieu et donc de pressions environnementales, une question est de savoir si les espèces ont adopté la même stratégie adaptative ou bien, si différentes stratégies existent. Avec l'aide de mes collaborateurs, nous pensons définir 3 sous-groupes de proaselles, chaque groupe ayant son jeu de paramètres pour les espèces hypogées. Ces 3 groupes seront définis empiriquement, par la connaissance qu'ont mes collaborateurs sur ce jeu de données.
Dernier point, la couche A est celle qui résume à elle seule les plus fortes différences entre espèces hypogées et épigées. Il est indispensable d'effectuer une analyse plus fine de ces résultats : peut-on relier les différences de préférences d'AA à des facteurs environnementaux tels que la limitation dans la quantité de ressources des milieux souterrains ? Les résultats présentés ici montrent ces différences mais n'autorisent pas à conclure quant aux explications menant à cela. Il faut définir des groupes d'AA par leur nature chimique par exemple et voir si certains groupes chimiques sont favorisés ou évités chez les espèces hypogées.

CHAPITRE 8

Perspectives d'utilisation de SENCA

Dans cette partie, nous avons aussi présenté deux possibilités d'extensions de SENCA, soit avec des paramètres rendant compte du biais de conversion génique et de l'effet de l'hypermutabilité des CpG tels qu'ils existent chez les Mammifères ; soit avec un paramètre rendant compte d'une variation d'intensité de sélection de l'usage du code. Ces deux approches sont très complexes à mettre en oeuvre car il n'existe pas forcément encore de cadre théorique clair. En ce qui concerne le chapitre 5 SENCA+bgc+CpG, nous nous sommes inspirés du cadre proposé par Duret & Galtier (2009) qui modélise le gBGC entre les codons synonymes 4 fois dégénérés sans CpG. Dans le chapitre 6 sur SENCA^e, je n'ai pas trouvé d'approche préexistante que j'aurai pu intégrer. Nous avons aussi vu dans le chapitre 7 que SENCA permet de répondre à diverses questions biologiques. SENCA a aussi déjà été utilisé pour étudier :

1. l'évolution réductive du génome de Prochlorococcus. Cette espèce de cyanobactérie a en effet subit une réduction d'à peu près 30% de son génome par rapport à celui de Synechococcus. Batut (2013) a utilisé SENCA afin d'estimer les différences de pressions de sélection entre les branches de l'arbre phylogénétique de Prochlorococcus et les branches de Synechococcus. SENCA a permis de mettre en évidence que les espèces AT-riches ont une composition nucléotidique principalement déterminée par la couche N alors

que les espèces GC-riches sont dominées par un effet de la couche C.

2. l'évolution des compositions génomiques chez les algues vertes *Ostreococcus* (eucaryote unicellulaire). La composition en GC varie entre 0.5 et 0.65 entre les différentes espèces d'*Ostreococcus*. Mathieu Groussin a observé qu'il existe une corrélation linéaire entre la composition en AA et le contenu en GC, sauf pour une espèces, *O. marinus*. SENCA permettrait dans cette étude de tester si *O. marinus* a subit soit (1) un changement de patron de mutation rapide par rapport aux autres espèces d'*Ostreoccocus* sans qu'il y ait encore eu de changement d'AA soit (2) un changement de BUC fort qui a totalement modifié la corrélation qui existait entre GC et contenu en AA.

3. la co-évolution du BUC du virus de l'immunodéficience humaine (VIH-1) aavec le BUC humain. Le VIH-1 est un lentivirus qui possède un fort biais de composition nucléotidique (plus de 36% de A). En conséquence, le BUC du VIH-1 est très différent de celui de l'Homme (son hôte) alors qu'il utilise les ARNt humains pour traduire ses protéines virales. Avec Marc Bailly-Bechet et Laurent Guéguen, nous souhaitons donc regarder s'il existe une sélection sur le BUC du VIH pour correspondre au BUC humain malgré le fort patron mutationnel qui enrichit la souche virale en A.

Ainsi, ces 3 études prouvent que SENCA est un modèle qui permet d'étudier diverses situations. Ceci est fortement encourageant pour affiner les extensions SENCA+bgc+CpG et SENCA^e qui permettraient une généralisation de son utilisation. Nous avons aussi développé une extension de SENCA avec des distances entre acides aminés (voir en Introduction, p.27). Je n'ai pas détaillé cette extension car nous n'avons encore effectué ni test de simulations ni analyse sur des données réelles. De manière générale, cette extension repose sur la constatation que notre paramétrisation de la couche A est une hypothèse fortement simplificatrice puisque, (1) il existe des distances physico-chimiques ou phylogénétiques entre AA et (2) les sites n'ont pas le même profil. Nous avons pour le moment proposé un paramètre de distance entre AA rendant compte de ces distances. Par ailleurs, puisque certains sites sont contraints en termes de fonction, il serait intéressant de discuter avec Nicolas Lartillot et Nicolas Rodrigue qui développent des modèles avec des profils de substitutions d'AA site-spécifiques tels que Rodrigue et al. (2010) mais qui ne considèrent pas, pour le moment, le BUC. Dans la dernière partie, je vais toujours parler d'usage du code mais d'un point de vue de génomique comparative et sans utiliser de modèle d'évolution de codons. Je cherche à comprendre l'origine de la variation de l'usage du code chez l'Homme.

Quatrième partie

Pourquoi l'usage des codons synonymes varie entre différentes catégories fonctionnelles chez l'Homme ?

CHAPITRE 9

Avant-propos

L'Article "Why does synonymous codon usage vary among different functional categories of human genes?" de cette partie présente une étude de génomique comparative sur les variations d'usage des codons synonymes dans les gènes humains. Il existe deux grandes catégories de modèles qui expliquent les variations d'usage des codons synonymes dans le vivant : (1) les modèles adaptatifs avec l'hypothèse de sélection traductionnelle et (2) les modèles non-adaptatifs avec le biais mutationnel et le biais de conversion génique (gBGC), voir le paragraphe 2.2, p. 33 pour plus de détails. Chez l'Homme, l'influence relative de ces processus sur la variation d'usage des codons synonymes est encore débattue (voir Duret (2002) pour une revue) et de nombreux articles estiment que le gBGC est à l'oeuvre (Galtier et al., 2001; Galtier & Duret, 2007; Duret & Galtier, 2009). Nous nous sommes intéressés à ce sujet suite à la publication de Gingold et al. (2014). Dans cet article, les auteurs étudient le niveau d'expression des gènes d'ARN de transfert (ARNt) dans deux types cellulaires différents (les cellules en prolifération – normales ou tumorales – et les cellules différenciées) à l'aide de microarrays ou de carte génomique des modifications d'histones autour des gènes d'ARNt. Par ailleurs, ils étudient les gènes selon leur catégorie fonctionnelle (i.e. Gene Ontology ou GO, voir l'encadré pour une définition). Les auteurs considèrent les GO "Biological Process" et définissent

deux sous-groupes de gènes : les GO liés à la prolifération et ceux liés à la différentiation. Ils observent que (1) l'abondance en anticodons des ARNt change entre les états cellulaires de prolifération et ceux de différentiation mais aussi que (2) l'usage des codons synonymes entre les gènes associés spécifiquement à la prolifération ou à la différentiation est différent et positivement corrélé aux variations d'anticodons. Cette corrélation a été interprétée comme de la sélection traductionnelle : les différents types cellulaires expriment divers gènes dont l'usage des codons synonymes est co-adapté aux anticodons des ARNt de la dite cellule. Étonnamment, ils ne mentionnent ni le gBGC ni le biais mutationnel ce qui nous a conduit à nous intéresser à cette question.

Les GO sont des groupes de gènes définis par leur fonction et les relations entre ces fonctions. Il existe trois grandes classes GO :

1. les fonctions moléculaires ("Molecular Function") : activités moléculaires des produits des gènes (comme la catalyse),
2. les composants cellulaires ("Cellular Component") : lieu où les produits de gènes sont actifs (comme la paroi cellulaire ou le cytoplasme),
3. les processus biologiques ("Biological Process") : les réseaux et les événements moléculaires qui ont un début et une fin définis et dont les activités requièrent l'intervention de plusieurs gènes (comme la phosphorylation oxydative ou l'induction de la mort cellulaire).

Chaque terme GO appartient à une de ces grandes classes et, au sein d'une classe, les termes GO peuvent être emboîtés. Ainsi, un gène appartient généralement à plusieurs GO.

Dans l'article présenté ci-dessous, nous remettons en cause les interprétations de Gingold et al. (2014). En effet, deux articles Schmitt et al. (2014); Rudolph et al. (2016) contestent l'idée d'une corrélation adaptative entre les ARNt exprimés dans les cellules et l'ensemble des codons qui sont traduits dans ces mêmes cellules. Schmitt et al. (2014) montrent que l'expression des gènes d'ARNt au niveau individuel varie mais que le total d'ARNt cellulaire est quant à lui stable. Rudolph et al. (2016) montrent qu'il n'y a pas de covariation entre l'ensemble des ARNt et l'usage des codons synonymes entre les cellules en prolifération et celles en différentiation.

Mais alors, qu'est-ce qui explique la variation d'usage du code synonyme entre catégories fonctionnelles ? Gingold et al. (2014) n'évoquent pas les processus non-adaptatifs comme facteurs explicatifs des variations d'usage des codons synonymes. Nous allons ici tester si les différences observées d'usage des codons synonymes entre les gènes de prolifération et ceux de différentiation résultent non pas d'un processus adaptatif comme proposé par Gingold et al. (2014), mais du gBGC. Dans un premier temps, nous montrons que les variations d'usage des codons synonymes s'expliquent par des variations du contenu en GC en 3ème position des codons. Puis nous montrons que l'usage du code mesuré sur les acides aminés mono et multi isoaccepteurs (voir l'encadré pour une définition) est très corrélé, c'est à dire qu'il est dirigé par un processus qui affecte les deux sortes d'acides aminés et n'est donc pas lié à la sélection traductionnelle. Nous testons ensuite l'hypothèse du gBGC.

N'ayant pas directement accès à l'intensité du gBGC au sein du génome humain, nous analysons les relations qui existent entre catégories fonctionnelles, composition, taux de recombinaison et expression des gènes. Comme prédit par le modèle de gBGC, la composition régionale en base et le taux de recombinaison expliquent la plupart des variations d'usage des codons synonymes entre les catégories fonctionnelles. Nous montrons que le niveau d'expression des gènes lors de la méiose est inversement corrélé à l'usage des codons synonymes terminant par G ou C. Cela s'explique par l'influence de l'expression en méiose sur le taux de recombinaison, ce qui est donc en accord avec le modèle de gBGC. Par ailleurs, selon leur fonction, les gènes ont plus ou moins de chances d'être exprimés à la méiose, ce qui influence leur taux de recombinaison et l'intensité de la gBGC. À l'issu de cet article nous montrons que 81.3% de la variation d'usage du code synonyme au sein du génome humain s'explique par le gBGC. Cet article est soumis.

Les acides aminés mono/multi isoaccepteurs.

- les acides aminés mono-isoaccepteurs correspondent aux acides aminés ayant plusieurs codons synonymes et un unique ARN de transfert capable de traduire les différents codons. C'est le cas pour 4 acides aminés : phénylalanine Phe, asparagine Asp, histidine His et cystéine Cys.
- les acides aminés multi-isoaccepteurs correspondent aux acides aminés ayant plusieurs codons synonymes et qui sont décodés par plusieurs ARN de transfert.

Why does synonymous codon usage vary among functional categories of human genes?

Fanny Pouyet[1], Dominique Mouchiroud[1], Laurent Duret[1*], Marie Sémon[2*]

[1] Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, Université Claude Bernard, CNRS UMR 5558, 43 boulevard du 11 novembre 1918, F-69100, Villeurbanne, France [2] Laboratory of Biology and Modelling of the Cell, Université de Lyon, Université Claude Bernard, CNRS UMR 5239, INSERM U1210, 46 allée d'Italie, F-69007, Lyon, France

*Corresponding authors: marie.semon@ens-lyon.fr, laurent.duret@univ-lyon1.fr

Abstract

In humans, as in other mammals, synonymous codon usage (SCU) varies widely among genes. Interestingly, genes involved in cell differentiation or in proliferation display a distinct codon usage, suggesting that SCU might reflect functional constraints to optimize translation efficiency in these distinct cellular states. However, in mammals, SCU is also known to correlate with large-scale fluctuations of GC-content along chromosomes, affecting both coding and noncoding regions. This variation is caused by meiotic recombination, via the process of GC-biased gene conversion (gBGC). To disentangle the different factors driving SCU in humans, we analyzed the relationships between functional categories, base composition, recombination, and gene expression. We first demonstrate that variation in SCU is not linked to constraints on tRNA abundance, and is predominantly driven by large-scale variation in GC-content. In agreement with the gBGC model, differences in SCU among functional categories are explained by variation in intragenic recombination rate, which, in turn, is strongly negatively correlated to gene expression levels during meiosis. Our results indicate that variation in SCU among functional categories (including variation associated to differentiation or proliferation) does not result from selection on translation efficiency but from differences in levels of meiotic transcription, which interferes with the formation of crossovers and thereby affects gBGC

intensity within genes. Overall, the gBGC model explains 81.3% of the variance in SCU among genes. The strong heterogeneity of SCU induced by gBGC in mammalian genomes precludes any optimization of the tRNA pool to the demand in codon usage.

Keywords: Codon Usage; Biased gene conversion; Translational selection; Recombination; Meiosis

Introduction

Although synonymous codons encode the same amino acid, some are used more frequently than others. This preferential usage of a subset of synonymous codons is known as “codon usage bias”. In many species, including humans, this bias varies substantially among genes in the genome. Both adaptive and non-adaptive processes, which are not mutually exclusive, have been proposed to explain the existence of codon usage biases (Duret, 2002; Chamary et al, 2006; Plotkin and Kudla, 2011 for reviews). According to the main adaptive model, termed translational selection, synonymous codon usage (SCU) and abundance of tRNA are co-adapted to optimize the efficiency of translation (Ikemura, 1981; Kanaya et al, 2001; Drummond and Wilke, 2008; Hershberg and Petrov, 2008; dos Reis and Wernisch, 2009). Non-adaptive models propose instead that codon usage bias results from biases in neutral substitution patterns, driven by mutation or by GC-biased gene conversion (gBGC, Galtier et al, 2001; Chen et al, 2004; Sémon et al, 2006; Duret and Galtier, 2009). In humans, these processes have been long studied, but the relative influence of adaptive and non-adaptive processes on SCU is still a matter of debate (Duret, 2002; Chamary et al, 2006; Plotkin and Kudla, 2011 for reviews).

Recently, Gingold *et al.* (2014) compared synonymous codon usage among sets of human genes associated to different functional categories (as defined in the Gene Ontology) and observed a striking difference between sets of genes associated with cellular proliferation and those associated with differentiation. They also observed that the relative abundance of tRNA varies according to the

proliferative or differentiative state of cells, which was logically interpreted in term of translational selection: different cell types express specific sets of genes whose coding sequence is co-adapted with specific pools of tRNAs (Gingold et al, 2014).

However, this interpretation stands in contradiction with two other studies. First, expression levels of individual tRNA genes do indeed vary extensively between tissue types and developmental stages in mice. But when tRNA genes are grouped by isoacceptor families (which recognize the same codon) the resulting collective expression levels are stable throughout development and specify a constant pool of anticodons (Schmitt et al, 2014). Second, in continuation to this work, a recent study specifically contrasted cells undergoing proliferation and those undergoing differentiation, and found no covariation of tRNA pool and codon usage between these cells (Rudolph et al, 2016). Hence, neither result is consistent with the initial claim interpreting differences in codon usage bias between functional classes as a consequence of translational selection. Then, why does synonymous codon usage vary between genes associated to different functional categories?

It has long been known that in mammals, variation in synonymous codon usage between genes is linked to large-scale fluctuation of GC-content along chromosomes, the so-called isochores (Bernardi et al. 1985; Mouchiroud et al, 1988; Mouchiroud et al, 1991; Clay and Bernardi, 2011). There is strong evidence that isochores are the consequence of GC-biased gene conversion (gBGC), a form of segregation distortion that occurs during meiotic recombination and that favors the transmission of GC alleles over AT alleles (Duret and Galtier, 2009; Munch et al, 2014; Williams et al. 2015). The gBGC process leads to an increase in the GC-content in regions of high recombination rate, which affects both coding and non-coding regions, including synonymous codon positions (Galtier and Duret, 2007; Duret and Galtier, 2009, Glémin et al. 2015). Besides, it has been shown that the rate of recombination within genes is negatively affected by their level of expression in the germline (McVicker and Green, 2010). This suggests that the impact of gBGC on the synonymous codon usage of genes might depend on their pattern of expression.

To investigate the parameters responsible for variation in codon usage among functional categories, we analyzed the relationships between synonymous codon usage, GC-content, recombination rate and expression patterns of human genes. Our results are fully consistent with the hypothesis that synonymous codon usage is driven by gBGC, and not by translational selection, and that the differences observed among functional categories reflect variation in long-term intragenic recombination rates, resulting from differences in meiotic expression levels.

Results

Variation in codon usage among functional categories results from differences in GC-content

To better understand what distinguishes codon usage between sets of genes involved in cellular proliferation and differentiation, we started by investigating the main factors that would discriminate codon usage between functional categories in general. For this purpose, we grouped genes per functional category (687 biological processes, associated to more than 40 genes in the Gene Ontology database), and computed codon frequencies for each of these gene sets. Variation in codon usage among these GO gene sets was analyzed by Principal Component Analysis (PCA). In agreement with Gingold *et al.* (2014), the first principal component of this analysis explains 41% of the total variance, and segregates “proliferation” (7 categories) from “differentiation” (6 categories) GO categories (Figure S1A). To remove potential confounding effects of variation in amino acid composition between these categories, we also analyzed the relative synonymous codon usage (RSCU; the RSCU of a given codon corresponds to its frequency, normalized by the frequency of the corresponding amino-acid) for each functional category. The first principal component of the PCA analysis is even stronger, explaining 73% of the total variance of RSCU, and still separates categories related to proliferation (red dots) and differentiation (blue dots, Figure 1A). Thus, synonymous codon usage clearly varies between functional categories in general, and between proliferation and differentiation

in particular. What property of sequence composition underlies this difference? It is well known that synonymous codon usage is strongly correlated to GC content at third position of codons (termed GC3; Mouchiroud et al, 1988). Thus, we computed the average GC3 of each GO gene set. The GC3 of these functional categories vary widely (from 0.45 to 0.73) and is perfectly correlated to their coordinates on the first PCA axis (R^2 = 0.99; Figure 1B). Hence, variation in SCU between functional categories is fully explained by variation in GC3.

On average, in our dataset, each gene is associated to 9 GO biological processes. Many genes belong to more than one GO biological-process category, either because they have several functions (pleiotropy) or because these categories are nested from specific to broad functions. Hence, GO-terms are not independent. To avoid this redundancy, for the remainder of this study we switched from analyses at the level of GO gene sets to analyses at the level of individual genes (except when stated otherwise). Each gene was assigned with one of three categories based on their GO annotation: 1,008 genes associated with “proliferation”, 2,833 genes associated with “differentiation”, and 12,129 “other” genes unrelated to these key-words (see methods). The distribution of GC3 content over the entire dataset is bimodal (Figure 1C). For the subsets of genes associated to “proliferation” and “differentiation”, the two distributions of GC3 differ significantly from each other (T-test, p-value < 2.10^{-16}), and their peaks coincide with each of the two modes observed for the whole genome. Genes associated to “proliferation” are on average less GC-rich than genes associated to “differentiation” (mean GC3 are respectively 0.53 and 0.61 in the two subsets).

5

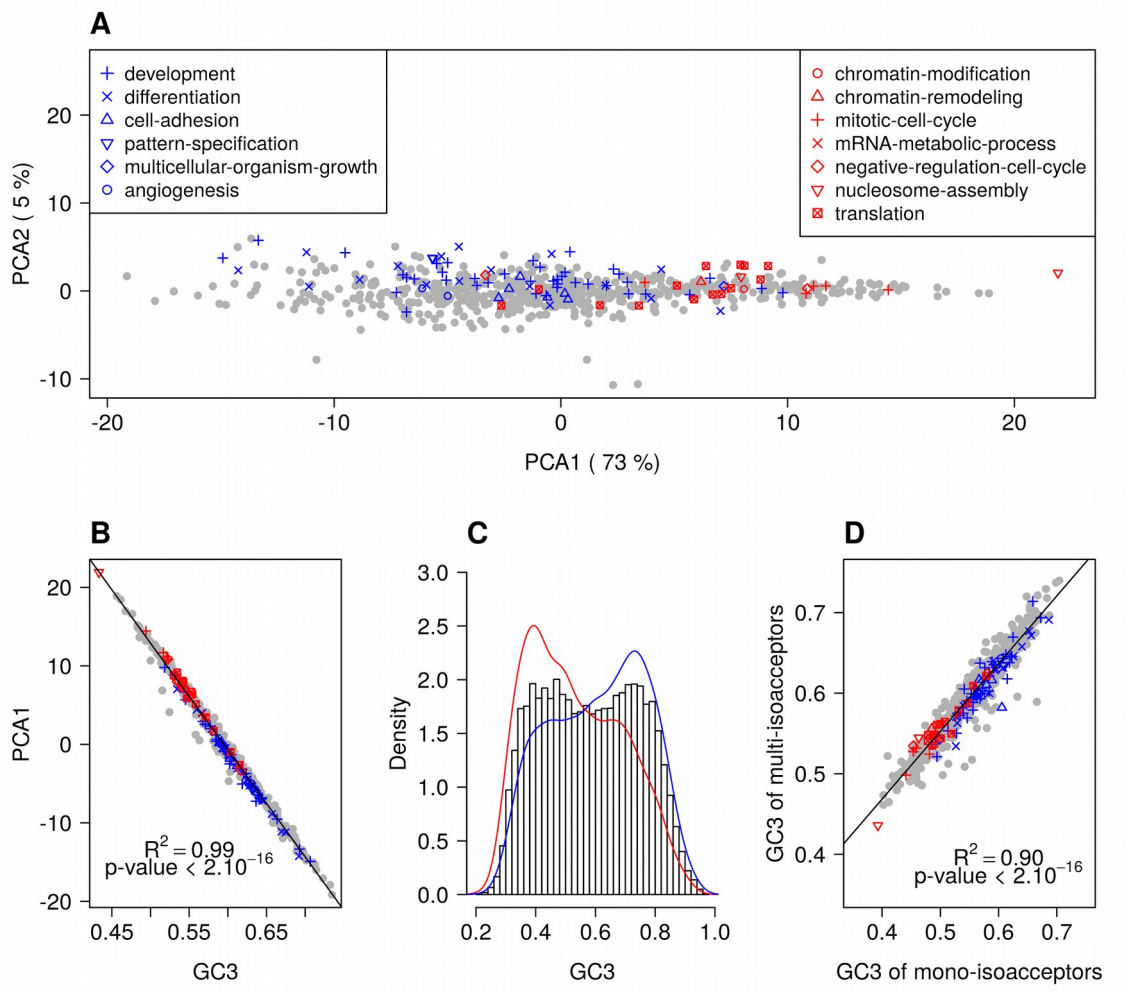


***Figure 1*: Variation in synonymous codon usage and in GC3 among functional categories.** *(A) Factorial map of the principal-component analysis of synonymous codon usage in GO functional categories in the human genome. Each dot corresponds to a GO gene set, for which the relative synonymous codon usage (RSCU) was computed. GO categories that are associated with "differentiation" or with "proliferation" are displayed respectively in blue and in red. (B) Correlation between the RSCU of GO gene sets (first PCA axis) and their average GC-content at third codon position (GC3). (C) Distribution of GC3 of human protein coding genes. The black histogram represents the distribution for whole dataset (15,970 genes). The blue curve (resp. the red curve) is a smoothed distribution of GC3 for "differentiation" genes (N=2,833) (resp. "Proliferation" genes, N=1,008) (D) Correlation between the GC3 of mono-isoacceptor amino-acids and multi-isoacceptor amino-acids. For each GO gene set, the average GC3 was computed separately for amino-acids decoded by multiple tRNA isoacceptors (N=14 multi-isoacceptor amino-acids), and for those decoded by one single tRNA isoacceptor (mono-isoacceptor amino-acids: Phe, Asp, His, Cys). Amino-acids encoded by a single codon (Met, Trp) were excluded.*

Variation in synonymous codon usage is not driven by translational selection.

We first investigated whether the observed variation in synonymous codon usage (i.e. variation in GC3) might be driven by translational selection. This model proposes that the relative usage of synonymous codons should co-vary with the abundance of their cognate tRNAs. A property of the tRNA gene repertoires allows us to test this hypothesis. The human genome contains 506 tRNA genes (decoding the 20 standard amino-acids), corresponding to 48 different tRNA isoacceptors (Chan and Lowe, 2009). Among the 18 amino acids having two or more synonymous codons, four are decoded by a single tRNA isoacceptor (mono-isoacceptor amino-acids: Phe, Asp, His and Cys), and the 14 other ones are decoded by several tRNA isoacceptors (multi-isoacceptors amino-acids).

For multi-isoacceptors amino-acids, the relative abundance of the different tRNA isoacceptors can vary among different cell types, and hence might covary with the relative synonymous codon usage of genes expressed in these cell types. For instance, let us consider Gln, which has two synonymous codons (CAG, CAA) that are decoded by two tRNA isoacceptors (respectively anticodons CTG and TTG). Let us consider a theoretical example of two cell types (say A and B) that differ in their relative tRNA abundance (CTG-tRNA being more abundant in A cells, and TTG-tRNA in B cells). According to the translational selection model, sets of genes that are over-expressed in A cells, should preferentially use the CAG codon whereas genes that are over-expressed in B cells, should preferentially use the CAA codon. However mono-isoacceptor amino-acids are, by definition, decoded by a single tRNA isoacceptor and the relative tRNA abundance cannot vary across cell types. Hence, according to the translational selection model, the relative synonymous codon usage for mono-isoacceptor amino-acids is not expected to vary among cell-specific gene sets. In other words, for mono-isoacceptor amino-acids, variation in synonymous codon usage among GO gene sets cannot be explained by co-adaptation with the tRNA pool.

To test whether variation in synonymous codon usage was driven by translational selection, we computed synonymous codon usage (GC3) in GO gene sets, separately for codons corresponding to mono-isoacceptor amino-acids and for codons corresponding to multi-isoacceptor amino-acids. We observed that the range of variation in GC3 is very similar for mono- and multi-isoacceptor amino-acids. Importantly, the two parameters are strongly correlated (R^2 = 0.90) (Figure 1D). This implies that GC3 variation is driven by a process that affects both mono-isoacceptor and multi-isoacceptor amino-acids, and hence that this process is not related to variation in tRNA abundance. This observation holds true for all functional categories, including those associated to differentiation or proliferation (red and blue dots in Figure 1D).

Impact of large-scale variation in genomic GC-content on synonymous codon usage.

So far, we have shown that genome-wide variation in synonymous codon usage is not driven by translational selection. Early studies, more than 30 years ago, have shown that variation in human synonymous codon usage is strongly correlated with large-scale fluctuations of GC-content along chromosomes (the so-called isochores), affecting both coding and noncoding regions (Bernardi et al. 1985, Mouchiroud et al, 1988; Mouchiroud et al, 1991; Clay and Bernardi, 2011). We therefore tested whether genes associated with “proliferation” were located in genomic regions with a lower GC-content than genes associated with “differentiation”. We observed, as expected, that the GC3 of genes correlates with the GC-content of their flanking regions (GC-flank, measured in 10 kb upstream and 10 kb downstream of the transcription unit; Figure 2A, R^2 = 0.48). This correlation is observed for all genes, including the subsets of genes associated with “proliferation” and “differentiation” (respectively R^2 = 0.49 and 0.46; Figure 2A). Thus, in agreement with the literature (Bernardi et al. 1985, Mouchiroud et al, 1988; Mouchiroud et al, 1991; Clay and Bernardi, 2011), our observations indicate that variation in SCU between genes is to a large extent attributable to their position in the genome. However, when the regional GC-content is controlled for, there remains a significant

difference in GC3 between gene categories: on average, for a given regional GC-content, there is a gap of 5% to 7% of GC3 between the categories “differentiation” and “proliferation” (Figure 2A). This implies that the difference in synonymous codon usage between these gene categories does not result from a preferential location in different isochores.

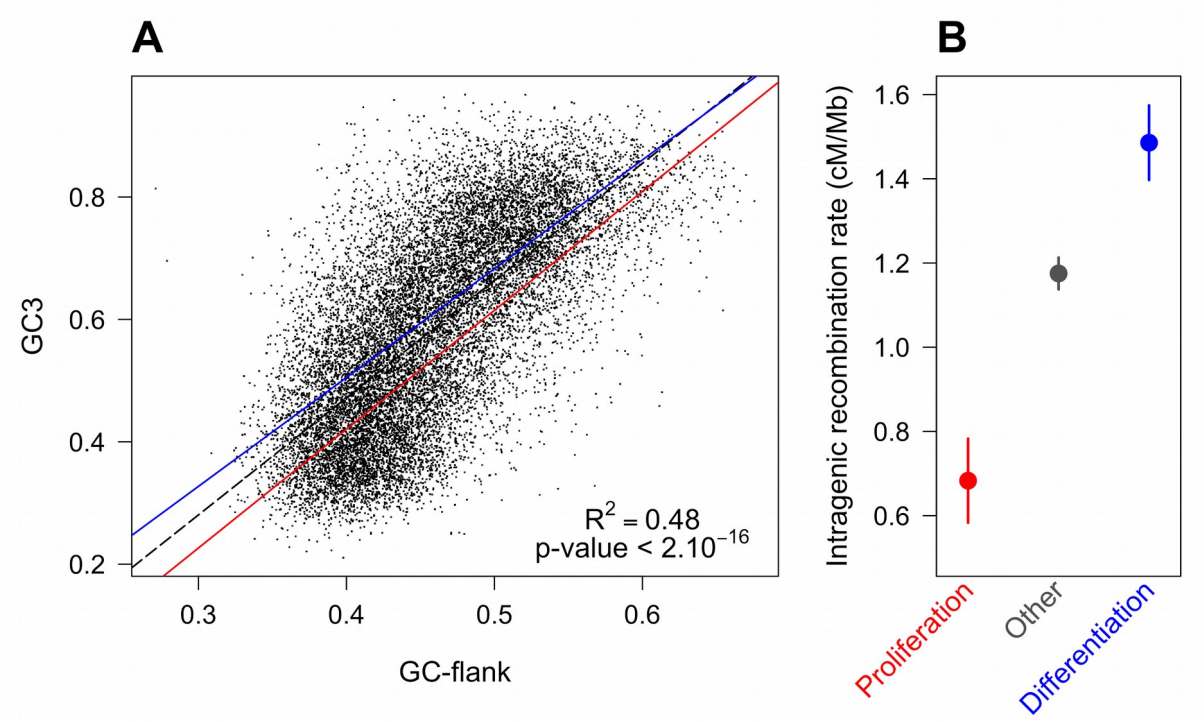


Figure 2: Difference in SCU between “proliferation” and “differentiation” genes is linked to variation in intragenic recombination rate, and not to their isochore context. *(A) Correlation between the GC3 of genes and the GC content of their flanking regions (GC-flank). Each dot corresponds to one gene. Regression lines were computed independently for “differentiation” genes (N=2,833, R^2 = 0.46, blue), “proliferation” genes (N=1,008, R^2 = 0.49, red) and other genes (N=12,129, R^2 = 0.49, dashed line). All p-values < 2.10^{-16}. (B) Average intragenic recombination rate in each functional categories. Error bars represent standard errors.*

Variation in synonymous codon usage among functional categories correlates with differences in intragenic recombination rate.

Previous studies have shown that the evolution of GC-content along chromosomes is driven by meiotic recombination, both on a broad (Mb) scale (Duret and Arndt 2008, Munch et al, 2014) and on a fine (kb) scale (Clément and Arndt, 2013; Pratto et al. 2014). There is now strong evidence that this correlation between GC-content and recombination is caused by the process of GC-biased gene

conversion (gBGC) which leads to increase the GC-content in regions of high recombination (Galtier et al, 2001; Galtier and Duret, 2007; Duret and Galtier, 2009; Munch et al, 2014; Pratto et al. 2014; Williams et al. 2015; Glémin et al. 2015). Recombination rate varies along chromosomes, and notably tends to be lower within genes than in flanking regions (Myers et al, 2005; McVicker and Green 2010). Interestingly, we observed that intragenic recombination rates (in cM/Mb) differ among the three sets of genes defined previously, and covary with their GC3: the average intragenic recombination rate is lower in "proliferation" genes compared to other genes, whereas it is higher in "differentiation" genes (Figure 2B; p-value of Kruskal-Wallis test $< 2.10^{-16}$ as for all pairwise Wilcoxon tests). These observations are therefore consistent with the hypothesis that differences in GC3 between "differentiation" and "proliferation" genes could also be driven by gBGC.

The difference in intragenic recombination rate between functional categories is explained by their expression level in meiosis.

Why do recombination rates vary across functional categories? Previous studies have shown that intragenic recombination rates vary according to gene expression patterns: genes that are expressed in many tissues tend to have lower intragenic crossover rates (Necsuela at al, 2009; Mc Vicker and Green, 2010). Mc Vicker and Green (2010) analyzed expression levels in many different samples, including both somatic tissues and meiotic or non-meiotic germ cells. Interestingly, they showed that the negative correlation between intragenic recombination rate and expression level is stronger in germ cells than in somatic tissue, and more specifically, stronger in meiotic cells than in other germ cells, most probably because gene expression in meiotic cells interferes with the formation of crossovers (Mc Vicker and Green, 2010).

To test whether the differences in intragenic recombination rates that we observed between "proliferation" and "differentiation" genes could be linked to their expression patterns, we analyzed published RNA-seq data sets, covering a broad range of samples: somatic or germ cells at different

stages of developing male and female embryo (20 different conditions; Guo et al, 2015); pachytene spermatocytes and round spermatids from adult males (Lesch et al. 2016), and differentiated adult tissues (26 somatic tissues plus testis, which contains a fraction of germ cells; Fagerberg et al, 2014). We first confirmed the negative relationship between intragenic recombination rate and gene expression level during meiosis (Figure 3A, S2A, S2D). We also confirmed that for both single cell data and bulk samples, the negative correlation between expression level and intragenic recombination rate is stronger in samples including germ cells than in samples including germ cells (Figure S3). This confirms that the intragenic crossover rate is affected specifically by expression in the germline.

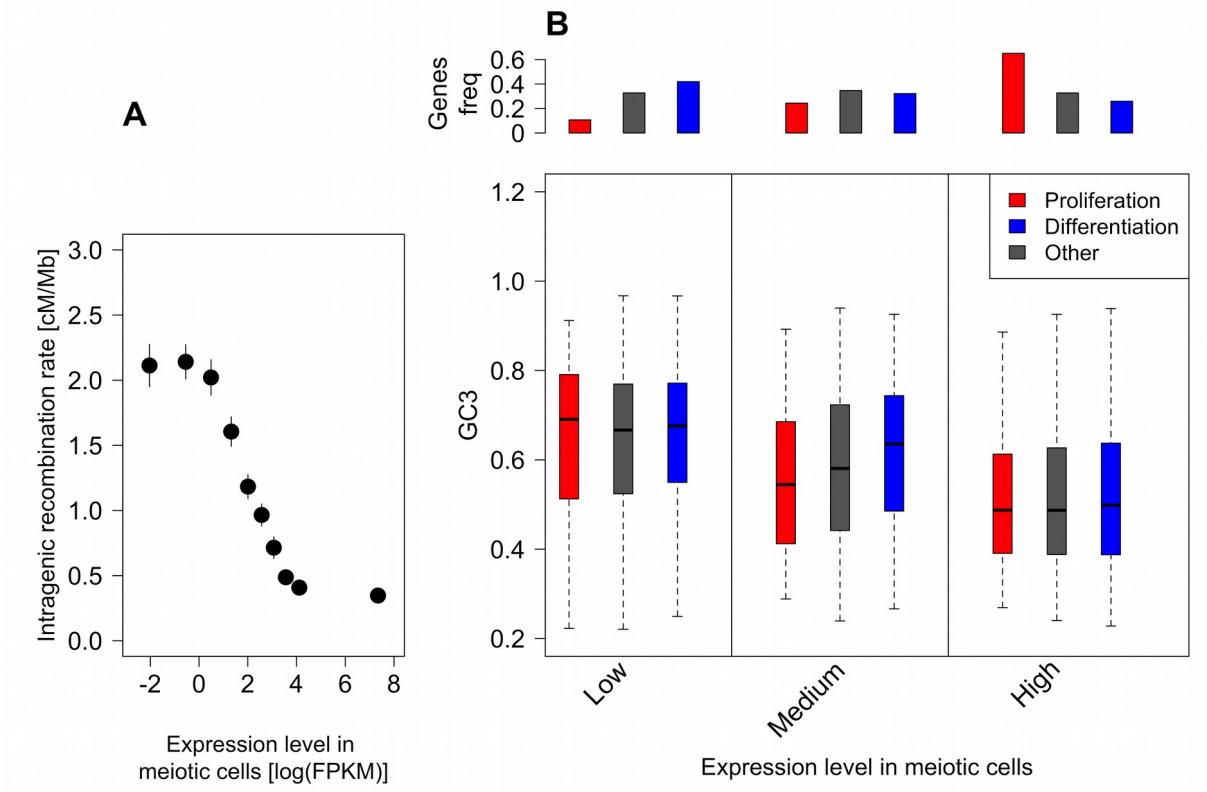


Figure 3: Variation in intragenic recombination rate and GC3 according to expression levels in meiotic cells. *(A) Genes were classified according to their sex-averaged expression level in meiotic cells into 10 bins of equal sample size. The mean intragenic recombination rate was computed for each bin. Error bars represent the standard error of the mean. Similar results were obtained when analyzing separately expression levels in female or male meiotic cells (Figure S2A, S2D). (B) Variation in GC3 according to meiotic expression levels. Genes were first binned into 3 classes of equal sample size according to their sex-averaged expression level in meiotic cells (low: < 3.07 FPKM; high: >22.68 FPKM: medium: the others), and then split into three sets according to their functional category: "proliferation" (red), "differentiation" (blue), and "other" genes (grey). Boxplots display the distribution of GC3 for each functional category within each expression bin. Above barplots display the distribution of genes among expression bins for each functional category.*

Many “proliferation” genes are involved in basic cellular functions, and hence, tend to be expressed at relatively high levels in many tissues and at all developmental stages. In particular, most of these genes are highly expressed in meiotic cells: 65% of “proliferation” genes are among the top 33% of genes with highest expression level (whereas only 11% are in the first tercile; Figure 3B). Conversely, only 26% of “differentiation” genes are highly expressed in meiotic cells, while 42% of are in the first tercile (Figure 3B).

This large proportion of “proliferation” genes with high meiotic expression levels can therefore explain why they tend to have relatively low intragenic recombination rate (Figure 2B), and hence, given the gBGC process, why they tend to have a lower GC3 (Figure 1C). To further test whether these differences in expression patterns could account for the difference in GC3 between “proliferation”, “differentiation” and “other genes”, we binned genes into three classes of increasing meiotic expression level. The distribution of GC3 is clearly shifted towards lower values for genes highly expressed at meiosis, compared to genes weakly expressed (average GC3 0.51 in the “high” category compared to 0.65 in the “low” category, p-value $<2.10^{-16}$) (Figure 3B). However, there is no significant difference in the distribution of GC3 between “proliferation” and “differentiation” within bins of low or high expression (p-value = 0.68 and 0.15 respectively). Hence, the striking difference in synonymous codon usage between these functional categories (Figure 1C) disappears once the level of expression during meiosis is controlled for (Figure 3B);

Thus, differences in expression levels at meiosis may be responsible for differences in synonymous codon usage among gene categories in human, through the following causative chain: (i) The set of “proliferation” genes is enriched in genes highly expressed in meiosis. (ii) Because high expression at meiosis decreases the rate of crossovers, intragenic recombination rates are lower in the “proliferation” set. (iii) In turn, reduced intragenic recombination diminishes the effect of gBGC on exon base composition, and hence GC3 is lower in the set “proliferation” compared to “differentiation”.

To check whether this cascade of effects fully recapitulates the difference in synonymous codon usage between “proliferation” and “differentiation”, we investigated whether differences in SCU between functional categories is driven by expression level in cells undergoing meiosis, rather than by expression level in another cell type or tissue. We examined the relationship between GC3 and expression levels in a broad panel of cell and tissue conditions (Figure 4). As predicted by our model, expression levels in germ cells, either from single cell samples or from testis (which contains germ cells) are better predictors of GC3 than expression in all other somatic tissues. Strikingly, the levels of expression in primary germ cells is, on average, twice as informative than expression in somatic cells taken at comparable stage of development (Figure 4B). Among all individual samples, the strongest correlation between GC3 and expression level was found in male meiotic cells (pachytene spermatocytes, R^2=6.3%, p-value<2.10^{-16}). Female meiotic cells (primordial germ cells, PGC 17 W) showed a similar correlation level (R^2=4.0%, p-value <2.10^{-16}). As expected, the correlation is even stronger with sex-averaged meiotic expression level (R^2=8.6%, p-value<2.10^{-16}). Hence, these results confirm that the cell type for which gene expression level is the best predictor of GC3 (and therefore SCU) corresponds to meiotic cells.

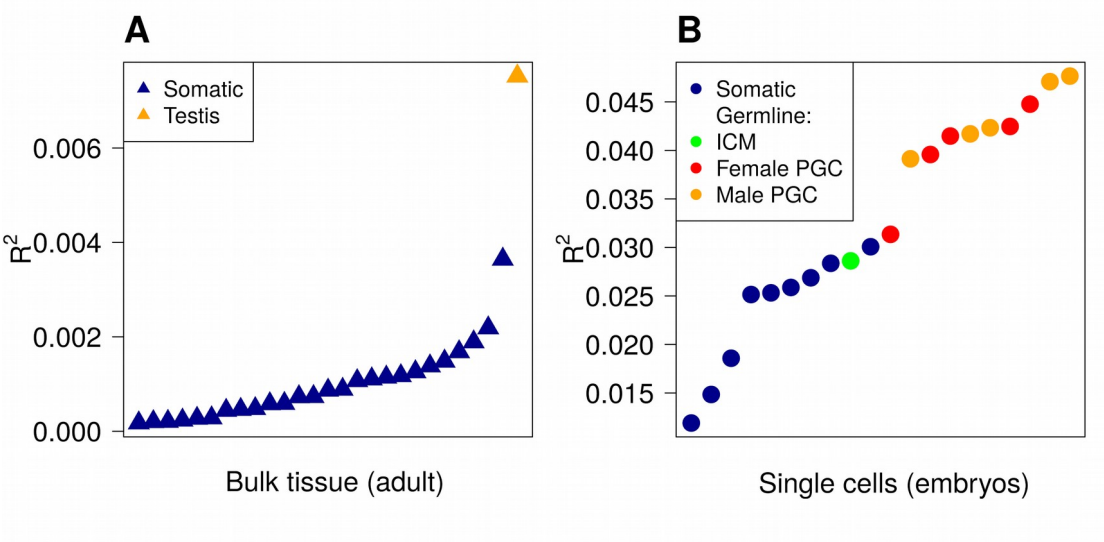


Figure 4: Correlation between expression level and GC3 in a panel of tissues and cell types. *(A) bulk adult tissues data (Fagerberg et al, 2014) and (B) early embryo single cell data (Guo et al, 2015). These two subsets were obtained via very different protocols, which prevents direct cross-comparisons. Samples are sorted by increasing correlation coefficient (R^2) between expression levels and GC3 (NB: all correlations are negative).*

Samples containing somatic cells are shown in blue; male germ cells in orange (testis or single cell) and female germ cells in red (PGC : primordial germ cells). The green point corresponds to cells from the inner cell mass (ICM) of the blastocysts, i.e. pluripotent cells from an early stage of development preceding the differentiation of germ cells.

GC-content of non-coding regions and meiotic expression explain more than 80% of the variation in synonymous codon usage of human genes.

Meiotic expression affects recombination rates along the entire gene (McVicker and Green 2010). Thus, the expression pattern is expected to affect gBGC intensity (and hence the GC-content) both in exons and in introns. Consistent with that prediction, the GC3 of human genes is strongly correlated to the GC-content of their introns (GCi, R^2=62.7%, p-value <2.10^{-16}). We build a linear model to quantify the relative contribution of the different parameters that covary with the GC3 of human genes (GCi, GC-flank, intragenic recombination rate, meiotic expression level, and "proliferation" or "differentiation" functional category). The analysis of variance demonstrates that GCi is by far the best predictor of GC3 (Table 1). GC-flank is largely redundant with GCi, whereas both intragenic recombination rate and gene expression level during meiosis significantly improve the model (by 3.9% and 1.3% respectively, Table 1, ANOVA, p-values < 2.10^{-16}). The integration of a categorical variable "differentiation" versus "proliferation" in the model significantly improves the model but its quantitative influence is minor (0.1%, p-value < 2.10^{-16},Table 1). Altogether, 68.2% of the variance in GC3 among human genes can be explained by the first four parameters (GCi, GC-flank, intragenic recombination rate, meiotic expression). Adding interaction terms to the linear model gives very similar results (70.4% variance explained, same levels of significance for all variables).

GC3 predictors		Pairwise R^2	p-value	Model R^2	F statistic	p-value
GCi		62.7%	$<2.10^{-16}$	62.7%	30232.4	$<2.10^{-16}$
+	GC-flank	48.1%	$<2.10^{-16}$	63.0%	126.8	$<2.10^{-16}$
+	Intragenic recombination rate	13.0%	$<2.10^{-16}$	66.9%	1453.3	$<2.10^{-16}$
+	Expression level in meiosis	8.8%	$<2.10^{-16}$	68.2%	875.7	$<2.10^{-16}$
+	Functional category	1%	$<2.10^{-16}$	68.3%	30.43	$<2.10^{-16}$

Table 1: Analysis of the variance of GC3 among individual genes. *Variables included in the linear model are : GC-content of introns (GCi), GC-content of flanking regions (GC-flank), intragenic recombination rate (log scale), sex-averaged meiotic gene expression level (log scale) and functional category ("differentiation", "proliferation" and "other"). Pairwise correlations (pairwise R^2) were computed between GC3 and each of the other variables. Correlations of the model (model R^2) were computed by adding variables sequentially.*

It should be noted that the number of codons in a gene is limited, and hence, a part of the variance in GC3 might simply result from stochastic sampling effects. To quantify this, we randomly sampled for each gene a number of sites (corresponding to its number of codons) in flanking regions (10 kb upstream and 10kb downstream). We then correlated the true GC-content of flanking regions, to the GC-content measured in this subset of sites (this process was repeated 100 times). The average correlation between the "true" and "sampled" GC content is 83.9% (Figure S4). In other words, only 83.9% of the variance in GC3 is explainable, the rest of the variance is caused by stochastic fluctuations due to the limited number of sampled sites. Thus, we conclude that at least 81.3% (= 68.2/83.9) of the explainable variance in GC3 of individual genes is explained by the GC-content of non-coding regions (GCi, GC-flank), intragenic recombination rate and meiotic expression level.

15

Discussion

In the human genome, gene sets that belong to different functional categories differ by their synonymous codon usage. Initially this pattern has been interpreted as evidence that the translation program was under tight control, notably to ensure a precise regulation of genes involved in cellular differentiation or proliferation (Gingold et al. 2014). According to this model, selection should optimize the match between the SCU of genes and tRNA abundances in the cells where they are expressed. However, the comparison of synonymous codon usage for amino-acids with single or multiple tRNA isoacceptors (Figure 1D) shows that the difference in SCU between functional categories does not result from constraints linked to tRNA abundance. In fact, variation in synonymous codon usage among functional categories is explained by one single dominant factor: the GC-content at third codon position (Figure 1B). The GC3 of human genes is strongly correlated to the GC-content of their introns and flanking regions (Table 1). This implies that variation in SCU results from a process that affects both coding and non-coding regions, and hence that it is not caused by translational selection.

Many lines of evidence indicate that large-scale variation in GC-content along chromosomes (isochores) is driven by the gBGC process, both in mammals and birds. First, there is direct evidence that recombination favors the transmission of GC-alleles over AT-alleles during meiosis (Odenthal-Hesse et al. 2014, Arbeithuber et al. 2015, de Boer et al. 2015, Williams et al. 2015, Smeds et al. 2016). Second, the analysis of polymorphism and divergence at different physical scales (from kb to Mb) showed that recombination induces a fixation bias in favor of GC alleles (Duret and Arndt 2008, Clément and Arndt 2013, Munch et al. 2014, Pratto et al. 2014, Weber et al. 2014; Glémin et al. 2015, Singhal et al. 2015). Third, the gBGC model predicts that the GC-content of a given genomic segment should reflect its average long-term recombination rate over tens of million years (Duret and Arndt 2008). Consistent with this prediction, analyses of ancestral genetic maps in the primate lineage revealed a very strong correlation between long-term recombination rates (in 1 Mb long windows) and stationary GC-content (R^2=0.64; Munch et al. 2014). The strong correlation between GC3 and GC-

flank therefore implies that variation in synonymous codon usage is primarily driven by large-scale variation in long-term recombination rate.

Besides these regional fluctuations, recombination rates also vary at finer scale. In particular, recombination rates tend to be reduced within human genes compared to their flanking regions (Myers et al. 2005), and this decrease depends on the level of expression of genes during meiosis (McVicker and Green, 2010; see also Figure 3A). Hence, the gBGC model predicts that the GC3 of a gene should depend not only of the long-term recombination rate of the region where it is located, but also on its specific pattern of expression. And indeed, we observed that the difference in synonymous codon usage between "proliferation" and "differentiation" genes is not due to their preferential location in different classes of isochores, but to the fact that "proliferation" genes tend to be expressed a high level in meiotic cells, and therefore to have a reduced intragenic recombination rate (Figure 2, 3).

To test whether this observation holds true for other functional categories, we measured the average GC3, intragenic crossover rate and meiotic expression level of each GO gene set. As predicted by the gBGC model, we observed a very strong correlation between GC3 and the average intragenic recombination rate of GO gene sets (R^2=0.51, Figure 5A). The variance in intragenic recombination rate, in turn, is very well explained by differences in meiotic expression levels among functional classes (R^2=0.46, Figure 5B). As mentioned previously, these correlations measured on gene concatenates should be interpreted with caution because the different points are not independent (a same gene can belong to different GO categories). However, this analysis clearly shows that a large fraction of the variance in SCU observed among GO gene sets can be explained by variation in gBGC intensity, caused by variation in intragenic recombination rates, driven by differences expression patterns (Figure 5C). In agreement with the gBGC model, the intragenic recombination rate correlates with the base composition of the entire gene, including introns (Figure 5D). This observation clearly invalidates the hypothesis that the observed differences in SCU among functional categories might be driven by selection on codon usage.

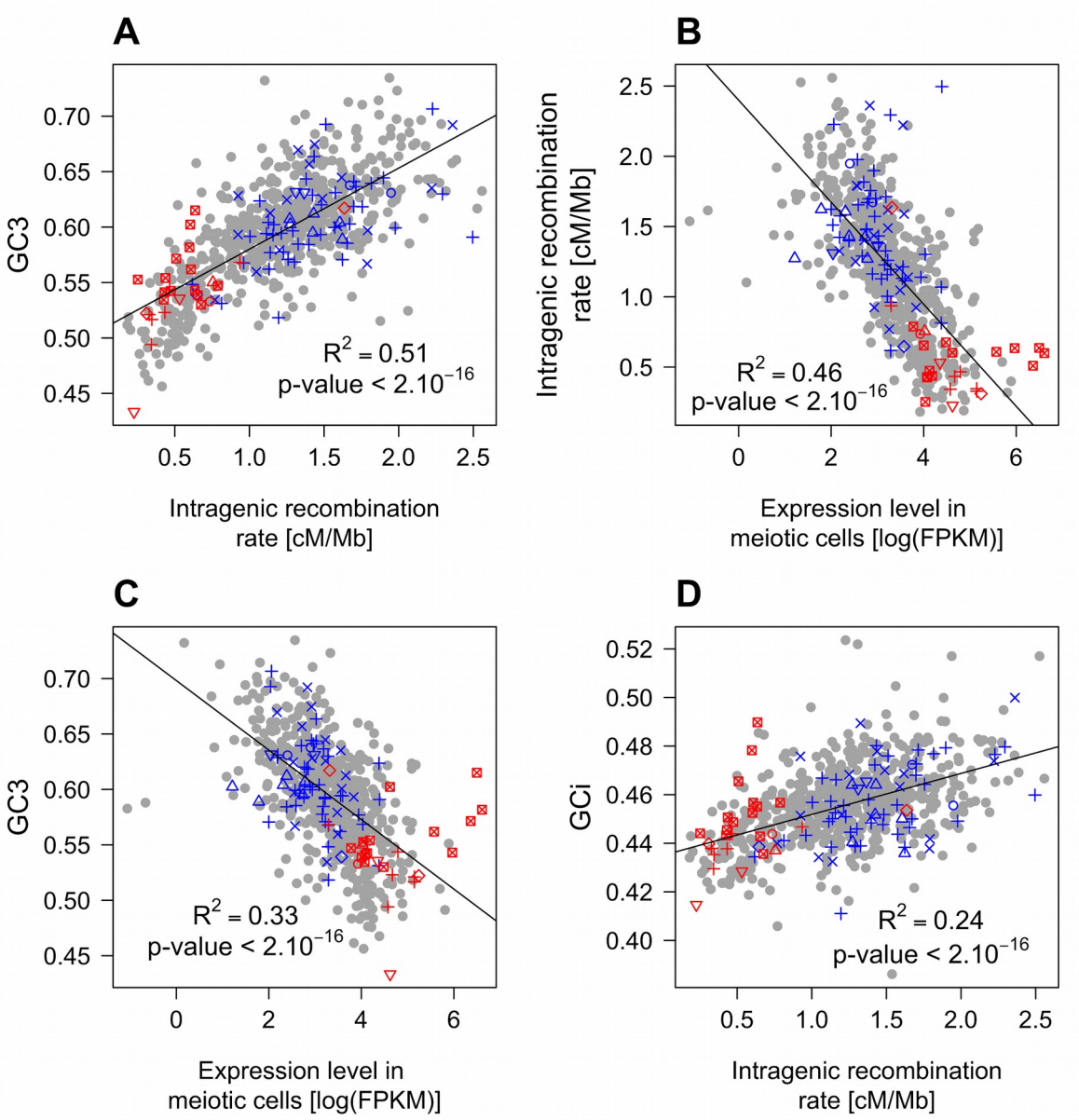


Figure 5: Relationships between GC-content, intragenic recombination rates and meiotic expression levels (sex-averaged) among functional gene categories. *Average values of these parameters were computed for each GO gene set. We then measured correlations between these parameters: (A) Mean GC3 vs. mean intragenic recombination rate. (B) Mean intragenic recombination rate vs. mean expression level in meiotic cells. (C) Mean GC3 vs. mean expression level in meiotic cells. (D) Mean intronic GC-content (GCi) vs. mean intragenic recombination rate. GO gene sets associated to "proliferation" (red) or "differentiation" (blue) are displayed as in Figure 1. Similar results were obtained when analyzing separately expression levels in female or male meiosis (Figure S2).*

The analysis of individual genes showed a much weaker correlation between GC3 and intragenic recombination rate (R^2=13.0%; Table 1) than that observed with gene sets (R^2=51%, Figure 5A). This

difference can be explained by the fact that fine-scale recombination landscapes evolve very rapidly (Auton et al. 2012) and hence, present-day genetic maps are poor predictors of long-term intragenic recombination rate. For instance, although human and chimpanzee diverged only ~7 million years ago, their recombination rates at the 10 kb scale are weakly correlated (R^2=10%) (Auton et al. 2012). In gene set analyses, intragenic recombination rates are averaged over a large number of genes, which leads to reduce the variance caused by measurement errors and temporal fluctuations, and hence leads to increase the correlation with GC3 (Figure 5A). In absence of accurate estimates of long-term intragenic recombination rate of individual genes, we analyzed four indirect predictors: GCi, GC-flank, present-day intragenic recombination rate and meiotic expression levels. As expected, GCi is by far the best predictor of GC3 (Table 1). According to the gBGC model, if GCi was a perfect predictor of the long-term recombination rate within exons, then the other parameters should not appear as significant predictors of GC3. However, there is evidence that recombination rates differ between exons and introns (Kong et al. 2010). Moreover, whereas the base composition of exons is almost exclusively driven by base substitutions, introns are also affected by deletions and insertions (notably of transposable elements). Thus the base composition of introns does not perfectly reflect the long-term intensity of gBGC within exons. On the other hand, patterns of gene expression are well conserved among mammals (Brawand et al. 2011). Thus, expression levels measured in humans are expected to be good predictors of long-term average meiotic expression level, and thereby to provide some information on long-term intragenic recombination. This can explain why meiotic expression level appears as an important additional predictor of GC3 (Table 1). Altogether, these four variables explain 81.3% of the explainable variance in GC3 of individual genes. In other words, the gBGC model can account for virtually all the variation in synonymous codon usage in the human genome.

It should be noted that co-variation between SCU and expression is generally considered as a typical signature of translational selection, and is often used to predict optimal codons (Duret 2002, Plotkin et al. 2004, dos Reis and Wernisch, 2009). However, as shown here, such correlations can also emerge as a result of a non-adaptive process. Given that gBGC is widespread in eukaryotes (Mancera et al. 2008,

Capra and Pollard 2011, Pessia et al. 2012, Williams et al. 2015, de Boer et al. 2015, Smeds et al. 2016), it appears essential to take this process into account to interpret variation in synonymous codon usage (and more generally in base composition) among genes.

There is clear evidence the usage of synonymous codons is under selective pressure in some metazoan species (such as drosophila or nematode), which implies that it has a significant impact on the fitness of organisms (for review, see Duret 2002, Chamary et al. 2006; Plotkin and Kudla 2011). It is *a priori* expected that codon usage should also affect translation efficiency (speed and accuracy) in mammals. However, our results show that selection on codon usage is not strong enough to counteract the impact of gBGC. In principle this does not exclude the hypothesis that the human genome might be subject to selection for translational efficiency: even if the GC-content of genes is driven by non-adaptive processes, there might be a selective pressure on the expression of tRNA genes to match the demand in synonymous codon usage. However, recent analyses of tRNA isoacceptors pools found no evidence for such variation (Schmitt et al, 2014; Rudolph et al. 2016). Moreover, we argue here that the peculiar base composition landscape induced by gBGC in the genomes of mammals and birds makes it impossible to match the tRNA pool to the demand in codon usage. Indeed, large-scale variation in recombination rates along the genome causes very strong variation in GC3 among genes, and this, whatever their functional category. In particular, "proliferative" genes, which are involved in basic cellular process, and are expressed at high levels in most tissues, show a very strong heterogeneity in GC3 (from 20% to almost 100%; Figure 1C). This implies that in any given cell, the set of highly expressed genes will show a very heterogeneous usage of synonymous codons. Hence, whatever the pool of tRNA available in that cell, there will be a large fraction of genes with a codon usage that does not match tRNA abundance. In other words, the heterogeneity of synonymous codon usage in mammalian genomes reflects a non-optimal situation, caused the gBGC process, in which it is not possible to adapt the tRNA pool to the demand in codon usage of the transcriptome of any cell type.

20

Material and Methods

Human protein coding genes

For each of the human protein coding genes in the Ensembl release 83 (Yates et al, 2016; assembly GRCh38.p5), we identified a canonical transcript as defined in http://www.ensembl.org/Help/Glossary?id=346 (PERL script available in supplementary material). Mitochondrial genes were excluded from this analysis. Sequences of the remaining 19,766 canonical transcripts together with exons coordinates, were downloaded through the BioMart query interface (Smedley et al, 2015)

Recombination rates

Intragenic crossover rates were measured using the HapMap genetic map (The International HapMap Consortium, 2007). We chose this genetic map, which is based on the analysis of linkage disequilibrium in human populations, because its resolution (~1 SNP per kb) is much higher than that of pedigree-based genetic maps (~1 SNP per 10 kb) (Kong et al, 2010).

Definition of functional categories

The GO Term Accessions and GO domain were retrieved from Ensembl version 83 for the 19,766 genes. We retrieved biological process GO terms, counted the number of genes associated to each GO term and kept the ones that include at least 40 genes, except GO:0005515 that is too general to be informative ("protein binding" GO set, which includes 14,542 genes). This led to a final list of 687 GO gene sets. For each gene set, we concatenated coding sequences to compute the total codon usage, the RSCU and GC-content, and we also computed the average intragenic recombination rate and average expression levels (see below).

Following the classification used by (Gingold et al, 2014), we further defined two broad functional categories: "proliferation" and "differentiation". GO terms containing the following keywords were

associated to “proliferation”: “Chromatin modification”, “chromatin remodeling”, “mitotic cell cycle”, “mRNA metabolic process”, “negative regulation of cell cycle”, “nucleosome assembly”, “translation”. GO terms containing the following keywords were associated to “differentiation”: “Development”, “differentiation”, “cell adhesion”, “pattern specification”, “multicellular organism growth”, “angiogenesis”. Please note that GO terms corresponding to negative effects were excluded where appropriate (e.g. “negative regulation of proliferation” was not included in the “proliferation” category). Complete lists of GO terms are available in supplementary material.

Analyses of individual genes

We also measured the codon usage of individual genes, to analyze covariations with their GC-content, expression levels and intragenic recombination rate. To limit noise in intragenic crossover rate estimates, we only retained genes longer than 5 kb (N=16,223 genes).

We defined three non-overlapping classes of genes according to their GO category: genes associated to at least one of the “proliferation” GO terms (N=1,008), genes associated to “differentiation” GO terms (N=2,833) and other genes (N=12,129). A group of 253 genes that were associated to both “proliferation” and “differentiation” GO terms were discarded from further analyses. The final dataset used in our analyses included 15,970 genes. In this dataset, there were 15,816 genes for which all parameters were available (a few genes had no introns, and hence no GCi, and a few genes were absent from the recombination rate data set).

Expression data

Gene expression levels were collected from three publicly available human RNA-seq experiment datasets. The first one includes 27 differentiated adult tissues (Fagerberg et al, 2014; Kryuchkova-Mostacci and Robinson-Rechavi, 2015; EBI accession number E-MTAB-1733). The second one is based on single-cell RNA-seq analysis, and includes 20 samples, corresponding to inner cell mass (ICM) of the blastocysts, and to primordial germ cells (PGC) and somatic cells, from male and female embryos at different development stages (4, 7 or 8, 10, 11 and 17 or 19 weeks) (Guo et al, 2015; GEO

accession number GSE63818). Female 17 weeks PGCs are entered in meiosis (Guo et al, 2015). This sample was therefore taken as representative of the transcriptome of meiotic cells in female. The third dataset corresponds to human male germ cells at pachytene spermatocytes (i.e. cells entering meiosis) and at round spermatids stages (post meiotic stage) (Lesch et al, 2016; GEO accession number GSE68507). Guo and Lesch datasets include several replicates for each sample. We therefore computed the average expression levels over all replicates for each sample. The sex-averaged meiotic expression level was estimated by computing the mean of expression levels in female 17 weeks PGCs (Guo et al, 2015) and male spermatocytes or spermatids (Lesch et al, 2016). The correspondence between gene expression datasets and codon usage tables was based on Ensembl gene identifiers (Fagerberg and Lesch datasets), or on gene names (Guo dataset). In total, our analyses of expression levels were based on 15,305 genes (665 genes were absent from the Guo dataset).

Statistical analysis

Unless stated otherwise, reported R^2 values correspond to Pearson correlation tests. R version 3.2.2 (R Core Team, 2015) was used with Base package for statistical tests and graphics, plus ade4 library (Dray and Dufour, 2007) for PCA analysis. The data and R scripts, which permit to reproduce the figures and tests presented here, are provided in supplementary material.

Acknowledgement

This work was supported by French National Research Agency (ANR) grant DaSiRe (ANR-15-CE12-0010-01/DaSiRe). FP received a doctoral scholarship from Ecole Normale Supérieure de Lyon (http://www.ens-lyon.eu/). We thank Gaël Yvert for initiating the discussion.

Disclosure declaration

The authors declare no competing financial interest.

References

Auton A, Fledel-Alon A, Pfeifer S, Venn O, Ségurel L, Street T, … McVean G. (2012) A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. *Science*.; 336(6078):193-98 doi:10.1126/science.1216872

Arbeithuber B, Betancourt A J, Ebner T, Tiemann-boege I. (2015) Crossovers are associated with mutation and biased gene conversion at recombination hotspots, *Proc Natl Acad Sci U S A 112(7), 2*109–2114. http://doi.org/10.1073/pnas.1416622112

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, … Rodier F. (1985). The mosaic genome of warm-blooded vertebrates. *Science*, 228, 953–958. doi:10.1126/science.4001930

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, ... Kaessmann H. (2011) The evolution of gene expression levels in mammalian organs. *Nature;* 478: 343–8. doi:10.1038/nature10532

Capra JA, Pollard KS. (2011) Substitution patterns are GC-biased in divergent sequences across the metazoans. *Genome Biol Evol*;3: 516–527. doi:10.1093/gbe/evr051

Chamary JV, Parmley JL, Hurst LD. (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet.*;7(2):98–108. doi:10.1038/nrg1770

Chan P P, Lowe T M. (2009). GtRNAdb: A database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Research*, *37*(SUPPL. 1), 93–97. http://doi.org/10.1093/nar/gkn787

Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH. (2004) Codon usage between genomes is constrained by genome-wide mutational processes. *Proc Natl Acad Sci U S A*;101(10):3480–5. doi: 10.1073/pnas.0307827100

Clay O K, Bernardi G. (2011). GC3 of Genes Can Be Used as a Proxy for Isochore Base Composition: A Reply to Elhaik et al. *Mol. Biol. Evol*, *1*(28), 21–23. http://doi: 10.1093/molbev/msq222

Clément Y, Arndt P F. (2013): Meiotic recombination strongly influences GC-content evolution in short regions in the mouse genome. *Mol Biol Evol.* 30: 2612–8. http://doi:10.1093/molbev/mst154

de Boer E, Jasin M, Keeney S. (2015) Local and sex-specific biases in crossover vs. noncrossover outcomes at meiotic recombination hot spots in mice. *Genes Dev.* 29: 1721–1733. doi:10.1101/gad.265561.115

dos Reis M, Wernisch L. (2009) Estimating translational selection in eukaryotic genomes. *Mol Biol Evol.*; 26(2):451–61. http://doi:10.1093/molbev/msn272

Dray S, Dufour A B. (2007): The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software.* 22(4): 1-20.

Drummond D A, Wilke C O (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*. ;134(2):341–52. http://doi:10.1016/j.cell.2008.05.042

Duret L. (2002). Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev, 12*(6), 640–649.doi:10.1016/S0959-437X(02)00353-2

Duret L, Arndt P F. (2008). The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genetics*, *4*(5). http://doi.org/10.1371/journal.pgen.1000071

Duret L, Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics*, *10*, 285–311. doi: 10.1146/annurev-genom-082908-150001.

Fagerberg L, Hallström B M, Oksvold P, Kampf C, Djureinovic D, Odeberg J, … Uhlén M. (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & Cellular Proteomics : MCP*, *13*(2), 397–406. http://doi.org/10.1074/mcp.M113.035600

Galtier N, Piganeau G, Mouchiroud D, Duret L. (2001). GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. *Genetics*, *159*(2), 907–911. doi:10.1371/journal.pgen.1004941.

Galtier N, Duret L. (2007). Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends in Genetics*, *23*, 273–277. http://dx.doi.org/10.1016/j.tig.2007.03.011

Gingold H, Tehler D, Christoffersen N R, Nielsen M M, Asmar F, Kooistra S M, … Pilpel Y. (2014).

A dual program for translation regulation in cellular proliferation and differentiation. *Cell*, 158(6), 1281–1292. doi:10.1016/j.cell.2014.08.011.

Glémin S, Arndt P F, Messer P W, Petrov D, Galtier N, Duret L. (2015). Quantification of GC-biased gene conversion in the human genome. *Genome Research*, *25*(8), 1215–28. http://doi.org/10.1101/gr.185488.114

Guo F, Yan L, Guo H, Li L, Hu B, Zhao Y, … Qiao, J. (2015). The transcriptome and DNA methylome landscapes of human primordial germ cells. *Cell*, *161*(6), 1437–1452. http://doi.org/10.1016/j.cell.2015.05.015

Hershberg R, Petrov DA. (2008) Selection on codon bias. *Annu Rev Genet.*;42:287–99. doi:10.1146/annurev.genet.42.110807.091442

Ikemura T (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its proteins; *J. Mol. Biol*, 146(1):1-21. doi:10.1016/0022-2836(81)90003-6

Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T. (2001) Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J Mol Evol*.;53(4–5):290–8. doi:10.1007/s002390010219

Kong A, Thorleifsson G, Gudbjartsson D F, Masson G, Sigurdsson A, Jonasdottir A, … Stefansson K. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467(7319), 1099–1103. doi:10.1038/nature09525

Kryuchkova-Mostacci N, Robinson-Rechavi M. (2015). Tissue-Specific Evolution of Protein Coding Genes in Human and Mouse. *PloS One*, *10*(6), 1–15. http://doi.org/10.1371/journal.pone.0131673

Lesch B J, Silber S J, McCarrey J R, Page D C. (2016) Parallel evolution of male germline epigenetic poising and somatic development in animals. *Nat Genet*;48(8):888-94. doi:10.1038/ng.3591

Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. (2008) High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*; 454: 479–85. doi:10.1038/nature07135

McVicker G, Green P. (2010). Genomic signatures of germline gene expression. *Genome Research*, *20*(11), 1503–1511. http://doi.org/10.1101/gr.106666.110

Mouchiroud D, Gautier C, Bernardi G. (1988). The compositional distribution of coding sequences and DNA molecules in humans and murids. *J. Mol. Evol*, 27, no. 4: 311-320.

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

Munch K, Mailund T, Dutheil J Y, Schierup M H. (2014). A fine-scale recombination map of the human-chimpanzee ancestor reveals faster change in humans than in chimpanzees and a strong impact of GC-biased gene conversion. *Genome Research*, *24*(3), 467–474. http://doi.org/10.1101/gr.158469.113

Myers S, Bottolo L, Freeman C, Mcvean G, Donnelly P. (2005). A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome, *Science* 321(5746), 321–324. http://doi.org/10.1126/science.1117196

Necsulea A, Sémon M, Duret L, Hurst L. D. (2009). Monoallelic expression and tissue specificity are associated with high crossover rates. *Trends in Genetics*, *25*(12), 519–522. http://doi.org/10.1016/j.tig.2009.10.001

Odenthal-hesse L, Berg I L, Veselis A, Jeffreys A J, May C A. (2014). Transmission Distortion Affecting Human Noncrossover but Not Crossover Recombination: A Hidden Source of Meiotic Drive. *PLoS Genet.*, *10*(2). http://doi.org/10.1371/journal.pgen.1004106

Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GAB. (2012) Evidence for Widespread GC-biased Gene Conversion in Eukaryotes. *Genome Biol Evol.*; 4: 675–82. doi:10.1093/gbe/evs052

Plotkin JB, Robins H, Levine AJ. (2004) Tissue-specific codon usage and the expression of human genes. *Proc Natl Acad Sci U S A.*;101: 12588–91. doi:10.1073/pnas.0404957101

Plotkin JB, Kudla G. (2011) Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet.*;12(1):32–42. doi:10.1038/nrg2899

Pratto F, Brick K, Khil P, Smagulova F, Petukhova G V, Camerini-Otero R D. (2014). DNA recombination. Recombination initiation maps of individual human genomes. *Science (New York, N.Y.)*, *346*(6211), 1256442. http://doi.org/10.1126/science.1256442

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rudolph K L M, Schmitt B M, Villar D, White R J, Marioni J C, Kutter C, Odom D T. (2016). Codon-Driven Translational Efficiency Is Stable across Diverse Mammalian Cell States. *PLoS Genetics*, *12*(5), e1006024. http://doi.org/10.1371/journal.pgen.1006024

Schmitt B M, Rudolph K L M, Karagianni P, Fonseca N A, White R J, Talianidis I, … Kutter C. (2014). High-resolution mapping of transcriptional dynamics across tissue development reveals a stable mRNA-tRNA interface. *Genome Research*, *24*(11), 1797–1807. http://doi.org/10.1101/gr.176784.114

Sémon M, Lobry JR, Duret L. (2006) No evidence for tissue-specific adaptation of synonymous codon usage in humans. *Mol Biol Evol.*;23(3):523–9. http://doi.org/10.1093/molbev/msj053

Singhal S, Leffler EM, Sannareddy K, Turner I, Venn O, Hooper DM,... Przeworski M. (2015) Stable recombination hotspots in birds. *Science*, ;350: 928–932. doi:10.1126/science.aad0843

Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, … Kasprzyk A. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research*, *43*(W1), W589–98. http://doi.org/10.1093/nar/gkv350

Smeds L, Mugal CF, Qvarnström A, Ellegren H. (2016) High-Resolution Mapping of Crossover and Non-crossover Recombination Events by Whole-Genome Re-sequencing of an Avian Pedigree. *PLOS Genet.*;12: e1006044. doi:10.1371/journal.pgen.1006044

The International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164), 851–61. http://doi:10.1038/nature06258

Weber CC, Boussau B, Romiguier J, Jarvis ED, Ellegren H. (2014) Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. *Genome Biol*;15: 549. doi:10.1186/s13059-014-0549-1

630 Williams A L, Genovese G, Dyer T, … Przeworski M. (2015). Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *eLife*, *2015*(4), 1–21. http://doi.org/10.7554/eLife.04637

Yates A, Akanni W, Amode M R, Barrell D, Billis K, Carvalho-Silva D, … Flicek P. (2016) Ensembl *Nucleic Acids Research*, *44*(D1), D710–D716. http://doi.org/10.1093/nar/gkv1157

Cinquième partie

Conclusion Générale

Ce manuscrit porte sur l'évolution de la composition des génomes et plus précisément l'évolution des séquences codant des protéines. De ce point de vue, j'ai employé des méthodes phylogénétiques et de génomique comparative pour étudier l'évolution du biais d'usage du code génétique. Ce thème de recherche passionne depuis la découverte du code génétique et de son usage biaisé avec notamment de nombreux articles fondateurs publiés dans les années 1980. Historiquement, le fait qu'à la fois les données soient très coûteuses donc peu abondantes et que les puissances de calcul soient limitées ne permettaient pas un usage systématique des modèles de codons pour étudier l'évolution des séquences. En effet, les modèles de codons emploient des matrices de grande taille (61x61) ce qui nécessite (i) de la puissance calculatoire et (ii) beaucoup de données. Aujourd'hui, et notamment depuis une petite dizaine d'années, l'augmentation significative des performances informatiques d'une part et l'avènement des données à faible coût (avec les nouvelles techniques de sequencage NGS – next generation sequencing) d'autre part redynamisent le domaine de recherche du développement de modèles de codons dont fait partie SENCA.

SENCA est un nouveau modèle de codons de type MutSel (avec des paramètres mutationnels et de sélection) qui modélise les mécanismes d'évolution du BUC entre espèces. D'un point de vue personnel, je pense que les modèles de type MutSel sont aujourd'hui extrêmement attractifs car (i) les capacités informatiques permettent de les mettre en oeuvre et (ii) il reste énormément de points à améliorer (notamment dans l'expression des termes de sélection). Les modèles de type MutSel ont comme origine le cadre théorique de génétique des populations et sont à l'interface entre ce domaine et celui de la phylogénétique. Cela m'a conduit à m'orienter vers une étude plus approfondie de la génétique des populations. Dans le futur, je souhaite regarder l'organisation génomique (humaine par exemple) d'un point de vue populationnel et voir comment les pressions évolutives agissent à courte échelle de temps. Dans le cadre d'une modélisation de l'évolution des gènes, il me semble que l'utilisation plus systématique de modèles complexes (de codons ou d'acides aminés) devient la règle. Ces modèles sont de plus en plus réalistes et permettent d'expliquer des histoires évolutives complexes. Il faut néanmoins garder à l'esprit que les usagers de tels modèles ne sont pas forcément du domaine et qu'il est nécessaire, en plus de fournir un modèle complet, de rendre disponible des outils, des règles, des explications pour analyser les résultats de modélisation.

Tout au long de ce manuscrit, j'ai étudié les variations de BUC à courte échelle de temps, soit au sein de l'Homme soit entre souches d'espèces bactériennes ou espèces d'un même clade ce qui justifiait l'emploi de modèles homogènes en temps. Il serait bien évidemment intéressant de regarder à plus grande échelle de temps et alors d'employer des modèles hétérogènes. Dans ce cas, il est aujourd'hui nécessaire de diminuer le nombre de paramètres du modèle en utilisant par exemple des profils d'acides aminés prédéfinis.

La partie III montre que plusieurs extensions sont possibles pour affiner le modèle déjà publié. SENCA^e est un projet extrêmement attractif et, à court terme, je compte le tester sur un véritable jeu de données, analyser le comportement d'`expr`. Il serait intéressant de pouvoir alors estimer et donc reconstruire le niveau d'expression ancestral des gènes. SENCA+bgc+CpG est un projet dont le comportement n'est pas tout à fait compris et il est indispensable d'en discuter avec d'autres chercheurs pour estimer les points positifs de ce modèle. Ainsi, j'ai déjà présenté les résultats préliminaires à Laurent Duret qui a proposé d'insérer l'effet CpG avant de prendre en compte le gBGC (ce que j'ai présenté). Avec Nicolas Lartillot, ils sont également en train de développer un modèle d'évolution de l'intensité du gBGC le long de séquences et les résultats de leur travaux en plus de discussions me permettraient aussi de mieux comprendre les miens. Je pense que l'effet gBGC n'est pas encore bien paramétré car même lorsque la couche C est bloquée à l'hypothèse nulle (i.e. pour chaque codon I $\phi(I) = 1/d_{AA_I}$), je retrouve des valeurs de `bgc` négatives pour les gènes AT-riches. Une autre approche pourrait donc plutôt de fixer à la fois la valeur du biais mutationnel à une valeur prise dans la littérature et les préférences de codons à l'hypothèse nulle. Cette approche se justifie entre autres grâce aux conclusions de la partie IV où nous avons montré que les variations d'usage des codons synonymes entre gènes chez l'Homme résulte simplement d'une variation d'intensité du gBGC. Dans ces perspectives sur SENCA+bgc+CpG, il me semble que le point clé pour comprendre ce que représente le paramètre `bgc` dans notre modèle est d'appréhender ce qu'il se passe dans les régions AT-riches.

En parallèle, je me suis intéressée à une autre méthode pour étudier les origines du BUC chez l'Homme (partie IV). La question est similaire à SENCA+bgc mais l'approche est différente puisqu'elle repose sur les variations du BUC au sein du génome humain. Notre résultat est que cette variation est expliquée par le processus gBGC à 81.5%. Ce résultat prouve de manière évidente que chez l'Homme en particulier et chez les mammifères en général, il est indispensable de modéliser le gBGC. Afin de proposer une modélisation réaliste, une solution

est de mettre en équation les mécanismes cellulaires connus du BGC (effet de la recombinaison, de la protéine PRDM9 etc.). A noter que l'intensité du gBGC évolue extrêmement rapidement puisque chez l'Homme et chez le Chimpanzé les hotspots de recombinaison sont peu corrélés ($R^2 = 10\%$). Bien que déjà étudié, j'aimerai personnellement regarder les variations de taux de recombinaison entre populations voire entre individus. Je souhaite regarder cela plus en détails grâce à l'augmentation significative de la précision des variations génomiques entre individus (cette année, de nombreux génomes humains ont été séquencés et sont disponibles) afin de quantifier précisément quelle part de variation existe au sein d'une espèce.

Bibliographie

En rouge, vous trouverez les pages où je cite la bibliographie.

H Akaike. Information theory and an extension of the maximum likelihood principle. *2nd Internat. Symp. on Inf. Theory*, pages 267–281, 1973. 48

H. Akashi. Synonymous codon usage in drosophila melanogaster : natural selection and translational accuracy. *Genetics*, 136(3) :927–935, 1994. 34

H Akashi and A Eyre-Walker. Translational selection and molecular evolution. *Curr. Opin. Genet. Dev.*, 8(6) :688–93, 1998. 34

M. Bailly-Bechet. Biais de codons et régulation de la traduction chez les bactéries et leurs phages., 2007. 33

B. Batut. Etude de l'évolution réductive des génomes bactériens par expériences d'évolution in silico et analyses bioinformatiques, 2013. 131

SK Behura and DW Severson. Coadaptation of isoacceptor trna genes and codon usage bias for translation efficiency in aedes aegypti and anopheles gambiae. *Insect Mol Biol.*, 20(2) : 177–187, 2011. 34

A. Belle, A. Tanay, L. Bitinka, R. Shamir, and E.K. O'Shea. Quantification of protein half-lives in the budding yeast proteome. *Proc Natl Acad Sci USA*, 35(103) :13004–130091, 2006. 43

G. Bernardi, B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival, and F. Rodier. The mosaic genome of warm-blooded vertebrates. *Science*, 228 :953–958, 1985. 36, 84, 85, 106

AP Bird and MH Taggart. Variable patterns of total DNA and rDNA methylation in animals. *Nucleic Acids Res.*, 8 :181–188, 1980. 89

J. Bodilis and S. Barray. Molecular evolution of the major outer-membrane protein gene (oprf) of pseudomonas. *Microbiology*, Pt 4(152) :1075–1088, 2006. 43, 44

M Botzman and H. Margalit. Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome Biol*, 12(10) :109, 2011. 35

D Brawand, M Soumillon, M Necsulea, P Julien, G Csárdi, P Harrigan, M Weier, A Liechti, A Aximu-Petri, M Martin Kircher, FW. Albert, U Zeller, P Khaitovich, F Frank Grützner, S Bergmann, R Nielsen, S Svante Pääbo, and H Kaessmann. The evolution of gene expression levels in mammalian organs. *Nature*, 478 :343–348, 2011. 95

M. Bulmer. The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129 (3) :897–907, November 1991. 33

K. P Burnham and D. R. Anderson. *Model Selection and Multimodel Inference : A Practical Information-Theoretic Approach*. Springer-verlag edition, 2002. 48

G Cannarrozzi, N.N Schraudolph, M Faty, P von Rohr, M.T Friberg, A.C. Roth, P Gonnet, P Gonnet, and Y Barral. A role for codon order in translation dynamics. *Cell*, 141(2) : 355–67, 2010. 35

A. Carbone, A. Zinovyev, and F. Képès. Codon adaptation index as a measure of dominating codon bias. *Bioinformatics*, 19(16) :2005–15, 2003. 35, 43, 44

CM Carlevaro, RM Irastorza, and F Vericat. Quaternionic representation of the genetic code. *Biosystems*, 141 :10–19, 2016. 28

O.K Clay and G Bernardi. Gc3 of genes can be used as a proxy for isochore base composition : A reply to elhaik et al. *Mol. Biol. Evol*, 1(28) :21–23, 2011. 37, 84

J.M. Comeron and M. Aguadé. An evaluation of measures of synonymous codon usage bias. *J. Mol. Evol.*, 3(47) :268–274, 1998. 43

DN Cooper and M Krawczak. Cytosine methylation and the fate of cpg dinucleotides in vertebrate genomes. *Human Genet.*, 83 :181–188, 1989. 89

F.H.C Crick. Codon-antidocon pairing : the wobble hypothesis. *J. Mol. Evol.*, 519 :548–555, 1966. 28

FHC Crick. Central dogma of molecular biology. *Nature*, 227 :561–563, 1970. 23, 24

FHC Crick, L Barnett, S Brenner, and R.J Watts-Tobin. General nature of the genetic code for proteins. *Nature*, 192 :1227–1232, 1961. 24, 25

PB Damgaard, A Margaryan, H Schroeder, L Orlando, E Willerslev, and ME Allentofta. Improving access to endogenous dna in ancient bones and teeth. *Sci Rep.*, 5 :11184, 2015. 48

V. Daubin and G Perrière. G+C3 structuring along the genome : a common feature in prokaryotes. *Mol. Biol. Evol.*, 20(4) :471–83, 2003. 34

M.O. Dayhoff, R. Schwartz, and B.C. Orcutt. A model of evolutionary change in proteins. atlas of protein sequence and structure. *Nat. Biomed. Res. Found.*, 5 :345–358, 1978. 28

H Dong, L Nilsson, and CG Kurland. Co-variation of trna abundance and codon usage in escherichia coli at different growth rates. *J Mol Biol.*, 260(5) :649–663, 1996. 35

WF Doolittle. Phylogenetic classification and the universal tree. *Science*, 284(5423) :2124–2128, 1999. 44

M. dos Reis, L. Wernisch, and R. Savva. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole escherichia coli k-12 genome. *Nucleic Acids Res*, 31 :6976–6985, 2003. 42

M. dos Reis, L. Wernisch, and R. Savva. Solving the riddle of codon usage preferences : a test for translational selection. *Nucleic Acids Res.*, 32(17) :5036–5044, 2004. 42

W Duchemin, V Daubin, and E. Tannier. Reconstruction of an ancestral yersinia pestis genome and comparison with an ancient sequence. *BMC Genomics*, 16 :S9, 2015. 48

L. Duret. Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.*, 12 (6) :640–9, December 2002. 37, 40, 135

L. Duret and N. Galtier. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics*, 10 :285–311, 2009. 89, 102, 104, 106, 131, 135

L Duret and M. Mouchiroud. Expression pattern and surprisingly gene length shape codon usage in caenorhabditis drosophila and arabidopsis. *PNAS*, 96 :4482–4487, 1999. 34

A Eyre-Walker and L Hurst. The evolution of isochores. *Nat. Rev. Genet*, 2 :549–555, 2001. 38, 83

J Felsenstein. Evolutionary trees from dna sequences : a maximum likelihood approach. *J. Mol. Evol*, 17(6) :368–376, 1981. 50

W. Fitch. Is there selection against wobble in codon-anticodon pairing? *Science*, 194(4270) : 1173–1174, 1976. 31, 32

W Fletcher and Z Yang. The effects of insertions, deletions and alignments errors on branch-site test to positive selection. *Mol Biol Evol.*, 27(10) :2257–2267, 2010. 46

BR. Francis. Evolution of the genetic code by incorporation of amino acids that improved or changed protein function. *J Mol Evol.*, 77(4) :134–138, 2013. 28

C. Francois. Évaluation des stratégies adaptatives des métazoaires aux faibles disponibilités en nutriments : couplage d'approches d'écologie isotopique et de transcriptomique chez des isopodes épigés et hypogés, 2015. 120, 121

SJ Freeland and LD. Hurst. The genetic code is one in a million. *J Mol Evol.*, 47(3) :238–48, 1998. 27

N. Galtier and L. Duret. Adaptation or biased gene conversion? extending the null hypothesis of molecular evolution. *Trends in Genetics*, 23 :273–277, 2007. 89, 102, 106, 135

N. Galtier, G. Piganeau, D. Mouchiroud, and L. Duret. Gc-content evolution in mammalian genomes : The biased gene conversion hypothesis. *Genetics*, 159(2) :907–911, 2001. 85, 135

N. Galtier, L. Duret, S. Glemin, and V. Ranwez. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet.*, 25(1) :1–5, 2009. 36, 89, 91, 92, 102, 106

S. Ghaemmaghami, WK. Huh, K. Bower, R.W. Howson, A. Belle, and N. Dephoure. Global analysis of protein expression in yeast. *Nature*, 425(6959) :737–741, 2003. 43

J. Gibert and L. Deharveng. Subterranean ecosystems : a truncated functional biodiversity. *BioScience*, 52(6) :473, 2002. 120

M.A. Gilchrist. Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns. *Mol Biol Evol*, 24(11) :2362–2372, 2007. 42

H. Gingold, D. Tehler, N R. Christoffersen, M M. Nielsen, F. Asmar, S M. Kooistra, N S. Christophersen, L. Lotte Christensen, M. Borre, K D. Sørensen, L D. Andersen, C L. Andersen, E. Hulleman, T. Wurdinger, E. Ralfkiær, K. Helin, K. Grønbæk, T. Orntoft, S M. Waszak, O. Dahan, J S. Pedersen, A H Lund, and Y. Pilpel. A dual program for translation regulation in cellular proliferation and differentiation. *Cell*, 158(6) :1281–92, 2014. 35, 106, 135, 136, 137

N Goldman and Z Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, 11(5) :725–36, 1994. 53

RA Goldstein and DD Pollock. Observations of amino acid gain and loss during protein evolution are explained by statistical bias. *Mol Biol Evol*, 23 :1444–1449, 2006. 90

M. Gouy and C. Gautier. Codon usage in bacteria : correlation with gene expressivity. *Nucleic Acids Res.*, 10(22) :7055–7074, 1982. 42

R Grantham, C Gautier, M Gouy, R Mercier, and A Pavé. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.*, 8(1) :r49–r62, 1980. 34

M. Gribskov, J. Devereux, and R.R Burgess. The codon preference plot : graphic analysis of protein coding sequences and prediction of gene expression. *Nucleic Acids Res*, 12(1 pt 2) : 539–549, 1984. 42, 43

L. Guéguen, S. Gaillard, B. Boussau, M. Gouy, M. Groussin, N C. Rochette, T. Bigot, D. Fournier, F. Pouyet, V. Cahais, A. Bernard, C. Scornavacca, B. Nabholz, A. Haudry, L. Dachary, N. Galtier, K. Belkhir, and J Y. Dutheil. Bio++ : Efficient Extensible Libraries and Tools for Computational Molecular Evolution. *Mol. Biol. Evol.*, 30(8) :1745–1750, 2013. 53, 62, 92, 94

A. L. Halpern and W. J. Bruno. Evolutionary distances for protein-coding sequences : modeling site-specific residue frequencies. *Mol. Biol. Evol.*, 15(7) :910–917, 1998. 56

M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2) :160–74, 1985. 50

S Henikoff and JG Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A.*, 89(22) :10915–9, 1992. 28

R. Hershberg and D A. Petrov. Selection on codon bias. *Annu. Rev. Genet.*, 42 :287–99, 2008. 111

R. Hershberg and D A. Petrov. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.*, 6(9) :e1001115, 2010. 36

F. Hervant and D. Renault. Long-term fasting and realimentation in hypogean and epigean isopods : a proposed adaptive strategy for groundwater organisms. *The Journal of Experimental Biology*, 205(14) :2079–2087, 2002. 120

PG Higgs. A four-column theory for the origin of the genetic code : tracing the evolutionary pathways that gave rise to an optimized code. *Biol Direct*, pages 4–16, 2009. 28

F. Hildebrand, A. Meyer, and A. Eyre-Walker. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet.*, 6(9) :e1001107, 2010. 111

A Hobolth, R Nielsen, Y Wang, F Wu, and SP Tanksley. Cpg and cpnpg analysis of protein-coding sequences from tomato. *Mol Biol Evol*, 23(6) :1318–1323, 2006. 89

R Holliday and GC Grigg. DNA methylation and mutation. *Mutat. Res*, 285 :61–67, 1993. 89, 102

J.P. Huelsenbeck. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and Systematics*, 28 :437–466, 1997. 48

LD Hurst, EJ Fell, and EPC Rocha. Protein evolution : causes of trends in amino-acid gain and loss. *Nature*, 442 :E11–E12, 2006. 90

S. Höhna, T. A. Heath, B. Boussau, M. J. Landis, F. Ronquist, and J. P. Huelsenbeck. Probabilistic graphical model representation in phylogenetics. *Mol. Biol. Evol.*, 32(4) :1097–1108, 2015. 53

K. Hüppop. Food-finding ability in cave fish (astyanax fasciatus). *International Journal of Speleology*, 16 :59–66, 1987. 120

K. Hüppop. Adaptation to low food. in encyclopedia of caves. *Elsevier, 2nd edition.*, pages 1–9, 2012. 120

T. Ikemura. Correlation between the abundance of escherichia coli transfer rnas and the occurence of the respective codons in its proteins. *J. Mol. Biol.*, 146(1) :1–21, 1981a. 34, 41

T. Ikemura. Codon usage and trna content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, 1(2) :13–34, 1985. 34, 40, 111

W. R. Jeffery. Regressive evolution in astyanax cavefish. *Annual Review of Genetics*, 43 :25–47, 2009. 120

I King Jordan, Fyodor A Kondrashov, Ivan A Adzhubei, Yuri I Wolf, Eugene V Koonin, Alexey S Kondrashov, and Shamil Sunyaev. A universal trend of amino acid gain and loss in protein evolution. *Nature*, 433(7026) :633–8, February 2005. doi : 10.1038/nature03306. 90

T. H. Jukes and C.R. Cantor. Evolution of Protein Molecules. pages 21–132. New York : academic Press, 1969. 49

TH Jukes. Codons and nearest-neighbor nucleotide pairs in mammalian messenger rna. *J Mol Evol*, 11(2) :121–127, 1978. 89

T. J. Kawecki, R. E. Lenski, D. Ebert, B. Hollis, I. Olivieri, and M. C. Whitlock. Experimental evolution. 10(27) :547–560, 2012. 48

M Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol*, 16(2) :111–120, 1980. 50

C. Kosiol, T. Vinaŕ, R.R. da Fonseca, M.J Hubisz, C.D. Bustamante, R Nielsen, and A Siepel. Patterns of positive selection in six mammalian genomes. *Plos Genetics*, 4(8) :e10000144, 2008. 94, 95

L Kubatko, P Shah, R Herbei, and MA Gilchrist. A codon model of nucleotide substitution with selection on synonymous codon usage. *Mol Phyl Evol.*, 2015. 56, 61

N Lartillot. Phylogenetic patterns of gc-biased gene conversion in placental mammals and the evolutionary dynamics of recombination landscapes. *Mol. Biol. Evol.*, 30 :489–502, 2013. 94, 95, 102, 106

Y. Lesecque. La conversion génique biaisée : origine, dynamique et intensité de la quatrième force d'évolution des génomes eucaryotes., 2014. 86, 88, 95

Y Lesecque, D Mouchiroud, and L Duret. Gc-biased gene conversion in yeast is specifically associated with crossovers : molecular mechanisms and evolutionary significance. *Mol. Biol. Evol.*, (30) :1409–1419, 2013. 88

Y. Lesecque, S. Glemin, N Lartillot, Mouchiroud D, and L Duret. The red queen model of recombination hotspots evolution in the light of archaic and modern human genomes. *PLoS genetics*, 10 :1004790, 2014. 87

JR. Lobry. Properties of a General Model of DNA Evolution Under No-Strand-Bias Conditions. *Journal of Molecular Evolution*, 40 :326–330, 1995. 51

E. Mancera, R. Bourgon, A. Brozzi, W. Huber, and L. M. Steinmetz. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, 454 :479–485, 2008. 87

GAB Marais and L Duret. Synonymous codon usage accuracy of translation and gene length in caenorhabditis elegans. *J. Mol. Evol.*, 52 :275–280, 2001. 36

R Marquez, S Smit, and R Knight. Do universal codon-usage patterns minimize the effects of mutation and translation error? *Genome Biol.*, 6(11) :R91, 2005. 35

SE Massey. Genetic code evolution reveals the neutral emergence of mutational robustness, and information as an evolutionary constraint. *Life (Basel)*, 5(2) :1301–1332, 2015. 28

G. A. McVean and J. Vieira. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. *Genetics*, 157(1) :245–257, 2001. 55

K Misawa and RF Kikuno. Evaluation of the effect of cpg hypermutability on human codon substitution. *Gene*, 431 :18–22, 2009. 89, 90, 92

K Misawa, N Kamatani, and RF Kikuno. The universal trend of amino acid gain–loss is caused by cpg hypermutability. *J Mol Evol*, 67 :334–342, 2008. 89

C. Morvan, F. Malard, E. Paradis, T. Lefébure, L. Konecny-Dupré, and C. J. Douady. Timetree of aselloidea reveals species diversification dynamics in groundwater. *Systematic Biology*, 62 : 512–522, 2013. 121

D Mouchiroud, G D'Onofrio, B Aïssani, G Macaya, C Gautier, and G Bernardi. The distribution of genes in the human genome. *Gene*, 100 :181–187, 1991. 38

CF Mugal, JB Wolf, and I. Kaj. Why time matters : codon evolution and the temporal dynamics of dn/ds. *Mol Biol Evol.*, 1(31) :212–231, 2014. 54

CF Mugal, PF Arndt, Holm L, and H Ellegren. Evolutionary consequence of dna methylation on the gc content in vertebrates genomes. *Genes*, 5 :441–447, 2015. 83, 89, 90

S. V. Muse and B. S. Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.*, 11(5) :715–724, 1994. 53

T. Nagylaki. Evolution of a finite population under gene conversion. *Proc. Natl. Acad. Sci. U.S.A.*, 80 :6278–6281, 1983. 90

MW Nirenberg and P Leder. RNA codewords and protein synthesis. The effect of trinucleotides upon the binding of srna to ribosomes. *Science*, 145 :1399–1407, 1964. 26

MW Nirenberg and JH Matthaei. The dependence of cell-free protein synthesis in e. coli upon naturally occurring or synthetic polyribonucleotides. *PNAS*, 47 :1588–1602, 1961. 25

S. Nishimura, T.M Jacob, and H.G Khorana. Synthetic deoxyribopolynucleotides as templates for ribonucleic acid polymerase : the formation and characterization of a ribopolynucleotide with a repeating trinucleotide sequence. *PNAS*, 52(6) :1494–1501, 1964. 26

J A. Novembre. Accounting for background nucleotide composition when measuring codon usage bias. *Mol. Biol. Evol.*, 19(8) :1390–4, 2002. 40

E.M Novoa and L Ribas de Pouplana. Speeding with control : codon usage, trnas, and ribosomes. *Trends in Genetics*, 28(11) :574–581, 2012. 34, 35

P.A. Nuñez, H Romero, M.D. Farber, and EPJ Rocha. Natural selection for operons depends on genome size. *Genome Biol. Evol.*, 5(11) :2242–2254, 2013. 117

M Jr Ochoa and IB Weinstein. Amino acid coding in a subsellular system derived from the l1210 mouse ascites leukemia. *PNAS*, 52 :470–477, 1964. 26

T. L. Poulson and W. B. White. The cave environment. *Science*, 165(3897) :971–981, 1969. 120

E P C. Rocha. Codon usage bias from tRNA's point of view : redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.*, 14(11) :2279–86, 2004. 34

N. Rodrigue, H. Philippe, and N. Lartillot. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc. Natl. Acad. Sci. U.S.A.*, 107(10) :4629–4634, Mar 2010. 56, 132

A Rokas and SB Carroll. Bushes in the tree of life. *PLOS Biology*, 4(11) :e352, 2006. 44

U.D. Roymondal and S.S. Sahoo. Predicting gene expression level from relative codon usage bias : an application to escherichia coli genome. *DNA Res.*, 16(1) :13–30, 2009. 42

K.L Rudolph, B.M Schmitt, D Villar, R.J White, J.C Marioni, D.T. Odom, and C Kutter. Codon-driven translational efficiency is stable across diverse mammalian cell states. *PLoS Genetics*, page in press, 2016. 136

V Scaiewicz, V Sabbía, R Piovani, and P Musto. Cpg islands are the second main factor shaping codon usage in human genes. *Bioch. and Biophys. Res. Comm.*, 343 :1257–1261, 2006. 89, 96

B.M Schmitt, K.L Rudolph, P Karagianni, NA Fonseca, RJ White, I Talianidis, DT Odom, JC Marioni, and C. Kutter. High-resolution mapping of transcriptional dynamics across tissue development reveals a stable mrna-trna interface. *Genome Res.*, 24(11) :1197–1807, 2014. 136

P M. Sharp and W H. Li. The Codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, 15(3) :1281–95, 1987. 34, 35, 41, 111

P M. Sharp, T M. Tuohy, and K R. Mosurski. Codon usage in yeast : cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.*, 14(13) :5125–43, 1986. 34, 39

P.M. Sharp, M. Stenico, J.F. Peden, and A.T. Lloyd. Codon usage : mutational bias, translational selection, or both? *Biochem Soc Trans.*, 21(4) :835–841, 1993. 33

D. Soares and M. L. Niemiller. Sensory adaptations of fishes to subterranean environments. *BioScience*, 63(4) :274–283, 2013. 120

S. J. Spielman and C. O. Wilke. The Relationship between dN/dS and Scaled Selection Coefficients. *Mol. Biol. Evol.*, 32(4) :1097–1108, 2015. 56

N Stoletzki and A Eyre-Walker. Synonymous codon usage in Escherichia coli : selection for translational accuracy. *Mol. Biol. Evol.*, 24(2) :374–81, 2007. 35, 111

N. Sugaya, M. Sata, H. Murakami, A. Imaizumi, S. Aburatani, and K. Horimoto. Causes for the large genome size in a cyanobacterium anabaena sp. pcc7120. *Genome Inform*, 1(15) : 229–238, 2004. 43, 44

F. Supek and K. Vlahovicek. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *BMC Bioinformatics*, 6 :182, 2005. 43

H. Suzuki, R. Saito, and M. Tomita. The 'weighted sum of relative entropy' : a new index for synonymous codon usage bias. *Gene*, 335 :19 – 23, 2004. 42

H. Suzuki, C.J Brown, L.J Forney, and E.M. Top. Comparison of correspondence analysis methods for synonymous codon usage in bacteria. *DNA Res*, 6(15) :357–365, 2008. 43

J. W. Szostak, T. L. Orr-Weaver, R. J. Rothstein, and F. W. Stahl. The double-strand-break repair model for recombination. *Cell*, pages 25–35, 1983. 87

K. Tamura. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G + C-content biases. *Mol. Biol. Evol.*, 9(4) :678–687, 1992. 50

A U. Tamuri, M. dos Reis, and R A. Goldstein. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics*, 190(3) : 1101–1115, 2012. 56

S. Tavaré. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences, American Mathematical Society*, 17 :57–86, 1986. 51, 93

S Thorvaldsen. A mutation model from the principles of the genetic code. *IEEE/ACM Transactions on Comp. bio and Bioinformatics*, 2015. 28

A. O Urrutia and L.D Hurst. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is no evidence for selection. *Genetics*, 159(3) : 1191–1199, 2001. 42

M. P. Venarsky, B. M. Huntsman, A. D. Huryn, J. P. Benstead, and B. R. Kuhajda. Quantitative food web analysis supports the energy-limitation hypothesis in cave stream ecosystems. *Oecologia*, 176(3) :859–869, 2014. 120

T. von der Haar. A quantitative estimation of the global translational activity in logarithmically growing yeast cells. *BMC Syst Biol.*, 2 :87, 2008. 41

A.L. Williams, G Genovese, T Dyer, N Altemose, K Truax, G Jun, N Patterson, S.R. Myers, J.E. Curran, R Ravi Duggirala, J Blangero, D Reich, and M Przeworski. Non-crossover gene conversions show strong gc bias and unexpected clustering in humans. *eLIFE*, (4) :e04637, 2015. 88

K. Wolfe, P. Sharp, and W. Li. Mutation rates differ among regions of the mammalian genome. *Nature*, 337 :283–285, 1989. 85

F. Wright. The 'effective number of codons' used in a gene. *Gene*, 87(1) :23–29, 1990. 15, 39

Y Xu, P Ma, P Shah, A Rokas, Y Liu, and CH Johnson. Non-optimal codon usage is a mechanism to achieve circadian clock conditionality. *Nature*, 495(7439) :116–120, 2013. 35

Z Yang. On the varied pattern of evolution of 2 fungal genomes : a critique of hughes and friedman. *Mol Biol Evol.*, 23(12) :2279–2282, 2006. 52

Z Yang. Paml 4 : phylogenetic analysis by maximum likelihood. *Mol Biol Evol.*, 24(8) :1586–1591, 2007. 53

Z. Yang and R. Nielsen. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol.*, 46 :409–418, 1998. 53

Z. Yang and R. Nielsen. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol. Biol. Evol.*, 25(3) :568–79, 2008. 55, 62

V.B. Yap and T.P. Speed. Modeling DNA Base Substitution in Large Genomic Regions from Two Organisms. *J Mol Evol*, 58(1) :12–18, 2004. 51

H Ying and G Huttley. Exploiting cpg hypermutability to identify phenotypically significant variation within human protein-coding genes. *Genome Biol. Evol.*, 3 :938–949, 2011. 90, 92, 93

M Zaheri, L Dib, and N Salamin. A generalized mechanistic codon model. *Mol Biol Evol.*, 31 (9) :2528–2541, 2014. 54

APPENDIX A

Supplementary Material of: SENCA : a multi-layered codon model to study the origins and dynamics of codon usage paper

Supplementary Material

Proof of equation 1

In the case of $x \neq y$:

$$\frac{g(y,x)}{g(x,y)} = \frac{\frac{-log\left(\frac{y}{x}\right)}{1-\frac{y}{x}}}{\frac{-log\left(\frac{x}{y}\right)}{1-\frac{x}{y}}}$$

$$\frac{g(y,x)}{g(x,y)} = \frac{\frac{-(log(y)-log(x))}{\frac{x-y}{x}}}{\frac{-(log(x)-log(y))}{\frac{y-x}{y}}}$$

$$\frac{g(y,x)}{g(x,y)} = \frac{-1 * x * (y-x)}{y * (x-y)}$$

$$\frac{g(y,x)}{g(x,y)} = \frac{x}{y}$$

Considerations on the identifiability of SENCA

We consider a SENCA modeling based on a reversible nucleotidic model, with a given set of parameters. This model defines a generator Q.

We want to check if, when the nucleotidic model is changed, it not possible to reparametrize the C layer to obtain a model with the same generator Q.

We show that strong constraints are necessary on the initial codon preferences so that it would be possible, making it very unlikely that is occurs. And we show that this constraint is impossible in a simpler case where only the equilibrium frequency of the model is changed, where the equilibrium frequencies of G and C (as well as A and T) are equal.

We consider an amino acid with 2, 3 or 4 synonymous codons, denoted as $I_1 I_2 x$. In this redundancy class, we denote the codons as $C_x = I_1 I_2 x$.

Since the nucleotidic model is reversible, its substitution rates can be defined as $\mu_{xy} \propto \pi_y s_{xy}$ where $s_{xy} = s_{yx}$ and π_y is the equilibrium frequency of nucleotide y.

Then, if $q_{C_x C_y}$ is the SENCA substitution rate from codon C_x to codon C_y, we have:

$$q_{C_x C_y} \propto s_{xy} \pi_y \frac{-\log(\frac{\phi_x}{\phi_y})}{1 - \frac{\phi_x}{\phi_y}}$$

where ϕ_x is the simplified notation of the preference of codon C_x in the codon layer.

From this parametrization, the equilibrium frequency of these synonymous codons is : $q^*_{C_x} \propto \pi_x \phi_x$ because all have the same amino acid preference and first two nucleotides.

Now we consider a different nucleotidic model with equilibrium frequencies π'_x and exchangibility rates s'_{xy}, and we look for ϕ'_x preferences such that the resulting SENCA generator is the same.

First, the equilibrium frequencies should be the same: $q^*_{C_x} \propto \pi'_x \phi'_x$. So there is a constant K such that, $\forall x, \pi_x \phi_x = K \pi'_x \phi'_x$. Then, $\forall x, y$, $\frac{\phi'_x}{\phi'_y} = \frac{\pi_x \pi'_y}{\pi_y \pi'_x} \frac{\phi_x}{\phi_y} = \frac{\delta_x}{\delta_y} \frac{\phi_x}{\phi_y}$ where $\delta_x = \frac{\pi_x}{\pi'_x}$.

Second, substitution rates should be the same, which means that

$$q_{C_x C_y} \propto s'_{xy} \pi'_y \frac{-\log(\frac{\phi'_x}{\phi'_y})}{1 - \frac{\phi'_x}{\phi'_y}}$$

So there is a constant K' such that

$$\forall x, y, s_{xy}\pi_y \frac{-\log(\frac{\phi_x}{\phi_y})}{1-\frac{\phi_x}{\phi_y}} = K'\pi'_y s'_{xy} \frac{-\log(\frac{\phi'_x}{\phi'_y})}{1-\frac{\phi'_x}{\phi'_y}}$$

then

$$\forall x, y, s_{xy}\delta_y \frac{-\log(\frac{\phi_x}{\phi_y})}{1-\frac{\phi_x}{\phi_y}} = K' s'_{xy} \frac{-\log(\frac{\delta_x\phi_x}{\delta_y\phi_y})}{1-\frac{\delta_x\phi_x}{\delta_y\phi_y}}$$

so

$$\forall x, y, z, \frac{s_{xy}}{s_{zy}} \frac{-\log(\frac{\phi_x}{\phi_y})}{1-\frac{\phi_x}{\phi_y}} \cdot \frac{1-\frac{\phi_z}{\phi_y}}{-\log(\frac{\phi_z}{\phi_y})} = \frac{s'_{xy}}{s'_{zy}} \frac{-\log(\frac{\delta_x\phi_x}{\delta_y\phi_y})}{1-\frac{\delta_x\phi_x}{\delta_y\phi_y}} \cdot \frac{1-\frac{\delta_z\phi_z}{\delta_y\phi_y}}{-\log(\frac{\delta_z\phi_z}{\delta_y\phi_y})} \quad (1)$$

This means that there is a very strong necessary condition linking the ϕ_x and both nucleotidic models to build the same SENCA generator in both conditions. Moreover, a resulting constraint on the ϕ_x is that they are the same for all amino acids with redundancy 2, 3, and 4. So in any other case (which is actually the case from all real data), the model is identifiable.

In the case when the ϕ depend only depend on the nucleotide at the 3rd position, it is tedious to prove if the necessary condition is possible in the most general case. So we restrict to the models where $\pi_A = \pi_T$ and $\pi_C = \pi_G$, with only a change in the equilibrium frequency. In this case, if we denote $\theta = \pi_C + \pi_G$, then $\pi_C = \pi_G = \frac{\theta}{2}$ and $\pi_A = \pi_T = \frac{1-\theta}{2}$ and similarly $\pi'_C = \pi'_G = \frac{\theta'}{2}$ and $\pi'_A = \pi'_T = \frac{1-\theta'}{2}$.

Then

$\forall x \in \{A, T\}, \forall y \in \{C, G\}, \frac{\phi'_x}{\phi'_y} = \frac{(1-\theta)\theta'}{(1-\theta')\theta}\frac{\phi_x}{\phi_y} = h(\theta, \theta')\frac{\phi_x}{\phi_y}$

We consider in the following that there is a change in the equilibrium of the nucleotidic model, which means that $h(\theta, \theta') \neq 1$. From equation (1), with $x = A$, $y = C$ and $z = T$:

$$\frac{-\log(\frac{\phi_A}{\phi_C})}{1-\frac{\phi_A}{\phi_C}} \cdot \frac{1-\frac{\phi_T}{\phi_C}}{-\log(\frac{\phi_T}{\phi_C})} = \frac{-\log(h(\theta,\theta')\frac{\phi_A}{\phi_C})}{1-h(\theta,\theta')\frac{\phi_A}{\phi_C}} \cdot \frac{1-h(\theta,\theta')\frac{\phi_T}{\phi_C}}{-\log(h(\theta,\theta')\frac{\phi_T}{\phi_C})}$$

There are solutions to this equation if there exist x, y such that for the function $f_{a,b}(\alpha) = \frac{-\log(\alpha a)}{1-\alpha a} \cdot \frac{1-\alpha b}{-\log(\alpha b)}$ there exists $\alpha \neq 1$ such that $f_{a,b}(\alpha) = f_{a,b}(1)$.

We can prove that if $a > b$ (resp. $a < b$) $f_{a,b}$ is strictly decreasing (resp. increasing), so the above equality is possible only if $a = b$, for which $f_{a,a}(\alpha) = 1$. Then $\phi_A = \phi_T$.

Symmetrically, with $x = C$, $z = G$ and $y = A$ in equation (1), $\phi_C = \phi_G$.

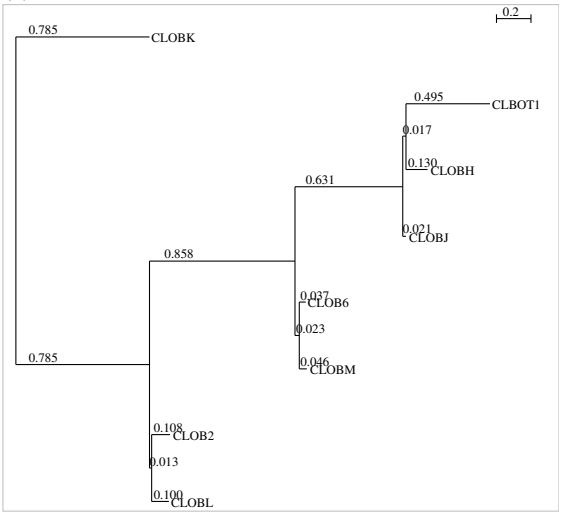
Now, in equation (1), with $x = A$, $y = T$ and $z = C$,

$$\frac{1 - \frac{\phi_C}{\phi_T}}{-\log(\frac{\phi_C}{\phi_T})} = \frac{1 - h(\theta', \theta)\frac{\phi_C}{\phi_T}}{-\log(h(\theta', \theta)\frac{\phi_C}{\phi_T})}$$

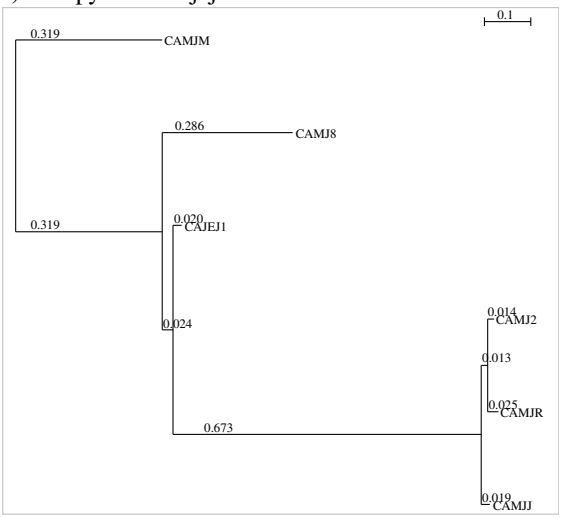
But function $g(a) = \frac{1-a}{-log(a)}$ is strictly monotonous, so since $h(\theta, \theta') \neq 1$ the above equality is impossible.

Phylogenetic trees

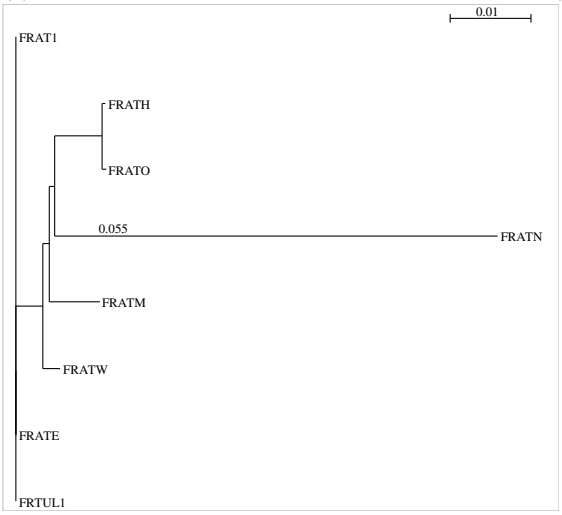
(a) Clostridium botulinum



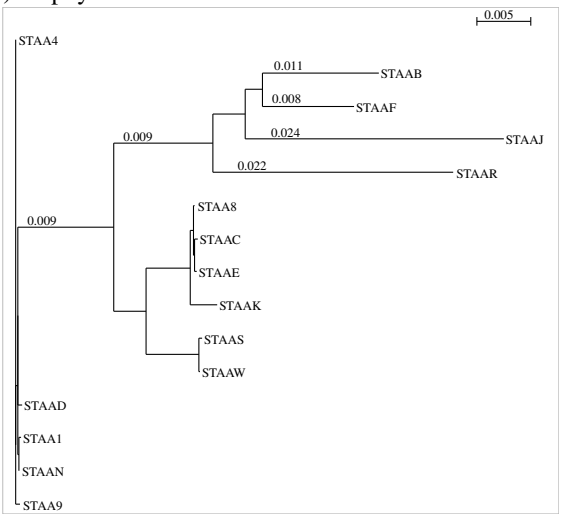
(b) Campylobacter jejunii



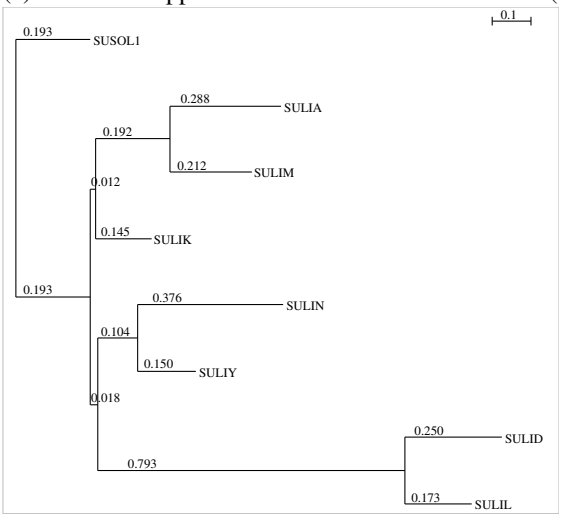
(c) Francisella tularensis



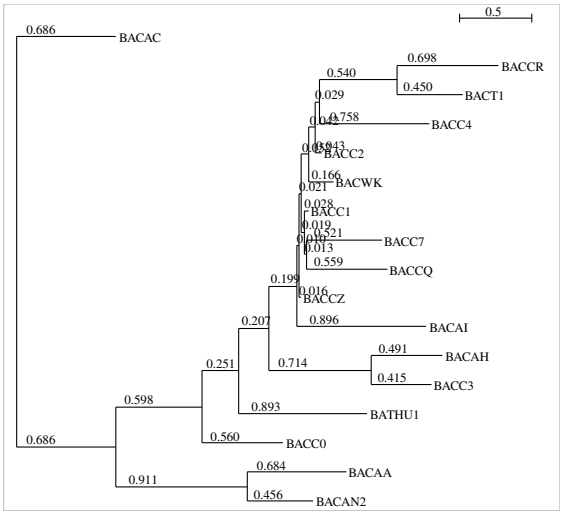
(d) Staphylococcus aureus



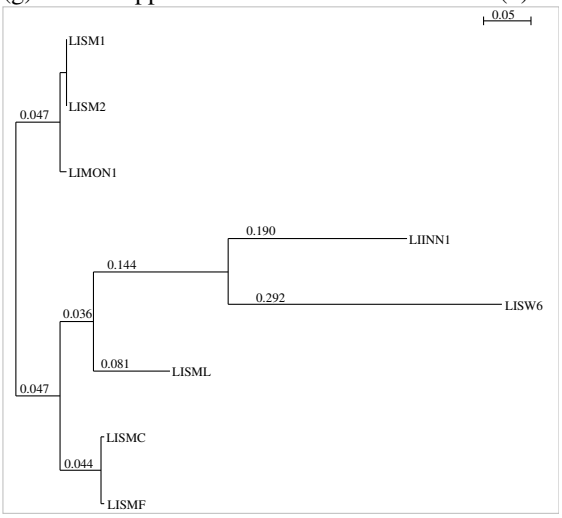
(e) Sulfolobus spp.



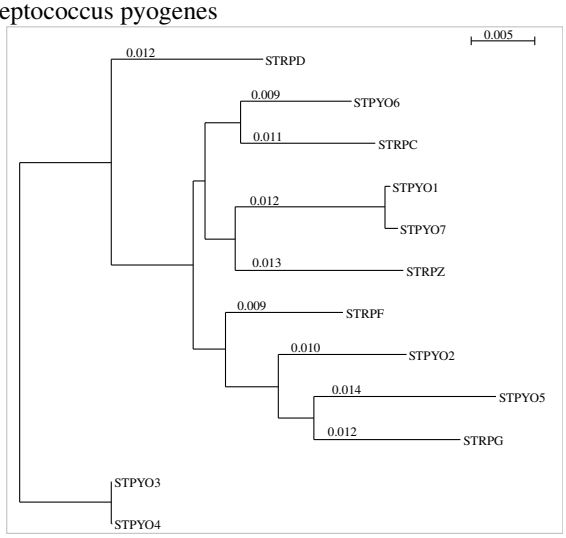
(f) Bacillus anthracis



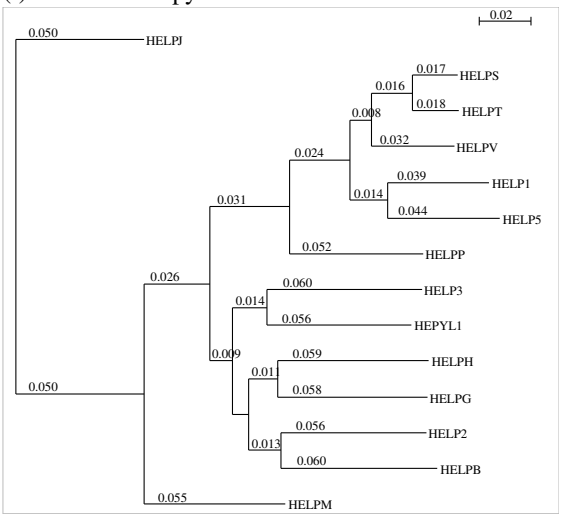
(g) Listeria spp.



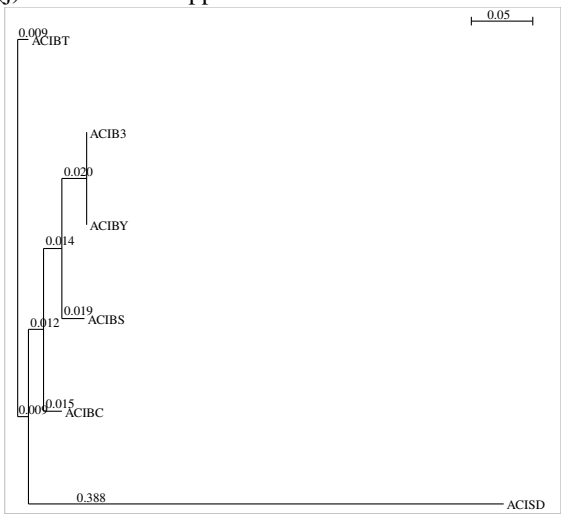
(h) Streptococcus pyogenes



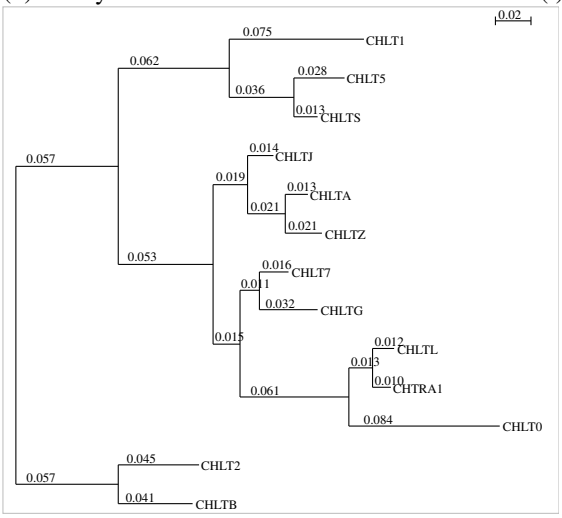
(i) Helicobacter pylori



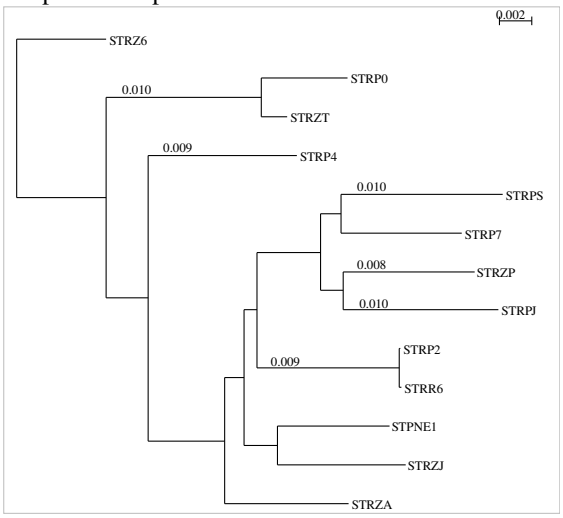
(j) Acinetobacter spp.



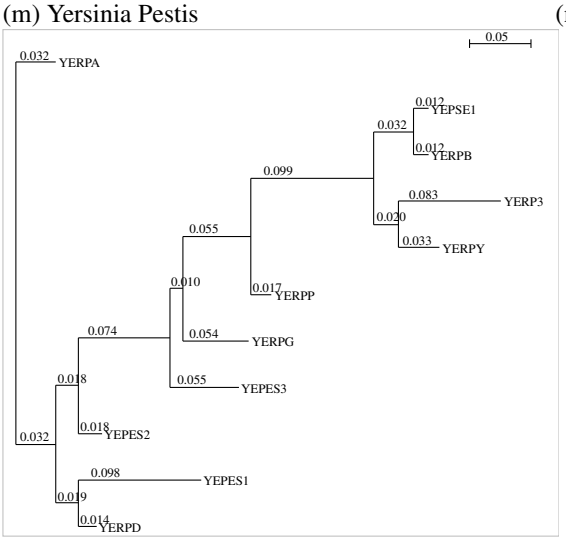
(k) Clamydia trachomatis



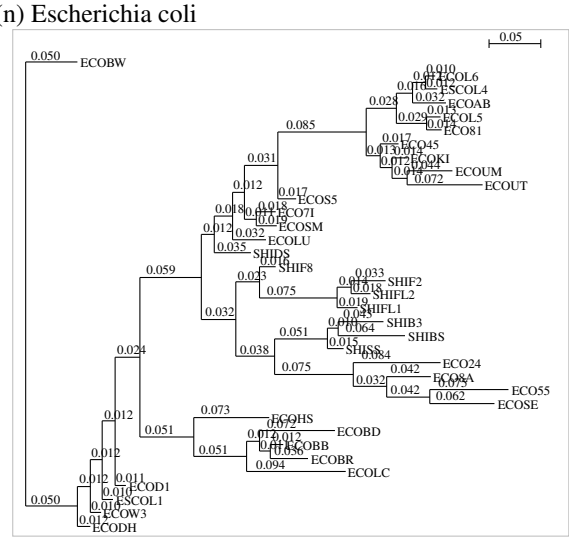
(l) Streptococcus pneumoniae



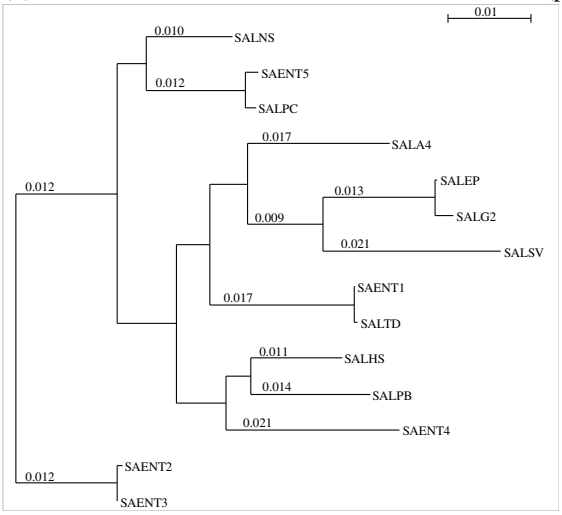
(m) Yersinia Pestis



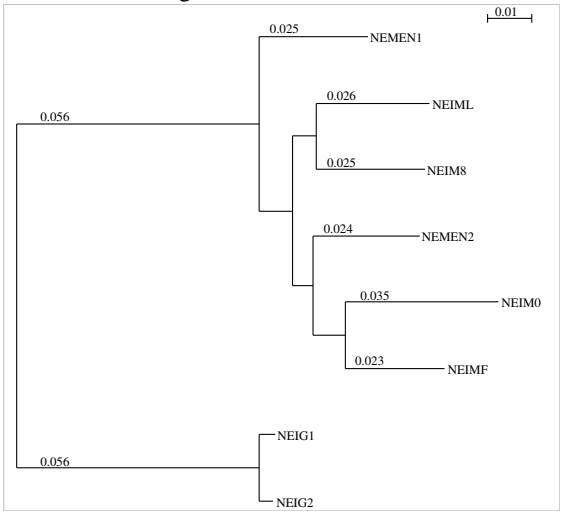
(n) Escherichia coli



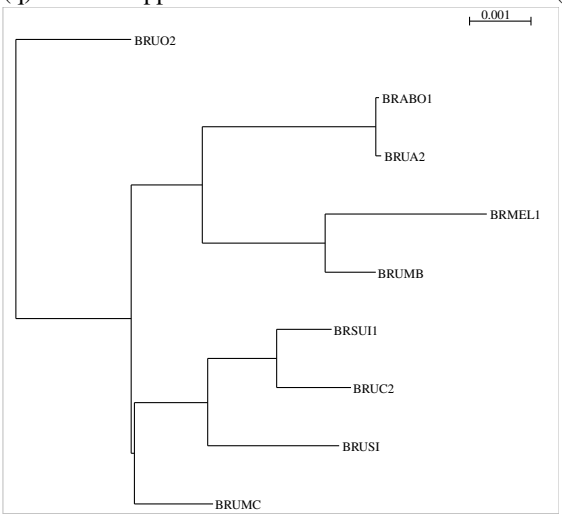
(o) Salmonella enterica



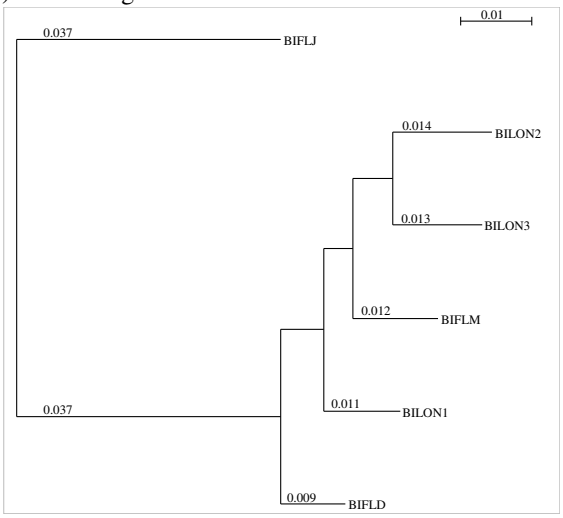
(p) Neisseria meningitidis



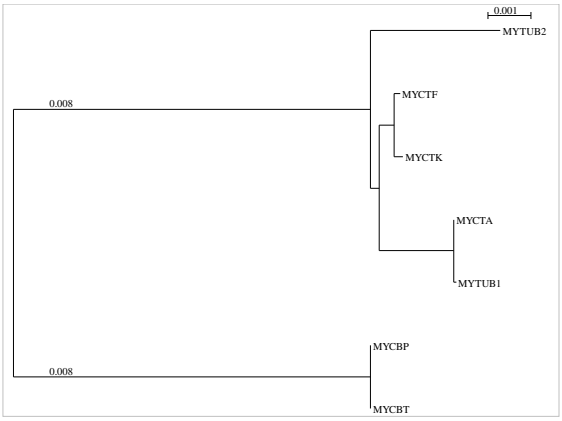
(q) Brucella spp.



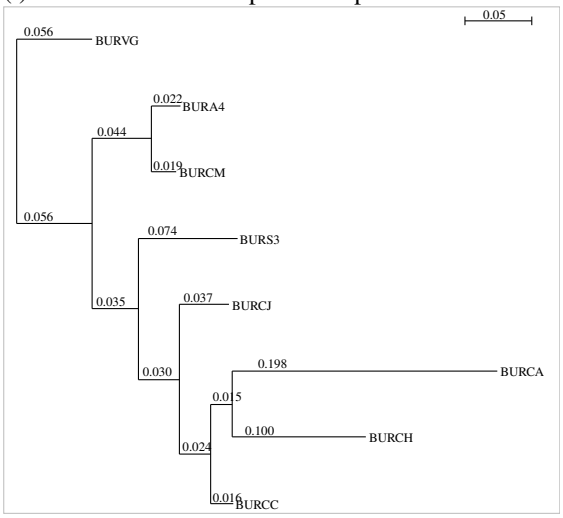
(r) Bifido longum



(s) Mycobacterium tuberculosis complex



(t) Burkholderia cenocepacia complex



(u) Burkholderia mallei

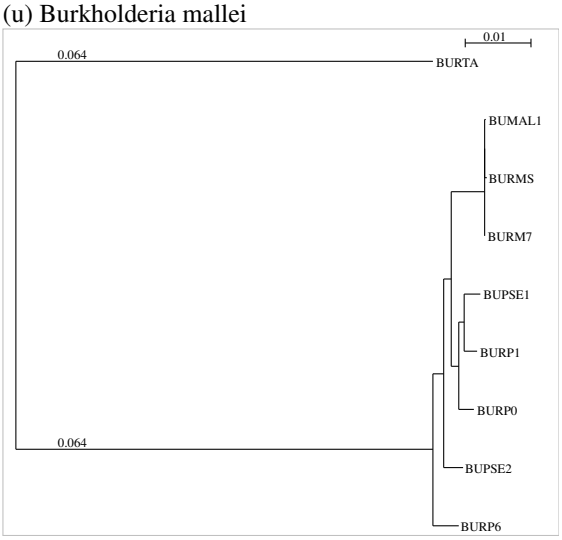


Figure S1: Trees used for intraspecific analyses.

Interspecies phylogenetic tree of Enterobacteria

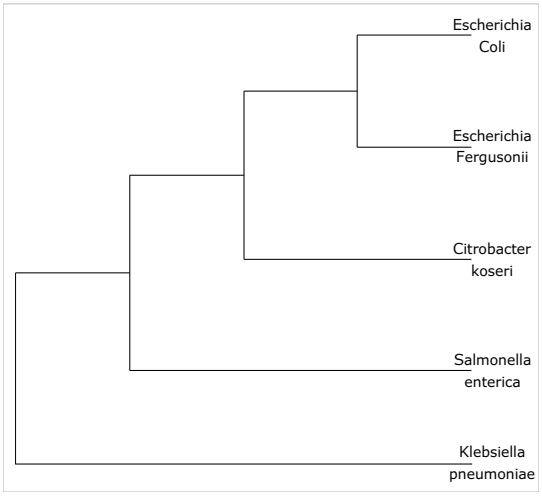


Figure S2: Enterobacteria species tree for the interspecies study – cladogram defined from the Hogenom Database reference species tree.

Table of maximum likelihood of SENCA and YN98

Table S1: Maximum likelihood values of SENCA, SENCA with one layer fixed to null, and YN98. LRT between SENCA and nested models, AIC and BIC values and number of sites and informative sites. See .txt file attached for full results. For the interspecies study, the number of parameters is increased by 19 (the A layer not being considered stationary).

Simulations

Tree used for simulations

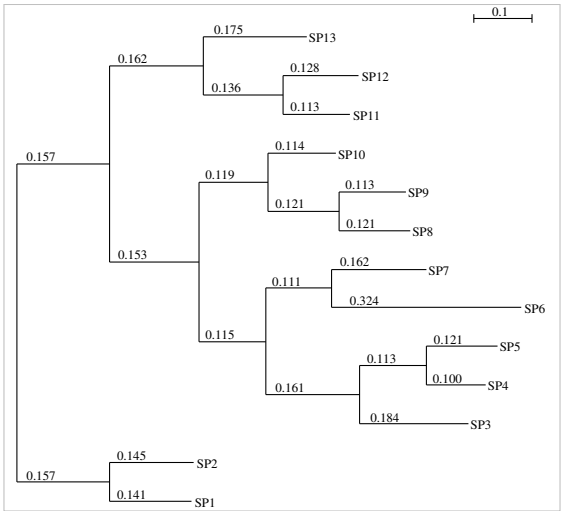


Figure S3: Tree used for simulations.

Simulations parameters values of AA preferences at equilibrium.

$\psi_{Ala} = 0.01754$, $\psi_{Arg} = 0.148254$, $\psi_{Asn} = 0.011366$, $\psi_{Asp} = 0.00557$, $\psi_{Cys} = 0.03446$, $\psi_{Gln} = 0.00517$, $\psi_{Glu} = 0.02047$, $\psi_{Gly} = 0.098252$, $\psi_{His} = 0.053056$, $\psi_{Ile} = 0.011673$, $\psi_{Leu} = 0.068634$, $\psi_{Lys} = 0.071684$, $\psi_{Met} = 0.010036$, $\psi_{Phe} = 0.018666$, $\psi_{Pro} = 0.0992$, $\psi_{Ser} = 0.101184$, $\psi_{Thr} = 0.072456$, $\psi_{Trp} = 0.019267$,$\psi_{Tyr} = 0.046754$, $\psi_{Val} = 0.086312$.

Simulation results with different N models

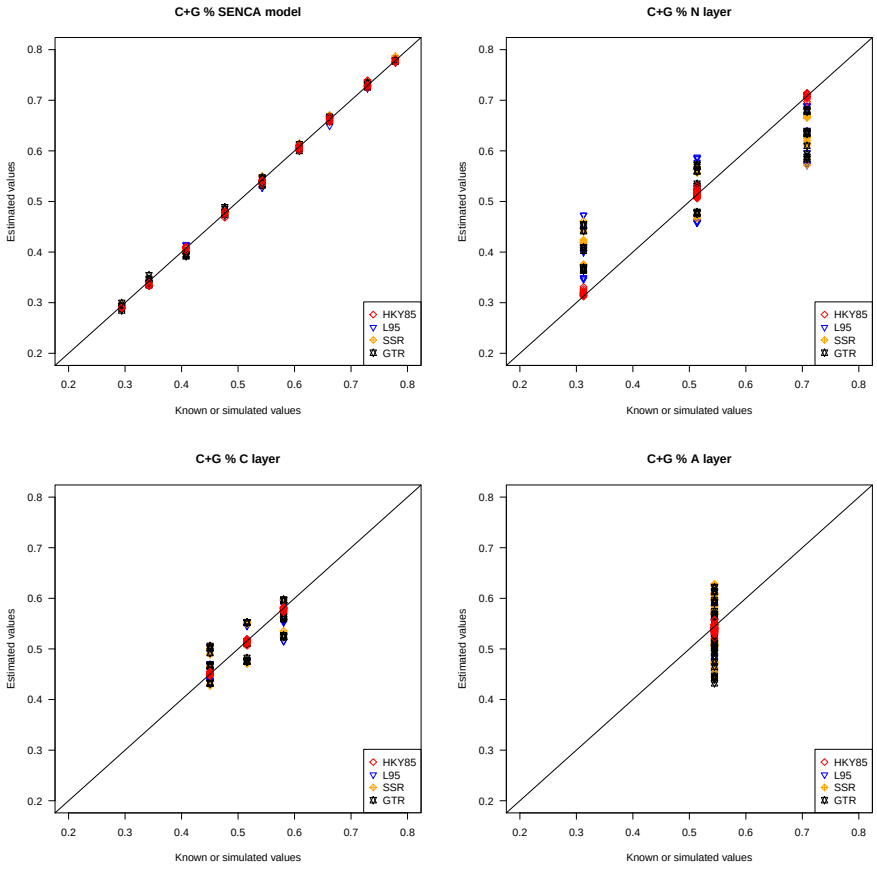
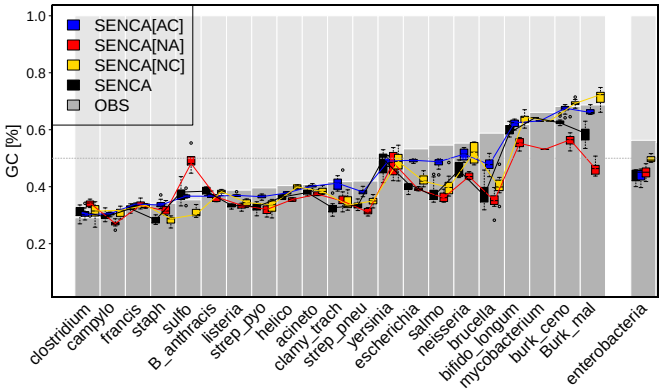


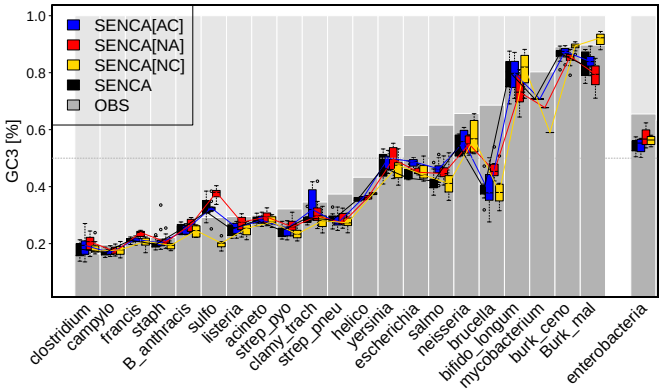
Figure S4: Simulations values for HKY85 (red dots), L95 (blue dots), SSR (orange dots) nucleotidic model and GTR (black dots).

Results with one layer fixed to the null hypothesis

(a) GC estimates



(b) $GC3$ estimates



(c) ENC estimates

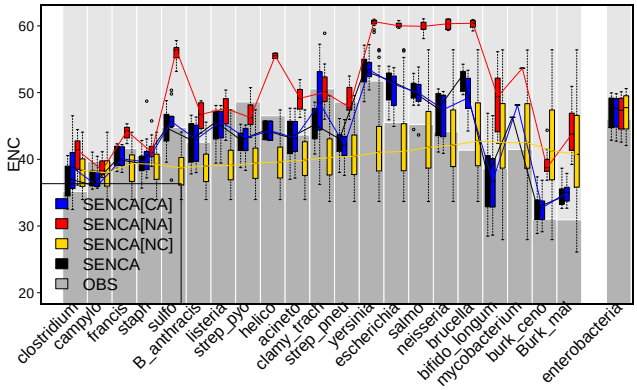
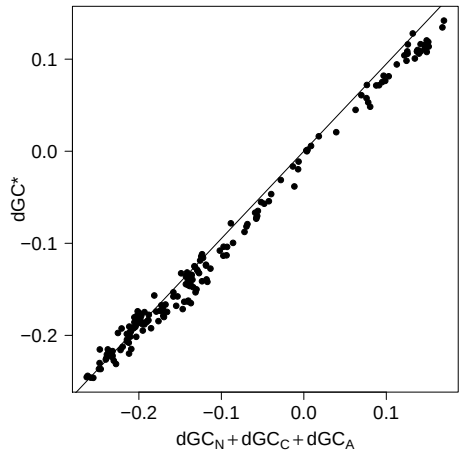


Figure S5: Results at equilibrium of different estimations of $GC^\star$, $GC3^\star$ and $ENC^\star$ with SENCA model and the model with one layer fixed at the time. In blue, $SENCA_{[AC]}$ stands for SENCA with no optimization of the N layer, in red, $SENCA_{[NA]}$ without the C layer optimization and in yellow, $SENCA_{[NC]}$ without the optimization of the A layer. In $SENCA_{[NC]}$, the amino acids have a fixed preference of 1/20, a biologically meaningless value, leading the C and N layers to compensate this strong assumption by being highly biased. Hence, $ENC^\star$ is more or less constant and highly biased around 0.4.

Figures on $GC^\star$

(a) Linear regression of $dGC^\star$



(b) Linear regression of $dGC3^\star$

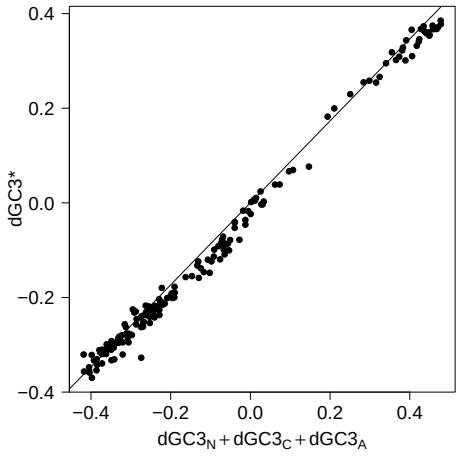


Figure S6: Linear regression between $dGC^\star$ contents. Figure (a) shows the linear regression between $dGC^\star$ and $dGC^\star_N + dGC^\star_C + dGC^\star_A$. Slope of the linear model is 0.95, $R^2 = 0.996$, p-value $< 10^{-16}$. Figure (b) shows the linear regression between $dGC3^\star$ and $dGC3^\star_N + dGC3^\star_C + dGC3^\star_A$. Slope of the linear model is 0.87, $R^2 = 0.997$, p-value $< 10^{-16}$.

Figures on $ENC^{\star}$

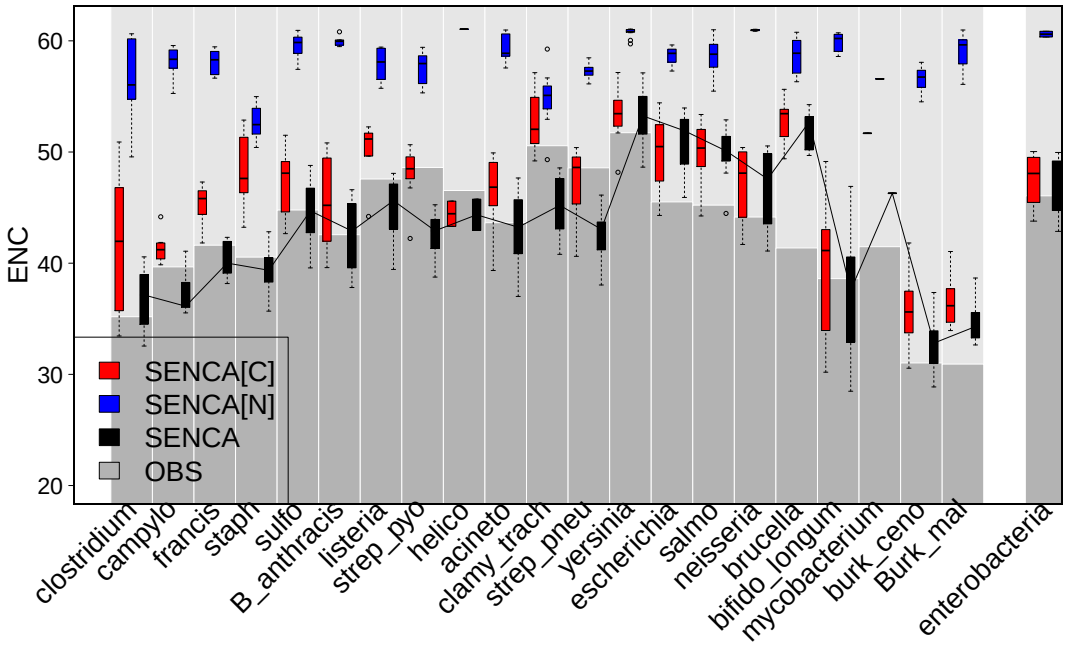


Figure S7: $ENC^{\star}$ of SENCA and of the partial equilibrium layers. SENCA[N] corresponds to $ENC^{\star}_{N}$ i.e. if only the N layer is taken into account (in blue) while SENCA[C] corresponds to $ENC^{\star}_{C}$ i.e. if only the C layer is taken into account (in red).

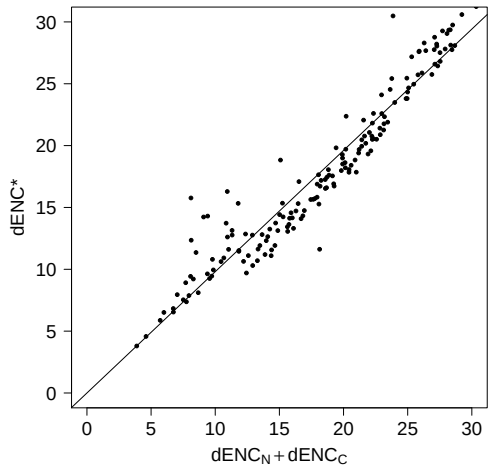


Figure S8: Linear regression between $dENC^{\star}$ and $dENC^{\star}_{N} + dENC^{\star}_{C}$, $R^2 = 0.964$, slope is 0.98.

APPENDIX B

Supplementary Material of: Why does synonymous codon usage vary among different functional categories of humans genes? paper

Supplementary Material

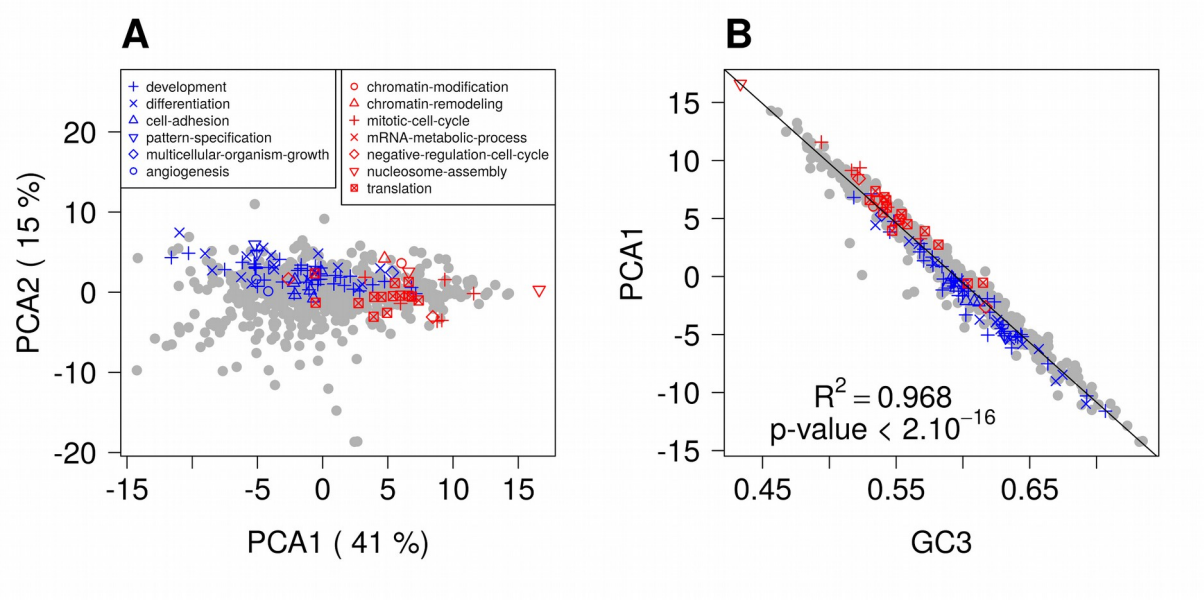


***Figure S1*: Variations in codon usage and in GC3 among functional categories.** *(A) Factorial map of the principal-component analysis of codon usage in GO functional categories in the human genome. Each dot corresponds to a GO gene set, for which the codon usage was computed. GO categories that are associated with "differentiation" or with "proliferation" are displayed respectively in blue and in red. (B) Correlation between the codon usage of GO gene sets (first PCA axis) and their average GC-content at third codon position (GC3).*

30

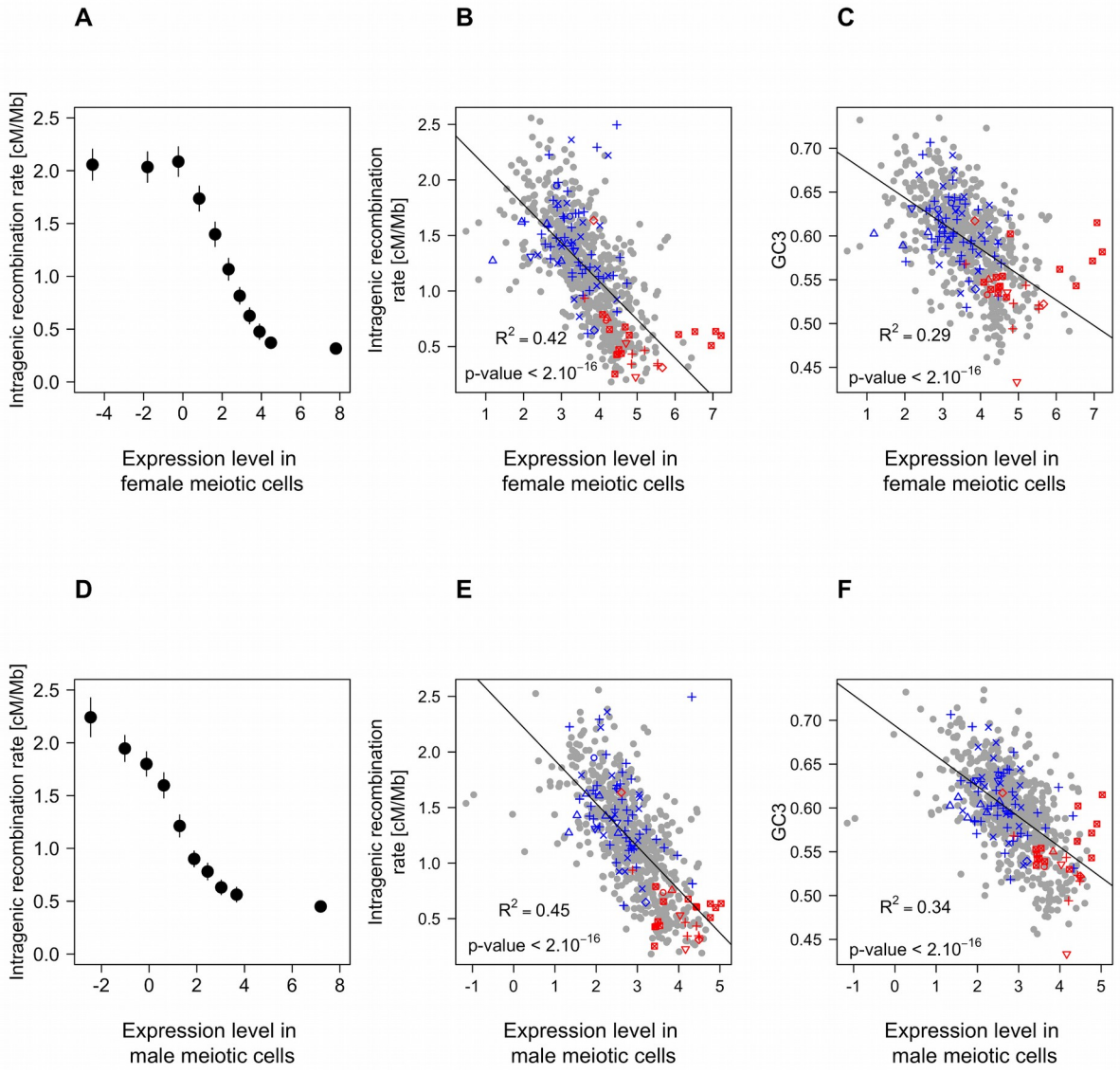
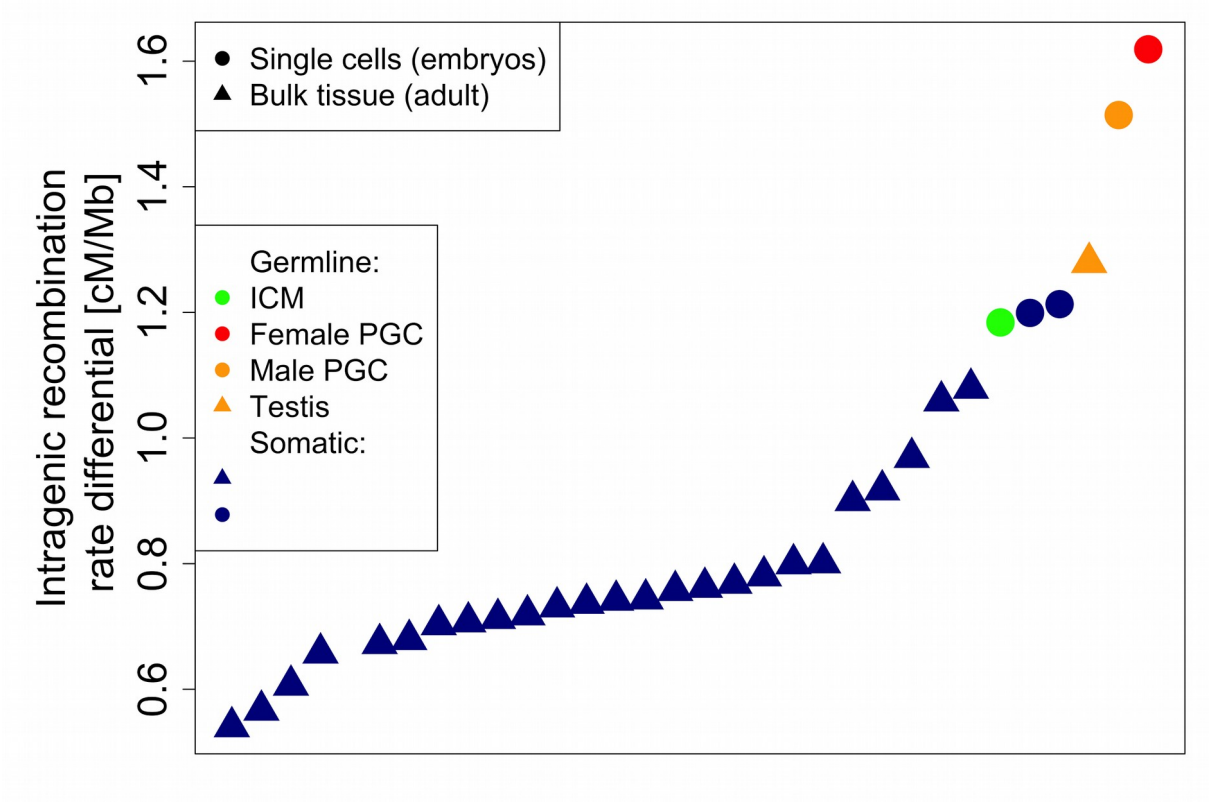
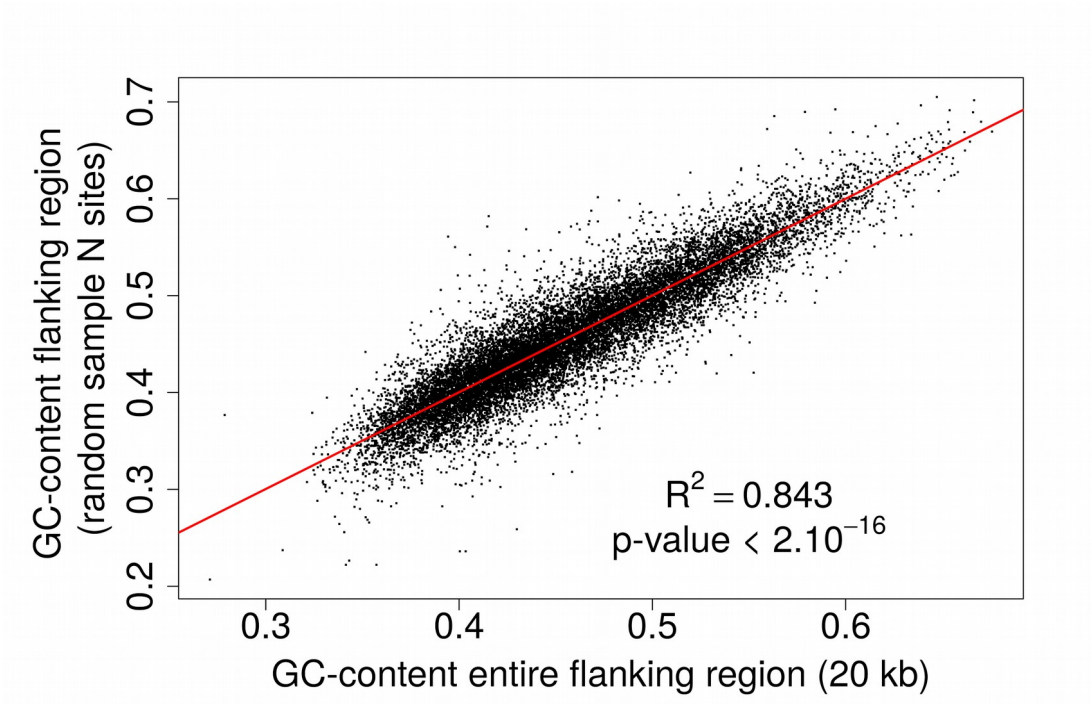


Figure S2: Relationships between expression levels in female or male meiotic cells and GC3 and intragenic recombination rates. *(A, B, C) Same as Figure 3A, 5B and 5C, but with expression level measured by single-cell analysis of female primordial germ cells at 17 weeks (Guo et al, 2015). In panel A, the first bin corresponds to genes whose transcripts were not detected in the RNA-seq experiments (N=1,704; we arbitrarily set their expression level at 0.02 FPKM, as a pseudocount to allow log calculation). The other genes (N=14,266) were grouped into 10 bins of equal sample size. (D, E, F) Same as Figure 3A, 5B and 5C, but with expression level measured in male meiotic cells (Lesch et al, 2016). Expression levels are expressed in log(FPKM).*



650 ***Figure S3**: **Differential intragenic recombination rate between lowly and highly expressed genes in adult tissues and in individual embryonic cells**. This differential is computed as the difference between the mean intragenic recombination rates of lowly expressed genes (10% most lowly expressed for bulk tissue data or non-expressed genes for single cells data) and the mean of the 10% most highly expressed genes. Dots are ordered by increasing differential values. Rounded dots*
655 *correspond to data from individual embryonic cells (Guo et al, 2015) and triangles to adult tissues (Fagerberg et al, 2014). Dark blue dots: somatic adult tissues and somatic embryonic cells are in dark blue. Orange dots: male testis tissue and primordial germ cells (between 4 and 19 weeks). Red dot: female primordial germ cells (between 4 and 17 weeks). Green dot: inner cell mass ICM of the blastocysts.*



660 ***Figure S4**: **Impact of random sampling noise on the variance in GC3**. For each gene, we computed the GC-content in its flanking regions (10 kb upstream and 10kb downstream) and in a subset of N sites (where N is the number of codons in the gene), randomly sampled from the same regions. We repeated this process 100 times. The graph displays one example (among the 100 replicates) of the correlation between the GC-content of the entire region and the GC-content observed in a subsample*
665 *of N sites. The average correlation over the 100 replicates is R^2=0.839.*