Analyse et modélisation des dépendances entre sites voisins dans l'évolution des séquences d'ADN

Leonor Palmeira

N d'ordre 111-2007 Année 2007

Thèse

Présentée

devant l'Université Claude Bernard - Lyon 1

pour l'obtention

du Diplôme de Doctorat

(arrêté du 7 août 2006)

et soutenue publiquement le
13 juillet 2007

par

Maria Leonor Gonon Rodrigues Palmeira

Analyse et modélisation des dépendances entre sites voisins dans l'évolution des séquences d'ADN.

Directeur de thèse : Jean Lobry

Jury :	Laurent Guéguen,	Examinateur
	Jean Lobry,	Directeur
	Didier Piau,	Président
	Eduardo Rocha,	Rapporteur
	Sophie Schbath,	Rapporteuse
	Arndt Von Haeseler,	Examinateur

UNIVERSITÉ CLAUDE BERNARD-LYON 1

Président de l'Université	**M. le Professeur L. COLLET**
Vice-Président du Conseil Scientifique	M. le Professeur J. F. MORNEX
Vice-Président du Conseil d'Administration	M. le Professeur J. LIETO
Vice-Président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur D. SIMON
Secrétaire Général	**M. G. GAY**

SECTEUR SANTÉ

Composantes

UFR de Médecine Lyon R.T.H. Laënnec	Directeur : M. le Professeur D. VITAL-DURAND
UFR de Médecine Lyon Grange-Blanche	Directeur : M. le Professeur X. MARTIN
UFR de Médecine Lyon-Nord	Directeur : M. le Professeur F. MAUGUIERE
UFR de Médecine Lyon-Sud	Directeur : M. le Professeur F.N. GILLY
UFR d'Ontologie	Directeur : M. O. ROBIN
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : M. le Professeur F. LOCHER
Institut Techniques de Réadaptation	Directeur : M. le Professeur MATILLON
Département de Formation et Centre de Recherche en Biologie Humaine	Directeur : M. le Professeur P. FARGE

SECTEUR SCIENCES

Composantes

UFR de Physique	Directeur : M. le Professeur A. HOAREAU
UFR de Biologie	Directeur : M. le Professeur H. PINON
UFR de Mécanique	Directeur : M. le Professeur H. BEN HADID
UFR de Génie Electrique et des Procédés	Directeur : M. le Professeur A. BRIGUET
UFR de Sciences de la Terre	Directeur : M. le Professeur P. HANTZPERGUE
UFR de Mathématique	Directeur : M. le Professeur M. CHAMARIE
UFR d'Informatique	Directeur : M. le Professeur M. EGEA
UFR de Chimie Biochimie	Directeur : Mme. le Professeur H. PARROT
UFR STAPS	Directeur : M. le Professeur R. MASSARELLI
Observatoire de Lyon	Directeur : M. le Professeur R. BACON
Institut des Sciences et des Techniques de l'Ingénieur de Lyon	Directeur : M. le Professeur J. LIETO
IUT A	Directeur : M. le Professeur M. C. COULET
IUT B	Directeur : M. le Professeur R. LAMARTINE
Institut de Science Financière et d'Assurances	Directeur : M. le Professeur J. C. AUGROS

À celles et ceux qui me liront, même en partie,
cette thèse est pour vous.

Remerciements

Je souhaite tout d'abord remercier Jean Lobry et Laurent Guéguen, pour m'avoir guidée tout au long de cette thèse et pour avoir su m'apporter des réponses précises aux questions parfois vagues que je me suis posées. Merci à eux pour cet encadrement en duo, quelquefois sûrement difficile. Je tiens à les remercier pour leur disponibilité et pour leur soutien sans faille, ainsi que pour leurs nombreuses suggestions qui ont su conduire mon travail toujours vers de nouvelles directions.

Je tiens ici à remercier l'ensemble des membres de mon jury, et notamment mes deux rapporteurs Eduardo Rocha et Sophie Schbath, qui ont fait un travail de relecture excellent et très méticuleux qui m'a permis d'améliorer de nombreux points de mon manuscrit. Le temps que m'a accordée Sophie Schbath pour répondre à mes questions sur les statistiques sur les dinucléotides a été pour moi très enrichissant, et je l'en remercie. Le travail de soutien que Didier Piau a fourni tout au long de cette thèse m'a permis de réaliser une très large partie des analyses ici présentées et je l'en remercie. Je remercie finalement Arndt von Haeseler d'avoir accepté de s'aventurer dans un jury francophone malgré les difficultés linguistiques.

Cette thèse est, par ailleurs, le fruit d'un travail effectué au sein des équipes Baobab et BGE. Je tiens à remercier chacun de leurs membres pour leur soutien quotidien. Les Baobab de naissance (Marie-France, Éric, Christian, Christelle, Vincent L., Vincent N., Claire), de naturalisation (Sandrine, Manu, Patricia, Marília, Paulo, Vicente) et d'adoption (Pierre), parmi lesquels j'ai vécu durant la première moitié de ma thèse, ainsi que l'ensemble des BGE (plus particulièrement Anne-Muriel, Anouk, Bastien, Claire, Dominique, Gabriel, Guy, Jef, Laurent, Manolo, Simon, Sophie, Sylvain, Vincent L., Vincent D., Yann), qui m'ont accueillie dans les nouveaux locaux du PRABI au deuxième étage. Je remercie ici Christian –aux multiples casquettes –en tant que directeur du PRABI.

Je tiens fortement à remercier l'ensemble des autres membres du laboratoire, et en particulier Émilie et toute l'équipe Fleury, qui s'est toujours montrée accueillante face à chacune de mes intrusions dans ses locaux.

Je tiens par ailleurs à remercier l'ensemble des services techniques d'avoir mis à ma disposition des ressources qui m'ont été d'une grande aide pour le développement de mes analyses et pour la vie au quotidien dans le laboratoire. Je remercie tout particulièrement le service de secrétariat, ainsi que le service informatique du LBBE, et le centre de calcul de l'IN2P3, sans qui ce travail n'aurait pas été possible.

J'en profite ici pour remercier mes amis GNU, GPL, Debian, LaTeX, Python, R (et j'en oublie) pour avoir su me montrer la voie de la liberté et du partage.

J'espère porter haut les valeurs de cet apprentissage dans chacun des gestes de ma vie quotidienne, car l'aventure ne s'arrête pas là.

Je remercie du fond du coeur ma maman, mon papa, mon petit frère et ma grande soeur, qui ont toujours été à mes côtés, et à qui je dois tout. À Linda, je dois un soutien sans faille et de tous les instants. Je vous remercie tous humblement.

Je tiens à remercier ici mes amis. Cloé, pour avoir enduré l'installation de Linux sur chacun des ordinateurs de l'appartement ; Beatriz, pour tous les souvenirs devant, dedans et autour des salles de concert ; toutes les grenouilles[1]*, et Olivier en particulier, pour avoir su apaiser mes oreilles avec leurs douces comptines ; Maud, Émilie, Vincent et Camille pour tous les bons moments qu'on partage ensemble depuis le DEA ; Rachid, Saïd et Alexis, pour la bande de joyeux drilles qu'on a formé et que je n'oublierai pas ; Clémentine, Jean et Benoît pour tous les moments passés et futurs à refaire le monde ; et bien entendu, toute mon équipe de foot, en particulier Gaëtane, Élodie, Nadège, Sophie, Lydie et Jean-Claude, pour tous les entraînements, matchs et pots.*

J'en oublie, ne m'en voulez pas. Mon cerveau est ainsi qu'il se remémore plus facilement les événements et les rencontres proches que lointaines. Pour toutes celles et ceux que j'ai honteusement oubliés, mais qui vivent en moi à travers qui je suis, merci.

[1] http ://www.frogg-music.com/

Table des matières

Liste des figures

Liste des notations

G+C	contenu relatif en bases G et C par rapport à la longueur de la séquence	5
XpY	deux nucléotides successifs X et Y sur un brin d'ADN : le 'p' représente ici la liaison phosphodiester liant deux nucléotides successifs sur un brin d'ADN	7
X|Y	deux nucléotides complémentaires sur les deux brins d'ADN : le '|' représente ici la lisaion hydrogène liant deux nucléotides complémentaires sur les deux brins d'ADN	7
f_{XY}	la fréquence en dinucléotide XpY	9
f_X	la fréquence en nucléotide X	9
ρ_{XY}	$f_{XY}/(f_X \times f_Y)$	9
ω_{XY}	une mesure associée au dinucléotide XpY,	11
$E(\omega_{XY})$	l'espérance de la mesure ω_{XY} sous le modèle nul choisi	11
$Var(\omega_{XY})$	la variance de la mesure ω_{XY} sous le modèle nul choisi.	11
n	la longueur totale de la séquence,	12
π_X	la fréquence à l'équilibre du nucléotide X sous le modèle markovien,	12
π_{XY}	la fréquence à l'équilibre du dinucléotide XpY sous le modèle markovien,	12
$P(X,Y)$	la probabilité de transition (au sens markovien) de X vers Y,	13
ϵ^1_{XY}	une indicatrice qui vaut 1 ssi X = Y.	13
$\rho_{n^{3-1}_{XY}}$	la statistique ρ pour les dinucléotides chevauchant deux codons,	14
n^{3-1}_{XY}	le nombre de dinucléotides chevauchant deux codons.	14
n_1	le nombre de codons se terminant par la lettre X,	15
n_2	le nombre de codons commençant par la lettre Y,	15
n_3	le nombre de codons commençant par la lettre Y et se terminant par la lettre X	15

n_{cod}	le nombre total de codons dans la séquence.	15
$f_i(t)$	la fréquence du nucléotide i au temps t.	44
α	le taux de transition	47
β	le taux de transversion.	47
θ	le contenu en G+C à l'équilibre.	48
$\mathcal{Q}_{i,j_1}$	l'intervalle $]0, \mathbf{Q}_{i,j_1}]$	73
$\mathcal{Q}_{i,j_2}$	l'intervalle $]\mathbf{Q}_{i,j_1}, \mathbf{Q}_{i,j_1} + \mathbf{Q}_{i,j_2}]$	73
$\mathcal{Q}_{i,j_3}$	l'intervalle $]\mathbf{Q}_{i,j_1} + \mathbf{Q}_{i,j_2}, \mathbf{Q}_{i,j_1} + \mathbf{Q}_{i,j_2} + \mathbf{Q}_{i,j_3}]$	73
$\mathcal{Q}_{i,i}$	l'intervalle restant $]\mathbf{Q}_{i,j_1} + \mathbf{Q}_{i,j_2} + \mathbf{Q}_{i,j_3}, M]$	73
$\mathbf{G}^i$	la matrice des taux de substitution du nucléotide i lorsque la substitution est influencée par le nucléotide à gauche de celui-ci, et qui contient les taux $\mathbf{G}^i_{g,j}$ de substitution du nucléotide i par le nucléotide $j \neq i$ lorsque i possède le nucléotide g à sa gauche.	75
$\mathbf{D}^i$	la matrice des taux de substitution du nucléotide i lorsque la substitution est influencée par le nucléotide à droite de celui-ci, et qui contient les taux $\mathbf{D}^i_{d,j}$ de substitution du nucléotide i par le nucléotide $j \neq i$ lorsque i possède le nucléotide d à sa droite.	75
M	le taux maximal parmi les M_i	76
M_i	le taux de substitution maximal susceptible d'affecter le nucléotide i, pris comme la somme de l'ensemble des taux de substitution possibles (simples, liés au nucléotide à gauche, liés au nucléotide à droite)	76
$\mathcal{Q}_i$	l'intervalle $]0, \sum_{j\neq i} \mathbf{Q}_{i,j}]$	76
$\mathcal{G}_i$	l'intervalle $]\sum_{j\neq i} \mathbf{Q}_{i,j}, \sum_{j\neq i} \mathbf{Q}_{i,j} + \sum_{j\neq i}\sum_g \mathbf{G}^i_{g,j}]$	76
$\mathcal{D}_i$	l'intervalle $]\sum_{j\neq i} \mathbf{Q}_{i,j} + \sum_{j\neq i}\sum_g \mathbf{G}^i_{g,j}, \sum_{j\neq i} \mathbf{Q}_{i,j} + \sum_{j\neq i}\sum_g \mathbf{G}^i_{g,j} + \sum_{j\neq i}\sum_d \mathbf{D}^i_{d,j}]$	76
$\mathcal{Z}_i$	l'intervalle restant $]\sum_{j\neq i} \mathbf{Q}_{i,j} + \sum_{j\neq i}\sum_g \mathbf{G}^i_{g,j} + \sum_{j\neq i}\sum_d \mathbf{D}^i_{d,j}, M]$	76
Δ	défini par $1 - (4\pi_{CG} + 4\pi_{TA} + 8\pi_{CA})$, nul à l'équilibre du modèle	83
$\hat{f}_{XYZ}$	un estimateur de la fréquence du trinucléotide XYZ,	86
Y	les pyrimidines (C et T)	87
R	les purines (A et T).	87
YpR	un dinucléotide formé d'une pyrimidine suivie d'une purine	87
R/Y	la matrice des taux de substitution simple pour lesquelles les deux pyrimidines (respectivement les deux purines) sont traitées de manière identique lors des transversions qu'elles peuvent subir	87
r_w	le taux de substitution de CpG|CpG vers TpG|CpA,	88
r_s	le taux de substitution de TpA|TpA vers CpA|TpG.	88

α^*	$\alpha_w + \alpha_s$	88
β^*	$\beta_w + \beta_s$	88
σ^*	$\alpha^* + \beta^*$	88
r	le taux de substitution r_w de CpG|CpG vers TpG|CpA.	90
$\overline{XY}$	le dimère de pyrimidine de type cyclobutane formé entre entre une pyrimidine X adjacente à une pyrimidine Y	126

Introduction

Les travaux de George Beadle et Edward Tatum sur les mécanismes génétiques contrôlant le métabolisme de la moisissure du pain *Neurospora* ont permis d'établir l'hypothèse 'un gène/une protéine' (Beadle & Tatum, 1941), qui entraînerait par la suite le développement – tant en biologie moléculaire, qu'en génétique ou plus tard en bioinformatique – de l'étude des gènes et des protéines comme porteurs de l'information génétique. Les travaux ultérieurs de Frederick Sanger et Hans Tuppy, sur la structure des protéines, au cours desquels ils furent les premiers à obtenir la séquence protéique de l'insuline, montrèrent que les protéines sont des molécules ordonnées constituées d'une séquence d'acides aminés (Sanger & Tuppy, 1951). On en déduisit alors que les gènes qui fabriquent ces protéines devaient très probablement eux aussi avoir une structure ordonnée permettant de porter l'information de la séquence protéique. Résoudre le problème du séquençage de l'ADN devint ainsi une extension naturelle du travail de Frederick Sanger sur le séquençage des protéines et le conduisit à des techniques de séquençage applicables à l'ADN (Sanger *et al.*, 1977). C'est grâce à ce travail qu'il put produire, en 1977, la première séquence complète d'un organisme, celle du phage Φ-X174.

Depuis ces travaux pionniers, les techniques de séquençage sont devenues de plus en plus rapides et fiables, et leur automatisation a permis le séquençage complet de très nombreux organismes. Les données disponibles publiquement dans les banques de données de séquences croissent en effet de manière extrêmement rapide. De manière générale, celles-ci ont permis le développement de l'**évolution moléculaire**, en tant que discipline visant à décrire les mécanismes régissant l'évolution du support de l'information génétique, notamment par l'étude des différences entre séquences homologues.

Il semble clair que, bien que les gènes, protéines et génomes séquencés soient stockés sous forme de séquences de lettres[2], les séquences biologiques représentent des molécules, qui possèdent une structure spatiale complexe dont la fonction est

[2]La séquence de lettres d'une protéine représente la suite d'acides aminés qui la composent ; la séquence de lettres d'un gène, ou d'un génome, représente la suite d'acides nucléiques qui le composent.

dirigée par les interactions que celles-ci établissent entre elles, et avec d'autres molécules du milieu. Les interactions nécessaires au bon fonctionnement de ces molécules se situent à différents niveaux, et l'on peut citer l'exemple clef des interactions nécessaires entre positions d'une séquence d'ARN pour le repliement spatial de la molécule. Il existe, par ailleurs, des dépendances entre positions d'une séquence à une échelle plus petite, comme par exemple l'influence des bases avoisinant une cytosine, sur la probabilité de substitution de celle-ci : l'exemple des substitutions sur les dinucléotides CpG en est un des exemples les plus marquants.

Ces interactions biologiques sont connues et largement documentées et étudiées, toutefois la modélisation de l'évolution de séquences homologues – que ce soit dans les méthodes d'alignement de séquences (Smith & Waterman, 1981), dans les mesures de distance entre deux séquences (Tajima & Nei, 1984), dans les méthodes d'inférence phylogénétiques par parcimonie (Jin & Nei, 1990), distance (Li, 1981; Lake, 1994) ou maximum de vraisemblance (Felsenstein, 1981) – s'est faite traditionnellement en posant comme hypothèse principale que les sites évoluent indépendamment les uns des autres. Cette hypothèse possède de nombreux avantages mathématiques, mais malheureusement aucune justification biologique, et les exemples de violation de cette hypothèse sont nombreux.

Mon travail de thèse se situe dans cette problématique. Il cherche dans un premier temps à déterminer comment estimer le poids des dépendances entre sites voisins dans les séquences biologiques, et à établir dans quelle mesure il est important de prendre en compte ces dépendances dans les modèles d'évolution de séquences. Dans un deuxième temps, il cherche à définir des modèles d'évolution de séquences prenant en compte des dépendances entre sites voisins et à développer des outils d'analyse permettant d'étudier les séquences biologiques grâce à ces nouveaux modèles.

Je présenterai dans le **premier chapitre** quelques mesures utilisées pour la détermination de la composition en bases des génomes. Je m'attarderai ensuite sur différentes statistiques pour la mesure de sur- et sous-représentation en dinucléotides dans les séquences biologiques. Certaines de ces statistiques, basées sur des modèles de permutation ont été implémentées et incorporées au paquet SeqinR du logiciel de statistiques R.

Dans le **deuxième chapitre**, j'exposerai d'abord la notion de biais d'usage du code génétique et quelques hypothèses qui ont pu être avancées pour l'expliquer. Ensuite, je m'attarderai sur l'étude de l'effet de l'environnement sur la composition en dinucléotides des génomes à travers l'exemple de l'effet des rayons ultra-violets. Je montrerai alors qu'avec la multiplication des données de génomes complets, une large étude de la question est possible et j'amènerai une réponse claire à la controverse longtemps non résolue de l'effet des rayons ultra-violets sur la composition en bases des génomes.

Dans le **troisième chapitre**, je présenterai le cadre classique de la modéli-

sation de l'évolution de séquences biologiques, qui fait l'hypothèse que les sites évoluent indépendamment les uns des autres. Puis je présenterai l'ensemble des hypothèses sous-jacentes à ces modèles et à leurs applications. Je présenterai quelques possibilités, proposées dans la littérature, pour se défaire de ces hypothèses. Pour finir, je me pencherai sur l'hypothèse d'indépendance entre sites, qui nous intéresse tout particulièrement. Je montrerai que le maintien de cette hypothèse a une conséquence néfaste sur la qualité des méthodes d'inférence phylogénétique.

J'introduirai dans le **quatrième chapitre** un modèle général d'évolution de séquences prenant en compte des dépendances entre sites voisins. Je proposerai ensuite plusieurs approches pour l'étude de ce modèle : étude par simulations, étude par approximation, étude analytique exacte. L'approche exacte nous permettra de développer des méthodes d'estimation des taux de substitution du modèle. J'appliquerai alors ces résultats à l'étude du chromosome 21 d'*Homo sapiens* et discuterai des avancées produites par le développement analytique de ce modèle.

L'ensemble de ce manuscrit est disponible en ligne à l'adresse suivante :
`http://biomserv.univ-lyon1.fr/~palmeira/repro/these`
Les graphiques présentés ici y peuvent être reproduits en ligne, et les données issues de ce travail y sont disponibles.

Chapitre 1

Composition en bases des génomes : mesures

Le terme de **composition en bases** des génomes désignera ici, dans un sens large, les mesures ayant trait autant à la composition en nucléotides, qu'à la composition en dinucléotides ou en mots des génomes. Ce terme désignera la composition en bases du **génome nucléaire – ou chromosomique** d'un organisme, par opposition aux génomes de ses éventuels organites ou plasmides.

Je présenterai ici d'abord quelques mesures et critères utilisés pour l'analyse de la composition en bases des génomes, puis je me pencherai sur la question de l'analyse statistique des biais en dinucléotides et présenterai le détail de plusieurs statistiques dans le cadre de cette analyse.

1.1 Vue générale sur les génomes

1.1.1 Le contenu en G+C

Le contenu en G+C est probablement une des **premières mesures historiques** du contenu en bases des génomes (Sueoka, 1962). Le contenu en G+C est en effet accessible très simplement à travers la mesure de la température de fusion de la double-hélice d'ADN, puisque les bases G et C liées par trois liaisons hydrogènes nécessitent une température de fusion plus importante par rapport aux bases A et T, liées par deux liaisons hydrogènes. À l'heure du séquençage en masse des génomes de modèles biologiques connus, le contenu en G+C est resté l'une des mesures les plus fréquentes pour rendre compte de la composition en bases d'un organisme, car il reste très facilement accessible par simple calcul sur une séquence :

$$G + C = \frac{n_G + n_C}{\sum_X n_X}$$

où n_X représente le nombre de nucléotides X dans la séquence, et $X \in \{A, C, G, T\}$.

La constatation de la variabilité de ce contenu entre organismes a amené à se pencher sur le lien entre celui-ci et différentes variables. Certaines de ces variables permettraient de mettre en évidence des pressions de sélection responsables de la mise en place de tel ou tel contenu en G+C. Je présenterai certains de ces résultats dans le chapitre suivant.

1.1.2 Les corrélations à longue portée

Le **caractère fractal d'une séquence**, déterminé par des **corrélations à portée plus ou moins longue**, s'observe à différentes échelles sur les séquences d'ADN. Ainsi, on observe une périodicité de période 3-bases dans l'ensemble des organismes vivants. Cette périodicité est la signature des codons, présents dans les régions codant pour des protéines. De la même manière, une périodicité toutes les 10-11 bases est aussi observée dans l'ensemble des organismes vivants et est généralement interprétée comme la signature du repliement de l'hélice d'ADN. Des corrélations à plus longue portée, de l'ordre de 200-400 bases, identifiées uniquement dans les séquences d'organismes eucaryotes, semblent être la signature du repliement de l'ADN autour d'un nucléosome ou d'un dinucléosome (Audit *et al.*, 2004). Cet ensemble d'échelles semble constituer le repliement hiérarchique de la fibre chromatinienne.

Certains travaux cherchent à modéliser ces corrélations à longue portée de manière à déterminer les mécanismes capables de les générer et de les maintenir. Messer *et al.* (2005) montrent notamment qu'il est possible de générer et de maintenir des corrélations à longue portée par de simples mécanismes de duplication, substitution et insertion-déletion. Toutefois, ces modèles imposent que la taille de la séquence soit croissante pour que les corrélations à longue portée puissent être maintenues car ce sont les processus de duplication qui sont principalement responsables du maintien de cette composante. Si la taille de la séquence est constante, les processus de substitution et d'insertion aléatoire tendent à faire disparaître les corrélations. Cette contrainte sur l'augmentation constante de la taille des séquences, qui semble peu réaliste d'un point de vue biologique, laisse supposer que d'autres mécanismes sont probablement à l'œuvre dans le maintien de cette organisation.

1.1.3 Les variations régionales : représentations graphiques

Historiquement, la variabilité du contenu en bases a d'abord été mise en évidence entre organismes (Sueoka, 1962; Muto & Osawa, 1987). Des données plus étendues et plus précises ont permis ensuite d'analyser la variabilité intra-génomique des organismes.

a. Les fenêtres glissantes

La manière la plus intuitive d'illuster les variabilités régionales du contenu en bases des génomes est d'utiliser une **fenêtre de mesure** d'une taille donnée, puis de faire glisser cette fenêtre le long du génome pour obtenir des mesures moyennes le long de celui-ci.

Cette représentation permet, par exemple, de visualiser des zones plus ou moins homogènes, comme la présence d'isochores dans les génomes de vertébrés (Bernardi *et al.*, 1985). L'utilisation des méthodes basées sur des fenêtres glissantes a ainsi permis de mettre en évidence des relations fortes entre un grand nombre de variables. Notamment, une étude récente de Versteeg *et al.* (2003) sur des fenêtres glissantes non chevauchantes de 20kb le long du génome humain a permis de mettre en évidence une relation entre contenu en G+C, contenu en gènes, longueur des introns, patrons d'expression, isochores, ilôts CpG. La figure 1.1 montre bien comment une méthode de fenêtres glissantes permet une représentation visuelle de certaines caractéristiques importantes de l'organisation des génomes.

b. Les promenades sur l'ADN

L'idée des fenêtres glissantes a probablement donné naissance aux méthodes dites des 'promenades sur l'ADN' (DNA walk). Il s'agit d'effectuer des **mesures cumulatives**, où l'on mesure de manière additive une variable donnée le long de l'ADN (Peng *et al.*, 1992), et qui peut se traduire par une représentation graphique simple.

Les promenades sur l'ADN mesurent la variation locale d'une certaine mesure de composition en base. Généralement, il s'agit de mesurer le contenu relatif en C par rapport à G, et en A par rapport à T, ce qui a, par exemple, permis de mettre en évidence la localisation de l'origine de réplication de certains chromosomes bactériens (voir figure 1.2) (Lobry, 1996). L'ensemble de ces représentations graphiques ont entraîné, depuis, le développement de méthodes de segmentation permettant d'effectuer des découpages de grandes séquences d'ADN en régions à contenu plus ou moins homogène pour une mesure donnée.

1.2 Analyse statistique des biais en dinucléotides dans les génomes

J'appelle **dinucléotide** deux nucléotides successifs sur un brin d'ADN. Je les noterai par la suite XpY : le 'p' représente ici la liaison phosphodiester liant deux nucléotides successifs sur un brin d'ADN, et permettra de les distinguer de deux nucléotides appariés entre les deux brins d'ADN au niveau d'une position. Les **paires de nucléotides** appariés face à face sur les deux brins d'ADN seront

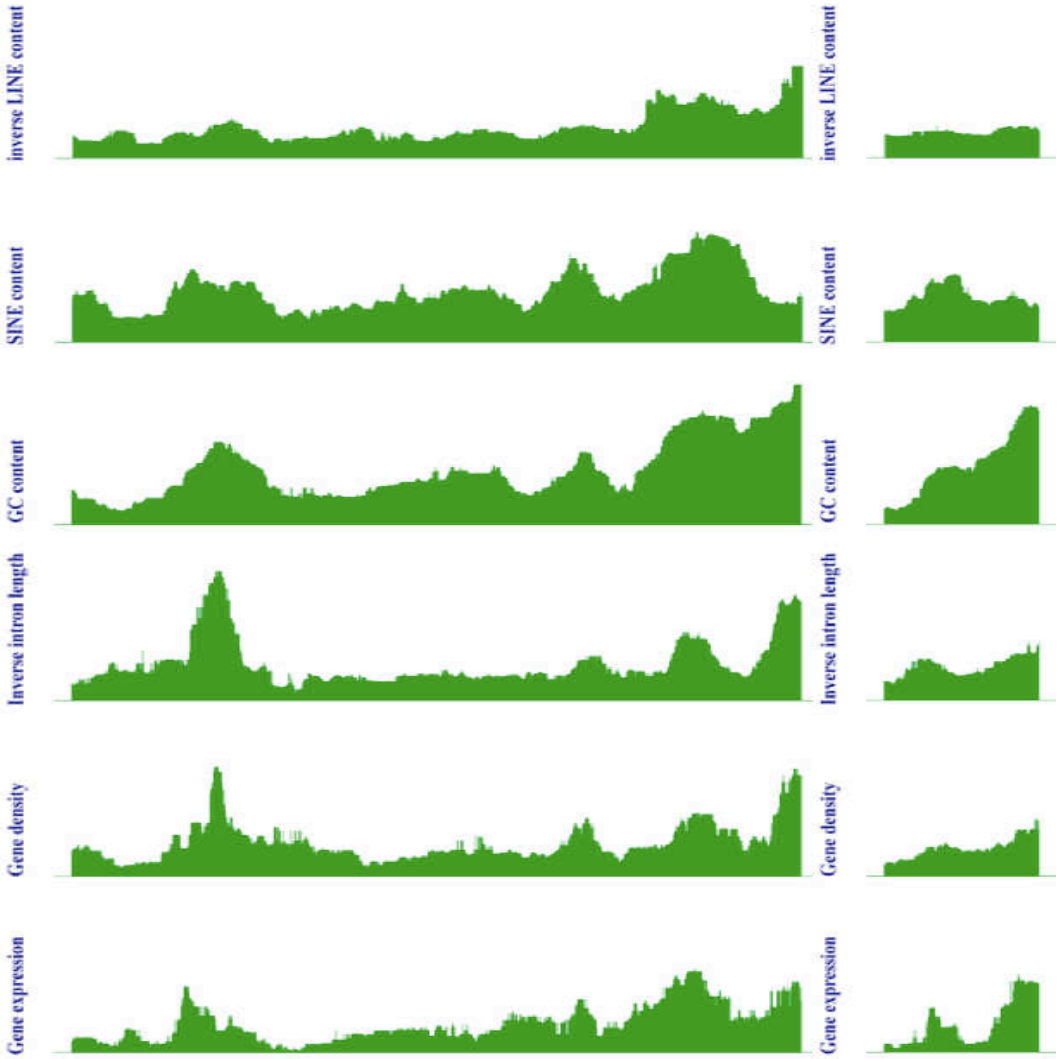


Figure 1.1 – Profils d'expression de gènes, de densité en gènes, de l'inverse de la longueur des introns, du contenu en G+C, de la densité en SINEs et de l'inverse de la densité en LINEs sur le chromosome 9 (à gauche) et sur le chromosome 21 (à droite) d'*Homo sapiens*, mesurés sur une fenêtre glissante de 49 gènes. Figure issue de Versteeg *et al.* (2003)

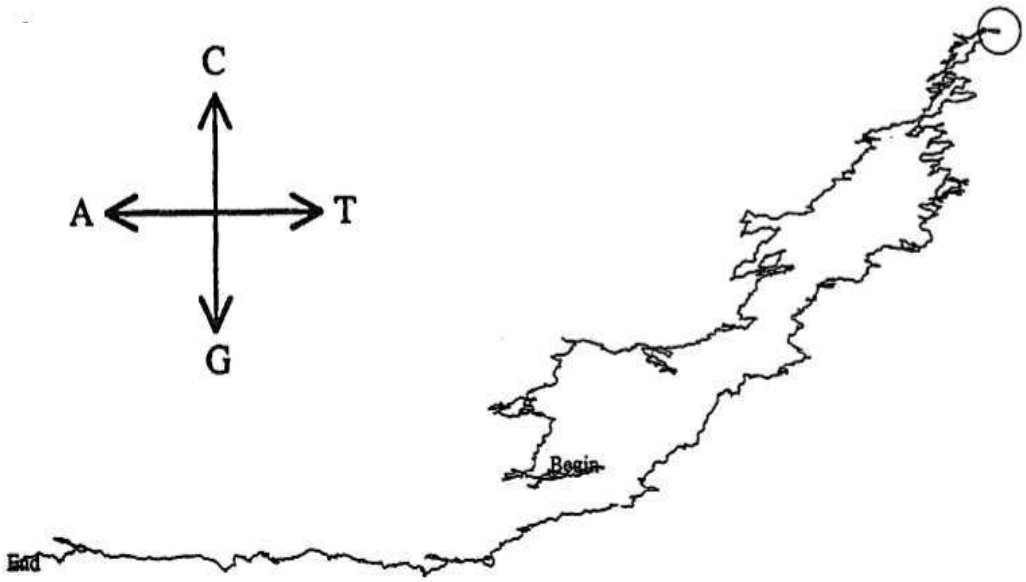


Figure 1.2 – Promenade sur l'ADN du génome de la bactérie firmicute *Mycoplasma genitalium*. Figure issue de Lobry (1996).

notés X|Y. Ceci s'avère important pour les dinucléotides ambigus tels que CpG, où G est le complémentaire de C, et où il est important de distinguer entre le dinucléotide CpG et l'appariemment C|G.

1.2.1 La statistique *rho*

La statistique ρ, issue des tests classiques en statistiques non paramétriques, est présentée par Burge *et al.* (1992) et reprise dans de nombreux travaux de Karlin (Karlin & Cardon, 1994) pour mesurer la sur- et sous-représentation en mots de deux lettres. Elle s'écrit comme suit :

$$\rho_{XY} = \frac{f_{XY}}{f_X \times f_Y}$$

où f_{XY} représente la fréquence en dinucléotide XpY et f_X représente la fréquence en nucléotide X dans la séquence d'intérêt. Le modèle sous-jacent suppose que le dinucléotide est formé aléatoirement par l'association des deux nucléotides qui le compose ($\rho_{XY} = 1$ sous l'hypothèse nulle). Un écart à cette valeur caractérise une sur- ou sous-représentation du dinucléotide XpY (respectivement $\rho_{XY} > 1$ ou $\rho_{XY} < 1$).

On s'attend à ce que la statistique ρ, mesurée sur une séquence générée aléatoirement ne montre aucune sur- ou sous-représentation. La loi de ρ n'est pas connue *a priori*, mais si on calcule cette statistique sur 1000 séquences générées avec une loi uniforme dans l'alphabet $\{A, C, G, T\}$, on peut ajuster une loi nor-

male centrée en 1 à la distribution des ρ calculés sur les 1000 séquences générées (voir Fig. 1.3).

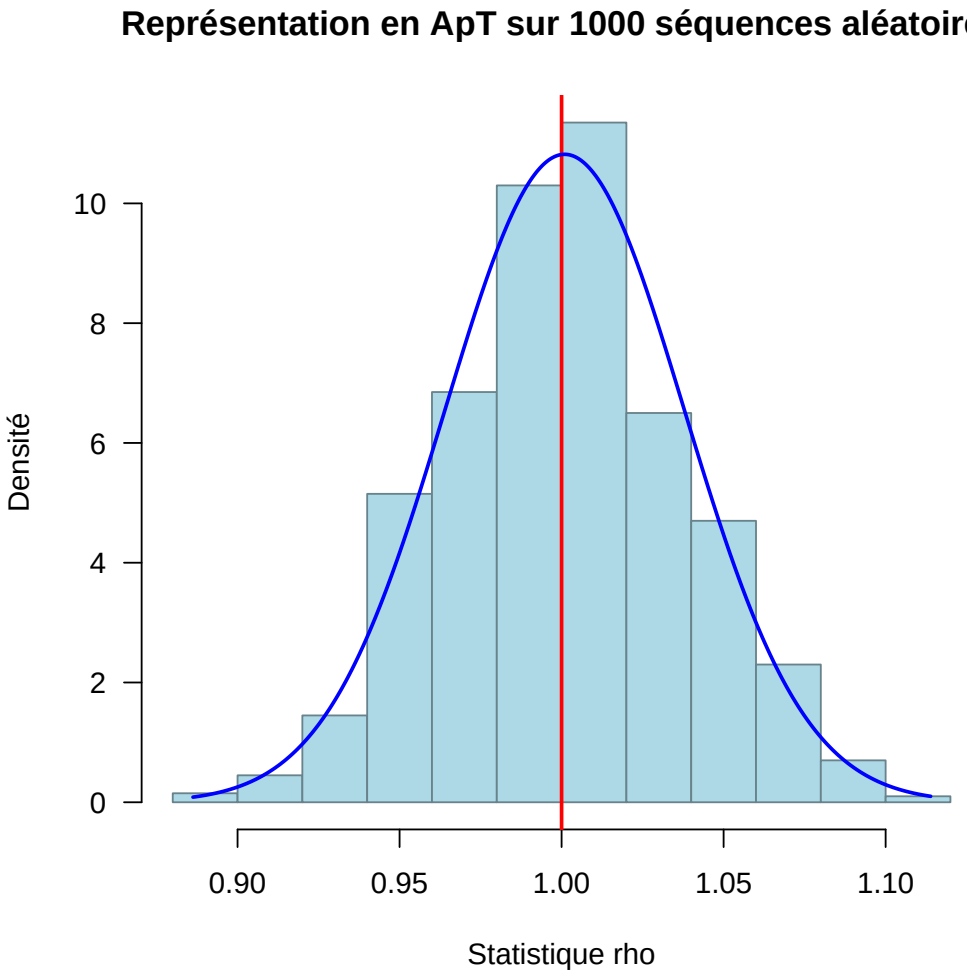


Figure 1.3 – Distribution de la statistique ρ calculée sur 1000 séquences de 6000 bases, générées avec une loi uniforme dans l'alphabet $\{A, C, G, T\}$. La droite verticale rouge est centrée en 1. La courbe en bleue ajuste la loi normale d'espérance égale à la moyenne des ρ, et de variance égale à la variance des ρ.

Il s'agit d'une des mesures les plus utilisées par les biologistes pour rendre compte de la sur- et sous-représentation d'un dinucléotide dans un génome donné (Duret & Galtier, 2000; Ponger *et al.*, 2001; Arndt *et al.*, 2003; The Honeybee Genome Sequencing Consortium, 2006; Oakes *et al.*, 2007). Elle est utilisée en particulier pour estimer la sur- et sous-représentation en dinucléotides CpG dans le cadre de la détection de la méthylation des CpG, et est aussi généralement appelée CpGo/e, puisqu'elle compare une fréquence observée ('o' pour 'obser-

ved') et une fréquence attendue sous l'hypothèse d'association aléatoire des deux nucléotides ('e' pour expected).

L'un des problèmes de cette statistique, est que le modèle nul est un modèle fixé extrêmement simple, qui est invalidé pour un très grand nombre de séquences biologiques, notamment pour les séquences codantes. Dans ce cas, le rejet du modèle nul utilisé dans la statistique ρ devient trivial, et les sur- et sous-représentations mesurées ne sont que le fruit des contraintes fortes s'exerçant sur ces séquences. Le problème est donc que ρ peut être différent de 1 sans que cela n'apporte une information claire sur les mécanismes évolutifs responsables de cette mesure.

1.2.2 La statistique z-score

La statistique z-score est bien plus souple que la statistique ρ car elle est définie de manière à ce que différents modèles puissent être utilisés pour la détermination de sur- et sous-représentations. Cette statistique permet une mesure plus fine de la sur- et sous-représentation dans les séquences selon le modèle nul choisi.

Le z-score est défini comme suit :

$$z_{score} = \frac{\omega_{XY} - E(\omega_{XY})}{\sqrt{Var(\omega_{XY})}}$$

où ω_{XY} représente une mesure associée au dinucléotide XpY, $E(\omega_{XY})$ représente l'espérance de la mesure ω_{XY} sous le modèle nul choisi et $Var(\omega_{XY})$ représente la variance de la mesure ω_{XY} sous le modèle nul choisi.

Lorsque le théorème central de la limite[1] peut être appliqué, cette statistique suit asymptotiquement une loi normale centrée réduite. L'application de différents modèles nuls permet alors de mesurer différentes composantes dans la sur- et sous-représentation d'un dinucléotide. En effet, lorsqu'on veut connaître la sur- ou sous-représentation d'un dinucléotide sur une séquence, on veut savoir si cette mesure s'écarte d'une mesure attendue. Le modèle nul représente cette mesure attendue, et on peut inclure dans le modèle des caractéristiques connues des séquences, de manière à déterminer si un dinucléotide est plus ou moins fortement représenté dans une séquence, étant données les caractéristiques que l'on connait déjà sur cette séquence.

On peut considérer qu'il existe deux grandes classes de modèles pour l'étude des sur- et sous-représentations de mots dans les séquences : les modèles markoviens et les modèles de permutation. Je ne présenterai ici que le cas où la séquence d'étude peut être considérée homogène.

[1] en traduction de 'der zentrale Grenzwertsatz' issu de la dénomination donnée par Pólya (1920)

a. Modèles markoviens

Les chaînes de Markov, dont la connaissance théorique et technique est très développée, sont une modélisation probabiliste extrêmement utilisée dans le cadre de la modélisation de séquences biologiques. Cette modélisation considère que la séquence d'intérêt a été générée par une **chaîne de Markov où la suite de lettres correspond à la séquence de la chaîne**. L'idée générale issue de la théorie markovienne étant que chaque position dans la séquence dépend du "passé" de la séquence à une échelle donnée fixe. Dans le cadre très simple d'une chaîne de Markov d'ordre 1, le "passé" correspondra à la base située directement en amont de la position à laquelle on s'intéresse. Je détaillerai dans le chapitre 3 un autre de ses grands champs d'application dans le domaine de l'analyse de séquences.

Dans le cadre de cette modélisation, de très nombreux développements statistiques ont été réalisés pour la mesure de sur- ou sous-représentation de mots donnés (Robin *et al.*, 2003), pour la détection de mots sur- et sous-représentés inconnus *a priori* (Schbath, 1997), pour la mesure de la répartition de mots le long d'une séquence (Robin *et al.*, 2003) et, plus récemment, pour la comparaison de la significativité statistique de la sur- ou sous-représentation d'un même mot entre deux séquences (Robin *et al.*, 2007).

- *Modèle de Markov d'ordre 1*

Lorsqu'on définit ω_{XY} comme le comptage n_{XY} du nombre de dinucléotides XpY présents dans la séquence observée, la loi exacte de ce comptage sous un modèle markovien d'ordre 1 s'obtient sous la forme d'une relation de récurrence. On notera que cette relation n'est pas restreinte aux mots de longueur deux, et est valable quelle que soit la longueur du mot. Sous le modèle markovien d'ordre 1, le calcul du z-score se fait donc à partir des écritures suivantes de la loi exacte du comptage en dinucléotides, issues du cadre général sur les mots de longueur k décrit par Schbath (1997) et Robin *et al.* (2003) :

$$E(n_{XY}) = (n-1)\pi_{XY}$$

$$\begin{aligned} Var(n_{XY}) &= (n-1)\pi_{XY}(1-\pi_{XY}) \\ &+2\pi_{XY}(n-2)(\epsilon^1_{XY}P(x,y)-\pi_{XY}) \\ &+2(\pi_{XY})^2\sum_{t=1}^{n-3}(n-t-2)\frac{1}{\pi_X}(P^t(Y,X)-1) \end{aligned}$$

où n représente la longueur totale de la séquence, π_X représente la fréquence à l'équilibre du nucléotide X sous le modèle markovien, π_{XY} représente la fréquence à l'équilibre du dinucléotide XpY sous le modèle markovien, $P(X,Y)$ représente

la probabilité de transition (au sens markovien) de X vers Y, et ϵ^1_{XY} représente une indicatrice qui vaut 1 ssi X = Y. L'estimation de la probabilité $P(X,Y)$ peut se faire par maximum de vraisemblance : $\widehat{P(X,Y)} = \frac{n_{XY}}{\sum_{Y_i} n_{XY_i}}$, où $Y_i \in \{A,C,G,T\}$.

Lorsque l'ordre du modèle augmente, la loi exacte du comptage reste toujours calculable, mais son développement en devient très fastidieux, et on lui préfère généralement le calcul d'une loi approchée que je ne détaillerai pas ici. Plus de détails, ainsi que des formules plus générales, peuvent être obtenus dans Robin *et al.* (2003).

b. Modèles de permutation

Les modèles de permutation sont des modèles où une séquence attendue selon l'un de ces modèles est le résultat d'une **permutation – sous contraintes – de la séquence étudiée**. Étant donnée la plus faible connaissance théorique de ces modèles due à leur complexité combinatoire, leur utilisation dans le cadre de la statistique sur les mots dans les séquences est beaucoup plus restreinte. Toutefois, en comparaison aux modèles markoviens, ils offrent la possibilité de fixer le nombre d'occurrences d'un mot de manière exacte, et non de manière approchée (car asymptotique), ce qui est particulièrement intéressant lorsqu'on veut traiter des séquences de faible taille. En outre, ils permettent de gérer des contraintes qui sont difficiles, voire impossibles, à modéliser dans un contexte markovien. Par ailleurs, dans le cadre de la mesure de sur- et sous-représentations de dinucléotides, le développement de ces modèles reste encore simple, et de nombreux travaux nous permettent une écriture exacte de la statistique z-score. C'est pour ces raisons que, dans les analyses relatives à cette thèse, je les ai préférés aux modèles markoviens.

- *Modèle de permutation de bases*

Le cas le plus simple des modèles de permutation est le cas des **permutations de bases**. Une séquence générée sous ce modèle est une séquence possédant le même contenu en bases que la séquence d'intérêt. Elle peut être obtenue par tirage sans remise dans une urne contenant l'ensemble des bases de la séquence étudiée.

Dans le cadre de la statistique z-score sur les dinucléotides, on montre facilement que l'écriture avec les comptages n_{XY} du nombre de dinucléotides est équivalente à l'écriture suivante, qui est une simple réduction et normalisation de la statistique ρ_{XY} présentée précédemment :

$$z_{score} = \frac{\rho_{XY} - E(\rho_{XY})}{\sqrt{Var(\rho_{XY})}}$$

où $E(\rho_{XY})$ et $Var(\rho_{XY})$ sont l'espérance et la variance de ρ_{XY} sous le modèle nul. Ces deux mesures peuvent être calculées suivant l'approximation suivante pour des grandes séquences :

$$\widehat{E(\rho_{XY})} = 1$$
$$\widehat{Var(\rho_{XY})} = \frac{(1-f_X)(1-f_Y)}{nf_Xf_Y}$$

où n est le nombre total de nucléotides dans la séquence, et f_X la fréquence du nucléotide X dans la séquence.

Dans la suite, je noterai cette statistique z-base.

Comme nous l'avons précisé, il s'agit ici d'une écriture approchée, l'écriture exacte étant disponible dans Schbath (1995). L'utilisation de cette approximation, dans les applications que nous présentons dans le chapitre suivant, ne modifie pas les conclusions que nous obtenons par rapport à l'utilisation de l'écriture exacte. L'annexe A présente ainsi la relation entre l'écriture exacte et l'écriture approchée tout en montrant que l'utilisation de cette approximation ne modifie pas les conclusions de cette thèse.

- *Modèle de permutation de codons*

Dans le cas d'une séquence codante (CDS), pour laquelle on connaît certaines contraintes, tel que le biais d'usage du code, on peut vouloir utiliser un autre type de modèle de permutations. Le modèle de **permutations des codons** permet en effet de mesurer les sur- et sous-représentations dans une séquence étant donné le biais d'usage du code. Une séquence générée sous ce modèle est une séquence possédant le même contenu en codons que la séquence d'intérêt.

On peut alors montrer que le calcul du z-score précédent sous ce modèle peut être réduit au calcul du z-score sur les dinucléotides qui chevauchent deux codons :

$$z_{score} = \frac{\rho_{n_{XY}^{3-1}} - E(\rho_{n_{XY}^{3-1}})}{\sqrt{Var(\rho_{n_{XY}^{3-1}})}} = \frac{n_{XY}^{3-1} - E(n_{XY}^{3-1})}{\sqrt{Var(n_{XY}^{3-1})}}$$

où $\rho_{n_{XY}^{3-1}}$ représente la statistique ρ pour les dinucléotides chevauchant deux codons, et où n_{XY}^{3-1} représente le nombre de dinucléotides chevauchant deux codons.

L'espérance et la variance présentées ici peuvent être calculées selon l'écriture proposée par Gautier *et al.* (1985) :

$$E(n_{XY}^{3-1}) = \frac{n_1n_2 - n_3}{n_{cod}}$$
$$Var(n_{XY}^{3-1}) = E(n_{XY}^{3-1}) - (E(n_{XY}^{3-1}))^2$$

$$+\frac{1}{n_{cod}(n_{cod}-1)}[(2n_3(n_1+n_2-n_1n_2-1)$$

$$+n_1n_2(n_1-1)(n_2-1))]$$

où n_1 représente le nombre de codons se terminant par la lettre X, n_2 représente le nombre de codons commençant par la lettre Y, n_3 représente le nombre de codons commençant par la lettre Y et se terminant par la lettre X et n_{cod} représente le nombre total de codons dans la séquence.

Dans la suite, je noterai cette statistique z-codon.

c. Comparaison de modèles de permutation

J'ai implémenté la statistique ρ ainsi que les statistiques du z-score avec un modèle nul de permutation des bases (z-base) et avec un modèle nul de permutation des codons (z-codon) dans la bibliothèque de fonctions SeqinR [2] du logiciel de statistiques R. Ces fonctions sont venues enrichir cette bibliothèque d'analyse de séquences biologiques de quelques outils de statistiques non-paramétriques pour l'estimation des sur- et sous-représentations en dinucléotides. Plus de détails sont disponibles sur la vignette de la bibliothèque (Charif *et al.*, 2007) ou dans Charif & Lobry (2006). Cette bibliothèque de fonctions m'a permis par la suite de répondre à la question longtemps controversée concernant la relation entre exposition aux rayons ultra-violets et composition en bases chez les bactéries (Singer & Ames, 1970; Bak *et al.*, 1972), qui sera développée en détail dans le chapitre suivant.

Avant de présenter les résultats obtenus grâce à l'utilisation de ces statistiques, je vous propose une brève analyse des trois statistiques suivantes :

- ρ
- z-base
- z-codon

sur un jeu de données biologiques simple : l'ensemble des 480 séquences géniques (CDS) annotées du petit génome de la bactérie firmicute *Mycoplasma genitalium* (numéro d'accession Genbank : L43967). Chacune de ces trois statistiques est calculée de manière indépendante sur chacun des CDS annotés, et j'effectue une analyse globale des résultats obtenus.

Les deux premières statistiques sont extrêmement proches et devraient nous procurer la même information sur les séquences étudiées. La troisième statistique par contre, en tenant compte des codons, nous permet de tenir compte du biais d'usage du code, et d'obtenir une information concernant la sur- et sous-représentation de mots dans les séquences étant donné ce biais.

[2] http ://pbil.univ-lyon1.fr/software/SeqinR/seqinr_home.php

- *Deux statistiques quasi identiques ?*

L'un des problèmes liés à la statistique ρ réside dans la définition d'un seuil de significativité. En effet, cette statistique suit une loi normale dépendante de la séquence, car la variance de cette loi est fonction de la longueur de la séquence, et des fréquences en nucléotides (voir page 14). Le seuil de significativité n'est donc pas constant, et dépend de chaque séquence et de chaque dinucléotide. En pratique, il faut donc connaître la variance de ρ pour le dinucléotide étudié et pour la séquence étudiée, pour pouvoir conclure sur la significativité de la valeur observée, ce qui peut être obtenu par simulations. Dans l'usage, et en particulier dans l'utilisation de la statistique CpGo/e[3], cette démarche n'est quasiment jamais suivie, et de nombreuses études considèrent empiriquement que des valeurs inférieures à 0.78 ou supérieures à 1.23 peuvent être considérées des valeurs extrêmes (Karlin & Cardon, 1994; Duret & Galtier, 2000; Ponger *et al.*, 2001; Arndt *et al.*, 2003; The Honeybee Genome Sequencing Consortium, 2006; Oakes *et al.*, 2007).

Parallèlement à cela, la statistique z-base est une simple ré-écriture de la statistique ρ de manière à obtenir une statistique centrée-réduite sous le modèle nul. Cette ré-écriture permet donc de définir un seuil de significativité général indépendant de la séquence et du dinucléotide étudiés et permet ainsi une comparaison aisée des valeurs de la statistique entre dinucléotides dans une même séquence et entre séquences différentes.

J'ai donc effectué le calcul de ces deux statistiques sur l'ensemble des CDS de *Mycoplasma genitalium*. Ainsi, on voit bien sur la figure 1.4 que, en prenant un seuil identique de significativité sur l'ensemble de la distribution des ρ calculés sur des CDS différents, et sur l'ensemble des dinucléotides, certains dinucléotides seront considérés sur- ou sous-représentés, alors qu'ils ne le sont pas lorsque l'on regarde le z-base. Prenons l'exemple des dinucléotides TpT et TpG : lorsqu'on ne regarde que la statistique ρ, on voit que ces deux dinucléotides semblent avoir la même faible sur-représentation, alors que le z-base montre clairement que TpT est plus sur-représenté que TpG.

- *Faut-il gérer le biais d'usage du code ?*

La différence essentielle entre le z-base et le z-codon réside dans le fait que la deuxième statistique prend en compte le biais d'usage du code dans le modèle nul, et permet d'effacer cette composante, et de mesurer la sur- et sous-représentation présente dans une séquence sous ce biais d'usage du code.

Toutefois, la corrélation mesurée entre z-base et z-codon (voir figure 1.5), bien que parfois faible, montre que certains mécanismes sont indépendants du biais d'usage du code. Lorsque la corrélation est forte, le biais d'usage du code ne fait qu'exprimer les mêmes sur- et sous-représentations que celles à l'œuvre sur

[3]équivalente à la statistique ρ sur le dinucléotide CpG

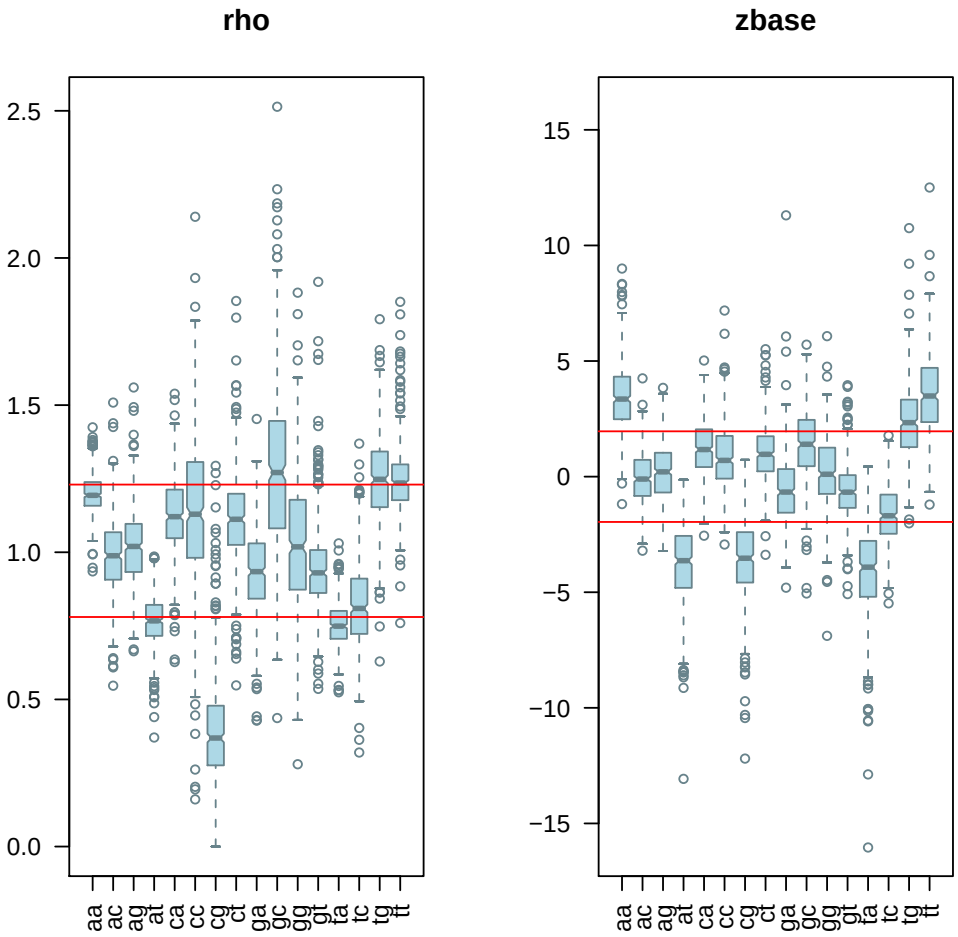


Figure 1.4 – La distribution des statistiques ρ et z-base, calculées sur tous les CDS de *Mycoplasma genitalium*. Les limites de significativité ont été tracées en rouge : pour le ρ, les limites sont placées empiriquement à 0.78 et à 1.23, pour le z-base, elles ont été placées à 5% de significativité.

la séquence lorsqu'on efface ce biais. Lorsque la corrélation est faible, par contre, le biais d'usage du code est plus important pour ce dinucléotide.

On note, par ailleurs, que la pente de la droite de régression est toujours inférieure à 1, ce qui indique que le calcul du z-codon diminue globalement la significativité de sur- ou sous-représentation du dinucléotide par rapport au z-base. Ceci est tout à fait naturel puisque le modèle nul du z-codon incorpore une contrainte supplémentaire, qui n'était pas incorporée dans le modèle du z-base.

Sur la majeure partie des CDS, la tendance à la sur- ou sous-représentation est donc maintenue dans les deux mesures. Pourtant quelques séquences, situées dans le deuxième ou quatrième cadran (voir par exemple, le cas du dinucléotide GpT dans la figure 1.5), montrent une tendance inversée entre la significativité dans le z-base et dans le z-codon, ce qui suggère que le biais d'usage du code est particulièrement fort sur ces séquences. L'utilisation du z-codon semble donc particulièrement justifiée pour l'étude de ces séquences.

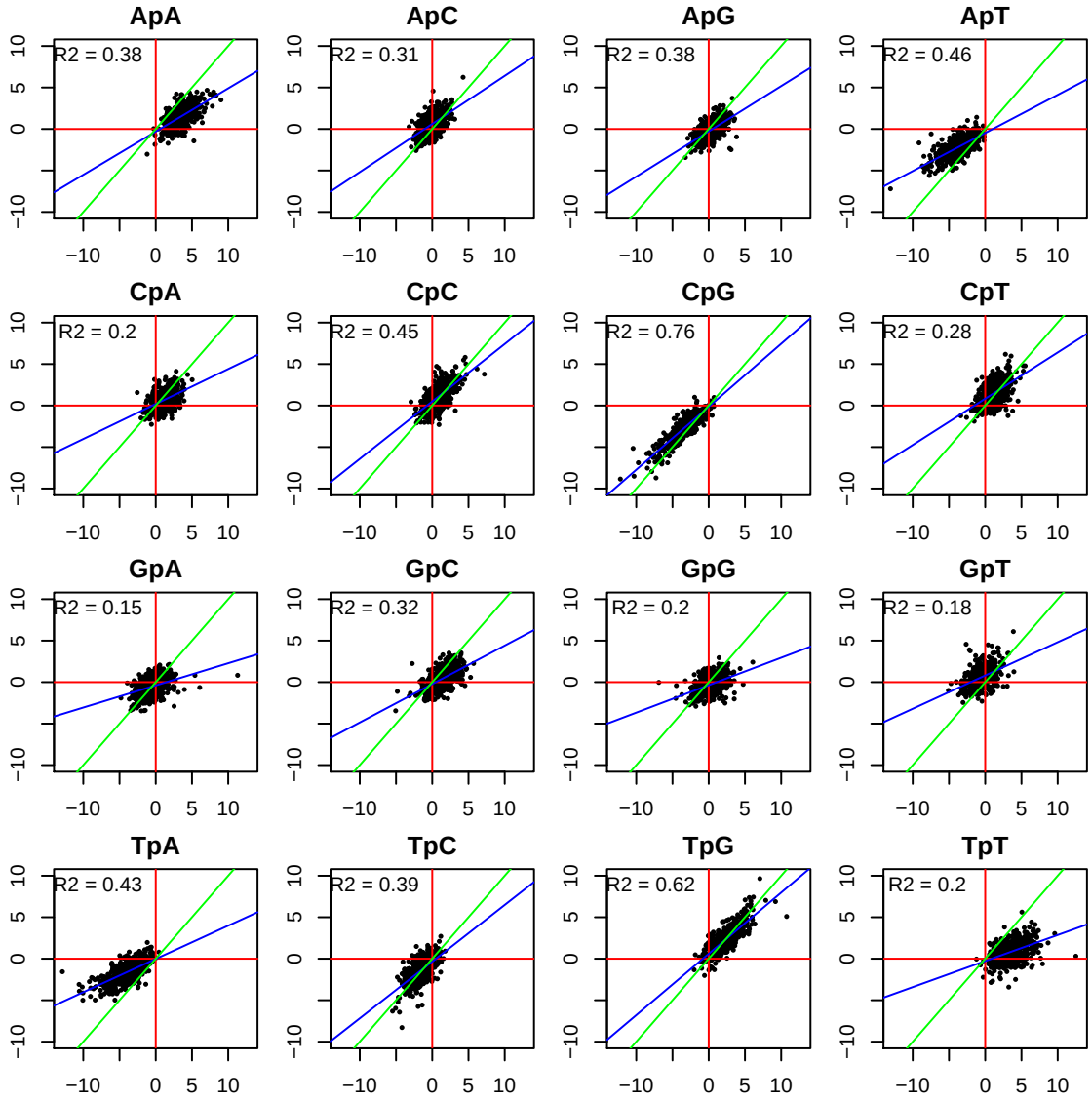


Figure 1.5 – Relation entre les statistiques z-base (axe x) et z-codon (axe y), calculées pour chacun des dinucléotides sur tous les CDS de *Mycoplasma genitalium*. La droite de corrélation est tracée en bleu, et la valeur du carré du coefficient de corrélation est donnée pour chacun des dinucléotides. La droite verte correspond à la première diagonale. Les droites en rouge indiquent la valeur nulle.

Chapitre 2

Composition en bases des génomes : mécanismes biologiques

Différents modèles peuvent être avancés pour expliquer la composition en bases observée dans les génomes. Ces modèles sont généralement présentés dos à dos, selon qu'ils soutiennent une vision plutôt **neutraliste** ou plutôt **sélectionniste** des processus évolutifs, c'est-à-dire selon qu'ils proposent des hypothèses liées à des processus de biais mutationnels ou à des processus sélectifs.

Je tâcherai ici de décrire certaines de ces hypothèses, et les résultats obtenus dans ce domaine, ce qui fera l'objet de la première section de ce chapitre. Je m'attarderai, dans la deuxième section de ce chapitre, sur le possible effet des rayons ultraviolets sur la composition en bases des micro-organismes, et montrerai que cet effet peut être considéré comme quasiment nul, contrairement à ce qui a longtemps été avancé. Pour finir, je présenterai l'effet de la méthylation des génomes sur la composition en bases de ceux-ci, en guise de préambule aux deux chapitres suivants.

2.1 Biais d'usage du code

Les séquences protéiques peuvent être résumées par leur séquence en acides aminés. Il existe 20 acides aminés naturels, et chaque séquence protéique est portée par l'ADN, où il n'existe que 4 bases. L'information protéique est donc codée dans l'ADN par une succession de triplets de bases : les **codons**. Le code génétique a à sa disposition 64 codons pour coder 21 caractères différents : les 20 acides aminés et la position de fin de séquence (STOP). Certains caractères peuvent donc être représentés par plusieurs codons : **le code génétique est redondant**. La figure 2.1 est une représentation circulaire du code génétique universel. Chaque couleur correspond à une des quatre bases ; le cercle interne

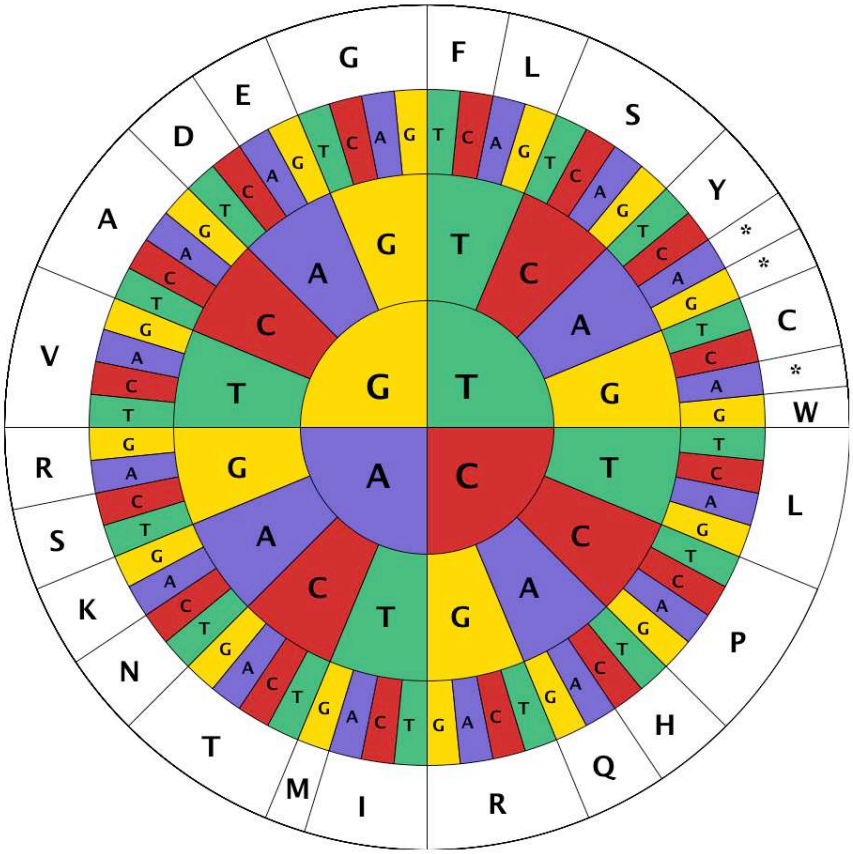


Figure 2.1 – Le code génétique standard – représentation circulaire générée par ISee (In Silico biology E-learning Environment, logiciel développé à l'INRIA Rhône-Alpes).

correspond à la première base de chaque codon, le cercle suivant à la deuxième base, le troisième à la troisième base ; le cercle externe correspond au caractère codé par chaque codon.

Cette représentation permet de voir qu'un même acide aminé peut être codé par plusieurs codons. Or, pour coder un acide aminé donné, chaque codon n'est pas utilisé à la même fréquence par les organismes. Certains codons sont préférés et utilisés à une fréquence bien plus forte que d'autres, et cette préférence peut varier non seulement selon les espèces, mais aussi selon les gènes d'un même organisme (Gouy & Gautier, 1982; Ikemura, 1981, 1982). Le **biais d'usage du code** désigne le fait que certains codons sont utilisés préférentiellement dans le codage des acides aminés.

Les premières mesures du biais d'usage du code ont été effectuées dans les années 1980 (Grantham *et al.*, 1980; Ikemura, 1985). Depuis, de nombreuses études se sont attachées à effectuer une analyse poussée des biais d'usage du code entre organismes et au sein des organismes. Un large volume de données est notamment disponible à travers la **Codon Usage Database**[1], initialement mise en place et réalisée par Ikemura (Nakamura *et al.*, 2000), qui compile les tables de biais d'usage du code pour l'ensemble des organismes présents dans GenBank.

L'ensemble de ces travaux est allé de pair avec une analyse des mécanismes capables de générer un biais d'usage du code. Les hypothèses proposées se rangeant généralement soit du côté d'une vision sélectionniste, soit d'une vision neutraliste des processus évolutifs. Voici un bref aperçu des hypothèses les plus marquantes.

2.1.1 Fiabilité et rapidité de la traduction

Depuis les travaux pionniers d'Ikemura (Ikemura, 1981, 1982, 1985), de nombreuses études ont mis en évidence une corrélation entre codons préférés, et la présence des ARN de transfert correspondants dans la cellule. Ceci autant chez des procaryotes que chez des eucaryotes. Cette corrélation semble d'autant plus forte au niveau des gènes fortement exprimés, qui semblent présenter un **biais d'usage du code optimisé pour une traduction rapide, et fiable** (Gouy & Gautier, 1982; Sharp & Li, 1986; Sharp & Matassi, 1994).

Chez les organismes unicellulaires, et les métazoaires invertébrés, on peut interpréter les biais d'usage du code par le niveau d'expression des gènes : les gènes fortement exprimés utilisent les codons optimaux qui correspondent aux ARNt les plus abondants. Il semble donc établi que le biais d'usage du code possède une relation forte avec le processus de traduction : l'utilisation de codons optimaux servant à accélerer la traduction et à la rendre plus fiable (Ren *et al.*, 2007), alors que l'utilisation de suites de codons rares semblerait liée à un ralentissement de la traduction pour un meilleur repliement de la protéine (Guisez *et al.*, 1993). Toutefois, chez les vertébrés, cette hypothèse reste beaucoup plus contestée. Bien

[1] http ://www.kazusa.or.jp/codon/

que certaines études aient avancé l'existence de biais d'usage du code différentiel selon le tissu d'expression des gènes (Plotkin *et al.*, 2004), des études récentes montrent que cette variabilité semble en réalité liée à la variabilité du contenu en G+C de la région portant ces gènes (Sémon *et al.*, 2006).

2.1.2 Adaptation à l'environnement

Un autre aspect de la relation génome-environnement sera développé en détail dans la section suivante, mais comme des hypothèses suggérant une influence de l'environnement sur la composition en bases des génomes sont très couramment avancées, voici un des exemples les plus largement étudiées dans le cadre du biais d'usage du code : le lien avec la température. En effet, une hypothèse parfois avancée est celle que le biais d'usage du code peut être en partie expliqué par des pressions de sélection liées à l'environnement. Les paires G|C, liées par trois liaisons hydrogène sont plus stables face à l'augmentation en température que les paires A|T, liées par deux liaisons hydrogène. Cette constatation amène à poser l'hypothèse suivante : le génome d'un organisme exposé à de très fortes températures a-t-il un génome plutôt riche en G+C pour résister plus longtemps à la dénaturation de ses molécules d'ADN ? Et en particulier, dans le cas du biais d'usage du code, y a-t-il un usage préférentiel des codons riches en G et en C ? Bien que cette hypothèse ait lontemps été controversée (Lobry & Chessel, 2003), des résultats récents montrent que le biais de G+C entre bactéries mésophiles et bactéries thermophiles ne semble pas lié à la température de croissance (Lobry & Necşulea, 2006).

L'une des seules relations entre génome et environnement qui ait pour l'instant été démontrée est le lien, indirect, entre le contenu en G+C et la présence d'oxygène dans l'habitat, par la démonstration d'une relation forte entre la mesure du contenu en G+C et la capacité métabolique des micro-organismes à vivre en présence d'oxygène. Malheureusement, aucun mécanisme clair n'a encore pu être avancé pour expliquer cette relation (Naya *et al.*, 2002).

2.1.3 Biais mutationnel

Il semblerait que le biais d'usage du code soit en réalité façonné par de nombreuses forces, dont le **biais mutationnel** serait une des forces majeures. On définit par biais mutationnel, le fait que la probabilité de muter vers G ou C, soit différente de celle de muter vers A ou T. Ce biais a des conséquences sur la composition en base des génomes de manière générale, et sur le biais d'usage du code en particulier. On considère de plus que ce biais mutationnel est étroitement lié à la machinerie de recombinaison et qu'il n'agit donc pas de manière uniforme le long des génomes. Une étude récente comparant introns et CDS (Chen *et al.*, 2004) montre qu'il s'agirait possiblement du mécanisme principal responsable de la génération d'un biais d'usage du code.

D'autres études montrent, par ailleurs, que l'orientation des gènes sur le brin codant ou non codant est le premier facteur discriminant de l'analyse du biais d'usage du code chez *Borrelia burgdorferi*, et que l'asymétrie des processus de réplication est une des sources majeures de biais dans l'usage du code (McInerney, 1998; Pimentel Cachapuz Rocha, 2000).

2.2 Pressions de l'environnement : exemple des UVs

L'étude de l'influence de l'environnement sur la composition globale en bases des génomes a fait l'objet de très nombreuses études, souvent controversées. L'ensemble des études portant, par exemple, sur la relation entre température et composition en bases des génomes en est un échantillon très représentatif (Galtier & Lobry, 1997; Vinogradov, 2003; Wang *et al.*, 2006a). Je ne m'attarderai pas ici sur l'ensemble des résultats découlant de ce corpus de travaux, mais sur une seule de ces controverses : celle de l'influence des rayons ultraviolets (UVs) sur la composition en bases des génomes, et tâcherai d'apporter l'ensemble des éléments permettant d'éclaircir cette question.

2.2.1 Les lésions dues aux UVs sur l'ADN

Il est depuis longtemps établi que l'ADN est la cible des rayons ultraviolets (UV), et que ceux-ci entraînent la formation de dommages spécifiques. On considère généralement que les UVc (100-290 nm) sont quasi entièrement absorbés par la couche d'ozone, et que leur effet sur l'ADN en milieu naturel est minimal. Les UVb (290-320 nm), eux, sont considérés particulièrement dangereux, car leur énergie peut être directement absorbée par l'ADN en induisant des modifications sur les bases pyrimidiques adjacentes. Ils sont en effet responsables de la formation de **dimères de pyrimidine de type cyclobutane** par la photoexcitation de pyrimidines adjacentes (Setlow, 1966) (voir figure 2.2). Ceux-ci sont les produits majeurs de l'attaque de l'ADN par les UVs ($\sim$ 75%), bien que les UVbs entraînent aussi en plus faible quantité la formation de **photoproduits pyrimidine (6-4) pyrimidone** ($\sim$ 25%), qui se convertissent parfois en leur **isomère de valence Dewar** (voir figure 2.3). La présence d'un de ces photoproduits, s'il n'est pas réparé, entraîne une distortion locale de la molécule d'ADN qui sera responsable du blocage des machineries de transcription et de réplication (Setlow, 1966; Singer & Ames, 1970; Sinha & Häder, 2002; Besaratinia *et al.*, 2005). De récentes études montrent que, contrairement à ce qui a été longtemps admis, les UVa (320-400 nm) peuvent eux aussi être responsables de la formation de dimères de pyrimidine de type cyclobutane (Mouret *et al.*, 2006).

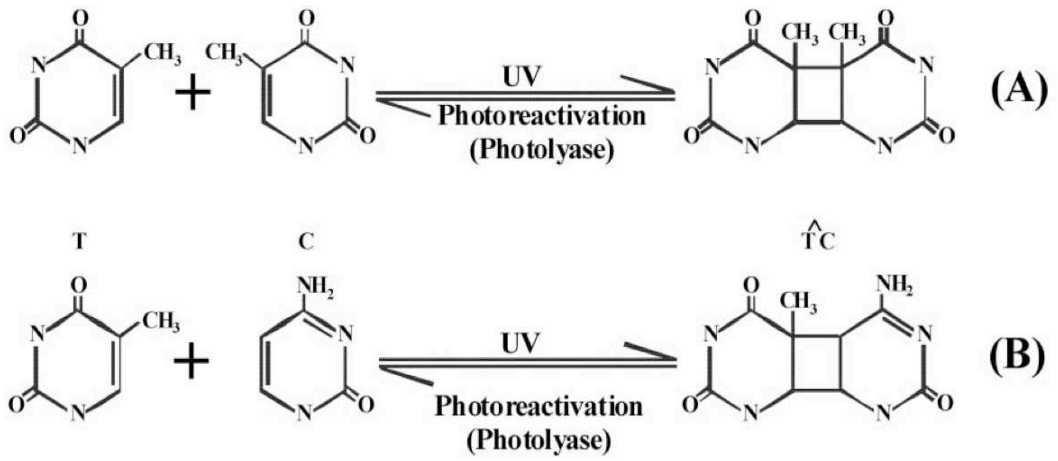


Figure 2.2 – Réaction de formation de dimères de pyrimidine de type cyclobutane entre (A) deux thymines adjacentes sur le même brin d'ADN et (B) une cytosine et une thymine adjacentes sur le même brin d'ADN par l'action des UVs. La réaction de photoréactivation, faisant intervenir une photoligase est indiquée dans le sens inverse (voir figure 2.4 pour plus de détail sur la réaction de photoréactivation). Figure tirée de Sinha & Häder (2002).

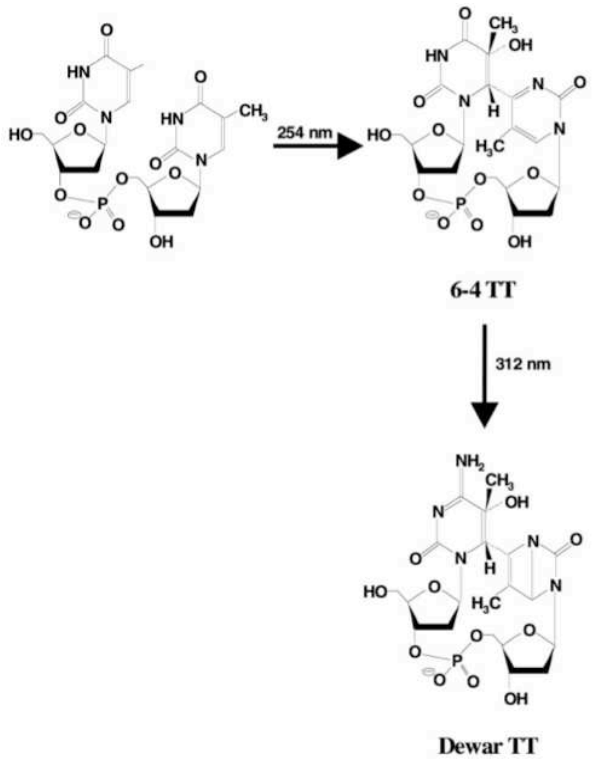


Figure 2.3 – Réaction de formation d'un photoproduit pyrimidine (6-4) pyrimidone et de son isomère de valence Dewar entre deux thymines adjacentes sur le même brin d'ADN par l'action des UVs. Figure tirée de Sinha & Häder (2002).

2.2.2 Les mécanismes de réparations des lésions dues aux UVs

Il existe deux mécanismes principaux de réparation des dommages causés par les UVs sur l'ADN (Kellogg & Paul, 2002) : la **réparation *dite* lumineuse ou photoréactivation** et la **réparation *dite* sombre ou réparation par excision de nucléotides**. Le premier mécanisme a été mis en évidence dans les trois royaumes du vivant, et est un processus spécifique qui nécessite la présence de lumière, puisqu'il met en œuvre une photoligase qui reconnait et répare spécifiquement les dimères de pyrimidine. La réparation par les photoligases se fait par le transfert d'un électron à travers le cofacteur FAD après excitation par la lumière (voir figure 2.4). Le deuxième mécanisme est un mécanisme non spécifique ne nécessitant pas la présence de lumière, et qui est impliqué dans les réparations de très nombreux types de lésions. Il s'agit d'un mécanisme beaucoup plus complexe, faisant intervenir de l'ATP, et un ensemble de plusieurs gènes (voir figure 2.5 pour une représentation schématique).

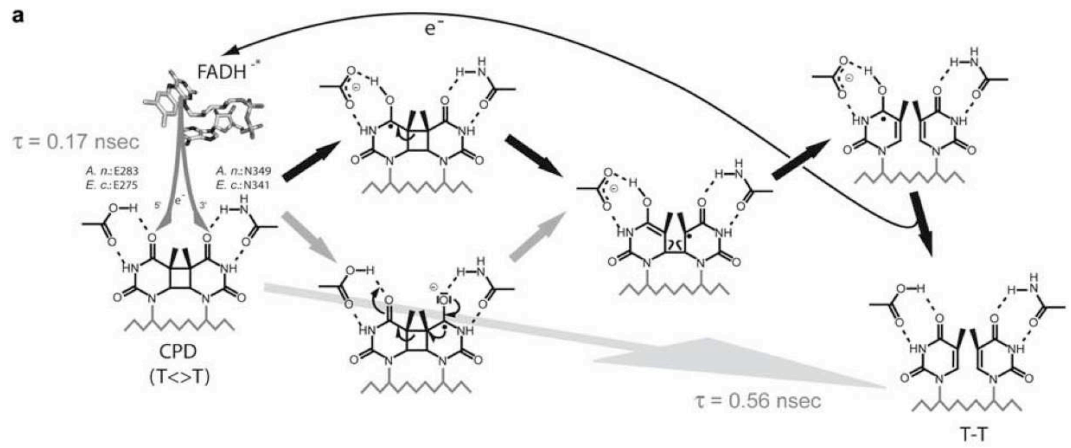


Figure 2.4 – Réaction de réparation d'un dimère de pyrimidine de type cyclobutane entre deux thymines adjacentes, par le système de réparation des photoligases faisant intervenir le cofacteur FAD, telle qu'elle est proposée par Essen & Klar (2006).

2.2.3 Y a-t-il un effet des UVs sur la composition des génomes ?

On note donc que les lésions provoquées par les UVs sur l'ADN ont lieu de manière prépondérante sur les bases pyrimidiques adjacentes. C'est pourquoi il est depuis longtemps considéré que ce mécanisme entraîne une **forte pression de sélection sur la composition en bases des génomes**, et en particulier sur leur contenu en G+C (Singer & Ames, 1970; Kellogg & Paul, 2002). En effet,

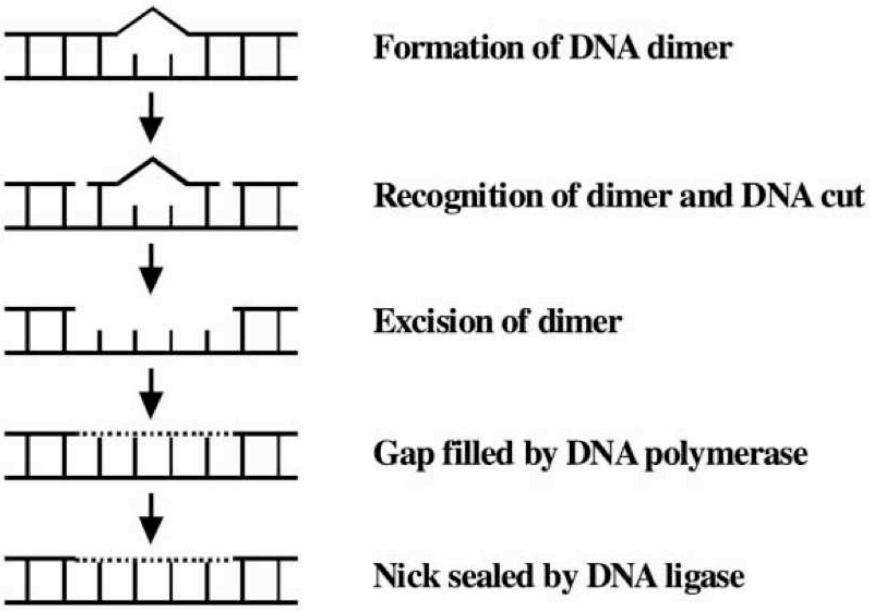


Figure 2.5 – Schéma du système de réparation par excision de nucléotides impliquant les cinq étapes suivantes : (1) formation d'un dimère par action des UVs, (2) reconnaissance du dimère et coupure dans l'ADN, (3) excision du dimère, (4) remplissage de l'intervalle par une polymérase de l'ADN, (5) fermeture de l'ouverture par une ligase de l'ADN. Figure tirée de Sinha & Häder (2002)

en considérant que les thymines sont les bases les plus fréquemment visées par les UVs – ceci sera discuté plus loin – il a été proposé qu'un fort contenu en G+C, en diminuant la fréquence des thymines, entraînerait une diminution de la fréquence en dinucléotides de thymine-pyrimidine – cibles des UVs, et que les forts contenus en G+C observés chez certains micro-organismes seraient une conséquence directe de la pression de sélection pour l'évitement de dommages causés par les UVs.

Singer & Ames (1970) montraient notamment que **le contenu en G+C des micro-organismes est corrélé positivement avec l'exposition aux UVs dans leur habitat naturel**. Toutefois je vais vous présenter plusieurs points criticables qui ont été soulevés à l'époque (Bak *et al.*, 1972) et plus récemment (Palmeira *et al.*, 2006) et qui remettent en cause la validité de ces résultats. Dans l'étude de Singer & Ames (1970), les micro-organismes étudiés sont regroupés par genre, si bien que (1) **le contenu en G+C est déterminé par genre** et la valeur pour chaque genre est la moyenne du contenu en G+C des espèces étudiées dans ce genre, et que (2) **l'exposition en habitat naturel est déterminée par genre**, alors que de nombreux genres possèdent des espèces colonisant des milieux parfois très différents. Par ailleurs, (3) **la mesure du contenu en G+C est une mesure indirecte qui s'avère mauvaise pour estimer la pression de sélection des UVs** sur la composition en bases des génomes (voir annexe B pour plus de détails).

Bien que l'étude historique de Singer & Ames (1970) aient été à l'époque contestée (Bak *et al.*, 1972), et bien qu'elle soit criticable sur de nombreux points, et notamment à la lumière d'une des études sur lesquelles elle se base (Setlow, 1966), aucune autre étude n'a été entreprise depuis, et de nombreux articles considèrent cette hypothèse comme testée et validée (Kellogg & Paul, 2002; Agogué *et al.*, 2005). Des études similaires ont notamment été étendues, par exemple chez les phages marins (Kellogg & Paul, 2002), où une corrélation positive entre contenu en G+C et coefficient d'inactivation par les UVs est mise en évidence. Pourtant l'existence de données sur des génomes complets, bactériens et viraux, permettrait de répondre à cette question de manière non équivoque en étudiant le contenu en dinucléotides de pyrimidine (puisque ce sont les cibles directes des UVs) et non plus le contenu en G+C.

2.2.4 Analyse de l'effet des UVs sur la composition des génomes

J'ai effectué trois études de manière à déterminer s'il existe des pressions de sélection dues aux UVs sur la composition en bases des génomes.

La première étude a consisté en une **analyse systématique de la fréquence en dinucléotides de pyrimidine sur l'ensemble des génomes bactériens entièrement séquencés**. Les prokaryotes semblent en effet particulièrement exposés à cette pression de sélection, car ils sont unicellulaires et la seule membrane

cellulaire offre une faible protection contre les UVs. J'ai ensuite effectué une **analyse spécifique de la fréquence en dinucléotides de pyrimidine sur trois souches adaptées à différentes expositions UVs du modèle biologique *Prochlorococcus marinus***. Ce micro-organisme marin possède en effet plusieurs souches qui se développent à différentes profondeurs dans la colonne d'eau, et il a été montré que ces souches possèdent des adaptations spécifiques à l'exposition lumineuse dans leur habitat (Moore *et al.*, 1998; van der Staay *et al.*, 2000; Rocap *et al.*, 2003; Coleman *et al.*, 2006). Par ailleurs, j'ai effectué une **analyse spécifique de la fréquence en dinucléotides de pyrimidine sur un large nombre de génomes complets de virus marins**. Les virus n'ont en effet aucune capacité intrinsèque de réparation et dépendent entièrement pour cela de la machinerie de réparation de leur hôte. En outre, les phages marins passent par une phase libre pendant laquelle ils sont transportés par les courants et peuvent séjourner plusieurs jours dans les premiers mètres d'eau en accumulant ainsi des dommages à l'ADN (Wilhelm *et al.*, 2003). Pour l'ensemble de ces études, la sur- et sous-représentation en dinucléotides a été mesurée par la statistique z-score présentée au chapitre précédent.

a. Analyse systématique des génomes bactériens complets

À partir de l'ensemble des 221 génomes complets de bactéries et d'archées disponibles sur la base de données de l'EBI Genome Reviews (téléchargés le 16 juin 2005), j'ai créé deux jeux de données : l'un contenant les séquences annotées comme 'CDS', que j'appellerai par la suite `sequences codantes`, l'autre contenant toutes les séquences n'étant pas annotées comme 'CDS', 'RRNA' (séquences codant pour des ARN ribosomaux) ou 'TRNA' (séquences codant pour des ARN de transfert), que j'appellerai par la suite `sequences intergeniques`.

Les résultats de cette analyse sont présentés sur les quatres graphiques de la figure 2.6. Chaque graphique correspond à un dinucléotide de pyrimidine, sur lequel chaque point représente – pour un chromosome bactérien donné – la moyenne du z-base calculé sur toutes les `sequences intergeniques` et la moyenne du z-codon calculé sur toutes les `sequences codantes` de ce chromosome. Certaines bactéries possèdent deux chromosomes, et sont donc représentées par deux points. Les limites de significativité à 5% de la loi normale centrée réduite sont indiqués par des droites en tirets pour permettre une lecture graphique de la significativité moyenne des z-scores obtenus.

On note tout d'abord (voir figure 2.6) qu'il n'y a pas de sous-représentation systématique des dinucléotides CpC, CpT, TpC ou TpT. Aucun des dinucléotides de pyrimidine n'est globalement sous-représenté, ce qui va clairement à l'encontre de l'hypothèse d'une forte pression de sélection pour l'évitement de ces dinucléotides. Par ailleurs, on note une bonne corrélation entre le contenu en dinucléotides dans les séquences intergéniques et dans les séquences codantes, ce qui montre que les séquences codantes et les séquences intergéniques sont soumises à des méca-

nismes généraux similaires. La sur-représentation assez étendue des dinucléotides TpT dans les génomes bactériens est assez étonnante et pourrait être associée à des périodicités en TpT et ApA responsables de l'enroulement de l'ADN (Tomita *et al.*, 1999).

On remarque que le dinucléotide CpC est sous-représenté pour les deux chromosomes de *Burkholderia mallei* et *Burkholderia pseudomallei*, qui sont des pathogènes couramment trouvés dans le sol et les eaux du sol. Cette ne semble pas être liée à une exposition particulièrement forte aux UVs, et pourrait bien être un trait spécifique de ce genre bactérien.

b. Analyse du modèle biologique *Prochlorococcus marinus*

Chacune des trois souches de *Prochlorococcus marinus* que j'ai étudiées est adaptée à une profondeur différente dans la colonne d'eau (Dufresne *et al.*, 2003; Rocap *et al.*, 2003) et est donc exposée à une intensité différente de radiations UVs. Dufresne *et al.* (2003) montrent que la souche SS120 est adaptée à une vie à 120 mètres de profondeur. Cette souche, et la souche MIT 9313 qui vit à une profondeur de 135 mètres, sont considérées adaptées à de faibles intensités lumineuses (Rocap *et al.*, 2003). La souche MED4 est une souche de surface, adaptée à une vie à 5 mètres de profondeur et est considérée comme adaptée à de fortes intensités lumineuses (Dufresne *et al.*, 2003).

Les intensités résiduelles en radiations UVs peuvent être estimées à partir du coefficient d'absorbance de l'eau pure (Quickenden & Irvin, 1980; Litjens *et al.*, 1999) (voir la figure 2.7 - partie gauche, page 33). À 260 nm, qui équivaut au pic d'absorption de l'ADN, l'intensité résiduelle est de 70% de l'intensité initale à 5 mètres de profondeur (souche MED4), de 0.0002% à 120 mètres de profondeur (souche SS120), et de 0.00007% à 135 mètres de profondeur (souche MIT 9313).

Les numéros d'accession et références des trois souches analysées sont les suivants : souche CCMP 1375 / SS120 / SARG (numéro d'accession GenBank AE017126) (Dufresne *et al.*, 2003), souche CCMP 1378 / MED4 (numéro d'accession GenBank BX548174) et souche MIT 9313 (numéro d'accession GenBank BX548175) (Rocap *et al.*, 2003).

Les résultats de cette analyse sont présentés sur les trois graphiques de droite de la figure 2.7. La partie gauche de la figure représente la perte d'intensité lumineuse selon la profondeur dans l'eau, en fonction de la longueur d'onde. Les profondeurs de l'habitat de chacune des souches y ont été tracées, et sont reliées à la partie droite de la figure. La partie droite de la figure représente trois graphiques, un pour chaque écotype de *Prochlorococcus marinus*, sur lesquels sont tracées les distributions – pour chaque dinucléotide de pyrimidine – des z-codon calculés sur tous les CDS du génome complet, où les limites de significativité à 5% ont été indiquées.

Cette figure (2.7) montre qu'il n'y a pas de différence systématique liée à l'exposition aux UVs entre les abondances relatives en dinucléotides de pyrimidine

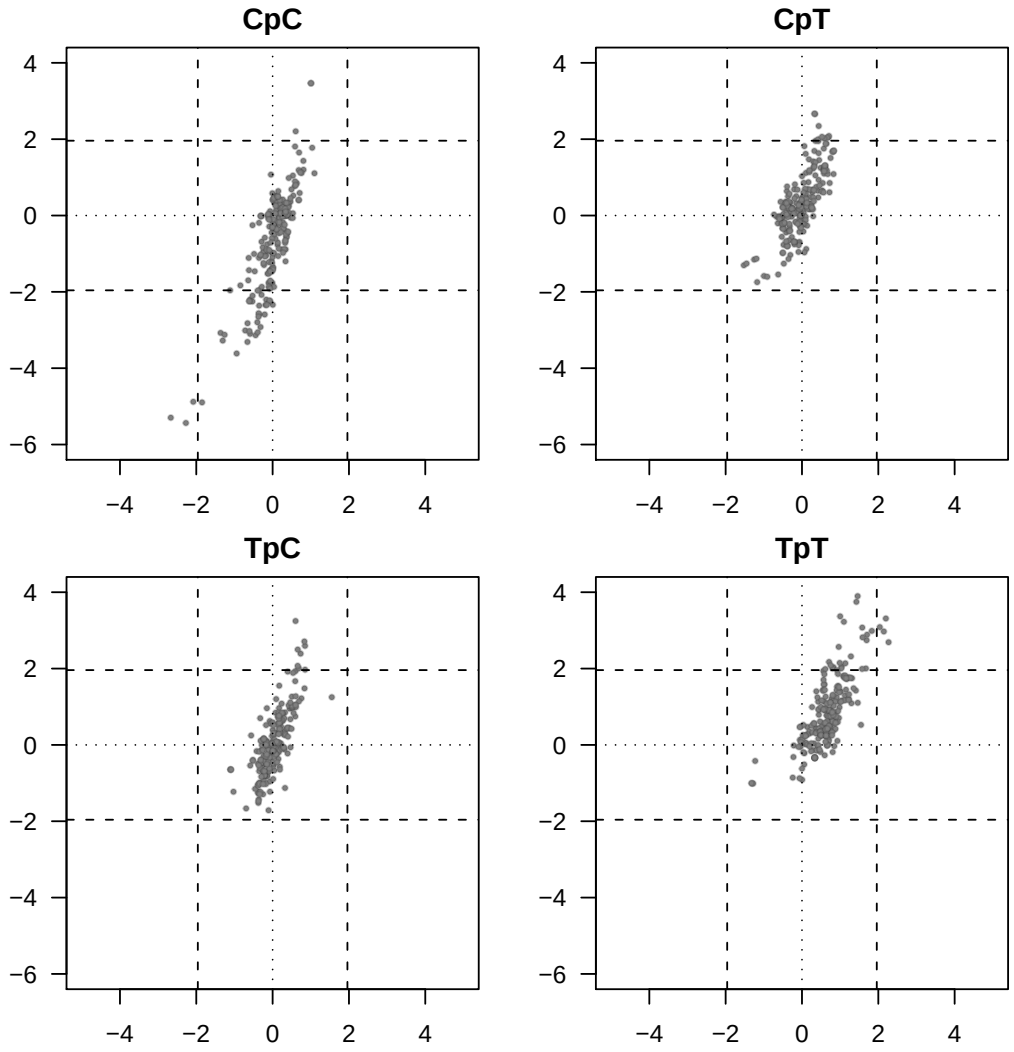


Figure 2.6 – Relation entre la moyenne (par chromosome bactérien) du z-base calculé sur l'ensemble des séquences intergéniques (axe des x) et la moyenne (par chromosome bactérien) du z-codon calculé sur l'ensemble des CDS (axe des y) pour chacun des quatre dinucléotides de pyrimidines. Les droites pointillées correspondent aux valeurs nulles, les droites en tirets correspondent aux limites à 5% de significativité de la loi normale centrée réduite.

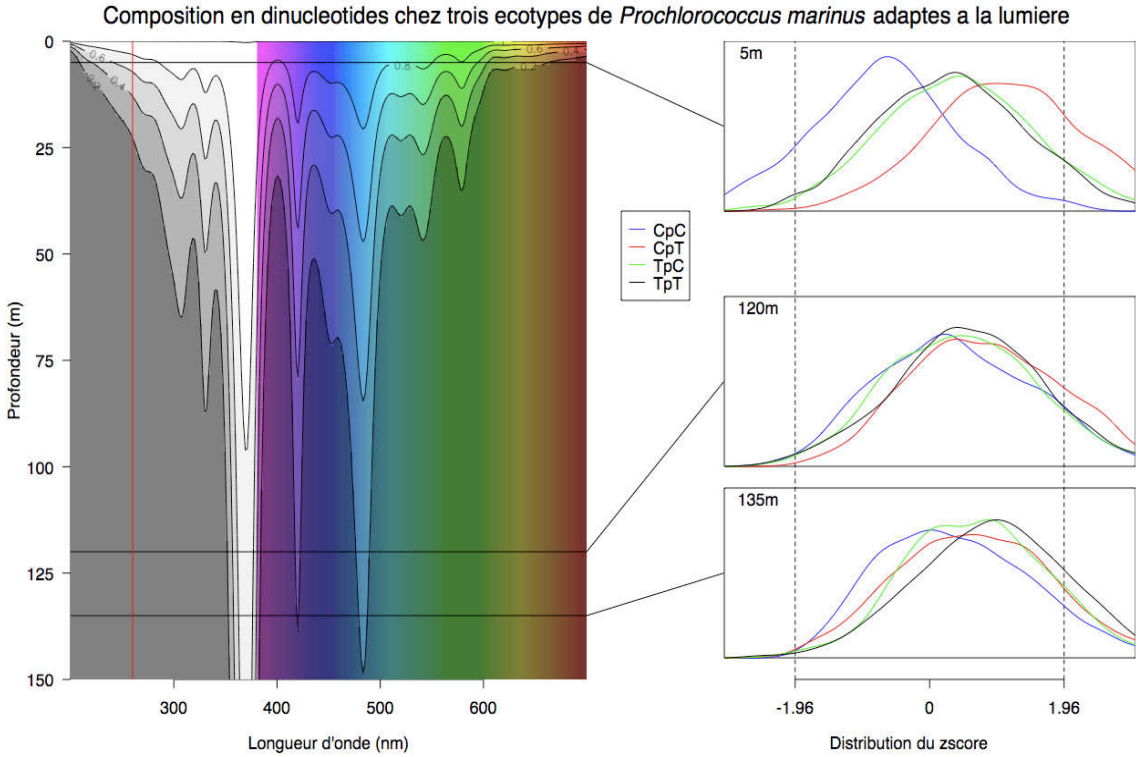


Figure 2.7 – Gauche : absorption de la lumière visible (dégradé coloré) et ultraviolette (dégradé gris) dans de l'eau pure (données compilées de Quickenden & Irvin (1980) et de Litjens *et al.* (1999)). Le pic d'absorption de l'ADN à 260 nm est tracé par une ligne rouge. Droite : la distribution de la statistique z-codon pour chaque dinucléotide de pyrimidine est tracée en face de la profondeur de l'habitat naturel de l'écotype de *Procholoroccus marinus* considéré (5m, 120m, 135m). Les limites à 5% de la loi normale centrée réduite sont tracés en lignes pointillées verticales pour référence. Les séquences des génomes des souches utilisées sont accessibles sur GenBank sous les numéros d'accession suivants : BX548175 (5m, habitat à forte luminosité), AE017126 (120m, habitat à faible luminosité) et BX548174 (135m, habitat à faible luminosité).

chez les trois souches de *Prochlorococcus marinus*. Pourtant, ces trois écotypes ont divergé depuis assez longtemps pour avoir évolué vers des contenus en G+C très différents (30.8% pour MED4 à 5 m de profondeur ; 36.4% pour SS120 à 120 m de profondeur ; 50.8% pour MIT9313 à 135 m de profondeur) (Rocap *et al.*, 2003; Dufresne *et al.*, 2003) et au moins deux de ces écotypes montrent des adaptations génomiques clairement différentes (Rocap *et al.*, 2003). Notre étude montre que, bien que les trois souches présentent des signes forts de divergence, l'éventuelle pression de sélection différentielle due aux radiations UVs n'a pas eu d'impact fort sur leur contenu en dinucléotides de pyrimidine.

c. Analyse de génomes complets de virus marins

Les phages marins semblaient être un modèle biologique de choix pour l'étude de l'impact des UVs sur la composition en bases des génomes. En effet, comme je l'ai précisé précédemment, ils possèdent une phase libre – lorsqu'ils quittent leur hôte – pendant laquelle ils sont transportés par les courants. Ce transport pourra les amener parfois à rester plusieurs heures, voire jours, à la surface de l'eau, notamment pour les phages infectant des bactéries se trouvant déjà proches de la surface. Toutefois, bien que de très nombreux génomes complets de phages aient déjà été séquencés, on ne trouve qu'une vingtaine de génomes complets de phages marins indiqués comme tels dans GenBank (Roseophage SIO1, Phi JL001, phiHSIC, 11b, GBSV1, deux phages de *Synechococcus*, trois phages de *Prochlorococcus marinus*, huit phages de *Vibrio parahaemolyticus*). De très nombreuses séquences virales marines sont en fait présentes dans les données de métagénomique, mais leur assemblage n'est souvent pas complet et il est alors difficile d'étudier des organismes spécifiques. Toutefois, ces séquences permettront probablement bientôt d'étudier les génomes de phages que l'on ne peut pas cultiver en laboratoire. En effet, une étude récente, qui a permis d'isoler des phages présents dans les sables de surface du désert du Sahara (Prigent *et al.*, 2005), aurait fourni à notre étude un jeu de données très intéressant. Toutefois, étant données les difficultés rencontrées pour les cultiver en laboratoire, les souches isolées n'ont pour le moment pas pu être séquencées et le séquençage de ces phages est en cours à travers des techniques de métagénomique (Michael DuBow, *comm. pers.*).

Notre analyse a donc été effectuée grâce à deux jeux de données : l'un contenant des phages que l'on peut considérer non-exposés aux UVs, et l'autre contenant des phages pouvant être exposés occasionnellement à de très fortes radiations ultraviolettes.

Le premier jeu de données de cette analyse consiste en six génomes complets de phages : trois infectant *Geobacillus stearothermophilus* et trois infectant *Geobacillus kaustophilus*, ayant été isolés à partir de boue récoltée dans la plus profonde fosse océanique du monde, la fosse des Mariannes ($\sim$ 10900 mètres de profondeur) (Takami *et al.*, 2004). À une telle profondeur, on peut estimer que l'exposition aux UVs de ces phages est négligeable.

Le deuxième jeu de données de cette analyse consiste en trois génomes complets de phages infectant différentes souches de *Prochlorococcus marinus*. Le podovirus P-SSP7 est spécifique à sa souche hôte et infecte une souche de *Prochlorococcus* adaptée à la lumière. P-SSM2 et P-SSM4 sont deux myovirus : alors que P-SSM4 infecte autant des souches de *Prochlorococcus* adaptées à la lumière que des souches adaptées à la vie en plus grande profondeur, P-SSM2 infecte trois souches de *Prochlorococcus* adaptées à des faibles quantités de lumière (Sullivan *et al.*, 2005). Pour ces trois phages, on peut considérer qu'ils peuvent subir de fortes expositions aux UVs lors de leur phase libre.

Les numéros d'accession GenBank et références des neuf génomes complets de phages que j'ai étudiés sont les suivants. Les phages GSA (numéro d'accession AB126615), GSB (numéro d'accession AB126616), GSC (numéro d'accession AB126617) – infectant *Geobacillus stearothermophilus* – et GKA (numéro d'accession AB126618), GKB (numéro d'accession AB126619), GKC (numéro d'accession AB126620) – infectant *Geobacillus kaustophilus* – ont été séquencés par Takami *et al.* (2004). Les trois phages de *Prochlorococcus marinus* sont P-SSP7 (numéro d'accession NC_006882), P-SSM2 (numéro d'accession NC_006883) et P-SSM4 (numéro d'accession NC_006884) et ont été initialement séquencés par Lindell *et al.* (2004).

Étant donnée la très grande proportion de codant dans les génomes de phages, et la grande compaction de ces génomes, j'ai préféré négliger les séquences intergéniques car les pressions de sélection sur celles-ci sont probablement très fortes et nous n'avons pas de méthode statistique adéquate pour les prendre en compte. J'ai donc calculé la distribution du z-codon sur l'ensemble des CDS de chaque phage et pour chacun des dinucléotides de pyrimidine. Les résultats de cette analyse sont présentés dans les neuf graphiques de la figure 2.8, où chaque graphique représente les quatre distributions de z-codon (pour CpC, CpT, TpC et TpT) calculé sur chaque CDS d'un phage. Les trois graphiques du haut correspondent aux trois génomes complets des phages infectant *Prochlorococcus marinus*. J'ai ici différencié ces trois phages par le type d'hôte qu'elles infectent en modifiant la couleur du fond du graphique. J'ai notamment indiqué par un fond blanc le phage P-SSP7, qui infecte une souche unique de *Prochlorococcus marinus* adaptée à la lumière. Les six graphiques du bas correspondent aux six génomes complets des phages infectant respectivement *Geobacillus stearothermophilus* (ligne intermédiaire) et *Geobacillus kaustophilus* (dernière ligne). Le fond de ces graphiques, d'un gris plus foncé, indique que ces phages ont été isolés à une grande profondeur, où la luminosité est quasi nulle.

Contrairement à ce qui est attendu, le z-codon sur les CDS du phage P-SSP7, qui infecte une souche de *Prochlorococcus marinus* vivant très proche de la surface de l'eau, ne montre pas une forte sous-représentation en dinucléotides de pyrimidine, et montre même une certaine sur-représentation en CpT pour certains CDS. De même pour les phages P-SSM2 et P-SSM4, on n'observe pas une sous-représentation forte en dinucléotides de pyrimidine, sauf pour TpT dans certains

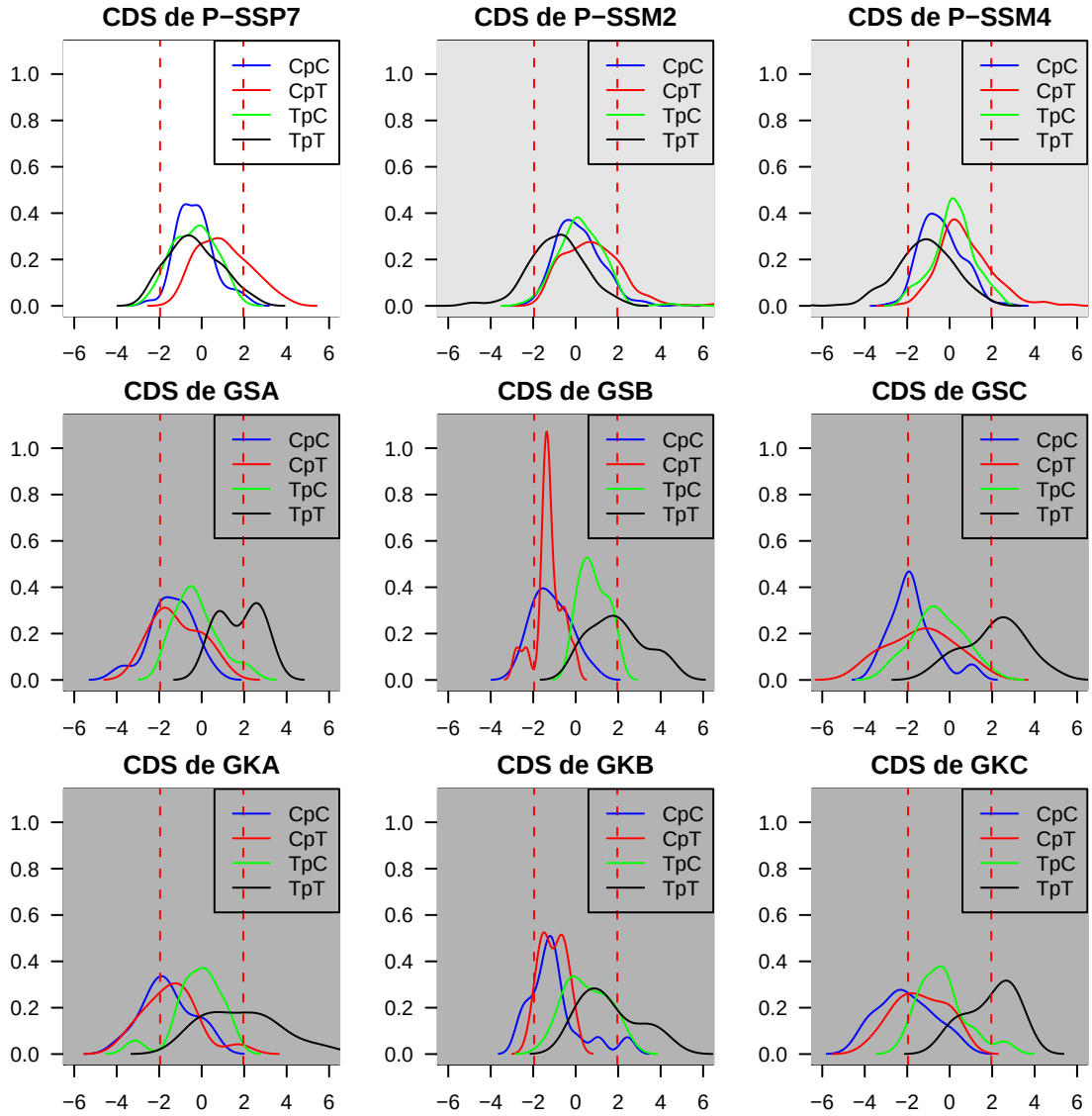


Figure 2.8 – Distribution de la statistique z-codon pour chaque dinucléotide de pyrimidine pour neuf génomes viraux différents. Chaque graphique correspond à un génome viral, et la couleur du fond du graphique indique la profondeur relative de leur habitat naturel : les trois graphiques du haut correspondent aux trois génomes complets des phages infectant *Prochlorococcus marinus*, les six graphiques du bas correspondent aux six génomes complets des phages infectant respectivement *Geobacillus stearothermophilus* (ligne intermédiaire) et *Geobacillus kaustophilus* (dernière ligne). Le fond blanc du graphique correspondant au phage P-SSP7 indique qu'il infecte une souche unique de *Prochlorococcus marinus* adaptée à la lumière. Le fond gris foncé des six graphiques du bas, indique que ces phages ont été isolés à une grande profondeur, où la luminosité est quasi nulle. Les droites en tirets correspondent aux limites à 5% de significativité de la loi normale centrée réduite.

CDS. En comparaison à ces observations, on note que le z-codon calculé sur les CDS des phages infectant *Geobacillus* montre de fortes sous-représentations en certains dinucléotides de pyrimidine (notamment CpC et CpT) alors que l'exposition aux UVs de ces phages peut être considérée comme quasi nulle.

On peut donc en conclure que les UVs ne sont pas directement responsables de la composition en dinucléotides de pyrimidine dans les génomes phagiques. Toutefois, on sait que les radiations UVs représentent le facteur le plus significatif impliqué dans la destruction de l'infectivité virale dans les eaux de surface et que les virus sont des dosimètres naturels de l'impact des UVs car ils ne possèdent aucune capacité de réparation intrinsèque, et dépendent entièrement de leur hôte pour réparer les dommages causés sur leur ADN (Wilhelm *et al.*, 2003). La clef de cette indépendance entre composition en dinucléotides de pyrimidine et exposition fréquente aux UVs réside probablement dans l'utilisation de la machinerie de l'hôte pour la réparation des dommages accumulés dans l'ADN pendant la phase libre des phages. Néanmoins, les dommages accumulés doivent probablement rester en faible nombre, car il a été montré une perte d'adhérence du phage à son hôte après des dommages générés par des UVs proches (Hartman & Eisenstark, 1982; Weinbauer, 2004). Cette perte d'adhérence nuirait alors à l'intégration du phage libre dans un hôte.

Les résultats que j'obtiens dans ces trois analyses (voir figures 2.6, 2.7 et 2.8) tendent à montrer que **les bactéries et les virus ont effectivement développé des stratégies et des mécanismes efficaces pour la prévention et/ou réparation des lésions causées par les UV** (Cleaver, 2006). Ce résultat est en accord avec des études récentes sur la résistance de bactéries marines à l'exposition à des radiations ultraviolettes (Agogué *et al.*, 2005), où l'on montre que le contenu en G+C n'est pas lié à la capacité de résistance ou de non-résistance aux UVs, et que d'autres mécanismes sont bien à l'œuvre.

2.3 Impact de la méthylation

Nous venons d'étudier, à travers l'exemple des UVs, l'influence d'une caractéristique de l'environnement sur la composition en bases des micro-organismes. Je vais maintenant vous présenter l'effet d'une caractéristique interne du génome, telle que la méthylation de certaines de ses bases, sur la composition en bases de celui-ci.

2.3.1 La méthylation

On désigne par méthylation toute transformation de la séquence d'ADN par ajout d'un groupement méthyl. La méthylation s'effectue très fréquemment par l'ajout d'un groupement méthyl sur le cinquième carbone d'une cytosine : on parle alors de la méthylation des cytosines (Costello & Plass, 2001). Il a, par ailleurs,

été mis en évidence que le contexte dans lequel se situe la cystosine influence sa méthylation (Costello & Plass, 2001).

Dans les génomes de vertébrés, et en particulier chez les mammifères et *a fortiori* chez *Homo sapiens* la méthylation de l'ADN a typiquement lieu dans le contexte d'un dinucléotide CpG (Bird, 1980), et l'on considère généralement que les cytosines sont essentiellement méthylées lorsqu'elles appartiennent à un dinucléotide CpG. Toutefois, d'autres contextes entraînent aussi une méthylation des cytosines qui en font partie. Il a été montré que les contextes symétriques où il y a présence d'une cytosine sur chacun des deux brins d'ADN, tel que les trinucléotides CpNpG, sont eux aussi favorables à la méthylation des cytosines, mais dans une mesure bien plus faible (Clark *et al.*, 1995). Certains contextes non symétriques peuvent eux aussi parfois être préférentiellement méthylés, comme c'est le cas des CpA chez *Drosophila melanogaster* (Costello & Plass, 2001).

Dans le génome d'*Homo sapiens*, autour de 5% des cytosines sont méthylées, et ce très préférentiellement dans un contexte CpG (Clark *et al.*, 1995; Colot & Rossignol, 1999). Chez *Drosophila melanogaster*, la méthylation est rarement associée à un dinucléotide CpG, mais plus couramment à un dinucléotide CpA (Costello & Plass, 2001). Chez *Apis mellifera*, la méthylation a préférentiellement lieu dans des contextes CpG, comme c'est le cas dans les génomes de mammifères, mais à un plus faible taux de méthylation que chez ces derniers (Wang *et al.*, 2006b).

On notera que la plupart des méthylations de cytosines sont associées à des contextes symétriques, ce qui permet trois types d'observations : selon que la méthylation a lieu sur les deux brins de l'ADN (méthylation complète), sur un seul des deux brins (hémiméthylation), ou qu'il n'y a pas de méthylation.

2.3.2 La réaction de méthylation

La réaction de méthylation des cytosines est catalysée par une famille d'enzymes capables de transférer un groupement méthyl à l'ADN : les ADN-méthyltransférases. Chez les mammifères, trois enzymes sont responsables des processus de méthylation : DNMT1, DNMT3a et DNMT3b. L'élimination d'une seule de ces enzymes est létal chez la souris (Costello & Plass (2001) et références citées), ce qui suggère une fonction extrêmement importante de ce groupe de gènes.

L'enzyme DNMT1 catalyse principalement le maintien de la méthylation car elle est capable de reconnaître et de reméthyler les sites hémi-méthylés. Cette fonction de maintien de la méthylation est particulièrement importante suite à la réplication : la réplication étant semi-conservative, le brin néo-synthétisé n'est pas méthylé, et cette enzyme permet sa méthylation. C'est pourquoi on peut dire que ce caractère épigénétique est héritable.

Le patron de méthylation de l'ADN chez *Homo sapiens* est effacé lors du stade blastocyste. Le génome est alors progressivement reméthylé jusqu'au stade gastrula, où les niveaux de l'adulte sont atteints. Cette méthylation *de novo* est

catalysée par les enzymes DNMT3a et DNMT3b. On remarquera que chez les mammifères, les gènes soumis à un 'genetic imprinting'[2] ne sont généralement pas déméthylés lors du stade blastocyste, et ne subissent donc pas une méthylation *de novo* (Costello & Plass, 2001).

2.3.3 Fonctions de la méthylation et non-méthylation

Les fonctions liées à la méthylation et à la non-méthylation de certains gènes ou régions du génome sont encore sujettes à de nombreux débats. Il a été suggéré différentes fonctions associables à la méthylation, telles que la **répression de gènes**, la **répression d'éléments transposables**, ou le **maintien de l'intégrité physique des chromosomes**. Il a été proposé, à partir des observations sur le génome d'*Homo sapiens* et sur d'autres génomes de mammifères, que la méthylation servirait à différencier le génome en zones actives et inactives, puisque la méthylation se retrouve préférentiellement sur des zones du génome que l'on considère inactives, malgré quelques exceptions (Costello & Plass, 2001).

On considère généralement que la méthylation d'éléments régulateurs tels que les promoteurs, les activateurs, ou les répresseurs inactive leur fonction (Costello & Plass, 2001). L'exemple le plus clair de répression par la méthylation est celui de la méthylation de gènes 'imprinted' et des gènes du chromosome X inactif. Chez les espèces diploïdes, alors que les deux copies de chaque gène sont généralement exprimées, certains gènes n'expriment qu'une seule des deux copies, et l'autre copie est rendue silencieuse par le 'genetic imprinting'. L'inactivation d'un des deux chromosomes X chez les mammifères femelles est d'ailleurs bien documentée comme étant la conséquence de la méthylation. Les mécanismes liés à ce phénomène ne sont pas encore totalement élucidés, mais les gènes 'imprinted' sont méthylés autour des ilôts CpG associés à ces gènes.

Il a aussi été suggéré que la méthylation permet la défense contre les éléments transposables, puisque de nombreux éléments transposables sont méthylés, et non fonctionnels (Costello & Plass, 2001). Une autre suggestion est que la méthylation sert au maintien de l'intégrité physique des chromosomes par la méthylation de régions pauvres en gènes, tel que l'hétérochromatine péricentrique (Costello & Plass, 2001).

Chez *Homo sapiens*, la méthylation anormale est souvent associée à des pathologies, telle que l'oncogenèse (Costello & Plass, 2001; Goh *et al.*, 2007), ou le syndrome du X fragile (Clark *et al.*, 1995).

2.3.4 Conséquences de la fragilité des cytosines méthylées

La conséquence de la méthylation des cytosines tient à deux choses : d'une part à la faible stabilité des cytosines méthylées, d'autre part au produit de la

[2]méthylation différentielle de l'allèle paternel et de l'allèle maternel

désamination des cytosines méthylées. En effet les cytosines méthylées sont moins stables que les cytosines non méthyléees et subsissent des désaminations spontanées à un taux plus élevé que les cytosines non méthyléees (Bird, 1980). Par ailleurs, lorsqu'une cytosine non méthylée subit une désamination spontanée, elle se tranforme en uracile (reconnaissable par les mécanismes de réparation comme ne faisant pas partie de la composition de l'ADN), et la réparation du mésappariement (par une uracile-ADN-glycosylase) est effectuée de manière à retrouver la séquence ancestrale. Par contre, les cytosines méthylées ont tendance à subir des désaminations spontanées qui les transforment en thymines. Les bases mal appariées ainsi produites sont reconnues par les mécanismes de réparation, mais qui ne peuvent déterminer quelle base vient de subir une mutation. La réparation intervient donc de manière à éliminer le mésappariement, mais selon qu'elle agit sur l'un ou l'autre des deux brins complémentaires, elle peut amener à la fixation de la mutation qui vient juste d'avoir lieu. Cette désamination spontanée sur les cytosines méthylées, qui sont très souvent dans un contexte CpG entraîne une mutation d'un CpG vers un TpG (CpA sur le brin complémentaire).

Dans les génomes de Vertébrés, on remarque que les dinucléotides CpG sont largement sous-représentés. Ceci est depuis longtemps interprété comme étant la conséquence des substitutions ayant lieu sur les dinucléotides CpG. En effet cette sous-représentation en dinucléotides CpG s'accompagne d'une sur-représentation en dinucléotides TpG et CpA, ce qui est une conséquence prévisible de la substitution des dinucléotides CpG méthylés en dinucléotides TpG (CpA sur le brin complémentaire).

Il existe par ailleurs des régions, sur le génome d'*Homo sapiens*, qui ne sont pas méthylées sur les cytosines des CpG, et qui possèdent une sous-représentation moins importante en dinucléotides CpG, et une sur-représentation moins importante en dinucléotides TpG et CpA. Ces régions, ayant une forte fréquence en dinucléotides CpG par rapport à l'ensemble de la séquence, sont dénommés **ilôts CpG**, et peuvent être détectés entre autres caractéristiques par ce patron en dinucléotides CpG, TpG et CpA différent.

2.3.5 La cinquième base de l'ADN ?

L'ensemble des projets de séquençage s'intéressent à la détermination des séquences constituant certains génomes d'intérêt. Ces séquences sont les séquences en quatre bases (deux pyrimidines, deux purines) portées par les génomes. Pourtant, on vient de voir que le rôle joué par certaines caractéristiques épigénétiques telles que la méthylation peut être très important. C'est pourquoi les cytosines méthylées sont souvent appelées la 'cinquième base' de l'ADN, car elles possèdent des caractéristiques différentes des autres cytosines, bien que sur de nombreux points, elles se comportent bel et bien comme des cytosines (transcription/traduction).

La méthylation d'une base, reste toutefois une caractéristique altérable d'un

génome, puisque nous avons vu que le patron de méthylation peut être modifié au cours du développement. Cette caractéristique n'étant pas fixe, on ne peut considérer une cytosine méthylée comme une réelle cinquième base de l'ADN. Par contre, il est important de savoir si une cytosine fait partie de la classe des cytosines non méthylées ou des cytosines méthylées. Il existe pour cela de nombreuses manières d'obtenir cette information (Azhikina & Sverdlov, 2005), et de nombreux projets à grande échelle se penchent actuellement sur ce problème.

2.3.6 Les projets d'épigénomique

Alors qu'il est important de connaître la position de ces cytosines méthylées, cette information est malheureusement très souvent absente des projets de séquençage. Plusieurs projets parallèles se sont mis en place récemment pour pallier à ce manque d'information. Le projet **ENCODE**[3] ('ENCyclopedia Of DNA Elements') envisage, par exemple, de détecter ces sites sur des séquences choisies à l'échelle du génome d'*Homo sapiens* et de les mettre à disposition sur les bases de données publiques. Quelques résultats récents sont d'ailleurs déjà disponibles (Hayashi *et al.*, 2006), mais l'étude se focalise sur l'analyse de la méthylation dans des tissus cancéreux, et ne peut donc pas servir à une analyse de la méthylation dans un cadre évolutif. D'autres groupes, comme celui de Weber *et al.* (2005) s'attaquent aussi à ce problème toujours à une résolution assez large (80kb) sur le génome d'*Homo sapiens*, mais le reproche reste le même : le patron de méthylation de cellules germinales saines n'est pas disponible. Dans la même lignée, mais en visant une résolution beaucoup plus précise, le **Human Epigenome Project**[4] est un projet du Wellcome Trust Sanger Institute pour l'identification de zones où l'on définira les degrés de méthylation à l'échelle de la cytosine (Eckhardt *et al.*, 2006).

Bien que le manque d'analyses sur des cellules germinales saines se fasse sentir, l'ensemble de ces projets laisse supposer qu'une partie de l'information sera bientôt disponible pour l'étude à grande échelle des variations de méthylation le long du génome d'*Homo sapiens*.

[3] http ://www.genome.gov/10005107
[4] http ://www.epigenome.org

Chapitre 3

Modèles d'évolution de séquences

Des modèles évolutifs décrivant les substitutions nucléiques sont utilisés dans une grande variété de domaines en biologie évolutive. Ils sont, par exemple, utilisés pour l'estimation de la distance évolutive entre deux séquences, pour estimer les taux de substitution sur une séquence, ou pour inférer un arbre phylogénétique. Ces modèles posent de nombreuses hypothèses, que l'on a cherché par la suite à relaxer. Je présenterai ici le **cadre général de la modélisation de l'évolution de séquences nucléiques**, quelques **applications classiques**, et les **hypothèses sous-jacentes** à ce type de modélisation. Pour finir, je mentionnerai quelques possibilités proposées dans la littérature pour **relaxer certaines de ces hypothèses**, et je présenterai quelques **conséquences néfastes du maintien de l'hypothèse d'indépendance entre sites**.

3.1 Comment modéliser l'évolution des séquences nucléiques ?

Les génomes sont façonnés autant par des événements à grande échelle, tels que les **transferts horizontaux**, les **réarrangements** (**inversions**, **recombinaisons**, **fusions**, **scissions**), les **duplications locales ou globales**, les **pertes de gènes**, que par des événements à plus faible échelle, tels que les **substitutions**, les **insertions**, les **déletions**. Toutefois, lorsqu'on s'intéresse à l'évolution d'une petite séquence (de l'ordre du gène, par exemple), on considère qu'elle n'évolue que par des mécanismes à faible échelle : les substitutions, les insertions, les duplications, et les déletions et que seule une partie de ces événements est donc suffisante pour décrire cette évolution. Dans certains domaines d'application, l'évolution des séquences biologiques n'est même modélisée que par un seul de ces mécanismes : les substitutions, et on considère que l'ensemble des autres mécanismes peut être négligé. Je ne discuterai pas de ce choix, mais je tâcherai de décrire ce cadre de modélisation et certaines des hypothèses sous-jacentes.

On peut définir les notions de **mutation** et de **substitution** de manière

non ambiguë : une mutation est un événement ponctuel, qui intervient au niveau d'une base et qui entraîne le remplacement de celle-ci par une autre. Une mutation peut, au cours de l'évolution, être maintenue ou éliminée ; une **substitution** est la différence (observée) – ceci sera discuté ultérieurement – entre deux séquences homologues et sur une position, lors d'un alignement, elle est par conséquent le résultat d'une mutation (ou de plusieurs mutations successives) et de la fixation de celle-ci, soit par sélection soit par dérive génétique.

3.1.1 Le modèle markovien

a. Dynamique du système

Dans le cadre de la modélisation de l'évolution d'une séquence d'ADN, la molécule est modélisée par la séquence de ses bases, et l'évolution d'une séquence d'ADN par un **processus stochastique de type markovien**. Chacune des quatre bases : A, T, C et G correspond à un état.

Selon ce modèle, on peut écrire la dynamique de la chaîne markovienne de la manière suivante :

$$\begin{cases} f_A(t+dt) &= f_A(t) + \mathbf{Q}_{TA} f_T(t)dt + \mathbf{Q}_{CA} f_C(t)dt + \mathbf{Q}_{GA} f_G(t)dt \\ & \qquad -(\mathbf{Q}_{AT} + \mathbf{Q}_{AC} + \mathbf{Q}_{AG}) f_A(t)dt \\ f_T(t+dt) &= f_T(t) + \mathbf{Q}_{AT} f_A(t)dt + \mathbf{Q}_{CT} f_C(t)dt + \mathbf{Q}_{GT} f_G(t)dt \\ & \qquad -(\mathbf{Q}_{TA} + \mathbf{Q}_{TC} + \mathbf{Q}_{TG}) f_T(t)dt \\ f_C(t+dt) &= f_C(t) + \mathbf{Q}_{AC} f_A(t)dt + \mathbf{Q}_{TC} f_T(t)dt + \mathbf{Q}_{GC} f_G(t)dt \\ & \qquad -(\mathbf{Q}_{CA} + \mathbf{Q}_{CT} + \mathbf{Q}_{CG}) f_C(t)dt \\ f_G(t+dt) &= f_G(t) + \mathbf{Q}_{AG} f_A(t)dt + \mathbf{Q}_{TG} f_T(t)dt + \mathbf{Q}_{CG} f_C(t)dt \\ & \qquad -(\mathbf{Q}_{GA} + \mathbf{Q}_{GT} + \mathbf{Q}_{GC}) f_G(t)dt \end{cases}$$

où $f_i(t)$ représente la fréquence du nucléotide i au temps t. La séquence évolue alors selon un pas de temps discret noté dt qui correspond au temps séparant deux événements de substitution. Les substitutions ont donc lieu de manière successive (et non simultanée), ce qui semble acceptable étant donnée la longueur des séquences et la faible probabilité des événements de substitution.

L'écriture de la dynamique de la chaîne markovienne ne fait pas intervenir les coefficients de type $\mathbf{Q}_{XX}$, qui sont pour cette raison considérés libres.

b. Écriture matricielle : utilisation de la matrice Q

Cette écriture peut être transformée en écriture matricielle en posant $\mathbf{F}(t)$ le vecteur des fréquences des bases au temps t :

$$d\mathbf{F}(t)/dt = \mathbf{F}(t)\mathbf{Q}$$

où la matrice carrée 4×4 des **taux de substitution**, notée $\mathbf{Q}$ est définie telle que ses coefficients $\mathbf{Q}_{ij}$ sont le taux instantané de substitution de la base i vers la base j. Les coefficients diagonaux, qui ne rentrent pas dans l'écriture de la dynamique du modèle sont, a priori, libres :

$$\mathbf{Q} = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{T} \\ \mathtt{C} \\ \mathtt{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{T} & \mathtt{C} & \mathtt{G} \end{array} \\ \begin{pmatrix} - & a & b & c \\ d & - & e & f \\ g & h & - & i \\ j & k & l & - \end{pmatrix} \end{array}$$

Toutefois, l'écriture matricielle impose que la somme des $\mathbf{Q}_{ij}$ soit nulle par ligne ($\mathbf{Q}_{ii} = -\sum_{j \neq i} \mathbf{Q}_{ij}$).

c. Écriture matricielle : utilisation de la matrice $\mathbf{P}(t)$

La matrice $\mathbf{Q}$ ne sera pas directement utilisée dans le calcul d'une distance entre deux séquences, ni dans le calcul de vraisemblance étant donné une topologie d'arbre qui seront présentés dans une section suivante (Felsenstein, 1981), il faut pour cela transformer la matrice $\mathbf{Q}$ des taux de substitution en une matrice des **probabilités de substitution**, que je noterai $\mathbf{P}(t)$, où les coefficients $\mathbf{P}_{ij}(t)$ sont la probabilité que la base i soit transformée en la base j après un temps t donné. La matrice $\mathbf{P}(t)$ est ensuite obtenue par la relation suivante :

$$\mathbf{P}(t) = e^{\mathbf{Q}t}$$

Il suffit pour cela de diagnoliser la matrice $\mathbf{Q}$ en écrivant : $\mathbf{Q} = \mathbf{U}\Lambda\mathbf{U}^{-1}$ où la matrice $\mathbf{Q}$ se décompose en la matrice $\mathbf{U}$ des vecteurs propres à droite de $\mathbf{Q}$ et la matrice diagonale Λ des valeurs propres λ_i de $\mathbf{Q}$. On peut alors écrire $\mathbf{P}(t) = e^{\mathbf{Q}t} = \mathbf{U}\Delta\mathbf{U}^{-1}$ où la matrice Δ est la matrice diagonale des $\delta_i = e^{\lambda_i t}$.

3.1.2 Calcul de la distribution stationnaire

La dynamique du système décrite précédemment, entraîne l'écriture suivante de la distribution, selon qu'on utilise la matrice $\mathbf{Q}$ ou la matrice $\mathbf{P}(t)$:

$$\mathbf{F}(t) = \mathbf{F}(0)e^{\mathbf{Q}t} \text{ ou } \mathbf{F}(t) = \mathbf{F}(0)\mathbf{P}(t)$$

La nature markovienne du processus entraîne l'**existence d'une distribution stationnaire unique** si et seulement si la chaîne est **irréductible** (chaque état i peut être modifié en n'importe quel état j) et **récurrente positive** (pour chaque état, l'espérance de la durée avant le retour vers cet état est finie). La chaîne converge vers la distribution stationnaire si elle est de plus **apériodique**. L'ensemble de ces propriétés est toujours vérifié dans le cadre du modèle markovien général présenté, la chaîne de Markov converge donc vers une distribution stationnaire unique π qui est solution de l'équation :

$$\pi = \pi \mathbf{P}(t) \text{ ou } \pi \mathbf{Q} = 0$$

La distribution stationnaire π est donc un vecteur propre à gauche normalisé, associé à la valeur propre 1 de $\mathbf{P}$, ce qui est aussi équivalent au vecteur propre associé à la valeur propre 0 de la matrice $\mathbf{Q}$ (Yang, 2006).

À titre d'exemple, voici quelques détails sur le modèle historique de Jukes & Cantor (1969) et, en prévision des modèles qui seront développés dans le chapitre suivant, sur les modèles de Kimura (1980) et Tamura (1992).

- *Le modèle de Jukes & Cantor (1969) : Jukes-Cantor*

Ce modèle, qui possède un seul paramètre libre peut s'écrire par la matrice $\mathbf{Q}$ suivante :

$$\mathbf{Q} = \begin{array}{c} \\ \texttt{A} \\ \texttt{T} \\ \texttt{C} \\ \texttt{G} \end{array} \begin{array}{c} \begin{array}{cccc} \texttt{A} & \texttt{T} & \texttt{C} & \texttt{G} \end{array} \\ \begin{pmatrix} - & \alpha & \alpha & \alpha \\ \alpha & - & \alpha & \alpha \\ \alpha & \alpha & - & \alpha \\ \alpha & \alpha & \alpha & - \end{pmatrix} \end{array}$$

On peut ainsi écrire la distribution stationnaire en nucléotides de ce modèle, qui est la distribution uniforme :

$$\pi_i = \frac{1}{4}$$

pour tout $i \in \{A, C, G, T\}$.

D'autre part, étant donné que ce modèle suppose l'indépendance entre sites, on peut écrire la distribution uniforme des seize dinucléotides comme la distribution uniforme du produit de la fréquence des deux nucléotides qui le composent :

$$\pi_{ij} = \frac{1}{16}$$

pour tout $i, j \in \{A, C, G, T\}^2$.

- *Le modèle de Kimura (1980) : K80*

Ce modèle, qui possède deux paramètres libres (l'un pour décrire les **transitions** – substitution d'une pyrimidine par une pyrimidine ou d'une purine par une purine, l'autre pour décrire les **transversions** – substitution d'une pyrimidine par une purine ou inversement) peut s'écrire par la matrice $\mathbf{Q}$ suivante :

$$\mathbf{Q} = \begin{array}{c} \\ \texttt{A} \\ \texttt{T} \\ \texttt{C} \\ \texttt{G} \end{array} \begin{array}{c} \begin{array}{cccc} \texttt{A} & \texttt{T} & \texttt{C} & \texttt{G} \end{array} \\ \begin{pmatrix} - & \beta & \beta & \alpha \\ \beta & - & \alpha & \beta \\ \beta & \alpha & - & \beta \\ \alpha & \beta & \beta & - \end{pmatrix} \end{array}$$

où α représente le taux de transition et β représente le taux de transversion.

On peut ainsi écrire la distribution stationnaire en nucléotides de ce modèle, qui est la distribution uniforme :

$$\pi_i = \frac{1}{4}$$

pour tout $i \in \{A, C, G, T\}$.

D'autre part, étant donné que ce modèle suppose l'indépendance entre sites, on peut écrire la distribution uniforme des seize dinucléotides comme la distribution uniforme du produit de la fréquence des deux nucléotides qui le composent :

$$\pi_{ij} = \frac{1}{16}$$

pour tout $i, j \in \{A, C, G, T\}^2$.

- *Le modèle de Tamura (1992) : T92*

Ce modèle, développé pour rendre compte des écarts du contenu en G+C à la valeur de 50%, possède trois paramètres libres : un pour les **transitions**, un autre pour les **transversions** et un autre pour le **contenu en G+C**. Sa matrice **Q** s'écrit :

$$\mathbf{Q} = \begin{array}{c} \\ \texttt{A} \\ \texttt{T} \\ \texttt{C} \\ \texttt{G} \end{array} \begin{array}{c} \begin{array}{cccc} \texttt{A} & \texttt{T} & \texttt{C} & \texttt{G} \end{array} \\ \begin{pmatrix} - & (1-\theta)\beta & \theta\beta & \theta\alpha \\ (1-\theta)\beta & - & \theta\alpha & \theta\beta \\ (1-\theta)\beta & (1-\theta)\alpha & - & \theta\beta \\ (1-\theta)\alpha & (1-\theta)\beta & \theta\beta & - \end{pmatrix} \end{array}$$

où α représente le taux de transition, β le taux de transversion et θ représente le contenu en G+C à l'équilibre.

On peut ainsi écrire la distribution stationnaire en nucléotides de ce modèle :

$$\pi_A = \pi_T = \frac{1-\theta}{2}$$

$$\pi_C = \pi_G = \frac{\theta}{2}$$

pour tout $i \in \{A, C, G, T\}$.

D'autre part, étant donné que ce modèle suppose l'indépendance entre sites, on peut écrire la distribution uniforme des seize dinucléotides à partir de la distribution des deux nucléotides qui le composent :

$$\pi_{AA} = \pi_{TT} = \pi_{AT} = \pi_{TA} = \frac{(1-\theta)^2}{4}$$

$$\pi_{CC} = \pi_{GG} = \pi_{CG} = \pi_{GC} = \frac{\theta^2}{4}$$

et $\pi_{XY} = \dfrac{(1-\theta)\theta}{4}$ pour les huit dinucléotides restants.

3.1.3 Estimation d'une distance évolutive entre deux séquences

On définit la **distance évolutive entre deux séquences homologues** comme le nombre moyen de substitutions par site depuis la divergence d'avec la séquence ancestrale. Cette distance ne peut être mesurée directement, puisque nous n'avons généralement jamais accès à la séquence ancestrale, mais elle peut être approchée. Historiquement, cette distance a d'abord été estimée par le pourcentage de différences observées entre deux séquences homologues ($\hat{d}$).

En effet, les événements de substitution sont rares, et le temps évolutif peut donc être, en première approximation, estimé par le nombre d'événements de substitution visibles entre deux séquences. Toutefois, cette estimation tend à sous-estimer la vraie divergence, car elle ne permet pas de prendre en compte les substitutions successives qui auraient pu se produire sur un même site (**substitutions multiples**). Elle n'est donc pas applicable à des séquences très divergentes, qui peuvent avoir subi plusieurs événements de substitution sur un même site.

Les modèles d'évolution de séquences peuvent alors être utilisés pour estimer une distance évolutive entre deux séquences, car la modélisation probabiliste permet d'estimer la proportion de substitutions multiples, et de produire une estimation corrigée.

On peut ainsi écrire facilement une mesure corrigée sous le modèle de Jukes & Cantor (1969) de la mesure observée $\hat{d}$:

$$d_{JC} = \frac{3}{4} log(1 - \frac{4}{3}\hat{d})$$

On note, sur la figure 3.1 que la correction de l'estimation du nombre de substitutions par site sous le modèle de Jukes & Cantor (1969) permet effectivement de prendre en compte les substitutions multiples, qui deviennent invisibles lorsqu'on considère les différences observées. D'autres mesures corrigées peuvent être écrites sous différents modèles, mais je ne les développerai pas ici.

3.1.4 Vraisemblance sous une topologie d'arbre donnée

La **phylogénie moléculaire** est une discipline qui vise à décrire l'évolution des gènes et à la lier à l'évolution des espèces à partir de données sur des séquences (nucléiques ou protéiques) présentes dans ces espèces. Les méthodes d'inférence

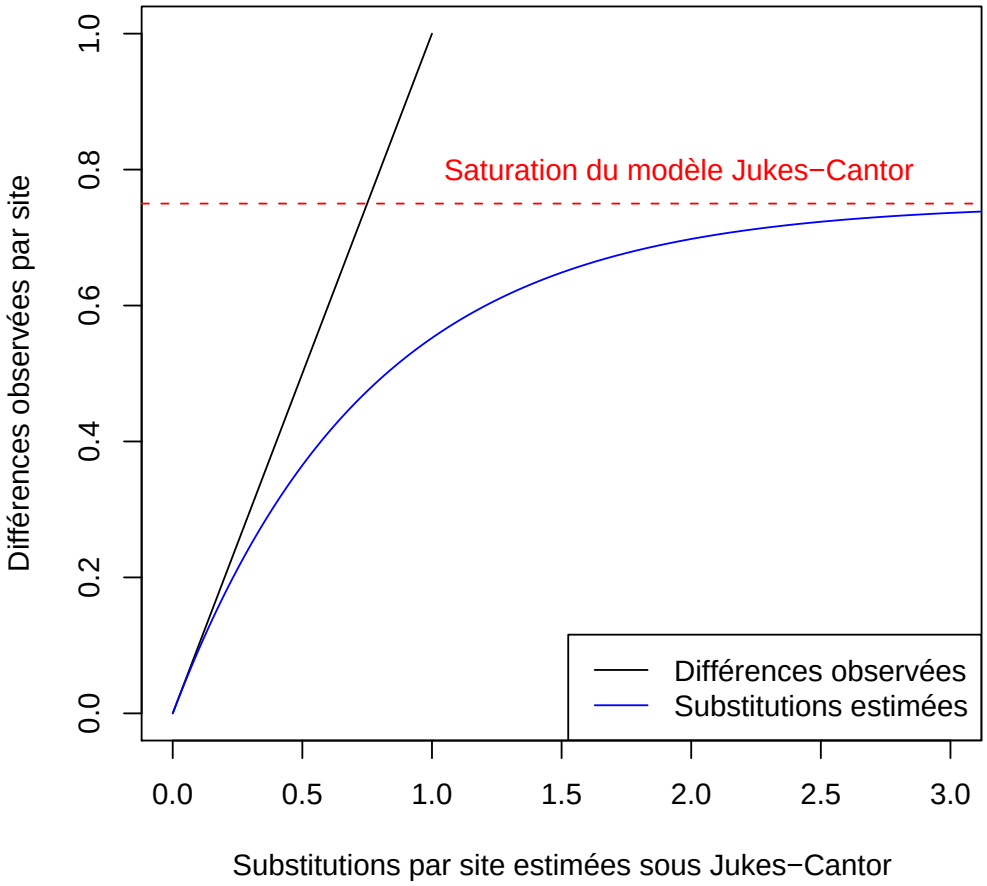


Figure 3.1 – Correction de l'estimation du nombre de substitutions par site entre deux séquences depuis leur divergence sous le modèle de Jukes & Cantor (1969). La droite des différences observées est tracée en noir, la courbe des distances corrigées sous le modèle de Jukes & Cantor (1969) est tracée en bleue. Ce modèle possède une saturation pour une divergence observée de 75% de la séquence, due à l'expression $log(1 - \frac{4}{3}\hat{d})$ et cette limite est tracée en rouge.

phylogénétique sont généralement classées en trois grandes catégories : les **méthodes dites de parcimonie**, les **méthodes basées sur des mesures de distance**, et les **méthodes basées sur un ajustement global (maximum de vraisemblance, méthodes bayésiennes)**. La première catégorie n'utilise pas de modèle évolutif explicite tel que nous les avons définis, et je ne m'attarderai donc pas sur cette classe de méthodes. La deuxième classe de méthodes consiste à établir une matrice de distances entre séquences prises deux à deux (à travers des méthodes proches de celle présentée à la sous-section précédente), puis à construire un arbre à partir de cette matrice de distances additives. La troisième classe de méthodes, pour laquelle je détaillerai quelque peu l'approche par maximum de vraisemblance, englobe les méthodes considérées les plus fiables pour construire des arbres phylogénétiques.

La **vraisemblance** y est définie comme la probabilité d'observer les données étant donnés les paramètres θ du modèle et la topologie d'arbre τ (ainsi que ses longueurs de branches) que l'on suppose avoir généré les données. Les données sont un ensemble D de séquences homologues alignées de longueur n. Cette vision de l'inférence phylogénétique a été initialement proposée par Felsenstein (1981) et est depuis devenue un standard des méthodes en phylogénie moléculaire.

En faisant l'hypothèse que les sites évoluent indépendamment les uns des autres, on peut écrire la vraisemblance de l'ensemble des données comme le produit des vraisemblances sur chacune des positions de l'alignement. Il s'agit ensuite de maximiser la vraisemblance en estimant les paramètres, et en parcourant l'espace des arbres.

$$L(\tau) = \prod_{k=1}^{n} L_k(\tau)$$

Pour des raisons d'implémentation, on utilise généralement plutôt la log-vraisemblance, qui est donc la somme sur l'ensemble des positions de la séquence :

$$l(\tau) = log(L(\tau)) = \sum_{k=1}^{n} log(L_k(\tau))$$

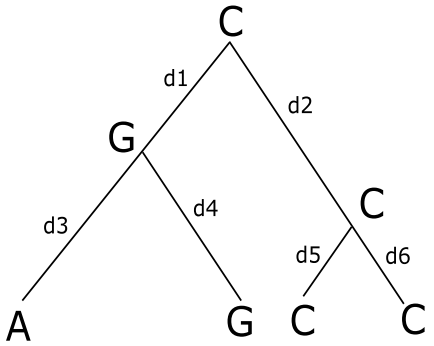
On écrit la vraisemblance sur une position k donnée comme la somme des probabilités de l'ensemble des scénarios d'évolution Υ aboutissant aux bases observées sur la position k, étant donnés θ et τ. Un scénario d'évolution correspond, pour une position k, et le long de l'arbre τ à la succession des bases ayant pu générer les s bases observées aux feuilles de l'arbre. Ainsi :

$$L_k(\tau) = \sum_{\Upsilon} L(\Upsilon|\tau, \theta)$$

Prenons un exemple simple pour illustrer ce calcul, et plaçons-nous sur la topologie T suivante. On s'intéresse à la probabilité du scénario E pour un site

donné, décrit par le graphique suivant. On suppose que l'évolution a eu lieu suivant un modèle représenté par la matrice $\mathbf{P(t)}$. On peut alors écrire, à partir de la matrice $\mathbf{P(t)}$ du modèle :

$$L(E|T,\theta) = P_{CG}(d_1).P_{GA}(d_3).P_{GG}(d_4).P_{CC}(d_2).P_{CC}(d_5).P_{CC}(d_6)$$



Cette méthode, très largement utilisée en phylogénie moléculaire, sera utilisée comme exemple dans la dernière section de ce chapitre, où nous traiterons de la conséquence de l'utilisation d'hypothèses inadéquates.

3.2 Hypothèses explicites et implicites

3.2.1 Présentation des hypothèses sous-jacentes

Le cadre de modélisation que je viens de présenter pose de nombreuses hypothèses. Les modèles classiques de Jukes & Cantor (1969), Kimura (1980), Tajima & Nei (1984), Hasegawa *et al.* (1985), Tamura (1992), Tamura & Nei (1993) et Rzhetsky & Nei (1995) posent tous les hypothèses suivantes : l'évolution d'une séquence est **modélisée uniquement par les processus de substitution** auxquels elle est soumise, ces processus sont **uniformes le long de la séquence** (tous les sites évoluent selon les mêmes taux de substitution), **homogènes au cours du temps** (les taux de substitution sont constantes au cours du temps et restent les mêmes sur différentes branches d'un arbre phylogénétique), **stationnaires** (la composition en bases des séquences est constante au cours du temps et correspond à l'équilibre du modèle) et **chaque site évolue de manière indépendante** par rapport aux autres sites de la séquence.

Ces hypothèses correspondent à des contraintes mathématiques nécessaires que l'on peut retrouver dans les écritures précédentes. Nous verrons dans la section suivante comment ces contraintes peuvent être relaxées et les conséquences mathématiques de la levée de ces contraintes.

L'**hypothèse d'uniformité le long de la séquence** réside dans l'application d'un processus markovien identique sur chacun des sites, puisque la matrice **Q** est indépendante de la position sur la séquence et est applicable sur l'ensemble des sites.

L'**hypothèse d'homogénéité temporelle** réside dans l'écriture des coefficients de la matrice **Q** du modèle markovien, que l'on note $\mathbf{Q}_{ij}$ (taux de substitution de la base i vers la base j) et qui sont indépendants du temps.

L'**hypothèse de stationnarité** permet une estimation simple de la fréquence à l'équilibre par les fréquences observées sur la séquence. On suppose que le processus est stationnaire, c'est-à-dire que la chaîne de Markov a atteint son équilibre et que les fréquences π d'équilibre sont atteintes et ne changent pas au cours du temps. Ceci permet des développements mathématiques, mais n'a pas de fondement biologique. Elle permet de mesurer une distance, et aussi de calculer une vraisemblance sous une topologie (voir ci-dessus).

L'**hypothèse d'indépendance entre sites**, généralement couplée à une hypothèse de distribution identique, réside dans le traitement de la séquence comme un lot de variables aléatoires indépendantes et identiquement distribuées. Elle est utilisée autant dans l'estimation de la distance entre deux séquences que dans l'estimation de la vraisemblance sous une topologie donnée, car elle permet de traiter tous les sites indépendamment les uns des autres, et d'obtenir une écriture faisant intervenir une somme (ou un produit) sur l'ensemble des sites.

3.2.2 Levée de ces hypothèses

a. Levée de l'hypothèse d'uniformité

La constatation de l'existence de structurations au sein des séquences nucléiques (séquences codantes, séquences régulatrices, pseudogènes, variations du contenu en G+C, présence d'ilôts CpG) peut entraîner la remise en question de l'hypothèse d'**uniformité** des processus le long des séquences. En effet, les différentes structures, à cette échelle, peuvent évoluer soit à des vitesses différentes, soit selon des patrons de substitution différents. Ceci amène à la modélisation de l'hypothèse de variation des taux de substitution entre sites.

Comme il n'est pas possible, pour des raisons de complexité, d'intégrer un trop grand nombre de paramètres dans ces modèles, voici deux des développements les plus utilisés :

- *L'hypothèse du covarion*

Le **covarion**[1] est une hypothèse proposée par Fitch & Markowitz (1970) puis développée en détail par Tuffley & Steel (1998) qui permet d'introduire des taux

[1] concomitantly variable codons

de substitution qui sont spécifiques aux différents sites, et qui ne sont plus uniquement fonction de l'identité de la base. Cette hypothèse découle des relations fonctionnelles entre acides aminés qui contraignent les relations entre codons au niveau de l'ADN. Le modèle du covarion suppose que, en termes évolutifs, les sites sont soit 'éteints' (ils n'évoluent pas), soit 'allumés' (ils peuvent évoluer).

- *Loi Gamma*

Or, d'un point de vue biologique, il semble plus judicieux d'estimer que les sites ne sont pas soumis à des pressions de substitution de type binaire mais d'intensité graduelle. C'est pourquoi Yang (1994) ainsi que Kelly & Rice (1996) modélisent la variation des taux de substitution par un continuum (distribution de probabilité) et incorporent une distribution du type Gamma (Yang, 1994; Kelly & Rice, 1996), ou une distribution du type log-normale (Kelly & Rice, 1996) au modèle. En décrivant la distribution de probabilité des taux de substitution entre sites à l'aide d'une distribution, on considère que la matrice de substitution en chaque site est multipliée par un facteur r tiré dans une distribution. La distribution la plus couramment utilisée est la loi Gamma, probablement de par sa flexibilité. La fonction de densité de cette distribution s'écrit en effet :

$$f(r;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} e^{-\beta r} r^{\alpha-1}$$

où on pose généralement $\alpha = \beta$, qui sont respectivement les paramètres de forme et d'échelle de la loi, de manière à avoir une distribution d'espérance égale à 1. La variance est alors égale à $1/\alpha$. Ce qui, pour différents paramètres, donne des allures de courbes très différentes (voir figure 3.2).

Yang *et al.* (1994) montrent que l'incorporation d'une distribution Γ telle que Yang (1994) le propose améliore grandement le modèle à 5 coefficients de Hasegawa *et al.* (1985), et cette amélioration est très largement répandue dans les méthodes d'inférence phylogénétique.

La levée de l'hypothèse d'uniformité permet donc de considérer que chaque site possède sa vitesse d'évolution propre. Toutefois, bien qu'il existe une variabilité entre positions le long d'une séquence, celle-ci est supposée constante au cours du temps. Autrement dit, la vitesse d'évolution est considérée identique, pour une même position sur deux séquences différentes car l'hypothèse d'homogénéité est maintenue.

b. Levée de l'hypothèse d'homogénéité

L'hypothèse d'homogénéité, tout comme celle d'uniformité, n'a aucun fondement biologique. Cette hypothèse est détruite dès lors qu'on imagine la modification des pressions de sélection agissant sur un site donné, sur un domaine protéique ou sur l'ensemble d'un gène au cours des temps évolutifs.

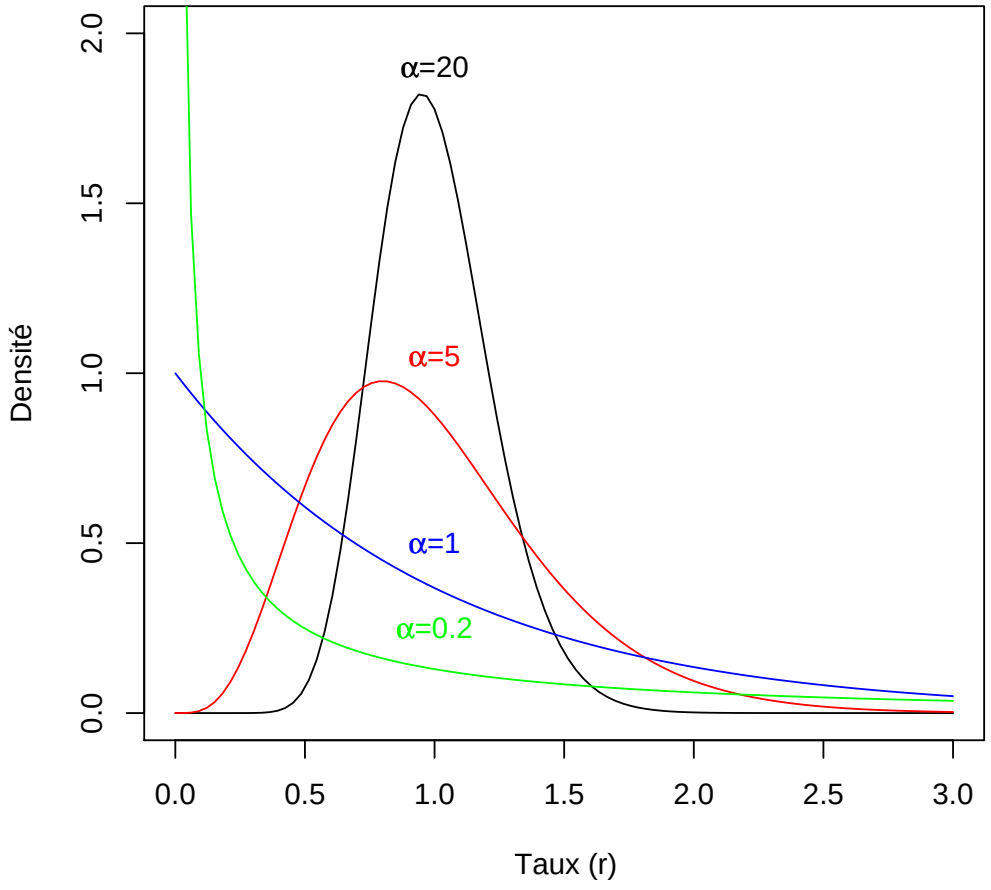


Figure 3.2 – Distribution gamma avec différents paramètres α, repris de Yang (1996). Le paramètre d'échelle (β)de la distribution est fixé égal au paramètre de forme (α), de manière à ce que l'espérance soit égale à 1. La variance est ainsi égale à $1/\alpha$.

De nombreux travaux se sont attelés à la levée de l'hypothèse d'homogénéité. On note notamment le travail de Galtier & Gouy (1998) et, qui a récemment été implémenté dans un programme d'inférence phylogénétique par maximum de vraisemblance (Boussau & Gouy, 2006). Ainsi, un modèle comme celui développé par Galtier (2001) élargit les modèles incorporant une loi Gamma (Yang & Kumar, 1996) et permet, en plus de la variation des taux de substitution entre sites par l'utilisation d'une loi Gamma – relaxation de l'hypothèse d'**uniformité** – d'intégrer une variation des taux de substitution entre branches d'un arbre phylogénétique – relaxation de l'hypothèse d'**homogénéité** – ceci en ne rajoutant qu'un seul paramètre supplémentaire.

3.3 Conséquences de l'inadéquation de l'hypothèse d'indépendance entre sites

Au cours des développements de la modélisation de l'évolution de séquences, certaines contraintes inhérentes au modèle ont été relaxées. Le modèle de Markov usuel suppose que les sites évoluent de manière **indépendante**, et que les processus sont **uniformes** et **homogènes**. Nous avons vu que l'hypothèse d'homogénéité peut être relaxée, et qu'il est possible d'intégrer une composante de variation temporelle permettant d'éliminer – du moins en partie – l'hypothèse d'homogénéité. Il nous reste à considérer l'hypothèse d'**indépendance** entre sites.

L'hypothèse d'indépendance suppose que les substitutions dépendent uniquement de la base subissant l'événement. Toutefois, et bien que les mécanismes responsables de leur formation et de leur maintien ne soient pas toujours connus, il existe au sein des génomes des séquences qui ne sont pas aléatoires (boîtes TATA, séquences répétées en tandem). Cette remarque amène à envisager l'hypothèse que les substitutions au sein d'une séquence puissent être dépendantes non seulement de la base subissant l'événement, mais aussi d'autres caractéristiques, comme par exemple des bases au voisinage de celle-ci.

Je vais donc maintenant m'intéresser à la conséquence de l'utilisation d'hypothèses inadéquates dans le cas de méthodes faisant intervenir des modèles probabilistes d'évolution. Je vais en particulier détailler la conséquence de l'utilisation de l'**hypothèse d'indépendance entre sites** dans le cas de l'inférence phylogénétique par maximum de vraisemblance. Les méthodes d'inférence phylogénétiques sont, en effet, largement utilisées, mais présentent certaines limites, essentiellement dues aux hypothèses sous-jacentes (explicites, ou implicites) qui ne décrivent pas toujours de manière correcte la réalité biologique du signal phylogénétique à traiter.

Avant de commencer cette analyse, faisons le point sur quelques notions générales sur l'estimation de la qualité en phylogénie moléculaire.

3.3.1 Notion de qualité en inférence phylogénétique.

a. Définitions.

L'inférence phylogénétique d'un arbre de séquences permettant d'obtenir des informations sur l'arbre des espèces porteuses est largement utilisée en biologie. Le choix d'une méthode d'inférence et des paramètres de celle-ci est un point crucial de l'analyse. Il existe de nombreuses manières de comparer des méthodes de reconstruction phylogénétique, nous allons ici décrire quelques critères de qualité qui nous permettent de classer les méthodes les unes par rapport aux autres, et qui peuvent être prises en compte lors du choix d'une méthode d'inférence.

Dans le cadre de l'inférence phylogénétique, on estime l'arbre phylogénétique vrai par un arbre phylogénétique inféré grâce à une méthode donnée d'inférence. L'estimateur est donc la méthode d'inférence phylogénétique. La valeur estimée est l'arbre phylogénétique inféré. La valeur vraie est l'arbre phylogénétique vrai qui a produit les données que l'on observe. Les données sont les séquences alignées aux feuilles de l'arbre.

- *Biais.*

Le biais d'un estimateur est défini comme l'**écart entre l'espérance de l'estimateur et la valeur vraie du paramètre estimé**. Lorsque ce biais est nul, l'estimateur est dit non biaisé. Pour un estimateur biaisé, on définit l'erreur systématique par la mesure de l'écart entre l'espérance de l'estimateur et la valeur vraie.

- *Consistance.*

Un estimateur consistant est un **estimateur qui converge en probabilité vers la valeur vraie du paramètre estimé lorsque la taille de l'échantillon augmente**. En réalité, il s'agit ici de la définition d'un estimateur faiblement consistant, mais ces deux notions sont généralement considérées équivalentes en phylogénie.

Ce critère est une condition nécessaire, mais pas suffisante pour définir un bon estimateur. En effet, Yang (2006) propose un exemple d'estimateur consistant mais biaisé : alors que l'estimateur usuel de la probabilité p associée à x succès dans un échantillon binomial de n essais est la proportion dans l'échantillon : $\hat{p} = x/n$ est consistant et sans biais, on peut définir un estimateur arbitraire $\hat{p} = (x - 1000)/n$, qui est lui aussi consistant, mais n'a pas de sens.

- *Robustesse.*

Un estimateur est considéré robuste s'il est **peu sensible à l'écart entre la population réelle observée et les hypothèses du modèle théorique sous-jacent**.

b. Stratégies.

Il existe deux grandes stratégies pour évaluer la qualité d'une méthode d'inférence phylogénétique. L'une est basée sur l'**étude théorique des méthodes d'inférence phylogénétique**, de manière à déterminer les caractéristiques statistiques de l'estimateur de l'arbre vrai (i.e. démonstration du caractère non biaisé d'un estimateur, démonstration de la consistance d'une méthode). L'autre est basée sur l'**étude empirique des méthodes d'inférence phylogénétique**. Lorsqu'il est possible d'effectuer une analyse théorique de la qualité d'une méthode, cette approche doit bien entendu être préconisée. Toutefois, il existe certains paramètres (tels que la robustesse d'une méthode) dont l'analyse théorique est difficile, et où l'étude empirique s'avèrera nécessaire.

Les études empiriques peuvent suivre différentes approches, selon que la phylogénie de référence est connue ou pas. La première de ces approches est expérimentale et il s'agit d'une des approches les plus intéressantes, mais probablement une des plus difficile à mettre en place, est l'**approche de phylogénétique expérimentale** (Hillis *et al.*, 1992; Hall, 2005). Il s'agit de faire évoluer une séquence (séquence virale, séquence d'ARN) in vitro pendant un certain temps. La phylogénie est ainsi connue, tout comme la séquence ancestrale, ainsi que les séquences feuilles, et permet une étude complète. En contrepartie, cette approche est difficile à mettre en place, et ne permet bien évidemment pas de balayer un large champ de paramètres différents, puisque peu de paramètres peuvent être contrôlés par l'expérimentateur. La deuxième de ces approches est une **approche de comparaison de l'arbre inféré à un arbre phylogénétique de référence**, et consiste à prendre des phylogénies bien établies de certains organismes comme phylogénie de référence, et à étudier si la méthode d'inférence utilisée retrouve l'arbre de référence. Une des critiques principales est celle de l'obtention de la phylogénie de référence. La troisième approche, très largement utilisée, est l'**approche par simulations de Monte-Carlo** (Saitou & Imanishi, 1989; Jin & Nei, 1991; Guindon & Gascuel, 2003). Cette approche est la seule qui permet d'analyser les méthodes d'inférence phylogénétiques en parcourant de manière assez exhaustive l'espace des paramètres. Les deux autres approches étant, en particulier, cantonnées à l'évaluation de l'inférence d'un très faible nombre de phylogénies différentes. Cette méthode est utilisée couramment pour valider des méthodes d'inférence. Elle a, en particulier été utilisée récemment par Guindon & Gascuel (2003) pour comparer la qualité des méthodes basées sur une approche par maximum de vraisemblance (voir figure 3.3).

Dans le cadre de notre question, que je vais maintenant développer, et qui vise à mesurer les conséquences de la violation de l'hypothèse d'indépendance entre sites sur la qualité des méthodes d'inférence phylogénétique, j'ai opté pour l'approche par simulations.

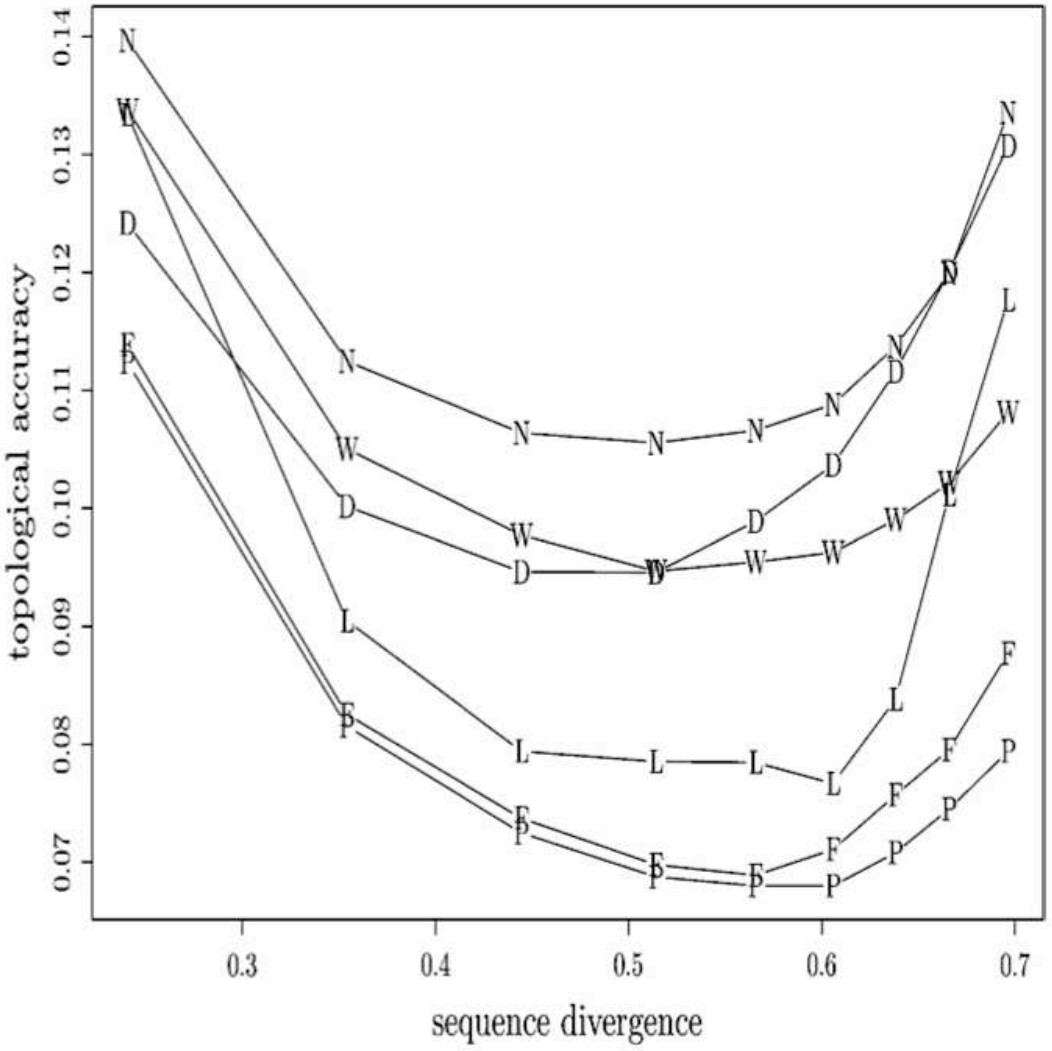


Figure 3.3 – Comparaison de différentes méthodes d'inférence phylogénétique, par simulation d'évolution de séquences sur un arbre connu puis reconstruction d'un arbre à l'aveugle. L'axe des ordonnées représente la distance de Robinson-Foulds[a] entre l'arbre vrai et l'arbre reconstruit, l'axe des abscisses représente la divergence entre séquences. Les différentes méthodes d'inférence sont représentées par des lettres : D pour DNAPARS – méthode de parcimonie –, N pour Neighbor-Joining, W pour Weighbor – méthodes de distances –, L pour NJML, F pour fastDNAml, P pour Phyml – méthodes au maximum de vraisemblance. Figure extraite de Guindon & Gascuel (2003).

[a]voir définition page 61

c. Bilan bibliographique

L'analyse de la qualité des méthodes d'inférence phylogénétique a fait l'objet de très nombreuses études. En effet, de par la vaste utilisation de ces méthodes, et de par le nombre d'hypothèses sous-jacentes qu'elles posent, il est primordial d'analyser la qualité de ces méthodes pour les valider.

Les premières études de l'efficacité des méthodes d'inférence remontent aux premières méthodes d'inférence phylogénétique. On peut citer Tateno *et al.* (1982) comme l'une des premières études de la qualité d'une inférence phylogénétique. Des études plus récentes, comme celle de Tillier & Collins (1995), qui s'appuie sur le même type de méthode pour la mesure de la qualité d'une inférence phylogénétique que Tateno *et al.* (1982), ont permis de comparer différentes méthodes d'inférence entre elles. Tillier & Collins (1995) considère notamment que les méthodes au maximum de vraisemblance sont de manière générale meilleures que les méthodes basées sur les distances, mais surtout que ces premières semblent très robustes à l'écart à l'hypothèse d'indépendance entre sites. Nous allons étudier cette question en détail.

3.3.2 Conséquences liées à la violation de l'hypothèse d'indépendance entre sites.

Je vais, dans cette dernière section, présenter la méthodologie choisie pour l'étude des conséquences liées à la violation de l'hypothèse d'indépendance entre sites sur la robustesse des méthodes d'inférence phylogénétique. Je décrirai ensuite les résultats obtenus et les perspectives qui se dégagent de ce travail.

a. Méthode d'analyse par simulations de la robustesse des méthodes d'inférence.

L'approche pour laquelle j'ai opté pour l'analyse de la robustesse des méthodes d'inférence phylogénétique a été très largement utilisée par le passé et présente l'avantage de permettre la maîtrise d'un très grand nombre de paramètres (Tateno *et al.*, 1982; Tillier & Collins, 1995; Guindon & Gascuel, 2003).

Soit un arbre phylogénétique à s feuilles, une séquence ancestrale, et un modèle évolutif : simulons l'évolution de la séquence le long de l'arbre jusqu'à obtention s séquences feuilles alignées. Appliquons ensuite sur l'alignement des s séquences feuilles, une méthode d'inférence phylogénétique choisie et comparons l'arbre initial à l'arbre reconstruit grâce à une méthode de comparaison.

- *Arbres phylogénétiques de référence*

Les arbres phylogénétiques utilisés lors des simulations effectuées sont le jeu de données de test proposé par Guindon & Gascuel (2003) pour l'analyse de l'algorithme Phyml. Il s'agit ici du jeu de données raciné initial, qui permet

d'affecter une séquence ancestrale à la racine de l'arbre pour simuler l'évolution de celle-ci le long de l'arbre.

Ce jeu de données comporte 5000 arbres phylogénétiques de 40 taxons chacun, générés par un processus aléatoire de spéciation standard décrit par Kuhner & Felsenstein (1994). La longueur moyenne des branches est de 0.06 substitutions par site, et la longueur totale des arbres est distribuée de manière uniforme dans [0.4 ;9] substitutions par site.

- *Séquences ancestrales*

J'ai choisi de générer des séquences ancestrales à contenu en G+C variable, de manière à obtenir des séquences ancestrales à contenu en dinucléotides CpG variable. J'ai donc généré des séquences de 500 bases indépendantes et identiquement distribuées, à contenu en G+C allant de 10% à 90%.

- *Modèles évolutifs*

Afin de tester l'hypothèse d'indépendance entre sites, j'ai choisi deux modèles de simulation. Un modèle simple, le modèle de Kimura (1980) (que je noterai K80) ; et un modèle incorporant le mécanisme de méthylation-désamination des dinucléotides CpG, que je noterai K80+CpG, et qui sera développé en détail dans le chapitre suivant. Pour le moment, il suffit de considérer que ce deuxième modèle est issu du modèle de Kimura (1980) et brise l'hypothèse d'indépendance entre sites, et va nous permettre de tester la robustesse de l'inférence phylogénétique face à l'écart à cette hypothèse.

- *Métrique de comparaison d'arbres*

Bien qu'un arbre phylogénétique soit autant défini par sa topologie que par les longueurs de ses branches, il n'existe pas encore de méthode satisfaisante de comparaison d'arbres phylogénétiques, et je me suis donc limitée à la mesure de Robinson & Foulds (1981). Il s'agit, en effet, d'une des métriques de distance topologique entre arbres les plus largement utilisées et je me suis servie du programme Robinson-Foulds[2], qui implémente l'algorithme défini par Makarenkov & Legendre (2001).

Cette mesure comptabilise le nombre minimum d'événements élémentaires – définis comme la scission et la fusion de noeuds – nécessaires pour passer d'un arbre à l'autre. Ce qui équivaut à dénombrer le nombre de bipartitions différentes entre les deux arbres. Une bipartition est définie comme les deux sous-arbres non-triviaux[3] issus de la destruction d'une branche interne. Cette valeur est ensuite normalisée par le nombre maximum de bipartitions différentes possibles entre deux arbres binaires à s feuilles, qui est égal à deux fois le nombre de branches

[2]http ://www.bio.umontreal.ca/casgrain/en/labo/robinson_foulds.html

[3]un sous-arbre trivial est une feuille

internes dans un arbre binaire à s feuilles : $2(s-3) = 2s-6$. La distribution de la mesure de Robinson-Foulds est discrète et contenue dans [0 :1] : la valeur 0 correspond à la distance séparant deux arbres topologiquement identiques, la valeur 1 à la distance séparant deux arbres n'ayant aucune branche interne en commun.

En résumé, j'ai effectué plusieurs lots de 100 simulations selon le protocole suivant, où, étant donné un modèle d'évolution et un contenu ancestral en G+C, j'ai :

1. tiré un arbre au hasard dans la collection de 5000 arbres,
2. construit une séquence ancestrale uniforme, contenant un G+C donné,
3. fait évoluer la séquence de la racine aux feuilles selon un modèle d'évolution donné (K80 et K80+CpG),
4. reconstruit un arbre phylogénétique par maximum de vraisemblance (Phyml),
5. comparé la distance entre l'arbre reconstruit et l'arbre vrai.

b. Résultats sur la violation de l'hypothèse d'indépendance entre sites

Après simulation sur 100 arbres sous le modèle de K80+CpG – et sous le modèle de K80 –, à partir de séquences ancestrales à contenu en G+C variable (de 10 à 90 %), et après reconstruction d'arbres phylogénétiques par PHYML, les résultats sont présentés sur les courbes en rouge de la figure 3.4.

On voit bien que lorsque l'effet de voisinage est faible (ici lorsque la séquence ancestrale est pauvre en G+C), la méthode d'inférence phylogénétique reste robuste à ce faible écart à l'hypothèse d'indépendance – voir annexe C pour plus de détails sur les variations en G+C. Lorsque le contenu ancestral en G+C augmente, le contenu des séquences ancestrales en dinucléotides CpG augmente. Ceci entraîne une violation de l'hypothèse d'indépendance d'autant plus forte, et se traduit clairement par une moins bonne inférence phylogénétique. En effet, la distribution des distances de Robinson-Foulds entre l'arbre vrai et l'arbre inféré, lorsque l'évolution a eu lieu sous K80+CpG, s'écarte de la valeur 0 (arbres identiques).

L'ordre de grandeur de ce biais est de l'ordre de la différence de performance entre certaines méthodes d'inférence phylogénétique (voir figure 3.3). Dans les cas où l'effet de voisinage est fort, comme ici lorsque la séquence ancestrale est riche en G+C, il faudra donc être prudent dans l'interprétation de résultats obtenus par des méthodes d'inférence phylogénétique pour lesquelles l'hypothèse d'indépendance entre sites est clairement violée.

Cette constatation de l'effet de la violation de l'hypothèse d'indépendance sur la qualité de l'inférence phylogénétique, pourrait être un indice que de nouvelles méthodes d'inférence phylogénétique sont nécessaires pour prendre en compte ces dépendances. Toutefois, il faut être prudent sur les possibilités de ces développements. En effet, d'un point de vue pratique, les méthodes d'inférence phylogénétiques arrivent en aval de tout un arsenal de méthodes qui supposent aussi cette indépendance évolutive entre les sites (alignement de séquences, nettoyage des sites mal alignés, mesure d'une distance entre séquences). La conception d'une méthode d'inférence phylogénétique qui incorporerait des dépendances entre sites nécessiterait donc de résoudre chacune de ces étapes en y incorporant de la dépendance entre sites.

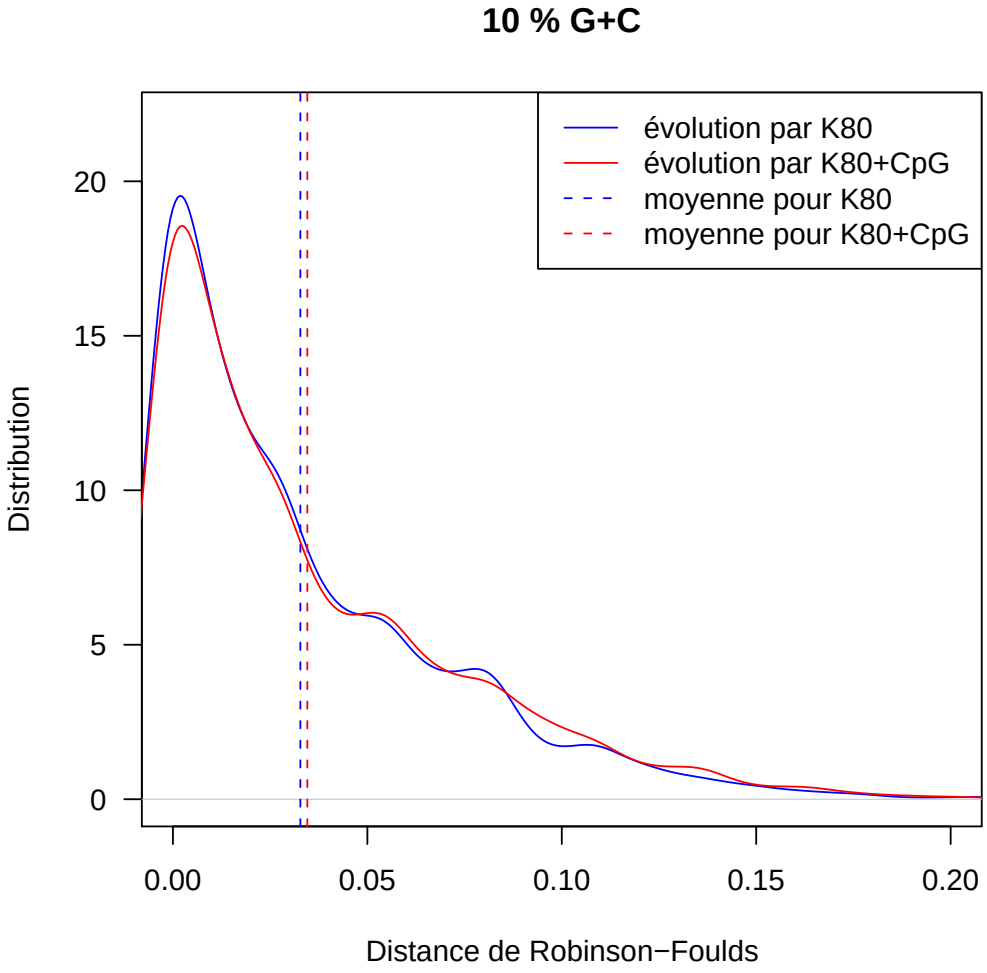


Figure 3.4 – (a) Distribution empirique, pour 100 arbres, de la distance de Robinson-Foulds entre l'arbre vrai et l'arbre reconstruit sur un alignement après simulation de l'évolution – avec dépendance entre sites – d'une séquence ancestrale à contenu en G+C de 10%. La moyenne de chaque distribution est représentée par une droite en tirets verticale.

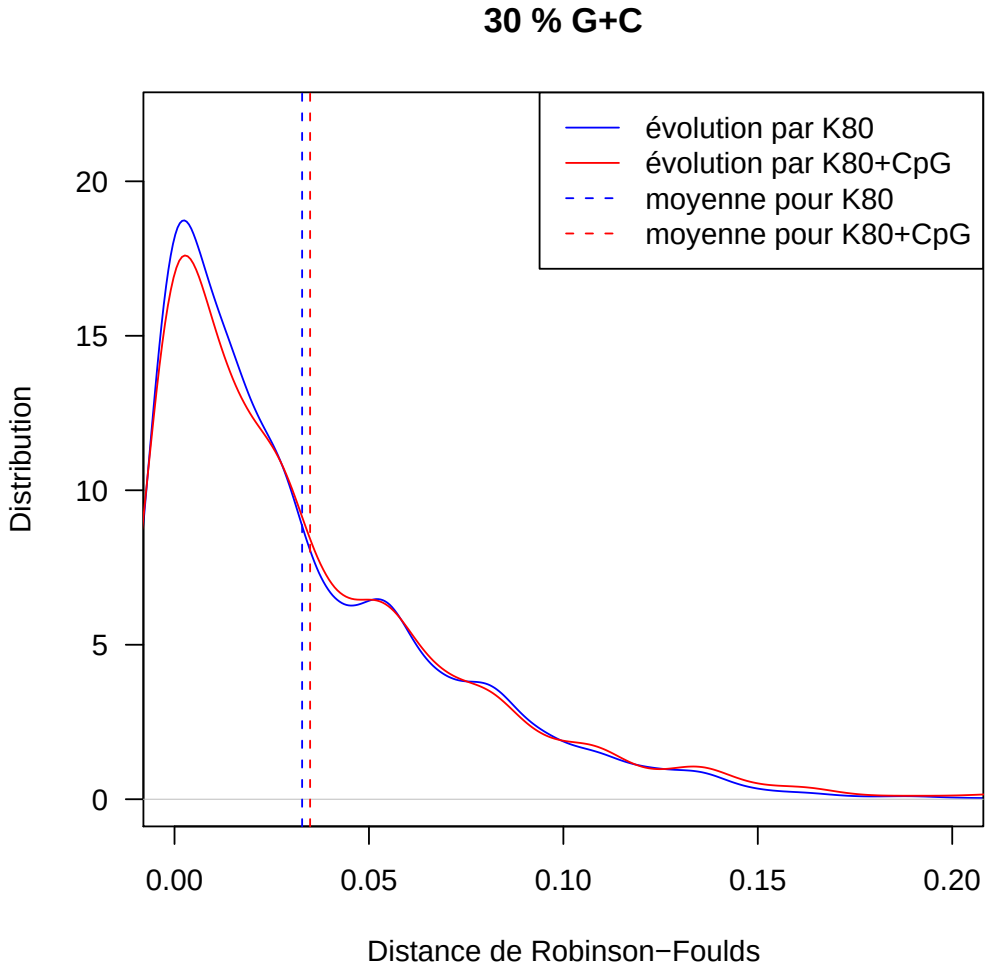


Figure 3.4 – (b) Distribution empirique, pour 100 arbres, de la distance de Robinson-Foulds entre l'arbre vrai et l'arbre reconstruit sur un alignement après simulation de l'évolution – avec dépendance entre sites – d'une séquence ancestrale à contenu en G+C de 30%. La moyenne de chaque distribution est représentée par une droite en tirets verticale.

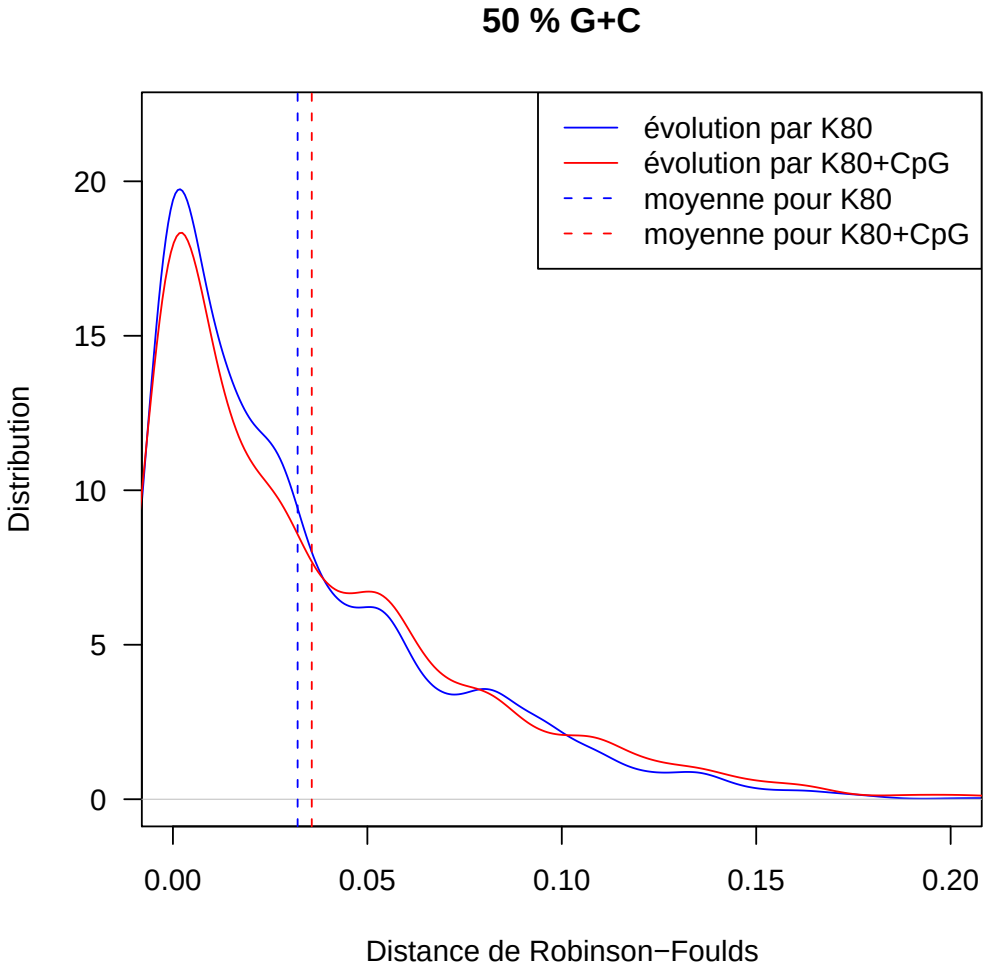


Figure 3.4 – (c) Distribution empirique, pour 100 arbres, de la distance de Robinson-Foulds entre l'arbre vrai et l'arbre reconstruit sur un alignement après simulation de l'évolution – avec dépendance entre sites – d'une séquence ancestrale à contenu en G+C de 50%. La moyenne de chaque distribution est représentée par une droite en tirets verticale.

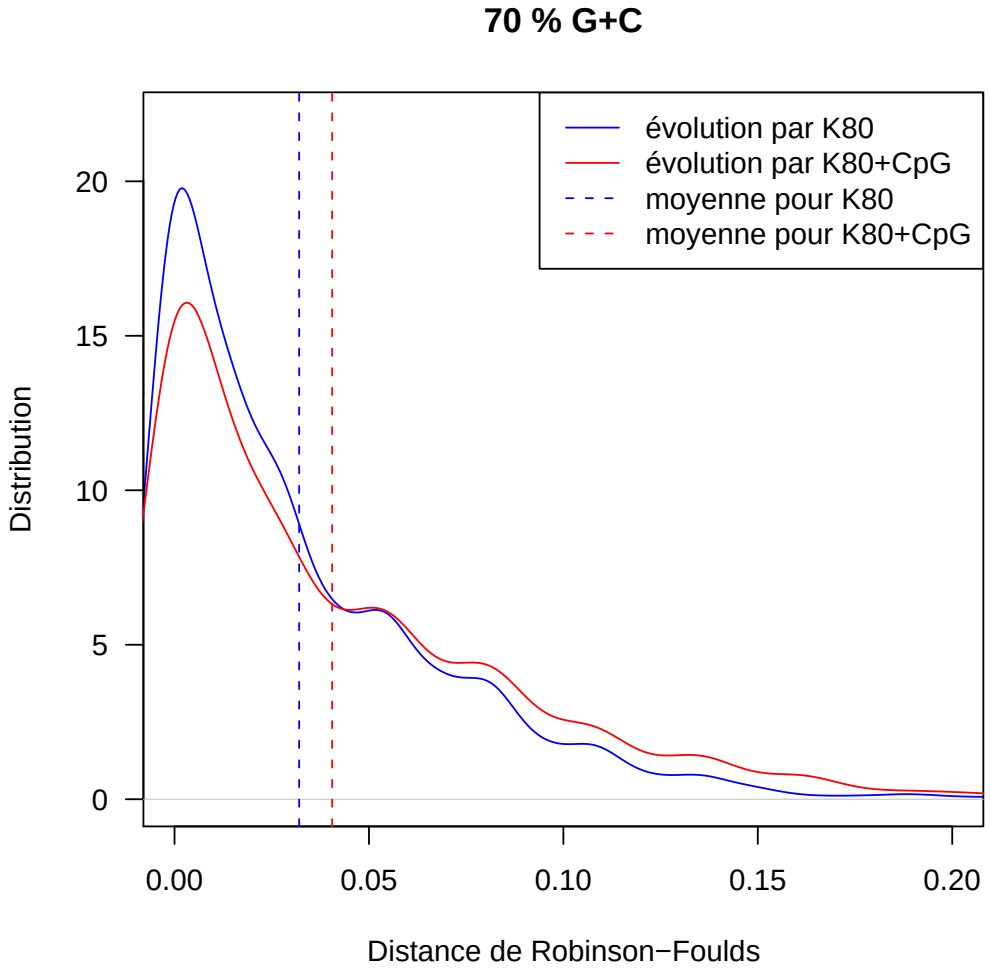


Figure 3.4 – (d) Distribution empirique, pour 100 arbres, de la distance de Robinson-Foulds entre l'arbre vrai et l'arbre reconstruit sur un alignement après simulation de l'évolution – avec dépendance entre sites – d'une séquence ancestrale à contenu en G+C de 70%. La moyenne de chaque distribution est représentée par une droite en tirets verticale.

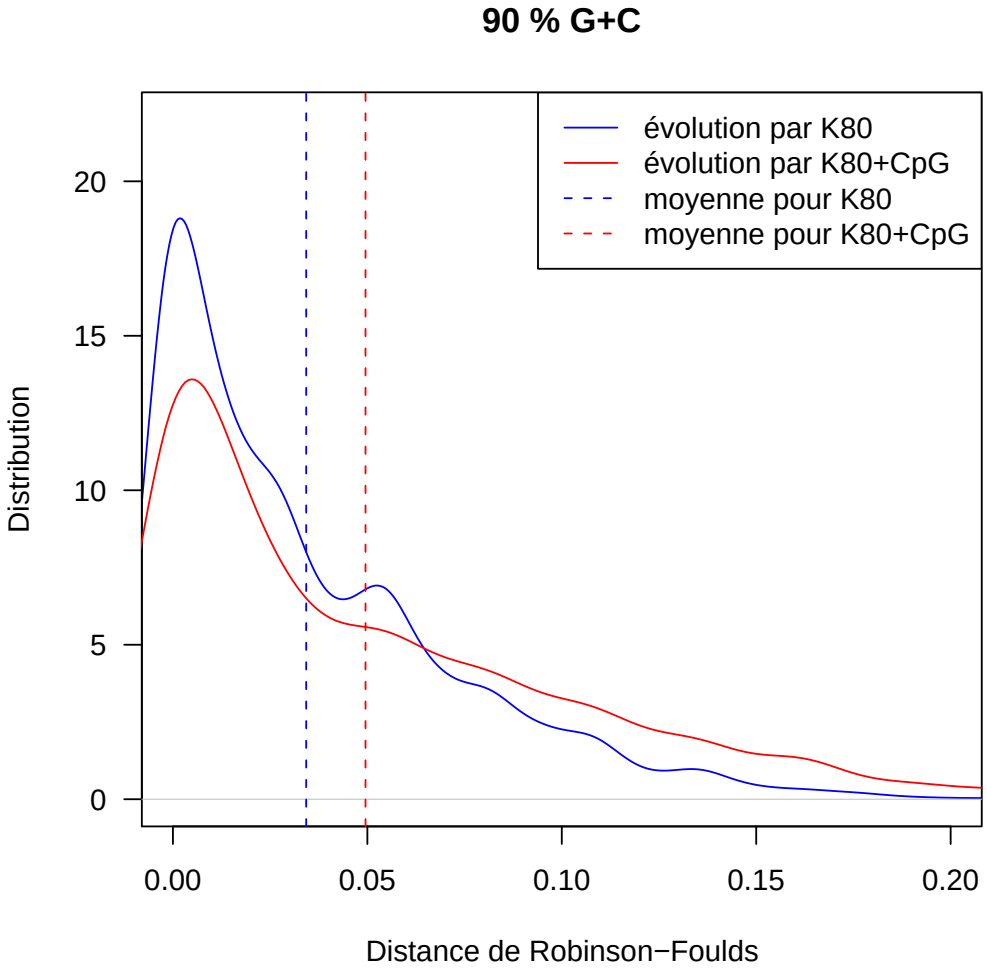


Figure 3.4 – (e) Distribution empirique, pour 100 arbres, de la distance de Robinson-Foulds entre l'arbre vrai et l'arbre reconstruit sur un alignement après simulation de l'évolution – avec dépendance entre sites – d'une séquence ancestrale à contenu en G+C de 90%. La moyenne de chaque distribution est représentée par une droite en tirets verticale.

Chapitre 4

Modèles avec dépendances entre sites voisins

Nous venons de voir que l'utilisation de l'hypothèse d'indépendance entre sites mène à des biais dans l'estimation des distances évolutives et dans l'inférence d'arbres phylogénétiques. Il est donc nécessaire de développer des méthodes et des outils permettant de prendre en compte cette caractéristique biologique dans l'analyse de l'évolution des séquences.

Je présenterai ici un **cadre général de modélisation incorporant la dépendance entre sites directement adjacents**. Je présenterai ensuite deux approches pour l'étude de ce modèle : une **approche par simulations de Monte-Carlo**, puis une **approche analytique**. Finalement, j'utiliserai les résultats obtenus par ces deux approches pour l'**analyse du chromosome 21 d'*Homo sapiens***.

4.1 Modèles avec dépendances entre sites

4.1.1 Présentation du cadre général

Je présente ici un cadre général de modélisation dérivé de la modélisation markovienne usuelle et des travaux de Duret & Galtier (2000), Arndt *et al.* (2003), Arndt (2007) et Bérard *et al.* (2005) sur l'incorporation de dépendances entre sites voisins.

Ainsi, on peut modéliser l'évolution d'une séquence biologique par la **combinaison d'une matrice de taux de substitution simple Q et de l'ensemble des substitutions faisant intervenir des dinucléotides**. On interdira par la suite les substitutions doubles, qui entraînent la substitution simultanée de deux nucléotides adjacents, bien que cette hypothèse ne soit pas nécessaire à la description du modèle général. La matrice $\mathbf{Q}$ peut donc être définie comme dans le cadre de la modélisation avec indépendance entre sites :

$$\mathbf{Q} = \begin{array}{c} \\ \texttt{A} \\ \texttt{T} \\ \texttt{C} \\ \texttt{G} \end{array} \begin{array}{c} \begin{array}{cccc} \texttt{A} & \texttt{T} & \texttt{C} & \texttt{G} \end{array} \\ \begin{pmatrix} - & a & b & c \\ d & - & e & f \\ g & h & - & i \\ j & k & l & - \end{pmatrix} \end{array}$$

On ajoute à cette matrice la description des substitutions liées au voisin direct à gauche ou au voisin direct à droite. Ainsi, en notant, par exemple, $\mathbf{V}$ la matrice carrée 16×16 décrivant les substitutions liées aux dinucléotides, on peut écrire la dynamique du modèle dans ce cas.

Prenons l'exemple d'un modèle combinant une matrice $\mathbf{Q}$ et une matrice $\mathbf{V}$ dont tous les coefficients sont nuls, sauf pour les taux de substitution entraînant la substitution de TpG par CpG (r_1) et la substitution de CpG par TpG (r_2). Je présente ici l'écriture de la dynamique de la fréquence en nucléotide C, l'ensemble de la dynamique du modèle s'écrivant de la même manière :

$$\begin{aligned} f_C(t+dt) \;=\; & f_C(t) + \mathbf{Q}_{AC} f_A(t)dt + \mathbf{Q}_{TC} f_T(t)dt + \mathbf{Q}_{GC} f_G(t)dt \\ & -(\mathbf{Q}_{CA} + \mathbf{Q}_{CT} + \mathbf{Q}_{CG}) f_C(t)dt \\ & +r_1 f_{TpG}(t)dt - r_2 f_{CpG}(t)dt \end{aligned}$$

4.1.2 Le problème du cône de dépendance

Alors que le cadre de la modélisation markovienne usuelle (1) garantit l'existence d'une distribution stationnaire et permet son écriture, (2) permet l'écriture d'une distance évolutive entre deux séquences, (3) permet l'écriture de la vraisemblance d'une séquence sous un scénario évolutif donné, notre nouveau cadre de modélisation entraîne des développements beaucoup plus complexes de ce type d'écritures. L'ensemble des problèmes soulevés par cette nouvelle modélisation relève principalement de ce qu'on peut appeler le **cône de dépendance**, que j'explicite ici dans le cas du calcul de la distribution stationnaire, mais qui peut être facilement généralisé.

Tout d'abord, alors que la modélisation markovienne usuelle garantit l'existence d'une distribution stationnaire, la modélisation incorporant des dépendances entre sites adjacents ne la garantit pas. Toutefois, si on suppose que cette distribution existe, alors elle doit s'écrire comme le point fixe de la dynamique des fréquences en nucléotides, dinucléotides, trinucléotides, ...

Pour ce qui est du **cône de dépendance**, alors que l'écriture de la dynamique sous le modèle markovien simple ne fait intervenir que les fréquences en nucléotides, que se passe-t-il dans le cas dans notre nouveau système ? Comme nous l'avons vu, l'écriture de la dynamique des fréquences en nucléotides fait intervenir les fréquences en dinucléotides au pas de temps précédent. En effet, l'écriture de la dynamique de la fréquence en C fait intervenir les fréquences en

TpG et CpG. On voit ainsi très bien que l'écriture de la dynamique des fréquences en dinucléotides fera intervenir des fréquences en trinucléotides au pas de temps précédent, et ainsi de suite. En remontant dans le passé, il s'agit de connaître tout un cône de dépendance, qui rend cette écriture impossible dans le cas général, et inversement, une lettre dans la séquence ancestrale a potentiellement une influence sur toutes les autres positions.

Je présenterai donc par la suite deux développements possibles, qui permettent de contourner ce problème. Je détaillerai dans un premier temps une **approche par simulations dans le cas général**, puis une **approche exacte dans le cas d'un sous-modèle**.

4.2 Étude du modèle par simulations de Monte-Carlo

Si on **suppose que le modèle admet une distribution stationnaire**, alors la méthode heuristique la plus simple à développer pour l'estimation de cette distribution, lorsqu'aucun calcul analytique n'est possible, est l'approche par simulations. Ainsi, une manière de résoudre le cône de dépendance est d'effectuer des simulations du modèle jusqu'à obtention d'un état que l'on considère stationnaire.

L'article de Schöniger & von Haeseler (1995) présente deux algorithmes efficaces de simulation de l'évolution de séquences dans le cas des modèles classiques – qui supposent l'indépendance entre les sites d'une séquence –, et plusieurs implémentations ont depuis été développées pour la simulation de l'évolution de séquences sous ces modèles. L'outil `evolver` du paquet PAML (Yang, 1997) ainsi que l'application Seq-Gen (Rambaut & Grassly, 1997) font partie des premiers outils à avoir été développés. L'application Seq-Gen est notamment basée sur l'un des deux algorithmes proposés par Schöniger & von Haeseler (1995) et implémente par ailleurs de l'hétérogénéité de taux de substitution entre sites (Rambaut & Grassly, 1997). Des travaux ultérieurs ont permis des complexifications des modèles de simulation, et notamment l'incorporation de mécanismes d'insertion et déletion (Cartwright, 2005; Rosenberg, 2005; Cantarel *et al.*, 2006; Strope *et al.*, 2007).

Dans le cadre de notre question générale, nous noterons le développement récent de SISSI pour la simulation de l'évolution de séquences en incorporant de l'information sur les interactions entre sites d'une séquence (Gesell & von Haeseler, 2006). Toutefois, le modèle de simulation implémenté par Gesell & von Haeseler (2006) impose de connaître en détail les interactions présentes entre chacun des sites d'une séquence. En effet, il nécessite l'écriture, pour chaque position de la séquence, d'une matrice spécifique incorporant l'information concernant les sites qui interagissent avec celle-ci. Ceci rend le modèle très contraignant du point de

vue de l'utilisateur, car l'écriture de l'ensemble de ces matrices n'est généralement pas gérable.

J'ai donc décidé de développer une **application simple qui permette de simuler l'évolution de séquences en incorporant des dépendances entre sites directement voisins**. Ceci, de manière à pouvoir étudier le comportement des modèles incoporant une dépendance entre sites dont l'étude analytique est complexe.

4.2.1 Détails sur l'implémentation

De manière à présenter l'algorithme de simulation que j'ai choisi pour implémenter l'évolution de séquences selon le modèle général présenté ci-dessus, je vais d'abord vous présenter les deux algorithmes proposés par Schöniger & von Haeseler (1995) dans le cadre du modèle avec indépendance entre sites. Le but de cet outil de simulation étant de permettre de simuler l'évolution d'une séquence d'ADN donnée pendant un temps où la séquence subira en moyenne un nombre de substitutions par site donné. Ceci permettra par exemple, de faire évoluer une séquence ancestrale le long d'un arbre phylogénétique donné, jusqu'à obtenir les séquences aux feuilles de l'arbre.

De manière plus formelle, il s'agit de simuler l'évolution d'une séquence ancestrale S pendant un temps t où la séquence subira d substitutions par site et deviendra la séquence S'. Ceci sous un modèle d'évolution donné.

a. Cas des modèles classiques

Schöniger & von Haeseler (1995) proposent deux algorithmes de simulation – que je détaille dans les encarts suivants – selon que l'on utilise la matrice des probabilités de substitution $\mathbf{P}(t)$ ou la matrice des taux de substitution $\mathbf{Q}$.

Le premier algorithme se déduit directement de la matrice $\mathbf{P}(t)$, et suppose clairement que les sites sont indépendants les uns des autres. Il est décrit dans la figure 4.1-a ci-dessous.

Données : Une séquence initiale S et un temps d'évolution t
Résultat : Une séquence S' ayant évolué pendant un temps t à partir de la séquence S ancestrale
pour chaque *position k de la séquence S* **faire**
  le nucléotide i occupant la position k est remplacé par j avec la probabilité $\mathbf{P}_{ij}(t)$;
fin

Figure 4.1 – (a) Algorithme d'évolution de séquence sous un modèle classique où tous les sites sont indépendants entre eux. Approche avec la matrice $\mathbf{P}(t)$ des probabilités de substitution.

Le deuxième algorithme, s'écrit au travers de la matrice $\mathbf{Q}$, et Schöniger & von Haeseler (1995) proposent de normaliser la matrice $\mathbf{Q}$ par le maximum M des sommes M_i des taux de substitution pouvant affecter chaque nucléotide :

$$M_i = \sum_{j \neq i} \mathbf{Q}_{i,j}$$

J'ai décidé, dans la présentation de cet algorithme, et pour plus de cohérence par rapport à ce qui va suivre, de ne pas effectuer cette normalisation de la matrice $\mathbf{Q}$. Ceci a pour conséquence de faire intervenir la valeur M, non pas en amont, mais directement dans la simulation. On montre facilement que ces deux approches sont équivalentes.

Ainsi, on s'intéressera dans la simulation à l'intervalle $\mathcal{M} =]0, M]$, composé des sous-intervalles disjoints, spécifiques de chaque nucléotide j vers lequel peut évoluer le nucléotide i. De manière plus formelle, les sous-intervalles sont définis comme suit (où j_1, j_2, j_3 sont différents de i) :

- Ω_{i,j_1} représente l'intervalle $]0, \mathbf{Q}_{i,j_1}]$
- Ω_{i,j_2} représente l'intervalle $]\mathbf{Q}_{i,j_1}, \mathbf{Q}_{i,j_1} + \mathbf{Q}_{i,j_2}]$
- Ω_{i,j_3} représente l'intervalle $]\mathbf{Q}_{i,j_1} + \mathbf{Q}_{i,j_2}, \mathbf{Q}_{i,j_1} + \mathbf{Q}_{i,j_2} + \mathbf{Q}_{i,j_3}]$
- $\Omega_{i,i}$ représente l'intervalle restant $]\mathbf{Q}_{i,j_1} + \mathbf{Q}_{i,j_2} + \mathbf{Q}_{i,j_3}, M]$

On définit le nombre de substitutions D ayant lieu sur une séquence donnée pendant un temps t, comme le produit de d, le nombre de substitutions par site et de n, la longueur de la séquence. Voici donc une version légèrement modifiée de l'algorithme de Schöniger & von Haeseler (1995) :

Données : Une séquence initiale S et un nombre de substitutions D
Résultat : Une séquence ayant subi D substitutions à partir de la séquence S ancestrale
tant que *le nombre D de substitutions n'est pas exactement atteint* **faire**
  on tire une position k (occupée par un nucléotide noté i) uniformément sur la séquence et on lui attribue une valeur p uniformément tirée entre 0 et M;
  si *p appartient à* $\mathcal{Q}_{i,j_1}$ **alors**
    i est remplacé par j_1 où $j_1 \neq i$;
  sinon si *p appartient à* $\mathcal{Q}_{i,j_2}$ **alors**
    i est remplacé par j_2 où $j_2 \neq i$;
  sinon si *p appartient à* $\mathcal{Q}_{i,j_3}$ **alors**
    i est remplacé par j_3 où $j_3 \neq i$;
  sinon
    i n'est pas remplacé;
  fin
fin

Figure 4.1 – (b) Algorithme d'évolution de séquence sous un modèle classique où tous les sites sont indépendants entre eux. Approche avec la matrice **Q** des taux de substitution.

b. Cas du modèle avec dépendances entre sites

L'approche faisant intervenir la matrice $\mathbf{P}(t)$ n'est plus possible si on veut prendre en compte des dépendances entre sites, car elle implique que les sites soient tous indépendants entre eux, pour pouvoir traiter l'évolution de tous les sites en une seule procédure. Par contre, l'approche faisant intervenir la matrice $\mathbf{Q}$ peut être étendue aux cas où on incorpore de la dépendance entre sites voisins. Il s'agit donc de simuler l'évolution de séquences sous le modèle présenté en début de chapitre. Je ne m'intéresserai donc ici qu'au cas où les dépendances se limitent au voisin direct à gauche, ou direct à droite d'un site donné. Par ailleurs, j'interdis les substitutions doubles, c'est-à-dire les substitutions de deux nucléotides voisins lors d'un seul événement de substitution. Ainsi, chaque substitution n'affecte qu'un seul nucléotide.

Certaines substitutions sont des substitutions simples, dues uniquement à l'identité du nucléotide i en une position, certaines sont influencées par l'identité du nucléotide à gauche de cette position, certaines sont influencées par l'identité du nucléotide à droite de cette position. Nous avons défini la matrice $\mathbf{Q}$ comme la matrice des taux de substitutions simples. À celle-ci, viennent s'ajouter les taux de substitutions dues aux voisins directs à gauche et à droite, que l'on peut décrire par les matrices suivantes :

- $\mathbf{G}^i$ représente la matrice des taux de substitution du nucléotide i lorsque la substitution est influencée par le nucléotide à gauche de celui-ci, et qui contient les taux $\mathbf{G}^i_{g,j}$ de substitution du nucléotide i par le nucléotide $j \neq i$ lorsque i possède le nucléotide g à sa gauche.

- $\mathbf{D}^i$ représente la matrice des taux de substitution du nucléotide i lorsque la substitution est influencée par le nucléotide à droite de celui-ci, et qui contient les taux $\mathbf{D}^i_{d,j}$ de substitution du nucléotide i par le nucléotide $j \neq i$ lorsque i possède le nucléotide d à sa droite.

Contrairement à l'algorithme proposé par Arndt *et al.* (2003), qui ne décrit pas l'évolution rigoureusement telle que le modèle la définit, l'algorithme suivant est proposé par Jean Bérard (*comm. pers.*) et s'inspire de l'approche présentée dans la figure 4.1-b pour simuler l'évolution de séquences avec dépendance entre sites. Tout d'abord, il s'agit de calculer le taux maximal M parmi les taux de substitution maximal susceptibles d'affecter un site occupé par un i :

$$M_i = \sum_{j \neq i} (\mathbf{Q}_{i,j} + \sum_g \mathbf{G}^i_{g,j} + \sum_d \mathbf{D}^i_{d,j})$$

Par la suite, on s'intéressera à l'intervalle $\mathcal{M} :]0, M]$, composé des sous-intervalles disjoints, spécifiques de chaque nucléotide i : Ω_i pour les substitutions

simples, $\mathcal{G}_i$ pour les substitutions dues au voisin à gauche, $\mathcal{D}_i$ pour les substitutions dues au voisin à droite, et $\mathcal{Z}_i$ pour les non-substitutions éventuelles.

De manière plus formelle, les sous-intervalles sont définis comme suit :

- $\mathcal{Q}_i$ représente l'intervalle $]0, \sum_{j\neq i} \mathbf{Q}_{i,j}]$
- $\mathcal{G}_i$ représente l'intervalle $]\sum_{j\neq i} \mathbf{Q}_{i,j}, \sum_{j\neq i} \mathbf{Q}_{i,j} + \sum_{j\neq i}\sum_g \mathbf{G}^i_{g,j}]$
- $\mathcal{D}_i$ représente l'intervalle $]\sum_{j\neq i} \mathbf{Q}_{i,j}+\sum_{j\neq i}\sum_g \mathbf{G}^i_{g,j}, \sum_{j\neq i} \mathbf{Q}_{i,j}+\sum_{j\neq i}\sum_g \mathbf{G}^i_{g,j}+\sum_{j\neq i}\sum_d \mathbf{D}^i_{d,j}]$
- $\mathcal{Z}_i$ représente l'intervalle restant $]\sum_{j\neq i} \mathbf{Q}_{i,j}+\sum_{j\neq i}\sum_g \mathbf{G}^i_{g,j}+\sum_{j\neq i}\sum_d \mathbf{D}^i_{d,j}, M]$

Chacun de ces sous-intervalles étant construit de manière analogue au cas avec indépendance entre sites, de manière à inclure l'ensemble des taux de substitutions de chaque catégorie d'événement.

Je décris l'algorithme de manière succinte, car il se déduit facilement de l'algorithme 4.1.

Données : Une séquence initiale S et un nombre de substitutions D
Résultat : Une séquence ayant subi D substitutions à partir de la séquence S ancestrale
tant que *le nombre D de substitutions n'est pas exactement atteint* **faire**
  on tire une position k (occupée par un nucléotide noté i) uniformément sur la séquence et on lui attribue une valeur p uniformément tirée entre 0 et M;
  si *p appartient à $\mathcal{Q}_i$* **alors**
    i est remplacé par $j \neq i$ selon la procédure classique;
  sinon si *p appartient à $\mathcal{G}_i$* **alors**
    suivant *le sous-intervalle dans lequel se situe p* **faire**
      i est remplacé par $j \neq i$;
    fin
  sinon si *p appartient à $\mathcal{D}_i$* **alors**
    suivant *le sous-intervalle dans lequel se situe p* **faire**
      i est remplacé par $j \neq i$;
    fin
  sinon
    i n'est pas remplacé;
  fin
fin

Figure 4.2 – Algorithme d'évolution de séquence sous un modèle où tous les sites ne sont pas indépendants entre eux. Approche avec la matrice **Q** des taux de substitution

4.2.2 Estimation de la distribution stationnaire

Grâce à l'implémentation de cet algorithme de simulation d'évolution de séquence, on peut simuler l'évolution d'une séquence jusqu'à obtention d'un état stationnaire et en **déduire empiriquement la distribution à l'équilibre du modèle**. Ceci, autant pour les fréquences en nucléotides que pour les fréquences en dinucléotides ou en mots de longueur plus grande.

Ainsi, en anticipation de la section 4.4, où l'on montrera qu'une sous-classe du modèle général avec dépendance entre sites permet une écriture analytique des distributions stationnaires, voici ce que des simulations de Monte-Carlo permettent d'obtenir sur ces modèles. À titre d'exemple, voici donc les résultats que l'on obtient sur la simulation d'une séquence linéaire sous deux de ces modèles – que j'appelle K80+CpG et T92+CpG et qui seront développés en détails dans la section 4.4.

La séquence initiale est une séquence de 50kb ne possédant que des A, et les fréquences en nucléotides et en dinucléotides au cours du temps d'évolution sont données par la figure 4.3. On note bien que ces fréquences convergent vers les fréquences stationnaires dont nous possédons l'écriture analytique (tracées en rouge sur la figure 4.3), et vers des fréquences stationnaires dont on peut déduire la valeur à partir de la simulation lorsque nous ne possédons pas cette écriture. Quelle que soit la séquence initiale utilisée pour les simulations, nous avons constaté que cet équilibre est toujours atteint, et que ces valeurs sont indépendantes de la séquence initiale.

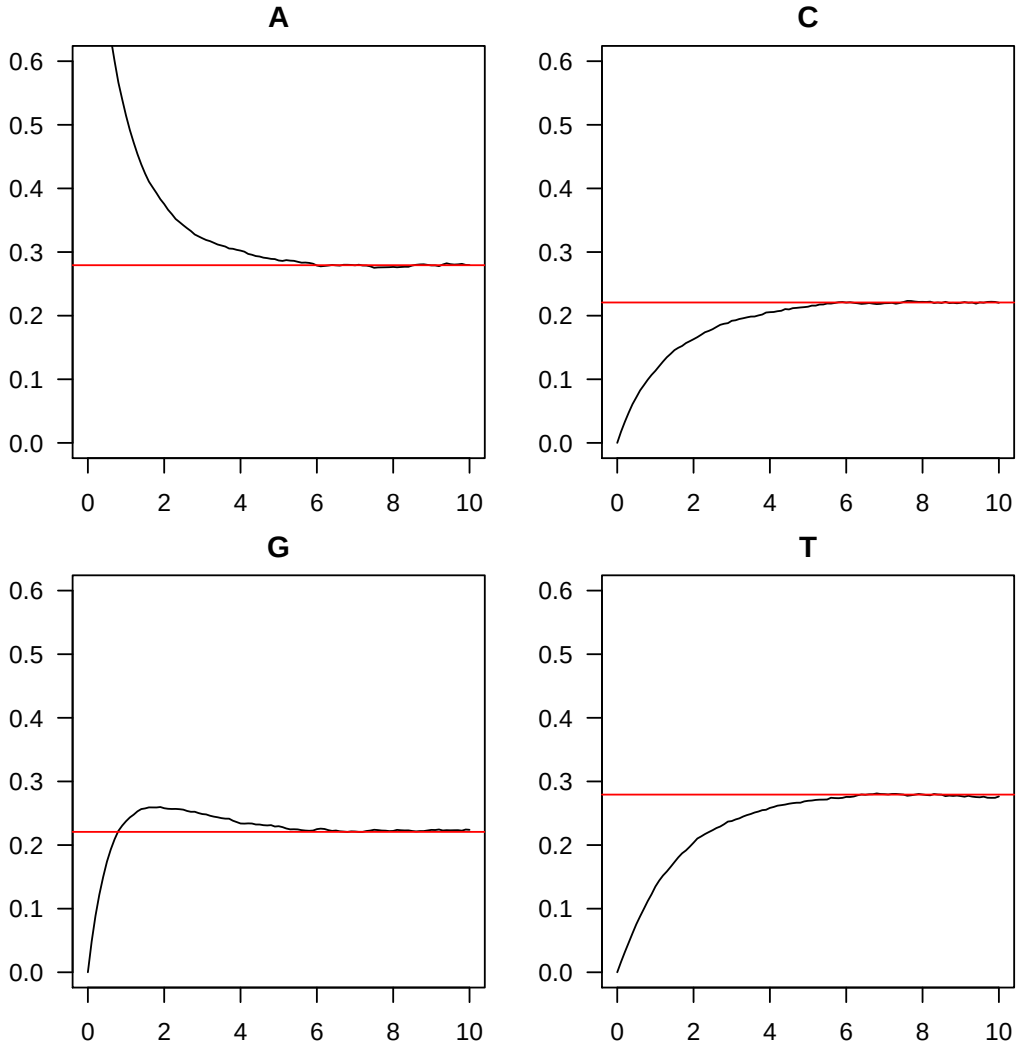


Figure 4.3 – (a) Mesure des fréquences en nucléotides (axe des ordonnées) sur une séquence de 50kb, initialement ne contenant que des A, au fur et à mesure de son évolution sous le modèle de K80+CpG ($\alpha = 3$, $\beta = 1$, $r = 10$), jusqu'à ce qu'elle ait subi, en moyenne, 10 substitutions par site (axe des abscisses). Les fréquences à l'équilibre, connues analytiquement, sont indiquées par une droite rouge.

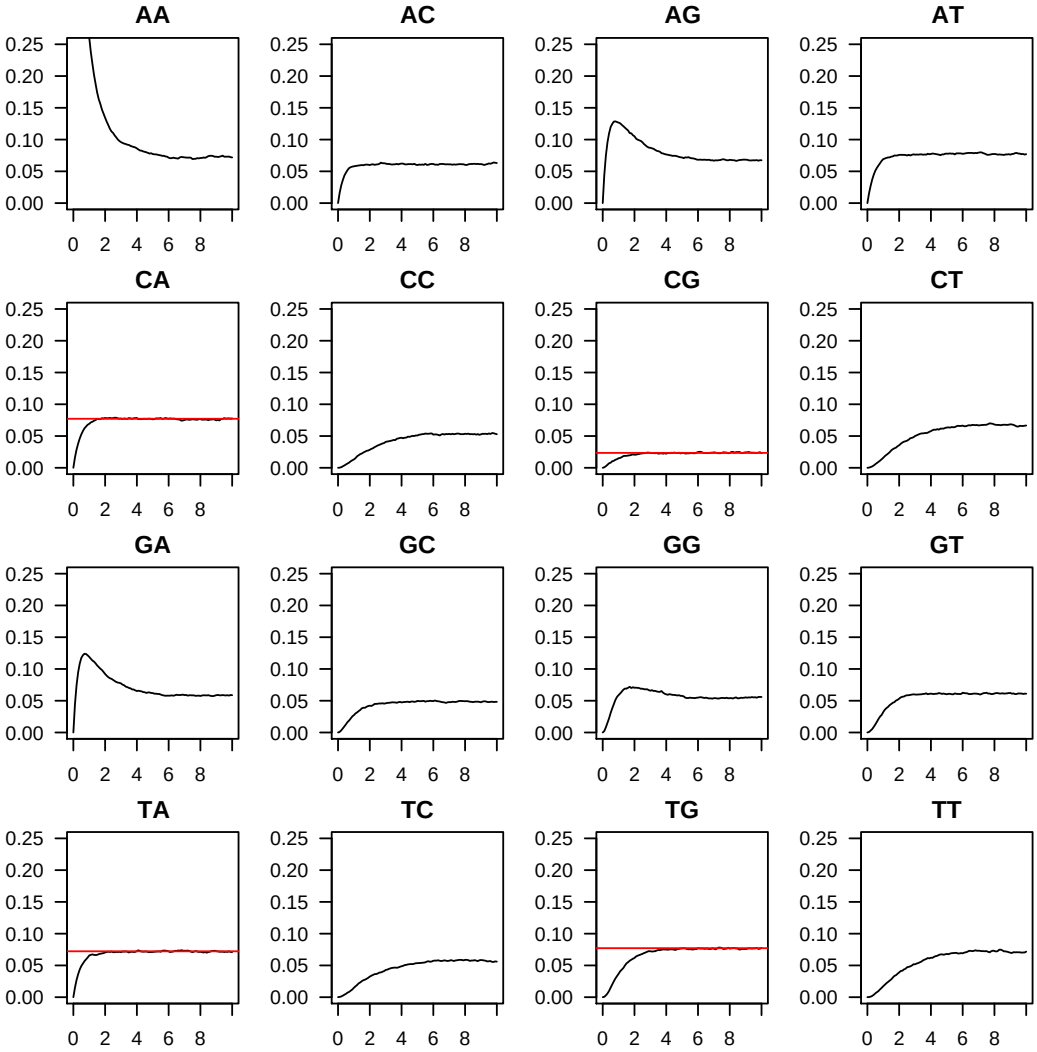


Figure 4.3 – (b) Mesure des fréquences en dinucléotides (axe des ordonnées) sur une séquence de 50kb, initialement ne contenant que des A, au fur et à mesure de son évolution sous le modèle de K80+CpG ($\alpha = 3$, $\beta = 1$, $r = 10$), jusqu'à ce qu'elle ait subi, en moyenne, 10 substitutions par site (axe des abscisses). Les fréquences à l'équilibre, lorsqu'elles sont connues analytiquement, sont indiquées par une droite rouge.

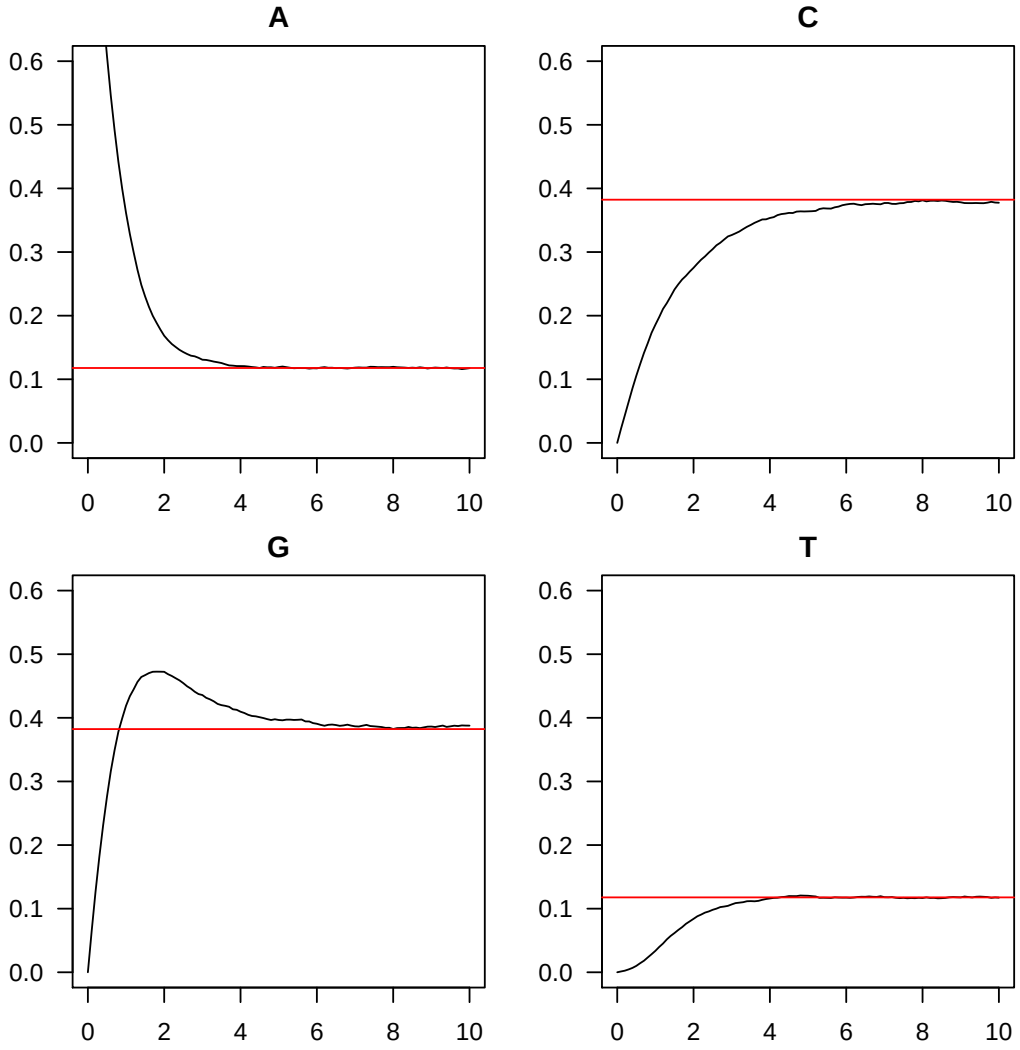


Figure 4.3 – (c) Mesure des fréquences en nucléotides (axe des ordonnées) sur une séquence de 50kb, initialement ne contenant que des A, au fur et à mesure de son évolution sous le modèle de T92+CpG ($\alpha = 3$, $\beta = 1$, $\theta = 0,95$, $r = 10$), jusqu'à ce qu'elle ait subi, en moyenne, 10 substitutions par site (axe des abscisses). Les fréquences à l'équilibre, connues analytiquement, sont indiquées par une droite rouge.

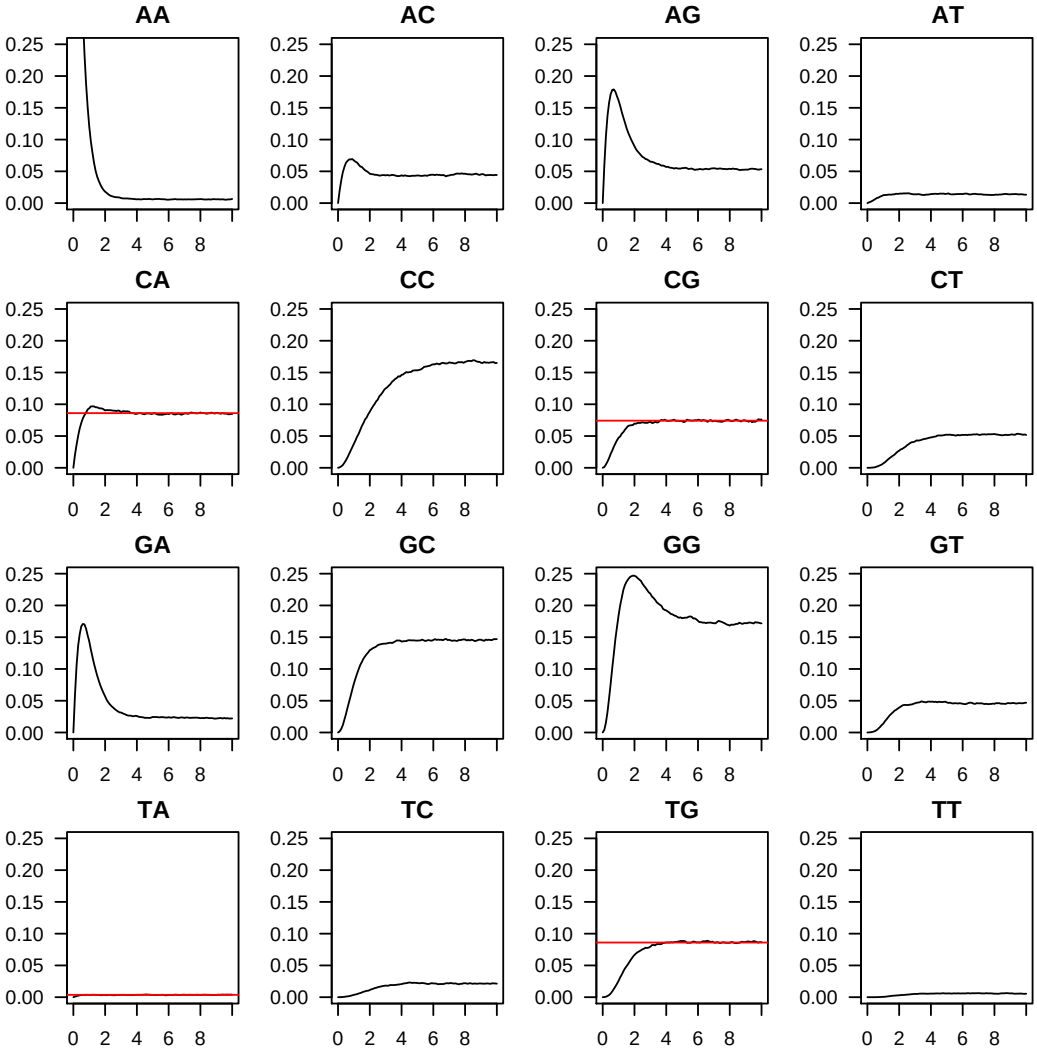


Figure 4.3 – (d) Mesure des fréquences en dinucléotides (axe des ordonnées) sur une séquence de 50kb, initialement ne contenant que des A, au fur et à mesure de son évolution sous le modèle de T92+CpG ($\alpha = 3$, $\beta = 1$, $\theta = 0,95$, $r = 10$), jusqu'à ce qu'elle ait subi, en moyenne, 10 substitutions par site (axe des abscisses). Les fréquences à l'équilibre, lorsqu'elles sont connues analytiquement, sont indiquées par une droite rouge.

4.2.3 Étude du comportement à l'approche de l'équilibre

Cette implémentation permet aussi d'étudier le comportement du modèle à l'approche de l'équilibre. Je me place de nouveau dans le cas des sous-modèles que je présenterai dans la section suivante 4.4. Ainsi, nous verrons un peu plus loin que, dans le cadre de ces modèles, on peut toujours écrire – à l'équilibre – $4\pi_{CG} + 4\pi_{TA} + 8\pi_{CA} = 1$. En posant $\Delta = 1 - (4\pi_{CG} + 4\pi_{TA} + 8\pi_{CA})$, observons le comportement du modèle T92+CpG à l'approche de l'équilibre.

Pour cela, simulons l'évolution d'une séquence initiale de taille variable, ne possédant que des A, et mesurons Δ au cours du temps d'évolution. Les résultats sont présentés sur la figure 4.4, où la longueur de la séquence est indiquée en abscisse et Δ en ordonnée. J'ai ici pris f_{CG}, f_{TA} et $(f_{CA} + f_{TG})/2$ comme estimateurs respectifs de π_{CG}, π_{TA} et π_{CA}, et chaque courbe correspond à un temps d'évolution donné.

Cette figure montre que, quelle que soit la taille de la séquence, Δ est constant et fonction de la distance évolutive qui sépare la séquence de l'équilibre du modèle. Par ailleurs, on note que plus la taille de la séquence augmente, moins la mesure de Δ est variable. J'ai pour cette raison choisi des fenêtres de 50kb pour l'analyse du chromosome 21 d'*Homo sapiens* que je vous présenterai en dernière section, de manière à minimiser les variations dues à la précision des mesures.

La figure 4.5 montre de surcroît que Δ décroit rapidement vers 0 au cours du temps évolutif. Cette deuxième figure permet de déterminer comment Δ peut servir de mesure de la distance d'une séquence donnée à l'équilibre du modèle, à partir des seules fréquences en dinucléotides mesurée sur une séquence.

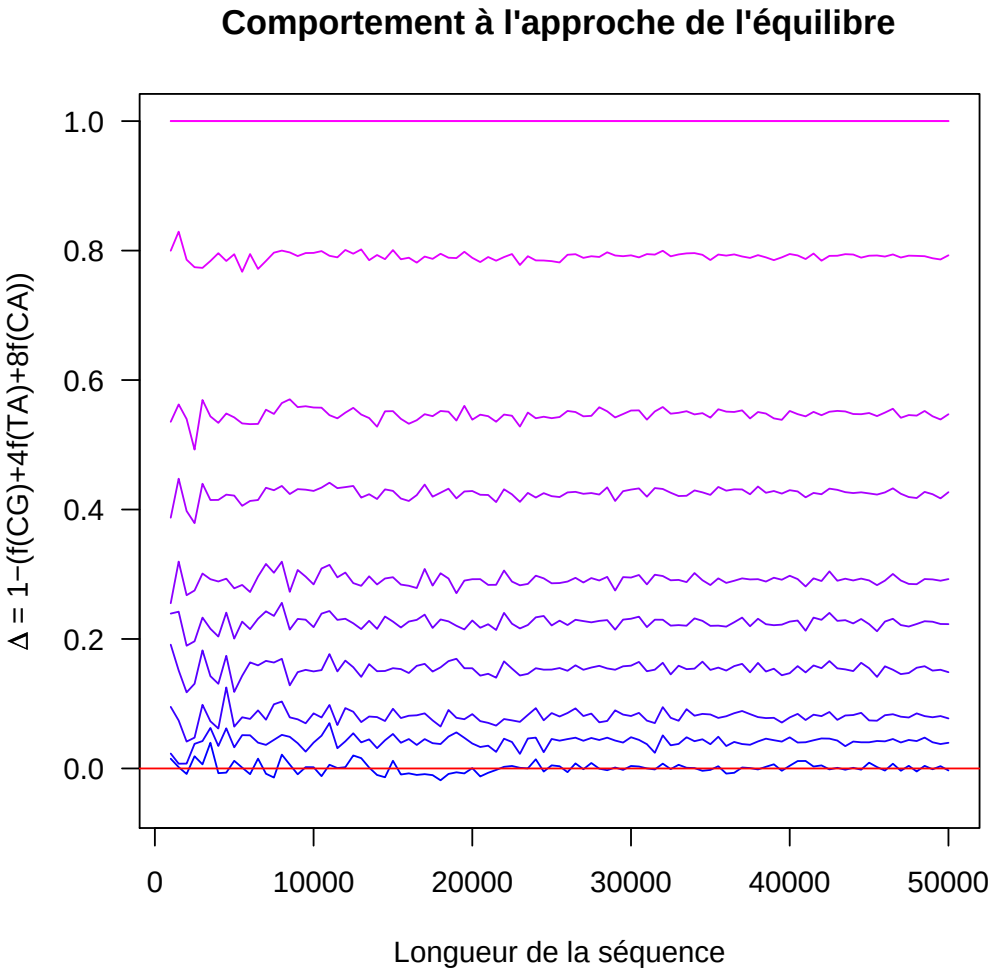


Figure 4.4 – $\Delta = 1 - (4\pi_{CG} + 4\pi_{TA} + 8\pi_{CA})$ mesuré sur des séquences de différentes longueurs et à une distance évolutive moyenne donnée de la séquence initiale. L'évolution est simulée, sous le modèle T92+CpG, sur une séquence initiale de longueur donnée et ne possédant que des A. De haut en bas, les courbes représentent la valeur de Δ sur : (1) la séquence initiale, (2) la séquence après 0,25 substitutions par site, (3) la séquence après 0,5 substitutions par site, (4) la séquence après 0,75 substitutions par site, (5) la séquence après 1 substitution par site, (6) la séquence après 1,25 substitutions par site, (7) la séquence après 1,5 substitutions par site, (8) la séquence après 2 substitutions par site, (9) la séquence après 2,5 substitutions par site, (10) la séquence après 5 substitutions par site. La droite tracée en rouge à 0 correspond à la valeur de Δ à l'équilibre du modèle T92+CpG.

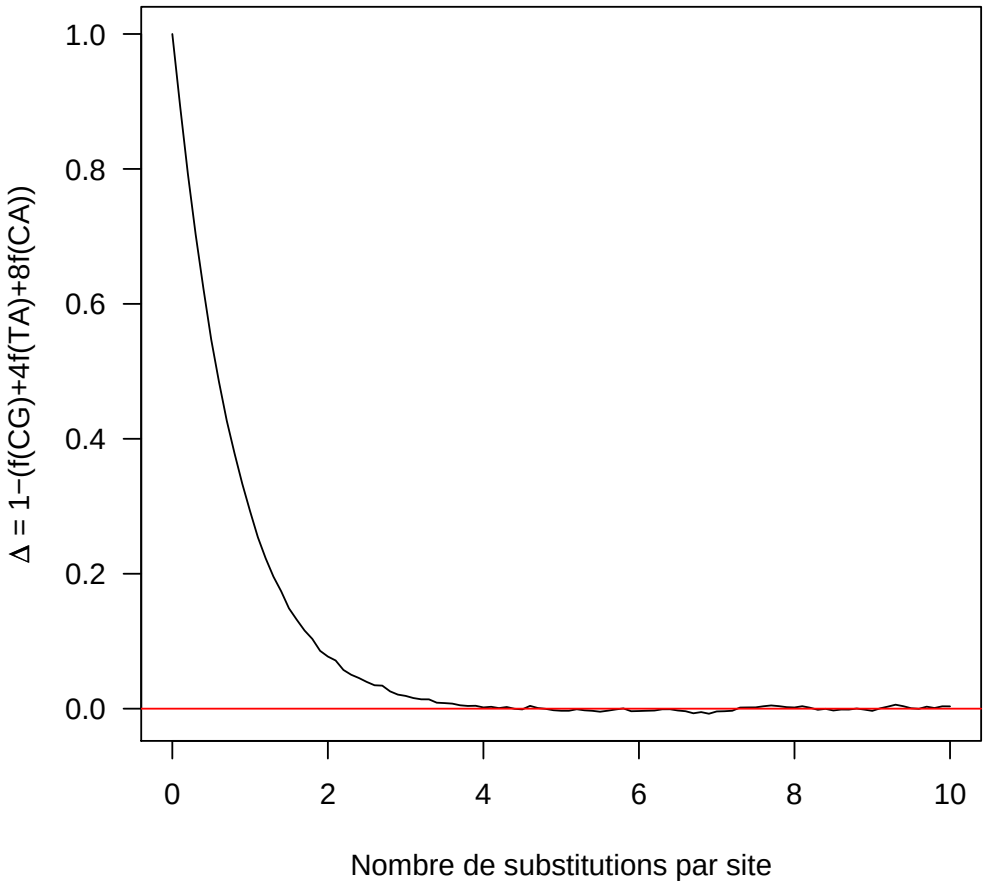


Figure 4.5 – $\Delta = 1-(4\pi_{CG}+4\pi_{TA}+8\pi_{CA})$ mesuré sur une séquence de 50kb et à une distance évolutive moyenne (axe des abscisses) donnée de la séquence initiale. L'évolution est simulée, sous le modèle T92+CpG, sur une séquence initiale de 50kb ne possédant que des A. La droite tracée en rouge à 0 correspond à la valeur de Δ à l'équilibre du modèle T92+CpG.

4.3 Étude du modèle par l'approximation du K-cluster

L'étude du modèle par simulations de Monte-Carlo possède les limitations de toute approche par simulations, et de nombreux travaux se sont donc penchés sur la question du développement d'écritures analytiques permettant l'étude exacte du modèle. C'est dans cette optique qu'a été introduite l'**approximation du K-cluster**, qui consiste à écrire les fréquences des trinucléotides en fonction des fréquences des dinucléotides qui les constituent. L'utilisation de cette approximation par Duret & Galtier (2000) pour l'estimation des fréquences en trinucléotides est, à ma connaissance, la première utilisation de cet estimateur pour la résolution du cône de dépendance. On écrit ainsi :

$$\hat{f}_{XYZ} = \frac{f_{XY} f_{YZ}}{f_Y}$$

où $\hat{f}_{XYZ}$ représente un estimateur de la fréquence du trinucléotide XYZ, f_{XY} et f_{YZ} les fréquences des dinucléotides XY et YZ, et f_Y la fréquence en nucléotide Y.

Cette approximation, introduite par (Duret & Galtier, 2000), tronque le cône de dépendance et a permis à Arndt *et al.* (2003) de poser un modèle avec dépendance entre sites voisins et d'écrire, grâce à cette approximation, les fréquences des dinucléotides XpY à l'équilibre comme un système de 16 équations quadratiques pouvant être résolues analytiquement. Le groupe de Arndt, a par la suite développé de nombreuses applications basées sur cette approximation – estimation de la distribution stationnaire, estimation des taux de substitution, écriture d'une vraisemblance approchée – que nous ne détaillerons pas ici (Arndt *et al.*, 2003; Arndt & Hwa, 2005; Arndt, 2007).

Toutefois, nous allons voir dans la section suivante, qu'il existe un cas particulier du modèle général avec dépendances entre sites, pour lequel le cône de dépendance est naturellement brisé par la nature-même des processus de substitution, et qui permet un développement analytique exact dans le cadre, notamment, des substitutions liées aux dinucléotides CpG.

4.4 Étude analytique du modèle

Cette section se base sur l'ensemble des résultats théoriques obtenus par Bérard *et al.* (2005). Pour de plus amples détails, en particulier pour l'ensemble des preuves, se référer à ce travail.

Bérard *et al.* (2005) s'intéressent à l'évolution d'une séquence biologique dans un cas particulier du modèle général présenté en début de chapitre. Ce cas particulier, qui englobe de nombreux modèles classiques tels que les modèles de Jukes

& Cantor (1969), de Kimura (1980), de Tamura (1992), mais encore de Hasegawa *et al.* (1985) et de Tamura & Nei (1993), et incorpore certaines substitutions faisant intervenir des dinucléotides (dont l'effet des CpG) entraîne une élimination naturelle du cône de dépendance, sans la nécessité de recourir à une approximation.

- *Élimination du cône de dépendance*

L'**élimination naturelle du cône de dépendance** est due à deux conditions sur le modèle général, l'une concernant les substitutions simples, l'autre concernant les substitutions influencées par le voisinage.

La condition sur les substitutions influencées par le voisinage est telle que l'on ne considère que les substitutions du type YpR→YpR, où Y représente les pyrimidines (C et T) et R représente les purines (A et T). Ces processus ne font intervenir que des transitions, et modifient les nucléotides tout en les maintenant dans la même classe chimique (pyrimidine ou purine).

Ensuite, la condition sur les substitutions simples, appelée aussi condition R/Y par Bérard *et al.* (2005) – qui correspond à la matrice de substitutions simples décrite par Rzhetsky & Nei (1995) – suppose un traitement identique des deux pyrimidines (respectivement des deux purines) lors des transversions qu'elles peuvent subir.

La matrice $\mathbf{Q}$ du modèle R/Y s'écrit donc comme suit :

$$\mathbf{Q} = \begin{array}{c} \\ \texttt{A} \\ \texttt{T} \\ \texttt{C} \\ \texttt{G} \end{array} \begin{array}{c} \begin{array}{cccc} \texttt{A} & \texttt{T} & \texttt{C} & \texttt{G} \end{array} \\ \begin{pmatrix} - & \beta_T & \beta_C & \alpha_G \\ \beta_A & - & \alpha_C & \beta_G \\ \beta_A & \alpha_T & - & \beta_G \\ \alpha_A & \beta_T & \beta_C & - \end{pmatrix} \end{array}$$

Bérard *et al.* (2005) modélisent alors l'évolution d'une séquence par la **combinaison d'une matrice des taux de substitution simple (de propriété R/Y) et de l'ensemble des substitutions faisant intervenir des dinucléotides YpR**. L'élimination du cône de dépendance entraîne alors l'écriture de la dynamique du modèle sous la forme d'un système fini d'équations qui converge vers un état stationnaire que l'on peut écrire analytiquement.

- *Propriétés générales du modèle*

De ce modèle, Bérard *et al.* (2005) prouvent plusieurs propriétés. Tout d'abord (1) que la fréquence exacte à l'équilibre de tous les mots peut être écrite à travers la résolution d'un système linéaire de taille finie, et (2) que la dynamique converge vers cet équilibre unique. Plus particulièrement, (3) les fréquences à l'équilibre des dinucléotides YpR sont solution d'un système linéaire 4×4, et (4) les fréquences à l'équilibre en nucléotides peuvent être écrits comme des fonctions affines des

fréquences à l'équilibre en dinucléotides YpR. Finalement, (5) les fréquences en pyrimidines et en purines s'écrivent toujours de la manière suivante :

$$f_T + f_C = \frac{\beta_T + \beta_C}{\beta_A + \beta_T + \beta_C + \beta_G}$$
$$f_A + f_G = \frac{\beta_A + \beta_G}{\beta_A + \beta_T + \beta_C + \beta_G}$$

4.4.1 Calcul de la distribution stationnaire

- *Cas général symétrique*

Bérard *et al.* (2005) donnent l'écriture explicite de la solution de ce système pour le cas particulier où la matrice R/Y est **symétrique**[1]. Étant donné que les coefficients diagonaux sont libres, une matrice symétrique de type R/Y peut alors s'écrire de la manière suivante :

$$\mathbf{Q} = \begin{array}{c} \\ \texttt{A} \\ \texttt{T} \\ \texttt{C} \\ \texttt{G} \end{array} \begin{array}{c} \begin{array}{cccc} \texttt{A} & \texttt{T} & \texttt{C} & \texttt{G} \end{array} \\ \begin{pmatrix} - & \beta_w & \beta_s & \alpha_s \\ \beta_w & - & \alpha_s & \beta_s \\ \beta_w & \alpha_w & - & \beta_s \\ \alpha_w & \beta_w & \beta_s & - \end{pmatrix} \end{array}$$

Puisqu'on considère que le modèle est symétrique, on considérera aussi qu'il y a symétrie des substitutions faisant intervenir des dinucléotides. On considérera donc que les taux de substitution de CpG vers CpA et vers TpG (inverse complémentaire de CpA) sont égaux et que les taux de substitution de TpA vers CpA et TpG sont, eux aussi, égaux.

On note donc que r_w représente le taux de substitution de CpG|CpG vers TpG|CpA, et que r_s représente le taux de substitution de TpA|TpA vers CpA|TpG.

Bérard *et al.* (2005) montrent alors que les fréquences à l'équilibre des dinucléotides YpR sont solution du système linéaire suivant.

$$\mathbf{M} \times \begin{pmatrix} \pi_{CG} \\ \pi_{CA} \\ \pi_{TA} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \beta_s(\alpha_s + \beta_s) \\ \beta_s(\alpha_w + \beta_w) + \beta_w(\alpha_s + \beta_s) \\ \beta_w(\alpha_w + \beta_w) \end{pmatrix}$$

Où $\mathbf{M}$ est la matrice suivante, en notant :

$$\alpha^* = \alpha_w + \alpha_s$$
$$\beta^* = \beta_w + \beta_s$$
$$\sigma^* = \alpha^* + \beta^*$$

[1] dans son sens biologique, *i.e.* le taux de substitution d'une base X vers une base Y est égal au taux de substitution de la base complémentaire X_c vers la base complémentaire Y_c

$$\begin{pmatrix} \sigma^*(\beta^* + \alpha_w + \beta_s) + r_w(\sigma^* + \beta_s) & \sigma^*(\beta_s - \alpha_s) & -\beta_s r_s \\ \sigma^*(\beta_w - \alpha_w) - r_w(2\beta_s + \alpha^*) & \sigma^*(\alpha^* + 3\beta^*) & \sigma^*(\beta_s - \alpha_s) - r_s(2\beta_w + \alpha^*) \\ -\beta_w r_w & \sigma^*(\beta_w - \alpha_w) & \sigma^*(\beta^* + \alpha_s + \beta_w) + r_s(\sigma^* + \beta_w) \end{pmatrix}$$

Dans le cas où la matrice R/Y n'est pas symétrique, les fréquences à l'équilibre des dinucléotides YpR sont aussi solution d'un système linéaire, voir Bérard *et al.* (2005).

La résolution de ce système linéaire donne l'écriture des fréquences à l'équilibre des dinucléotides de type YpR, que l'on peut écrire sous la forme :

$$\pi_{YpR} = \frac{D(YpR)}{4D}$$

▷ $D = D_0 + r_s D_s + r_w D_w + r_s r_w D_{sw}$

$$\begin{cases} D_0 = (\alpha^* + 3\beta^*)(\alpha^* + \beta^*)^2 \\ D_s = (\alpha^* + \beta^*)(\alpha^* + 3\beta^* + \alpha_w) + \beta^*(\beta_w + \alpha_w) \\ D_w = (\alpha^* + \beta^*)(\alpha^* + 3\beta^* + \alpha_s) + \beta^*(\beta_s + \alpha_s) \\ D_{sw} = 2(\alpha^* + 2\beta^*) \end{cases}$$

et où les $D(YpR)$ se construisent comme suit :

▷ $D(CG) = D_0(CG) + r_s D_s(CG)$

$$\begin{cases} D_0(CG) = (\alpha^* + 3\beta^*)(\alpha_s + \beta_s)^2 \\ D_s(CG) = (\alpha^* + \beta^*)\alpha_s + (\alpha_s + \beta_s)(\beta^* + 2\beta_s) \end{cases}$$

▷ $D(CA) = D(TG) = D_0(CA) + r_s D_s(CA) + r_w D_w(CA) + r_s r_w D_{sw}(CA)$

$$\begin{cases} D_0(CA) = (\alpha^* + 3\beta^*)(\beta_s + \alpha_s)(\beta_w + \alpha_w) \\ D_s(CA) = (\beta_w + \alpha_w)(\alpha^* + 2\beta^* + \beta_s) \\ D_w(CA) = (\beta_s + \alpha_s)(\alpha^* + 2\beta^* + \beta_w) \\ D_{sw}(CA) = \alpha^* + 2\beta^* \end{cases}$$

▷ $D(TA) = D_0(TA) + r_w D_w(TA)$

$$\begin{cases} D_0(TA) = (\alpha^* + 3\beta^*)(\beta_w + \alpha_w)^2 \\ D_w(TA) = (\alpha^* + \beta^*)\alpha_w + (\beta_w + \alpha_w)(\beta^* + 2\beta_w) \end{cases}$$

On en déduit alors les fréquences stationnaires en nucléotides, par :

$$\pi_C = \pi_G = \frac{\beta_s + \alpha_s}{2(\alpha^* + \beta^*)} - \frac{r_w \pi_{CG}}{\alpha^* + \beta^*} + \frac{r_s \pi_{TA}}{\alpha^* + \beta^*}$$

$$\pi_T = \pi_A = \frac{1 - \pi_C + \pi_G}{2}$$

Dans tous les cas, on montre que $\pi_{CG} + \pi_{TA} + 2\pi_{CA} = \frac{1}{4}$.

- *Modèle de Kimura (1980) + CpG : K80+CpG*

Le système est une complexification du modèle de Kimura (1980) présenté précédemment (noté par la suite K80+CpG), qui incorpore une différence entre les taux de transition[2] et les taux de transversion[3]. En reprenant la matrice présentée pour le modèle K80 et en lui rajoutant des contraintes sur les coefficients diagonaux, on obtient :

$$\mathbf{Q} = \begin{array}{c} \\ \texttt{A} \\ \texttt{T} \\ \texttt{C} \\ \texttt{G} \end{array} \begin{array}{c} \begin{array}{cccc} \texttt{A} & \texttt{T} & \texttt{C} & \texttt{G} \end{array} \\ \begin{pmatrix} - & \beta & \beta & \alpha \\ \beta & - & \alpha & \beta \\ \beta & \alpha & - & \beta \\ \alpha & \beta & \beta & - \end{pmatrix} \end{array}$$

où α représente le taux de transition et β le taux de transversion. Je me place dans le cas de la seule modélisation des substitutions de voisinage liées aux dinucléotides CpG, où l'on peut donc écrire $r_s = 0$ et noter que r représente le taux de substitution r_w de CpG|CpG vers TpG|CpA. dans le modèle K80+CpG.

On écrit alors les fréquences stationnaires des dinucléotides YpR de la manière suivante, si $4(3\beta + \alpha)(\alpha + \beta) + r(3\alpha + 7\beta) \neq 0$:

$$\pi_{CG} = \frac{(3\beta + \alpha)(\alpha + \beta)}{16(3\beta + \alpha)(\alpha + \beta) + 4r(3\alpha + 7\beta)}$$

$$\pi_{TG} = \pi_{CA} = \frac{2(3\beta + \alpha)(\alpha + \beta) + r(5\beta + 2\alpha)}{32(3\beta + \alpha)(\alpha + \beta) + 8r(3\alpha + 7\beta)}$$

[2]substitution d'une pyrimidine par une pyrimidine ou d'une purine par une purine

[3]substitution d'une pyrimidine par une purine ou inversement

$$\pi_{TA} = \frac{(3\beta+\alpha)(\alpha+\beta)+r(2\beta+\alpha)}{16(3\beta+\alpha)(\alpha+\beta)+4r(3\alpha+7\beta)}$$

et on peut écrire celle des nucléotides comme suit, si $\alpha+\beta \neq 0$:

$$\pi_C = \pi_G = \frac{1}{4} - r\frac{\pi_{CG}}{2(\alpha+\beta)}$$

$$\pi_A = \pi_T = \frac{1}{4} + r\frac{\pi_{CG}}{2(\alpha+\beta)}$$

Ceci implique que, sous ce modèle, les fréquences à l'équilibre en C et en G sont toujours inférieures à 0,25 lorsque r est positif, et que les fréquences à l'équilibre en A et en T sont toujours supérieures à 0,25 lorsque r est positif.

On note ici, que non seulement $\pi_{CG}+\pi_{TA}+2\pi_{CA} = \frac{1}{4}$, qui est valable dans tous les modèles issus du modèle R/Y + YpR, mais que de sucroît, $16\pi_{CA}+8\pi_C = 3$.

- *Modèle de Tamura (1992) + CpG : T92+CpG*

En reprenant la matrice présentée pour le modèle T92 et en lui rajoutant des contraintes sur les coefficients diagonaux, on obtient :

$$\mathbf{Q} = \begin{array}{c} \\ \texttt{A} \\ \texttt{T} \\ \texttt{C} \\ \texttt{G} \end{array} \begin{array}{c} \begin{array}{cccc} \texttt{A} & \texttt{T} & \texttt{C} & \texttt{G} \end{array} \\ \begin{pmatrix} - & (1-\theta)\beta & \theta\beta & \theta\alpha \\ (1-\theta)\beta & - & \theta\alpha & \theta\beta \\ (1-\theta)\beta & (1-\theta)\alpha & - & \theta\beta \\ (1-\theta)\alpha & (1-\theta)\beta & \theta\beta & - \end{pmatrix} \end{array}$$

où α représente le taux de transition et β le taux de transversion. Je me place dans le cas de la seule modélisation des substitutions de voisinage liées aux dinucléotides CpG, où on peut écrire $r_s = 0$ et noter que r représente le taux de substitution r_w de CpG|CpG vers TpG|CpA dans le modèle T92+CpG.

On écrit alors les fréquences stationnaires des dinucléotides YpR de la manière suivante, si $4(3\beta+\alpha)(\alpha+\beta)+4r(3\beta+\alpha+\theta(\alpha+\beta)) \neq 0$:

$$\pi_{CG} = \frac{\theta^2(3\beta+\alpha)(\alpha+\beta)}{4(3\beta+\alpha)(\alpha+\beta)+4r(3\beta+\alpha+\theta(\alpha+\beta))}$$

$$\pi_{TG} = \pi_{CA} = \frac{\theta(1-\theta)(3\beta+\alpha)(\alpha+\beta)+r\theta(\alpha+(3-\theta)\beta)}{4(3\beta+\alpha)(\alpha+\beta)+4r(3\beta+\alpha+\theta(\alpha+\beta))}$$

$$\pi_{TA} = \frac{(1-\theta)^2(3\beta+\alpha)(\alpha+\beta)+r(3\beta+\alpha+\theta(\beta-\alpha+2(3-\theta)\beta)}{4(3\beta+\alpha)(\alpha+\beta)+4r(3\beta+\alpha+\theta(\alpha+\beta))}$$

et on peut écrire celle des nucléotides comme suit, si $\alpha+\beta \neq 0$:

$$\pi_C = \pi_G = \frac{\theta}{2} - r\frac{\pi_{CG}}{\alpha+\beta}$$

$$\pi_A = \pi_T = \frac{1-\theta}{2} + r\frac{\pi_{CG}}{\alpha+\beta}$$

Ceci implique que, sous le modèle T92+CpG, les fréquences à l'équilibre en C et en G sont toujours inférieures à celles sous le modèle T92 lorsque r est positif, et que les fréquences à l'équilibre en A et en T sont toujours supérieures à celles sous le modèle T92 lorsque r est positif.

On note ici, que non seulement $\pi_{CG} + \pi_{TA} + 2\pi_{CA} = \frac{1}{4}$, qui est valable dans tous les modèles issus du modèle général, mais que de sucroît, on a $4\pi_C + \pi_{CG} = 3$.

4.4.2 Calcul des paramètres du modèle

a. Modèle de Kimura (1980) + CpG : K80+CpG

On cherche à estimer les paramètres $r/(\alpha+\beta)$ et α/β, qui correspondent au taux relatif de substitutions CpG par rapport aux substitutions simples et au rapport entre taux de transition et taux de transversion. Ceci de manière à, sous l'hypothèse de stationnarité, pouvoir estimer les paramètres du modèle à partir des seules fréquences obtenues à partir de la séquence.

On peut écrire, en reprenant l'écriture des fréquences à l'équilibre des nucléotides, que :

$$\frac{r}{\alpha+\beta} = \frac{1-4\pi_C}{2\pi_{CG}} = \frac{4\pi_T-1}{2\pi_{CG}}$$

Comme nous l'avons vu précédemment, ceci implique que les fréquences observées sur les séquences sont compatibles avec l'équilibre du modèle, et que donc, pour que r soit positif, les fréquences en C et en G ne soient pas supérieures à 0,25 et les fréquences en T et en A ne soient pas inférieures à 0,25.

Il suffit ensuite d'écrire un rapport de deux combinaisons linéaires de certaines des fréquences en nucléotides et dinucléotides écrites précédemment, pour obtenir le ratio $\dfrac{\alpha}{\beta}$ par élimination de la plupart des facteurs. Par exemple, si $r \neq 0$:

$$\frac{\alpha}{\beta} = \frac{24\pi_{CG}+56\pi_{CA}-5}{2-8\pi_{CG}-24\pi_{CA}} = \frac{6\pi_{CG}+2\pi_{CA}-6\pi_C+1}{2\pi_{CA}-2\pi_{CG}+4\pi_C-1}$$

Ces écritures peuvent aussi être atteintes par la manipulation du système linéaire, en y rajoutant l'égalité $\pi_{CG}+\pi_{TA}+2\pi_{CA} = \dfrac{1}{4}$.

Les distributions stationnaires peuvent, par exemple, être estimées par les fréquences observées sur la séquence.

b. Modèle de Tamura (1992) + CpG : T92+CpG

On cherche à estimer les paramètres θ et $r/(\alpha+\beta)$, qui correspondent au contenu en G+C à l'équilibre, au taux relatif de substitutions CpG par rapport aux substitutions simples et au rapport entre taux de transition et taux de transversion. Ceci de manière à, sous l'hypothèse de stationnarité, pouvoir estimer les paramètres du modèle à partir des seules fréquences obtenues à partir de la séquence.

On peut reprendre les écritures des fréquences à l'équilibre, ou repartir du système linéaire en y rajoutant les égalités

$$\pi_{CG} + \pi_{TA} + 2\pi_{CA} = \frac{1}{4}$$

$$\pi_C = \frac{\theta}{2} - r\frac{\pi_{CG}}{\alpha+\beta}$$

pour obtenir plusieurs écritures du paramètre θ. Par exemple :

$$\begin{aligned}\theta &= 1 + 2(\pi_C - 2\pi_{CA} - 2\pi_{CG}) - 2\sqrt{(\pi_C - 2\pi_{CA} - 2\pi_{CG})^2 + \pi_{TA}} \\ &= 4\pi_{CA} + 4\pi_{TA} + 2\pi_C - \sqrt{(4\pi_{CA} + 4\pi_{TA} + 2\pi_C - 1)^2 + 4\pi_{TA}}\end{aligned}$$

On remplace alors θ dans les fréquences à l'équilibre des nucléotides pour obtenir, par exemple :

$$\begin{aligned}\frac{r}{\alpha+\beta} &= \frac{4(\pi_{CA}+\pi_{TA}) - \sqrt{(4\pi_{CA} + 4\pi_{TA} + 2\pi_C - 1)^2 + 4\pi_{TA}}}{2\pi_{CG}} \\ &= \frac{2(\pi_{CA}+\pi_{TA}) - \sqrt{(\pi_C - 2\pi_{CA} - 2\pi_{CG})^2 + \pi_{TA}}}{\pi_{CG}}\end{aligned}$$

Les distributions stationnaires peuvent, par exemple, être estimées par les fréquences observées sur la séquence.

4.5 Application à l'étude du chromosome 21

Nous avons vu dans le chapitre 2 l'effet de la méthylation sur la composition en bases de l'ADN, et nous y avons explicité les détails du mécanisme lié à cette modification des cytosines dans l'ADN. Dans la section précédente, nous avons développé un modèle permettant de prendre en compte les substitutions liées aux cytosines méthylées dans un contexte de dinucléotide CpG. Ce modèle amène, entre autres, à l'écriture d'une estimation du rapport $r/(\alpha+\beta)$ le taux de substitution. Le but de cette dernière section est donc d'utiliser les résultats obtenus sur le modèle T92+CpG pour analyser la méthylation sur le génome d'*Homo sapiens*. Ce modèle semble plus adapté que le modèle K80+CpG, car il

permet de prendre en compte une variation dans le contenu en G+C, variation qui est largement connue et étudiée sur ce génome (Bernardi, 1989; Versteeg *et al.*, 2003; Fryxell & Moon, 2005).

Pour cela, je me suis intéressée à l'analyse du chromosome 21, pour lequel nous connaissons la structuration en contenu en G+C, et la présence d'ilôts CpG (voir figure 1.1). L'analyse que j'ai effectuée a consisté en une analyse du chromosome 21, sur des fenêtres glissantes non-chevauchantes de 50kb, de la version 17 de l'assemblage du génome d'*Homo sapiens* effectuée par l'UCSC[4] (University of California Santa Cruz).

Les paramètres mesurés ont été déduits des formules présentées dans la section précédente, et $(f_C+f_G)/2$, f_{CG}, f_{TA} et $(f_{CA}+f_{TG})/2$ ont servi comme estimateurs respectifs de π_C, π_{CG}, π_{TA} et π_{CA}.

4.5.1 Contenu en G+C

Le modèle T92+CpG possède un paramètre θ qui décrit le contenu en G+C à l'équilibre du modèle (voir section b. page 94). Sa valeur peut s'écrire comme une combinaison linéaire des fréquences mesurées sur la séquence, et nous pouvons comparer cette valeur au contenu en G+C mesuré directement sur la séquence.

La figure 4.6 montre que l'estimation du paramètre θ (en noir) suit la mesure du contenu en G+C (en bleu) sur les fenêtres glissantes, tout en étant systématiquement plus élevée. La figure 4.7, montre l'excellente corrélation entre ces deux paramètres (coefficient de corrélation = 0.995), et indique que le paramètre θ semble un bon estimateur du contenu en G+C dans les séquences. Toutefois, on note un décalage entre le contenu en G+C mesuré et la valeur estimée du paramètre θ. Ce décalage peut-il être dû à un écart à l'équilibre du modèle ?

4.5.2 Distance à l'équilibre du modèle

De manière à déterminer si la différence observée entre la variation du contenu en G+C et la variation de l'estimation du paramètre θ est due au décalage entre les séquences actuelles et l'équilibre du modèle T92+CpG, nous allons comparer cette différence à la distance à l'équilibre du modèle.

Prenons Δ[5] comme mesure de la distance à l'équilibre du modèle, puisque nous avons montré que Δ décroît exponentiellement à l'approche de l'équilibre (voir figure 4.5), et comparons cette valeur à la différence entre le contenu en G+C et le paramètre θ.

La figure 4.8 semble montrer une certaine relation entre ces deux mesures. Cette relation, qui semblait faible est confirmée par la figure 4.9, qui semble montrer que $\theta - (G + C)$ augmente exponentiellement avec Δ. Or Δ décroît

[4]http ://genome.ucsc.edu/

[5]$\Delta = 1 - (4\pi_{CG} + 4\pi_{TA} + 8\pi_{CA})$

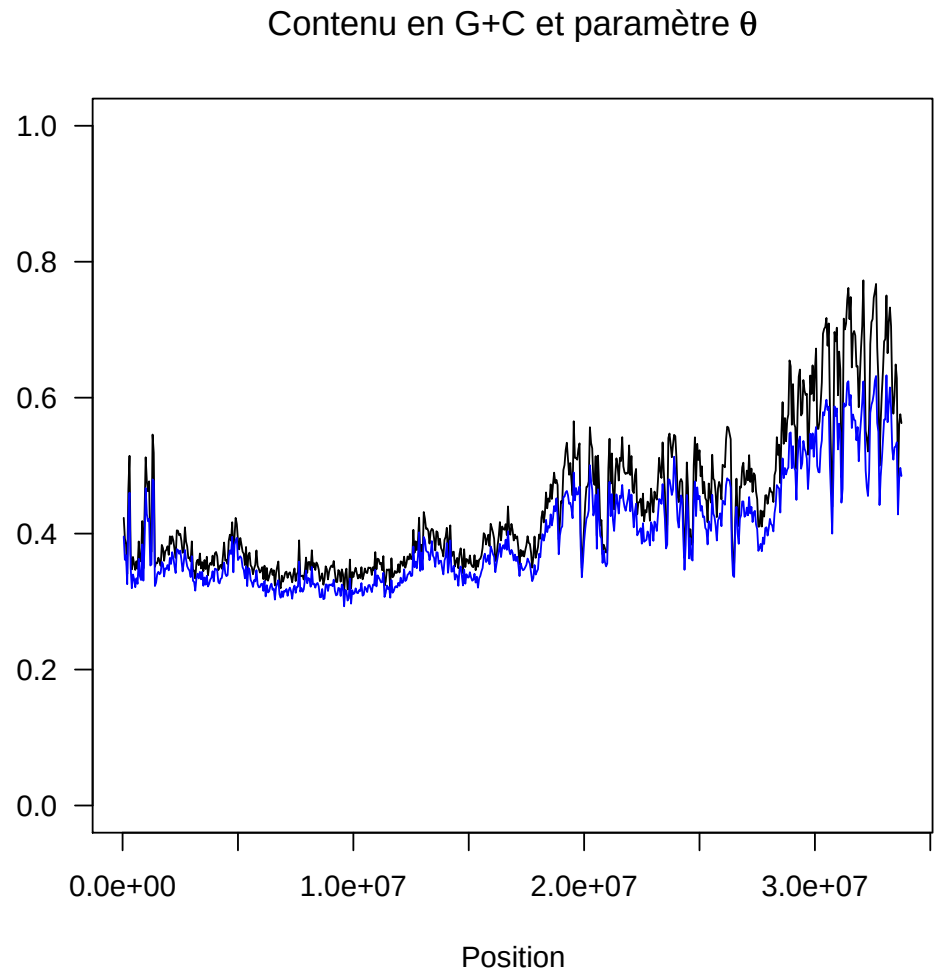


Figure 4.6 – Mesure du contenu en G+C (courbe bleue) et estimation du paramètre θ (courbe noire) sur des fenêtres glissantes de 50kb le long du chromosome 21.

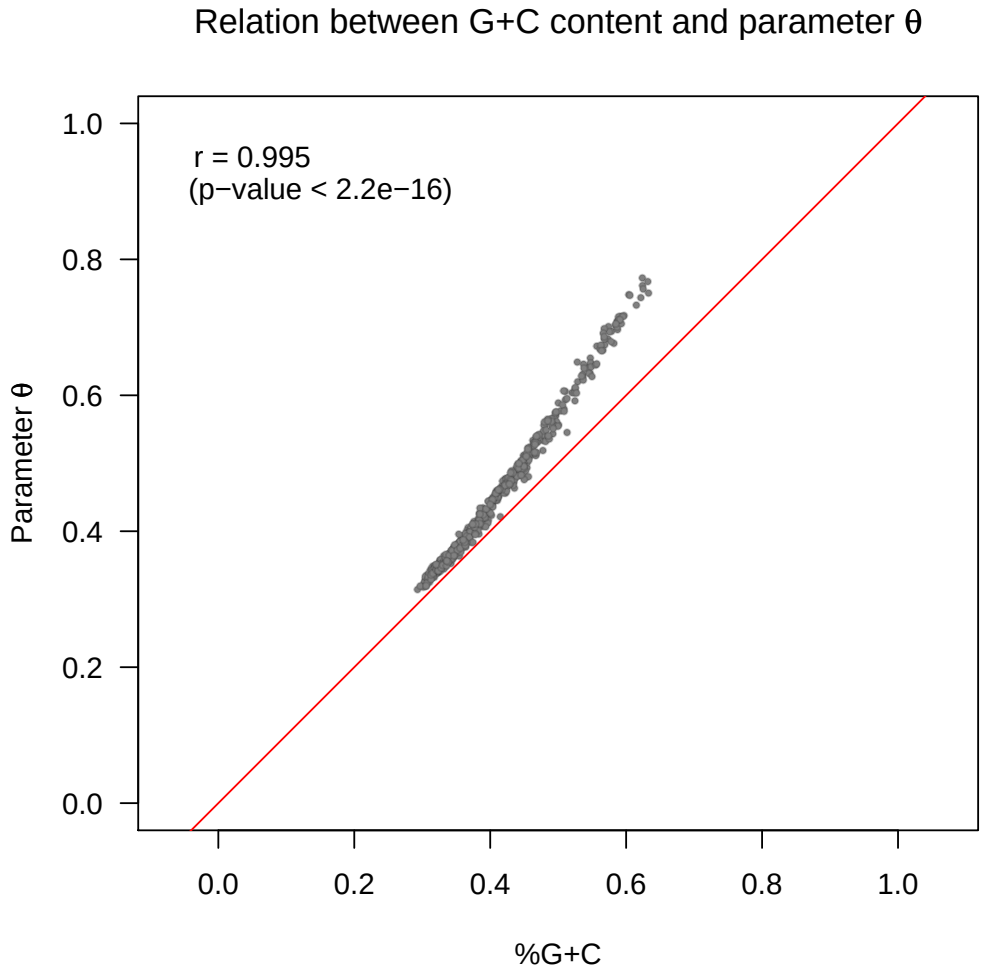


Figure 4.7 – Relation entre le contenu en G+C et l'estimation du paramètre θ sur des fenêtres de 50kb le long du chromosome 21. La première bissectrice est tracée en rouge.

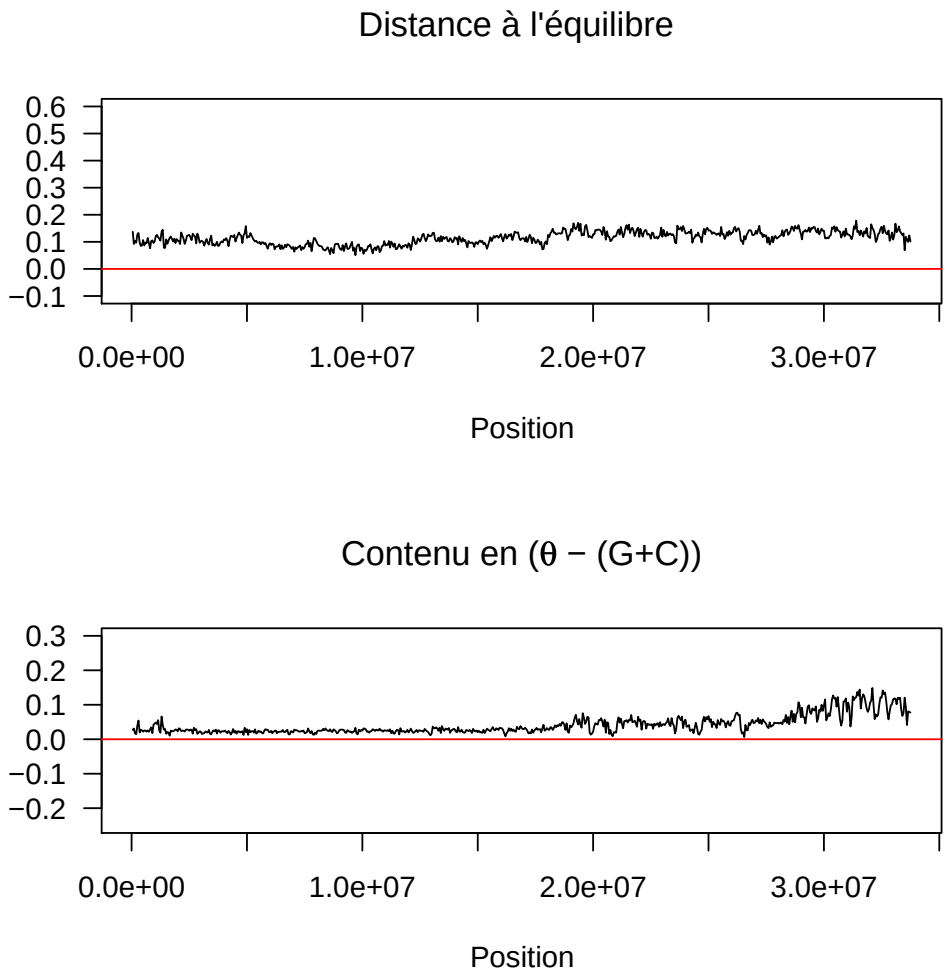


Figure 4.8 – Mesure de Δ sur des fenêtres de 50kb le long du chromosome 21, comme distance de la fenêtre à l'équilibre du modèel T92+CpG. Mesure de la différence $\theta - (G + C)$ sur des fenêtres de 50kb le long du chromosome 21. Les deux droites tracées à 0 en rouge correspondent à la valeur attendue à l'équilibre du modèle.

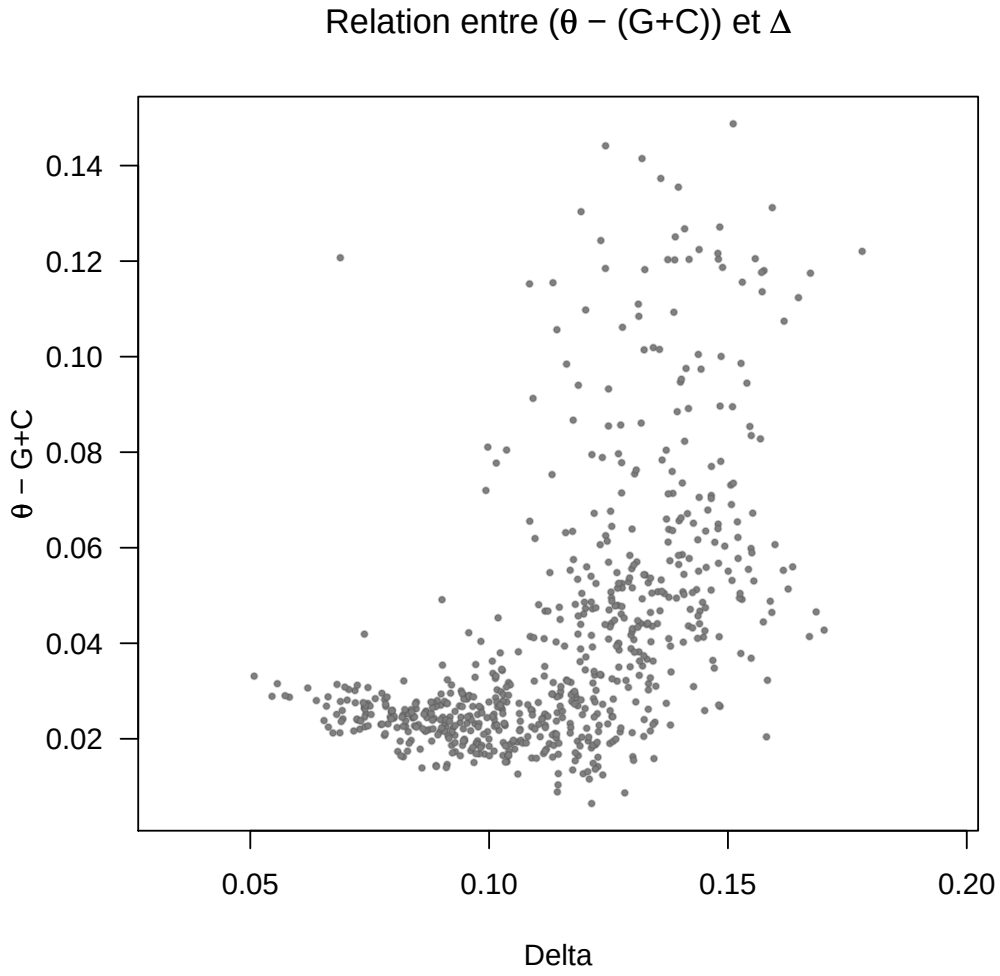


Figure 4.9 – Relation entre Δ et la différence $\theta - (G+C)$ sur des fenêtres de 50kb le long du chromosome 21.

exponentiellement à l'approche de l'équilibre, ce qui tendrait à montrer que le décalage entre θ et G+C peut effectivement être dû à la distance avec l'équilibre du modèle.

4.5.3 Estimation des taux de CpG

La figure 4.8 montre que la distance à l'équilibre de l'ensemble du chromosome 21 est assez constante, bien que la fin du chromosome se situe à une distance un peu plus élevée de l'équilibre que la première moitié du chromosome. Cette différence coïncide en partie avec l'augmentation du contenu en G+C dans la deuxième partie du chormosome 21, qui est visible autant sur nos résultats (voir figure 4.6) que sur les résultats de la littérature (voir figure 1.1).

Ceci nous permet donc d'appliquer l'estimation du rapport $r/(\alpha+\beta)$ présentée dans la section b., page 94. On observe ainsi (voir figure 4.10) que l'estimation du rapport $r/(\alpha+\beta)$ varie fortement le long du chromosome 21. Il diminue vers la fin du chromosome, et c'est précisément là qu'on sait qu'il y a beaucoup d'ilôts CpG.

Généralement, on estime que la mesure CpGo/e[6] est une bonne mesure du taux de méthylation d'une séquence, et donc indirectement du taux de substitution agissant sur les dinucléotides CpG. On observe que la mesure CpGo/e est fortement corrélée au contenu en G+C dans le chromosome (voir figure 4.11 - coefficient de corrélation = 0.766). Contrairement à la mesure CpGo/e, la mesure $r/(\alpha+\beta)$ ne possède qu'une faible corrélation avec le contenu en G+C comme nous pouvons le voir sur le nuage de points de la figure 4.13 (coefficient de corrélation = 0.266). Cette mesure est toutefois corrélée avec le contenu en CpGo/e (figure 4.12 - coefficient de corrélation = -0.683).

Ces résultats seraient-ils un indice que la mesure CpGo/e n'est pas une bonne mesure du taux de substitution sur les dinucléotides CpG, car fortement liée au contenu en G+C, et que l'estimation du rapport $r/(\alpha+\beta)$ tel que nous l'avons défini, est plus à même d'estimer ce taux ? Des analyses plus poussées sur la relation entre ces deux mesures et différentes variables, comme le degré de méthylation, le positionnement d'ilôts CpG, ... sont maintenant nécessaires.

[6] voir définition dans la section 1.2.1 du chapitre 1

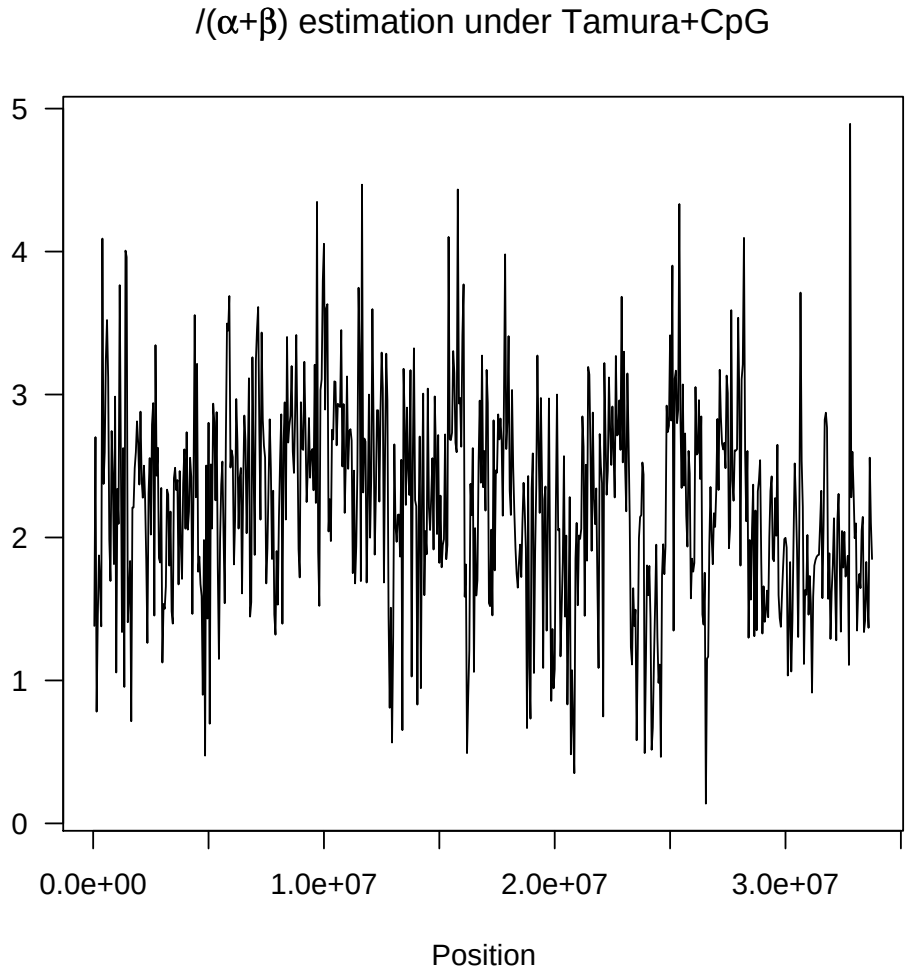


Figure 4.10 – Mesure de $r/(\alpha+\beta)$ sur des fenêtres de 50kb le long du chromosome 21.

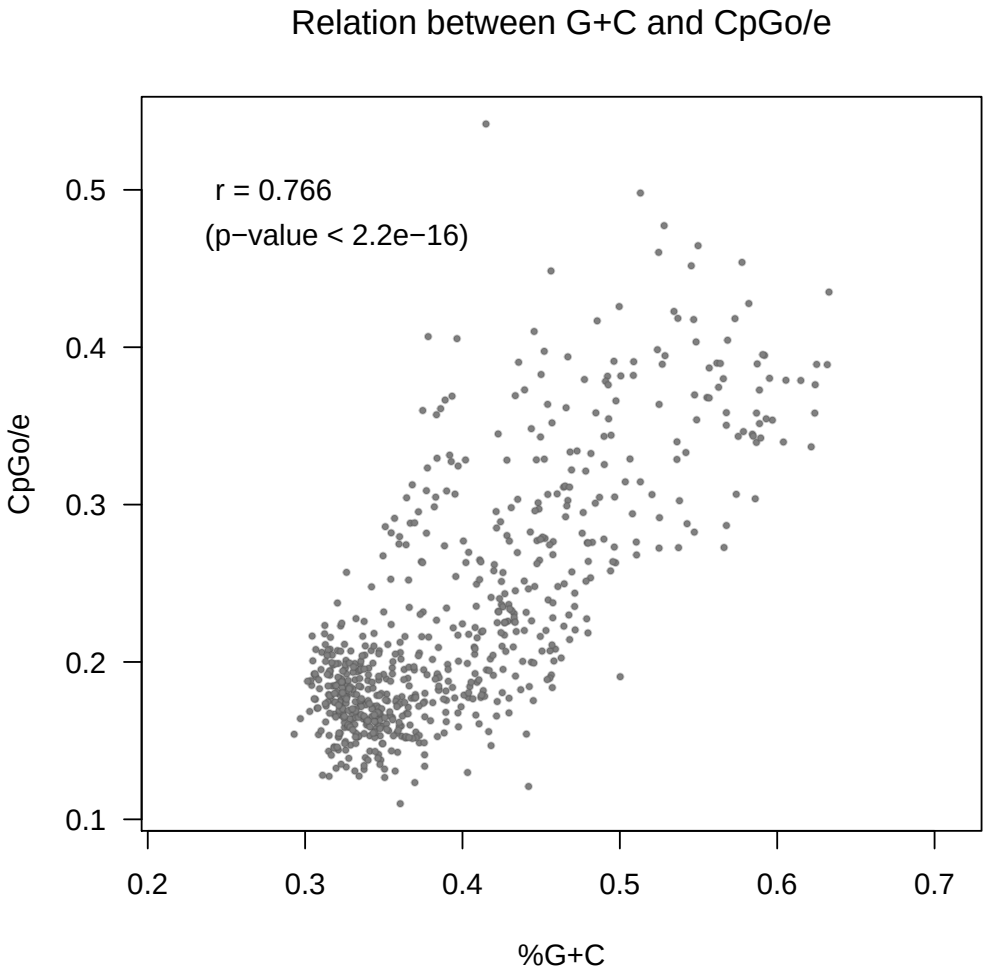


Figure 4.11 – Relation entre CpGo/e et contenu en G+C sur des fenêtres de 50kb le long du chromosome 21.

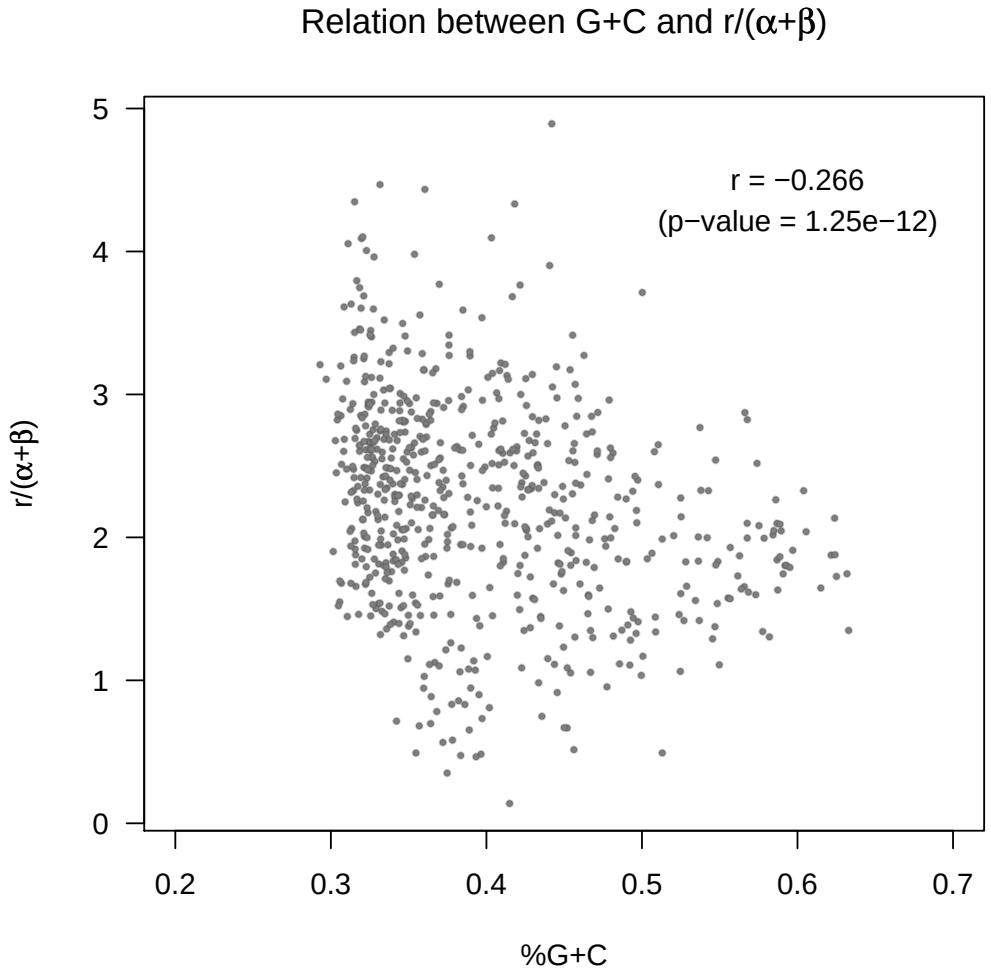


Figure 4.12 – Relation entre le contenu en G+C et $r/(\alpha + \beta)$ sur des fenêtres de 50kb le long du chromosome 21.

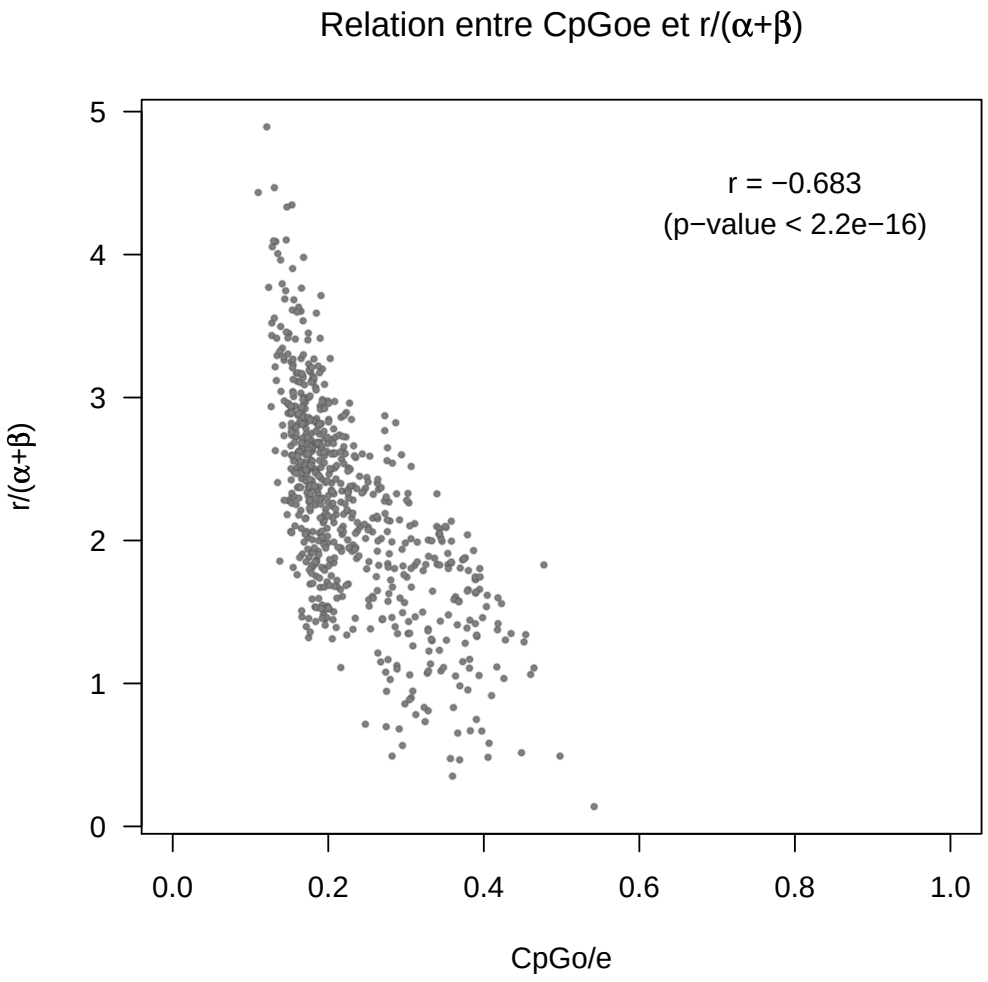


Figure 4.13 – Relation entre CpGo/e et $r/(\alpha+\beta)$ sur des fenêtres de 50kb le long du chromosome 21.

Conclusions et Perspectives

Les travaux que j'ai effectués durant ma thèse se sont articulés autour de l'**évaluation des dépendances entre sites voisins d'une séquence d'ADN** et de l'**incorporation de ces dépendances dans des modèles d'évolution de séquences**.

M'intéressant tout particulièrement aux substitutions liées à la méthylation des dinucléotides CpG, et aux substitutions liées aux dinucléotides de manière générale, je me suis d'abord intéressée aux méthodes de détection de sur- ou sous-représentation en dinucléotides. J'ai implémenté et incorporé certaines de ces méthodes au paquet SeqinR du logiciel de statistiques R. Cette implémentation m'a alors permis d'amener une réponse claire à la controverse longtemps non résolue de l'effet des rayons ultra-violets sur la composition en bases des génomes. J'ai en effet montré, sur l'ensemble des génomes complets bactériens, et sur deux exemples bien choisis de génomes bactériens et viraux qu'il n'existe pas de pression de sélection des rayons ultra-violets sur la fréquence en dinucléotides de pyrimidine dans les micro-organismes.

J'ai ensuite porté mon attention sur le problème du maintien de l'hypothèse d'indépendance entre sites voisins dans les modèles d'évolution de séquence en présentant le manque de robustesse des méthodes d'inférence phylogénétique par maximum de vraisemblance à la violation de cette hypothèse. Ceci m'a amené à l'introduction d'un modèle général d'évolution incorporant des dépendances entre sites, pour lequel j'ai implémenté un algorithme de simulation qui m'a permis une première étude par simulations du modèle général. Des travaux récents de Bérard *et al.* (2005) sur un cas particulier de ce modèle, incorporant entre autres les substitutions liées aux dinucléotides CpG, pour lesquels ils ont démontré qu'une écriture analytique était possible, m'ont permis de développer une méthode d'estimation des taux de substitution puis d'appliquer cette méthode sur le chromosome 21 d'*Homo sapiens*, et d'inférer la variation des taux de substitution des dinucléotides CpG le long de ce chromosome uniquement à partir des informations contenues dans la séquence.

Les perspectives qui se dégagent de ce travail sont de plusieurs ordres.

Tout d'abord, alors que l'utilisation de la statistique CpGo/e est encore très largement utilisée dans les études de génomique comparative pour déterminer le degré de méthylation d'un génome (Duret & Galtier, 2000; Ponger *et al.*, 2001; Arndt *et al.*, 2003; The Honeybee Genome Sequencing Consortium, 2006; Oakes *et al.*, 2007), j'ai pu montrer que cette statistique apporte peu d'information sur les mécanismes à l'œuvre sur les séquences. Les statistiques non-paramétriques implémentées dans le paquet SeqinR apportent une nette amélioration par rapport à cette statistique, et je ne peux que préconiser leur usage, voire l'usage de statistiques plus sophistiquées comme celles basées sur les chaînes de Markov dans la détermination de la sur- ou sous-représentation d'un mot.

D'autre part, cette thèse a apporté la preuve claire qu'il n'existe pas de pression de sélection des rayons ultra-violets sur la fréquence en dinucléotides de pyrimidine dans les micro-organismes. Ceci implique que les bactéries possèdent des mécanismes de protection et/ou de réparation efficaces leur permettant de s'affranchir de cette contrainte environnementale. Certains de ces mécanismes sont connus, d'autres restent à élucider. En parallèle, les mécanismes permettant aux virus de s'affranchir de cette contrainte restent encore plus mystérieux. Les virus possèdent-ils des mécanismes de protection ? se servent-ils de la machinerie de leur hôte pour réparer leur ADN endommagé ? comment alors la détournent-ils ? Le séquençage par métagénomique de virus et bactéries issues des sables de surface du désert du Sahara (Michael DuBow, *comm. pers.*) permettra bientôt de se pencher sur la question avec des données avec de larges données complètes de micro-organismes subissant de fortes pressions en radiations ultra-violettes.

Cette thèse a, par ailleurs, apporté la confirmation que le maintien de l'hypothèse d'indépendance entre sites dans les modèles d'évolution de séquences entraîne des biais dans les méthodes issues de ces modèles. Toutefois, l'incorporation de dépendances entre les sites dans les modèles d'évolution et dans les applications de ces modèles est un vaste problème. En effet, en prenant l'exemple des méthodes d'inférence phylogénétique, il faut soit repenser l'ensemble des étapes menant à la construction d'un arbre, et ré-écrire une alternative à chaque étape en autorisant des dépendances entre sites (alignement, distance entre séquences, ou inférence par maximum de vraisemblance), soit développer une méthode synthétique complète incorporant des hypothèses plus réalistes.

L'implémentation du modèle général incorporant des dépendances entre sites nous a permis d'étudier le comportement général du modèle à l'approche de l'équilibre, mais permettra surtout par la suite d'étudier le comportement de tous les modèles qui ne sont pas inclus dans le cas particulier de Bérard *et al.* (2005), et en particulier les modèles non symétriques, pour lesquelles une écriture analytique n'est pour l'instant pas envisageable. Pour ce qui est des modèles inclus dans le cas particulier de Bérard *et al.* (2005), il reste à chercher à écrire

une méthode d'estimation des paramètres du modèle dans le cas général, et non plus uniquement pour les deux cas particulier de K80+CpG et T92+CpG. Ceci permettrait d'avoir une approche complète pour cette classe de modèles, allant de l'écriture exacte de la distribution stationnaire à l'estimation des paramètres de subtitution à partir de la simple séquence.

Pour finir, les résultats obtenus sur le chromosome 21 permettent d'envisager une étude à grande échelle sur l'ensemble du génome d'*Homo sapiens* de manière à déterminer les taux de substitution de type CpG et leurs variations, ainsi que leur relation à un certain nombre de variables que l'on sait être liées à ce taux de substitution, comme le degré de méthylation, le contenu en G+C et en ilôts CpG ou le taux de recombinaison.

Références bibliographiques

Agogué H., Joux F., Obernosterer I. & Lebaron P. (2005). Resistance of Marine Bacterioneuston to Solar Radiation. *Applied and Environmental Microbiology*, **71**(9), 5282–5289. *Cité pages 29 et 37.*

Arndt P. F. (2007). Reconstruction of ancestral nucleotide sequences and estimation of substitution frequencies in a star phylogeny. *Gene*, **390**(1-2), 75–83. *Cité pages 69 et 86.*

Arndt P. F., Burge C. B. & Hwa T. (2003). DNA sequence evolution with neighbor-dependent mutation. *Journal of Computational Biology*, **10**(3-4), 313–322. *Cité pages 10, 16, 69, 75, 86 et 106.*

Arndt P. F. & Hwa T. (2005). Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics*, **21**(10), 2322–2328. *Cité page 86.*

Audit B., Vaillant C., Arneodo A., d'Aubenton Carafa Y. & Thermes C. (2004). Wavelet Analysis of DNA Bending Profiles reveals Structural Constraints on the Evolution of Genomic Sequences. *Journal of Biological Physics*, **30**, 33–81. *Cité page 6.*

Azhikina T. L. & Sverdlov E. D. (2005). Study of Tissue-Specific CpG Methylation of DNA in Extended Genomic Loci. *Biochemistry (Moscow)*, **70**(5), 596–603. *Cité page 41.*

Bak A. L., Atkins J. F., Singer C. E. & Ames B. N. (1972). Evolution of DNA Base Compositions in Microorganisms. *Science*, **175**(4028), 1391–1393. *Cité pages 15 et 29.*

Beadle G. W. & Tatum E. L. (1941). Genetic Control of Biochemical Reactions in Neurospora. *Proceedings of the National Academy of Sciences of the United States of America*, **27**, 499–506. *Cité page 1.*

Bérard J., Gouéré J.-B. & Piau D. (2005). Solvable models of neighbor-dependent nucleotide substitution processes. *e-arXiv math.PR/0510034. Cité pages 69, 86, 87, 88, 89, 105 et 106.*

Bernardi G. (1989). The isochore organization of the human genome. *Annu. Rev. Genet.*, **23**, 637–661. *Cité page 95.*

Bernardi G., Olofsson B., Filipski J., Zerial M., Salinas J., Cuny G., Meunier-Rotival M. & Rodier F. (1985). The Mosaic Genome of Warm-Blooded Vertebrates. *Science*, **228**(4702), 953–958. *Cité page 7.*

Besaratinia A., Synold T. W., Hsiu-Hua C., Chang C., Xi B., Riggs A. D. & Pfeifer G. D. (2005). DNA lesions induced by UV A1 and B radiation in human cells : Comparative analyses in the overall genome and in the p53 tumor suppressor gene. *Proceedings of the National Academy of the United States of America*, **102**(29), 10058–10063. *Cité page 25.*

Bird A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Researh*, **8**, 1499–1504. *Cité pages 38 et 40.*

Boussau B. & Gouy M. (2006). Efficient likelihood computations with non-reversible models of evolution. *Syst Biol.*, **55**(5), 756–768. *Cité page 56.*

Burge C., Campbell A. M. & Karlin S. (1992). Over- and under-representation of short oligonucleotides in DNA sequences. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 1358–1362. *Cité page 9.*

Cantarel B. L., Morrison H. G. & Pearson W. (2006). Exploring the Relationship between Sequence Similarity and Accurate Phylogenetic Trees. *Molecular Biology and Evolution*, **23**(11), 2090–2100. *Cité page 71.*

Cartwright R. A. (2005). DNA assembly with gaps (Dawg) : simulating sequence evolution. *Bioinformatics*, **21**(Suppl3), iii31–38. *Cité page 71.*

Charif D., Humblot L., Lobry J. R. & Palmeira L. (2007). SeqinR 1.0-7 : a contributed package to the project for statistical computing devoted to biological sequences retrieval and analysis. *Cité page 15.*

Charif D. & Lobry J. (2006). SeqinR 1.0-2 : a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In H. R. U. Bastolla, M. Porto & M. Vendruscolo, Eds., *Structural approaches to sequence evolution : Molecules, networks, populations*, volume NA of *Biological and Medical Physics, Biomedical Engineering*, p.ÑA. New York : Springer Verlag. *Cité page 15.*

Chen S. L., Lee W., Shapiro L. & McAdams H. H. (2004). Codon usage between genomes is constrained by genome-wide mutational processes. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(10), 3480–3485. *Cité page 24.*

Clark S. J., Harrison J. & Frommer M. (1995). CpNpG methylation in mammalian cells. *Nature Genetics*, **10**, 20–27. *Cité pages 38 et 39.*

Cleaver J. E. (2006). Cells have long experience of dealing with UVC light. *Nature*, **442**(7100), 244–244. *Cité page 37.*

Coleman M., Sullivan M., Martiny A., Steglich C., Barry K., DeLong E. & Chisholm S. (2006). Genomic islands and the ecology and evolution of Prochlorococcus. *Science*, **311**(5768), 1768 – 1770. *Cité page 30.*

Colot V. & Rossignol J. L. (1999). Eukaryotic DNA methylation as an evolutionary device. *Bioessays*, **21**(5), 402–411. *Cité page 38.*

Costello J. F. & Plass C. (2001). Methylation matters. *J Med Genet*, **38**(5), 285–303. *Cité pages 37, 38 et 39.*

Dufresne A., Salanoubat M., Partensky F., Artiguenave F., Axmann I., Barbe V., Duprat S., Galperin M., Koonin E., Le Gall F., Makarova K., Ostrowski M., Oztas S., Robert C., Rogozin I., Scanlan D., Tandeau de Marsac N., Weissenbach J., Wincker P., Wolf Y. & Hess W. (2003). Genome sequence of the cyanobacterium *Prochlorococcus marinus SS120*, a near minimal oxyphototrophic genome. *Proceedings of the National Academy of the United States of America*, **100**(17), 10020–10025. *Cité pages 31 et 34.*

Duret L. & Galtier N. (2000). The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact. *Molecular Biology and Evolution*, **17**(11), 1620–1625. *Cité pages 10, 16, 69, 86 et 106.*

Eckhardt F., Lewin J., Cortese R., Rakyan V. K., Attwood J., Burger M., Burton J., Cox T. V., Davies R., Down T. A., Haefliger C., Horton R., Howe K., Jackson D. K., Kunde J., Koenig C., Liddle J., Niblett D., Otto T., Pettett R., Seemann S., Thompson C., West T., Rogers J., Olek A., Berlin K. & Beck S. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genetics*, **38**(12), 1378–1385. *Cité page 41.*

Essen L. O. & Klar T. (2006). Light-driven DNA repair by photolyases. *Cellular and Molecular Life Sciences*, **63**, 1266–1277. *Cité page 27.*

Felsenstein J. (1981). Evolutionary Trees from DNA Sequences : A Maximum Likelihood Approach. *Journal of Molecular Evolution*, **17**, 368–376. *Cité pages 2, 45 et 51.*

Fitch W. M. & Markowitz E. (1970). An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics*, **4**(5), 579–593. *Cité page 53.*

Fryxell K. J. & Moon W.-J. (2005). CpG Mutation Rates in the Human Genome Are Highly Dependent on Local GC Content. *Molecular Biology and Evolution*, **22**(3), 650–658. *Cité page 95.*

Galtier N. (2001). Maximum-Likelihood Phylogenetic Analysis Under a Covarion-like Model. *Molecular Biology and Evolution*, **18**(5), 866–873. *Cité page 56.*

Galtier N. & Gouy M. (1998). Inferring pattern and process : maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Molecular Biology and Evolution*, **15**(7), 871–879. *Cité page 56.*

Galtier N. & Lobry J. (1997). Relationships Between Genomic G+C Content, RNA Secondary Structures, and Optimal Growth Temperature in Prokaryotes. *Journal of Molecular Evolution*, **44**(6), 632–636. *Cité page 25.*

Gautier C., Gouy M. & Louail S. (1985). Non-parametric statistics for nucleic acid sequence study. *Biochimie*, **67**, 449–453. *Cité page 14.*

Gesell T. & von Haeseler A. (2006). In silico sequence evolution with site-specific interactions along phylogenetic trees. *Bioinformatics*, **22**(6), 716–722. *Cité page 71.*

Goh L., Murphy S. K., Muhkerjee S. & Furey T. S. (2007). Genomic sweeping for hypermethylated genes. *Bioinformatics*, **23**(3), 281–288. *Cité page 39.*

Gouy M. & Gautier C. (1982). Codon usage in bacteria : correlation with gene expressivity. *Nucleic Acids Res*, **10**(22), 7055–7074. *Cité page 23.*

Grantham R., Gautier C., Gouy M., Mercier R. & Pavé A. (1980). Codon catalog usage and the genome hypothesis. *Nucleic Acids Research*, **8**(1), r49–r62. *Cité page 23.*

Guindon S. & Gascuel O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, **52**(5), 696–704. *Cité pages 58, 59 et 60.*

Guisez Y., Robbens J., Remaut E. & Fiers W. (1993). Folding of the MS2 Coat Protein in Escherichia coli is Modulated by Translational Pauses Resulting from mRNA Secondary Structure and Codon Usage : A Hypothesis. *Journal of Theoretical Biology*, **162**(2), 243–252. *Cité page 23.*

Hall B. G. (2005). Comparison of the Accuracies of Several Phylogenetic Methods Using Protein and DNA Sequences. *Molecular Biolomgy and Evolution*, **22**(3), 792–802. *Cité page 58.*

Hartman P. S. & Eisenstark A. (1982). Alteration of Bacteriophage Attachment Capacity by Near-UV Irradiation. *Journal of Virology*, **43**(2), 529–532. *Cité page 37.*

Hasegawa M., Kishino H. & Yano T. (1985). Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA. *Journal of Molecular Evolution. Cité pages 52, 54 et 87.*

Hayashi H., Nagae G., Tsutsumi S., Kaneshiro K., Kozaki T., Kaneda A., Sugisaki H. & Aburatini H. (2006). High-resolution mapping of DNA methylation in human genome using oligonucleotide tiling array. *Hum Genet*, **120**, 701–711. *Cité page 41.*

Hillis D. M., Bull J. J., White M. E., Badgett M. R. & Molineux I. J. (1992). Experimental phylogenetics : generation of a known phylogeny. *Science*, **255**(5044), 589–592. *Cité page 58.*

Ikemura T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes : a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *Journal of Molecular Biology*, **151**(3), 389–409. *Cité page 23.*

Ikemura T. (1982). Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. *Journal of Molecular Biology*, **158**(4), 573–597. *Cité page 23.*

Ikemura T. (1985). Codon Usage and tRNA Content in Unicellular and Multicellular Organisms. *Molecular Biology and Evolution*, **2**(1), 13–34. *Cité page 23.*

Jin L. & Nei M. (1990). Limitations of the evolutionary parsimony method of phylogenetic analysis. *Molecular Biology and Evolution*, **7**(1), 82–102. *Cité page 2.*

Jin L. & Nei M. (1991). Relative efficiencies of the maximum-parsimony and distance-matrix methods of phylogeny construction for restriction data. *Molecular Biology and Evolution*, **8**(3), 356–365. *Cité page 58.*

Jukes T. H. & Cantor C. R. (1969). *Evolution of protein molecules*, In H. N. Munro, Ed., *Mammalian Protein Metabolism*, volume 3, chapter 24, p. 21–132. Academic Press : New York. *Cité pages vi, 47, 49, 50, 52 et 86.*

Karlin S. & Cardon L. R. (1994). Computational DNA sequence analysis. *Annual Review of Microbiology*, **48**, 619–654. *Cité pages 9 et 16.*

Kellogg C. A. & Paul J. H. (2002). Degree of ultraviolet radiation damage and repair capabilities are related to G+C content in marine phages. *Aquatic Microbial Ecology*, **27**, 13–20. *Cité pages 27 et 29.*

Kelly C. & Rice J. (1996). Modeling Nucleotide Evolution : A Heterogeneous Rate Analysis. *Mathematical Biosciences*, **133**, 85–109. *Cité page 54.*

Kimura M. (1980). A Simple Method for Estimating Evolutionary Rates of Base Substitutions Through Comparative Studies of Nucleotide Sequences. *Journal of Molecular Evolution*, **16**, 111–120. *Cité pages iii, 47, 52, 61, 87, 90, 93 et 131.*

Kuhner M. K. & Felsenstein J. (1994). A Simulation of Phylogeny Algorithms under Equal and Unequal Evolutionary Rates. *Molecular Biology and Evolution*, **11**(3), 459–468. *Cité page 61.*

Lake J. A. (1994). Reconstructing Evolutionary Trees from DNA and Protein Sequences : Paralinear Distances. *Proceedings of the National Academy of the United States of America*, **91**, 1455–1459. *Cité page 2.*

Li W.-H. (1981). Simple method for constructing phylogenetic trees from distance matrices. *Proceedings of the National Academy of Sciences of the United States of America*, **78**(2), 1085–1089. *Cité page 2.*

Lindell D., Sullivan M. B., Johnson Z. I., Tolonen A. C., Rohwer F. & Chisholm S. W. (2004). Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(30), 11013–11018. *Cité page 35.*

Litjens R. A., Quickenden T. I. & Freeman C. G. (1999). Visible and near-ultraviolet absorption spectrum of liquid water. *Applied Optics*, **38**(7), 1216–1223. *Cité pages 31 et 33.*

Lobry J. R. (1996). A simple vectorial representation of dna sequences for the detection of replication origins in bacteria. *Biochimie*, **78**(5), 323–326. *Cité pages 7 et 9.*

Lobry J. R. & Chessel D. (2003). Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. *J Appl Genet*, **44**(2), 235–261. *Cité page 24.*

Lobry J. R. & Necşulea A. (2006). Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes. *Gene*, **385**, 128–136. *Cité page 24.*

Makarenkov V. & Legendre P. (2001). Optimal Variable Weighting for Ultrametric and Additive Trees and K-means Partitioning : Methods and Software. *Journal of Classification*, **18**, 245–271. *Cité page 61.*

McInerney J. O. (1998). Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(18), 10698–10703. *Cité page 25.*

Messer P. W., Arndt P. F. & Lässig M. (2005). Solvable sequence evolution models and genomic correlations. *Physical Review Letters*, **94**(13), 138103. *Cité page 6.*

Moore L. R., Rocap G. & Chisholm S. W. (1998). Physiology and molecular phylogeny of coexisting Prochlorococcus ecotypes. *Nature*, **393**(6684), 464–467. *Cité page 30.*

Mouret S., Baudouin C., Charveron M., Favier A., Cadet J. & Douki T. (2006). Cyclobutane pyrimidine dimers are predominant DNA lesions in whole human skin exposed to UVA radiation. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(37), 13765–13770. *Cité page 25.*

Muto A. & Osawa S. (1987). The Guanine and Cytosine Content of Genomic DNA and Bacterial Evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **84**(1), 166–169. *Cité page 6.*

Nakamura Y., Gojobori T. & T. I. (2000). Codon usage tabulated from international DNA sequence databases : status for the year 2000. *Nucleic Acids Research*, **28**(1), 292. *Cité page 23.*

Naya H., Romero H., Zavala A., Alvarez B. & Musto H. (2002). Aerobiosis Increases the Genomic Guanine Plus Cytosine Content (GC%) in Prokaryotes. *Journal of Molecular Evolution*, **55**(3), 260–264. *Cité page 24.*

Oakes C. C., La Salle S., Smiraglia D. J., Robaire B. & Trasler J. M. (2007). A unique configuration of genome-wide DNA methylation patterns in the testis. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(1), 228–233. *Cité pages 10, 16 et 106.*

Palmeira L., Guéguen L. & Lobry J. R. (2006). UV-Targeted Dinucleotides Are Not Depleted in Light-Exposed Prokaryotic Genomes. *Molecular Biology and Evolution*, **23**(11), 2214–2219. *Cité page 29.*

Peng C.-K., Buldyrev S. V., Goldberger A. L., Havlin S., Sciortino F., Simons M. & Stanley H. E. (1992). Long-range correlations in nucleotide sequences. *Nature*, **356**, 168–170. *Cité page 7.*

Pimentel Cachapuz Rocha E. (2000). *Analyse exploratoire des génomes bactériens*. PhD thesis, Université de Versailles Saint-Quentin-En-Yvelines. *Cité page 25.*

Plotkin J. B., Robins H. & Levine, Harold J. (2004). Tissue-specific codon usage and the expression of human genes. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(34), 12588–12591. *Cité page 24.*

Pólya G. (1920). Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem. *Mathematische Zeitschrift*, **8**, 171–181. *Cité page 11.*

Ponger L., Duret L. & Mouchiroud D. (2001). Determinants of CpG Islands : Expression in Early Embryo and Isochore Structure. *Genome Research*, **11**, 1854–1860. *Cité pages 10, 16 et 106.*

Prigent M., Leroy M., Confalonieri F., Dutertre M. & DuBow M. S. (2005). A diversity of bacteriophage forms and genomes can be isolated from the surface sands of the Sahara Desert. *Extremophiles*, **9**, 289–296. *Cité page 34.*

Quickenden T. I. & Irvin J. A. (1980). The ultraviolet absorption spectrum of liquid water. *The Journal of Chemical Physics*, **72**, 4416–4428. *Cité pages 31 et 33.*

Rambaut A. & Grassly N. C. (1997). Seq-Gen : an application for the Monte-Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences*, **13**(3), 235–238. *Cité page 71.*

Ren L., Gao G., Zhao D., Ding M., Luo J. & Deng H. (2007). Developmental stage related patterns of codon usage and genomic GC content : searching for evolutionary fingerprint by models of stem cell differentiation. *Genome Biology*, **8**, R35. *Cité page 23.*

Robin S., Rodolphe F. & Schbath S. (2003). *ADN, mots et modèles*. Collection Échelles. Belin. *Cité pages 12 et 13.*

Robin S., Schbath S. & Vandewalle V. (2007). Statistical tests to compare motif count exceptionalities. *Genome Biology*, **8**, 84. *Cité page 12.*

Robinson D. F. & Foulds L. R. (1981). Comparison of Phylogenetic Trees. *Mathematical Biosciences*, **53**, 131–147. *Cité page 61.*

Rocap G., Larimer F., Lamerdin J., Malfatti S., Chain P., Ahlgren N., Arellano A., Coleman M., Hauser L., Hess W., Johnson Z., Land M., Lindell D., Post A., Regala W., Shah M., Shaw S., Steglich C., Sullivan M., Ting C., Tolonen A., Webb E., Zinser E. & Chisholm S. (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature*. *Cité pages 30, 31 et 34.*

Rosenberg M. S. (2005). MySPP : Non-stationary evolutionary sequence simulation, including indels. *Evolutionary Bioinformatics Online*, **1**, 51–53. *Cité page 71.*

Rzhetsky A. & Nei M. (1995). Tests of Aplicability of Several Substitution Models for DNA Sequence Data. *Molecular Biology and Evolution*, **12**(1), 131–151. *Cité pages 52 et 87.*

Saitou N. & Imanishi T. (1989). Relative Efficiencies of the Fitch-Margoliash, Maximum-Parsimony, Maximum-Likelihood, Minimum-Evolution, and Neighbor-joining Methods of Phylogenetic Tree Construction in Obtaining the Correct Tree. *Molecular Biology and Evolution*, **6**(5), 514–525. *Cité page 58.*

Sanger F., Nicklen S. & Coulson A. R. (1977). DNA Sequencing with Chain-Terminating Inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**(12), 5463–5467. *Cité page 1.*

Sanger F. & Tuppy H. (1951). The Amino-acid Sequence in the Phenylalanyl Chain of Insulin. *Biochemical Journal*, **49**, 463–481. *Cité page 1.*

Schbath S. (1995). *Étude asymptotique du nombre d'occurrences d'un mot dans une chaîne de Markov et application à la recherche de mots de fréquence exceptionnelle dans les séquences d'ADN*. PhD thesis, Université René Descartes, Paris V. *Cité pages 14 et 122.*

Schbath S. (1997). An efficient statistic to detect over- and under-represented words in DNA sequences. *Journal of Computational Biology*, **4**(2), 189–192. *Cité page 12.*

Schöniger M. & von Haeseler A. (1995). Simulating efficiently the evolution of DNA sequences. *Computational Applied Biosciences*, **11**(1), 111–115. *Cité pages 71, 72 et 73.*

Setlow R. B. (1966). Cyclobutane-Type Pyrimidine Dimers in Polynucleotides. *Science*, **153**(734), 379–386. *Cité pages 25, 29, 126 et 127.*

Sharp P. M. & Li W. H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. *Journal of Molecular Evolution*, **24**(1-2), 28–38. *Cité page 23.*

Sharp P. M. & Matassi G. (1994). Codon usage and genome evolution. *Curr Opin Genet Dev*, **4**(6), 851–860. *Cité page 23.*

Singer C. E. & Ames B. N. (1970). Sunlight Ultraviolet and Bacterial DNA Base Ratios. *Science*, **170**(3960), 822–826. *Cité pages 15, 25, 27, 29 et 127.*

Sinha R. P. & Häder D.-P. (2002). UV-induced DNA damage and repair : a review. *Photochemical and Photobiological Sciences*, **1**, 225–236. *Cité pages 25, 26 et 28.*

Smith T. F. & Waterman M. S. (1981). Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, **147**, 195–197. *Cité page 2.*

Strope C. L., Scott S. D. & Moriyama E. N. (2007). indel-Seq-Gen : A New Protein Family Simulator Incorporating Domains, Motifs, and Indels. *Molecular Biology and Evolution*, **24**(3), 640–649. *Cité page 71.*

Sueoka N. (1962). On the genetic basis of variation and heterogeneity of DNA base composition. *Proceedings of the National Academy of Sciences of the U. S. A.*, **48**(4), 585–592. *Cité pages 5 et 6.*

Sullivan M. B., Coleman M. L., Weigele P., Rohwer F. & Chisholm S. W. (2005). Three *Prochlorococcus* cyanophage genomes : Signature features and ecological interpretations. *PLoS Biology*, **3**(5), e144. *Cité page 35.*

Sémon M., Lobry J. R. & Duret L. (2006). No Evidence for Tissue-Specific Adaptation of Synonymous Codon Usage in Humans. *Molecular Biology and Evolution*, **23**(3), 523–529. *Cité page 24.*

Tajima F. & Nei M. (1984). Estimation of Evolutionary Distance between Nucleotide Sequences. *Molecular Biology and Evolution*, **1**(3), 269–285. *Cité pages 2 et 52.*

Takami H., Nishi S., Lu J., Shimamura S. & Takaki Y. (2004). Genomic characterization of thermophilic *Geobacillus* species isolated from the deepest sea mud of mariana trench. *Extremophiles*, **8**, 351–156. *Cité pages 34 et 35.*

Tamura K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Molecular Biology and Evolution*, **9**(4), 678–687. *Cité pages iii, 47, 48, 52, 87, 91 et 94.*

Tamura K. & Nei M. (1993). Estimation of the Number of Nucleotide Substitutions in the Control Region of Mitochondrial DNA in Humans and Chimpanzees. *Molecular Biology and Evolution*, **10**(3), 512–526. *Cité pages 52 et 87.*

Tateno Y., Nei M. & Tajima F. (1982). Accuracy of estimated phylogenetic trees from molecular data. i. distantly related species. *Journal of Molecular Evolution*, **18**, 387–404. *Cité page 60.*

The Honeybee Genome Sequencing Consortium (2006). Insights into social insects from the genome of the honeybee Apis mellifera. *Nature*, **443**(7114), 931–949. *Cité pages 10, 16 et 106.*

Tillier E. & Collins R. (1995). Neighbor Joining and Maximum Likelihood with RNA Sequences : Addressing the Interdependence of Sites. *Molecular Biology and Evolution*, **12**(1), 7–15. *Cité page 60.*

Tomita M., Wada M. & Kawashima Y. (1999). ApA Dinucleotide Periodicity in Prokaryote, Eukaryote, and Organelle Genomes. *Journal of Molecular Evolution*, **49**, 182–192. *Cité page 31.*

Tuffley C. & Steel M. (1998). Modeling the Covarion Hypothesis of Nucleotide Substitution. *Mathematical Biosciences*, **147**, 63–91. *Cité page 53.*

van der Staay G. W., van der Staay S. Y. M., Garczarek L. & Partensky F. (2000). Rapid evolutionary divergence of Photosystem I core subunits PsaA and PsaB in the marine prokaryote Prochlorococcus. *Photosynth Res*, **65**(2), 131–139. *Cité page 30.*

Versteeg R., van Schaik B. D., van Batenburg M. F., Roos M., Monajemi R., Caron H., Bussemaker H. J. & van Kampen A. H. (2003). The Human Transcriptome Map Reveals Extremes in Gene Density, Intron Length, GC Content, and Repeat Pattern for Domains of Highly and Weakly Expressed Genes. *Genome Res.*, **13**(9), 1998–2004. *Cité pages 7, 8 et 95.*

Vinogradov A. E. (2003). DNA helix : the importance of being GC-rich. *Nucleic Acids Research*, **31**(7), 1838–1844. *Cité page 25.*

Wang H.-C., Susko E. & Roger A. J. (2006a). On the correlation between genomic G+C content and optimal growth temperature in prokaryotes : Data quality and confounding factors. *Biochem Biophys Res Commun*, **342**(3), 681–4. *Cité page 25.*

Wang Y., Jorda M., Jones P. L., Maleszka R., Ling X., Robertson H. M., Mizzen C. A., Peinado M. A. & Robinson G. E. (2006b). Functional CpG Methylation System in a Social Insect. *Science*, **314**(5799), 645–647. *Cité page 38.*

Weber M., Davies J. J., Wittig D., Oakeley E. J., Haase M., Lam W. L. & Schuebeler D. (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genetics*, **37**, 853–862. *Cité page 41.*

Weinbauer M. G. (2004). Ecology of prokaryotic viruses. *FEMS Microbiology Reviews*, **28**, 127–181. *Cité page 37.*

Wilhelm S. W., Jeffrey W. H., Dean A. L., Meador J., Pakulski J. D. & Mitchell D. L. (2003). UV radiation induced DNA damage in marine viruses along a latitudinal gradient in the southeastern Pacific Ocean. *Aquatic Microbial Ecology*, **31**, 1–8. *Cité pages 30 et 37.*

Yang Z. (1994). Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites : Approximate Methods. *Journal of Molecular Evolution*, **39**, 306–314. *Cité page 54.*

Yang Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology and Evolution*, **11**, 367–372. *Cité page 55.*

Yang Z. (1997). PAML : a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**(5), 555–556. *Cité page 71.*

Yang Z. (2006). *Computational Molecular Evolution.* Oxford Series in Ecology and Evolution. Oxford University Press. *Cité pages 46 et 57.*

Yang Z., Goldman N. & Friday A. (1994). Comparison of Models for Nucleotide Substitution Used in Maximum-Likelihood Phylogenetic Estimation. *Molecular Biology and Evolution*, **11**(2), 316–324. *Cité page 54.*

Yang Z. & Kumar S. (1996). Approximate Methods for Estimating the Pattern of Nucleotide Substitution and the Variation of Substitution Rates Among Sites. *Molecular Biology and Evolution*, **13**(5), 650–659. *Cité page 56.*

Annexes

A. Sur les écritures exactes et approchées du z-base

L'écriture de la statistique z-base que nous proposons page 14 est une écriture approchée, sur des grandes séquences, de l'écriture exacte. Nous développons ici la relation entre écriture exacte et écriture approchée, et nous montrons ensuite que l'utilisation de l'écriture approchée ne remet pas en cause les conclusions des travaux présentés dans le chapitre 2 de cette thèse.

En effet, nous nous intéressons à la statistique suivante :

$$z_{score} = \frac{\rho_{XY} - E(\rho_{XY})}{\sqrt{Var(\rho_{XY})}}$$

où $E(\rho_{XY})$ et $Var(\rho_{XY})$ sont l'espérance et la variance de ρ_{XY} sous le modèle de permutation des bases.

Écriture exacte

L'écriture en fréquence peut être transformée en une écriture en comptage :

$$\rho_{XY} = \frac{f_{XY}}{f_X \times f_Y} = \frac{n^2}{n-1}.\frac{n_{XY}}{n_X \times n_Y}$$

Or, Schbath (1995) consacre l'annexe A.5 de sa thèse à l'écriture de l'espérance et de la variance du comptage d'un mot dans une séquence générée sous le modèle de Markov d'ordre 0 conditionné par le comptage des lettres (qui est équivalent au modèle de permutation des bases). On constate que, sous ce modèle, le dénominateur $n_X \times n_Y$ est une constante, et que l'espérance et la variance de ρ_{XY} peuvent donc être facilement obtenus à partir de l'espérance et de la variance du comptage n_{XY}.

Ainsi, on peut se servir de l'écriture sur le comptage pour obtenir l'écriture de l'espérance de ρ_{XY} :

$$E(\rho_{XY}) = \frac{n^2}{(n-1)n_X n_Y} E(n_{XY})$$

et de la variance de ρ_{XY} :

$$Var(\rho_{XY}) = \frac{n^4}{(n-1)^2 n_X^2 n_Y^2} Var(n_{XY})$$

En différenciant le cas où $X \neq Y$ de celui où $X = Y$, nous obtenons à partir de Schbath (1995) :

$$E(\rho_{XY}) = \frac{n}{n-1}$$

$$E(\rho_{XX}) = \frac{n(n_X - 1)}{n_X(n-1)}$$

$$Var(\rho_{XY}) = \frac{n^2}{(n-1)^2}[\frac{n}{n_X n_Y} - 1 + \frac{n(n-4)(n_X-1)(n_Y-1)}{n_X n_Y (n-2)(n-3)}]$$

$$Var(\rho_{XX}) = \frac{n^2(n_X-1)}{n_X^2(n-1)^2}[1 - n_X + \frac{n}{n_X}[1 + \frac{2(n_X-2)}{n-1} + \frac{(n-4)(n_X-2)(n_X-3)}{(n-2)(n-3)}]]$$

Écriture approchée sur de grandes séquences

Espérance

Lorsque la séquence est grande, on constate que pour l'espérance :

$$\lim_{n\to\infty} E(\rho_{XY}) = \lim_{n\to\infty} E(\rho_{XX}) = 1$$

Ce qui équivaut bien à l'écriture proposée page 14 pour l'espérance.

Variance

Lorsque la séquence est grande, on constate que pour la variance :

$$\lim_{n\to\infty} Var(\rho_{XY}) = \frac{n}{n_X \times n_Y} \text{ et } \lim_{n\to\infty} Var(\rho_{XX}) = \frac{n}{n_X^2}$$

L'écriture proposée page 14 est la suivante :

$$\widehat{Var(\rho_{XY})} = \frac{(1-f_X)(1-f_Y)}{n f_X f_Y} = \frac{(n-n_X)(n-n_Y)}{n n_X n_Y} = \frac{n}{n_X n_Y} - \frac{1}{n_X} - \frac{1}{n_Y} + \frac{1}{n}$$

Et on constate que, lorsque la séquence est grande :

$$\lim_{n\to\infty} \widehat{Var(\rho_{XY})} = \frac{n}{n_X \times n_Y} \text{ et } \lim_{n\to\infty} \widehat{Var(\rho_{XX})} = \frac{n}{n_X^2}$$

Les deux écritures sont donc équivalentes pour de grandes séquences.

Impact de l'utilisation des écritures approchées sur les résultats de cette thèse

En reprenant l'exemple de *Mycoplasma genitalium* donné en fin de chapitre 1, la figure A.1 illustre bien la conséquence de l'utilisation de l'approximation du z-base plutôt que son écriture exacte.

De manière générale, les sur- ou sous-représentations observées avec l'écriture approchée deviennent moins significatives avec l'écriture exacte. On le constate

par exemple ici sur les dinucléotides ApT et TpA. On note donc que l'approximation augmente artificiellement la significativité des sur- ou sous-représentations mesurées et qu'il est donc important de veiller à utiliser l'écriture exacte lorsqu'on cherche à mettre en évidence une sur- ou sous-représentation.

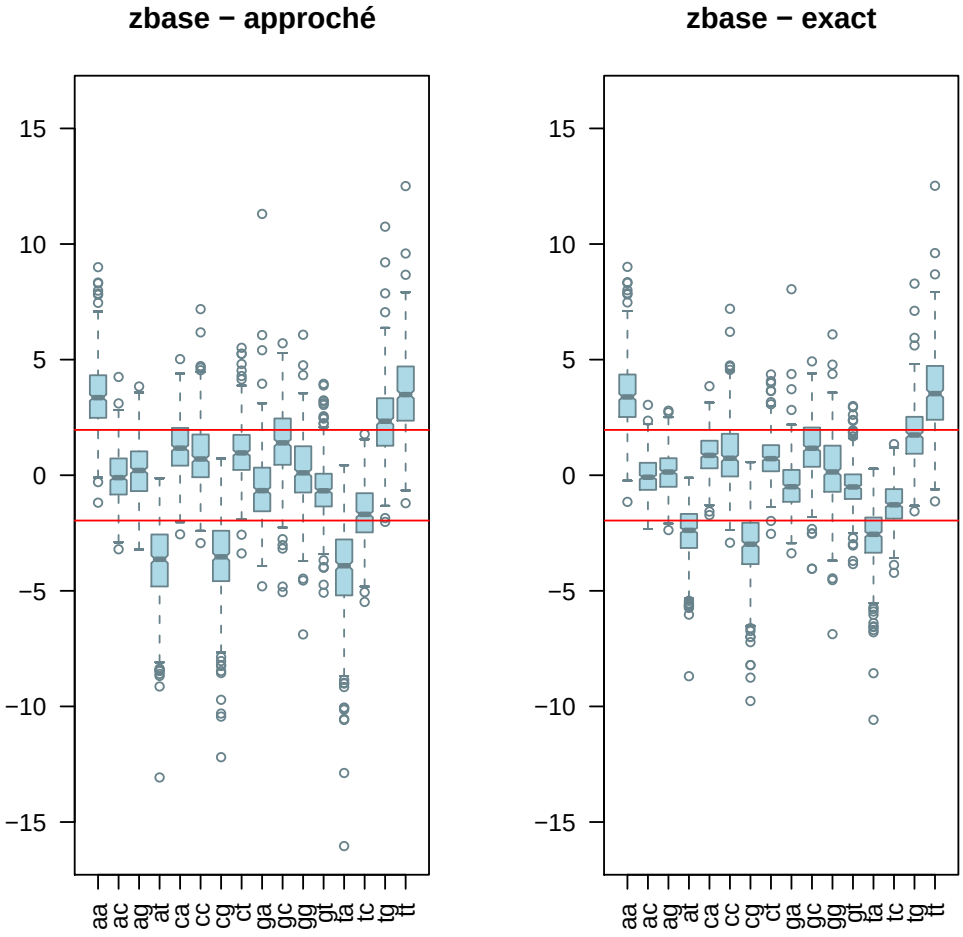


Figure A.1 – La distribution des statistiques z-base avec écriture approchée et écriture exacte, calculées sur tous les CDS de *Mycoplasma genitalium*. Les limites de significativité ont été tracées en rouge, elles ont été placées à 5% de significativité.

Toutefois, sur l'étude de l'impact des rayons ultra-violets que je présente au chapitre 2, c'est bien l'écriture approchée qui a été utilisée dans l'étude des zones intérgéniques sur l'ensemble des génomes complets de bactéries. Or, nous constatons sur l'exemple de *Mycoplasma genitalium* que l'utilisation de l'approximation sur le z-base entraîne une augmentation artificielle de la significativité de la sta-

tistique. Sur les résultats que nous présentons sur la figure 2.6 (page 32), nous obtenons justement des résultats peu significatifs, et c'est ce manque de significativité qui soutient nos conclusions. Ainsi, l'utilisation de l'écriture exacte du z-base n'aurait probablement que renforcé ces conclusions. Les autres résultats du chapitre 2 proviennent tous de l'utilisation du z-codon sur les CDS de différents génomes, et les conclusions ne sont en rien modifiées pour ces résultats-là.

B. Le contenu en G+C : une mesure indirecte et peu fiable de l'effet des UVs sur la composition en bases

Nous avonc avancé, page 29, que la mesure du contenu en G+C est une mesure indirecte qui s'avère mauvaise pour estimer la pression de sélection des UVs sur la composition en bases des génomes. Je vais développer ici les arguments sur lesquels se base cette affirmation. En effet, le contenu en G+C est une mesure indirecte, qui ne peut être un bon indicateur de l'impact des UVs sur les génomes que si plusieurs hypothèses sont respectées.

Premièrement, les dinucléotides TpT doivent être effectivement plus sensibles aux UVs que les dinucléotides TpC (ou CpT), eux-mêmes plus sensibles que les dinucléotides CpC, pour qu'un fort contenu en G+C puisse être considéré comme un avantage sélectif face à la pression des UVs. Ceci n'est pas un fait démontré, comme le montre le tableau B.1 (Setlow, 1966). Le tableau B.1 présente le pourcentage de dimères de pyrimidine observés sur l'ADN de trois espèces bactériennes différentes, après irradiation à 265 nm. On voit clairement, que les dinucléotides TpT, TpC et CpT sont effectivement plus sensibles chez *Haemophilus influenzae* et *Escherichia coli*, mais que ça n'est pas le cas chez *Micrococcus luteus*.

Contenu en G+C	Nombre total de dimères par nucléotide	Pourcentage en $\overline{CC}$	$\overline{CT}+\overline{TC}$	$\overline{TT}$
Haemophilus influenzae				
38	2.7×10^{-2}	5	24	71
Escherichia coli				
50	2.0×10^{-2}	7	34	59
Micrococcus luteus (synonyme de *M. lysodeikticus*)				
70	1.4×10^{-2}	26	55	19

Figure B.1 – Nombre total de dimères de pyrimidine de type cyclobutane, et fréquence de chacun des trois types de dimères ($\overline{CC}$, $\overline{CT}$ et $\overline{TC}$, $\overline{TT}$), tels qu'ils ont été mesurés sur l'ADN de trois espèces bactériennes après irradiation à 265 nm. Tableau tiré de Setlow (1966).

Deuxièmement, pour un même contenu en G+C, la fréquence des dinculéotides de pyrimidine dépend du degré d'agrégation du génome en question. On

peut, en effet, facilement imaginer deux extrêmes : soit un génome alternant successivement une pyrimidine, une purine, une pyrimidine, ... (noté $YR_{n/2}$) soit un génome constitué dans sa première moitié uniquement de pyrimidines, et dans sa deuxième moitié uniquement de purines, et ceci pour un même contenu en G+C (noté $Y_{n/2}+R_{n/2}$). Les fréquences en dinucléotides de pyrimidine seront alors bien différentes, et par conséquent les fréquences en photoproduits estimés seront elles aussi bien différentes.

En reprenant les sensibilités aux UV des dinucléotides CpC, TpT et TpC (et CpT) telles que mesurés sur trois espèces bactériennes par Setlow (1966), article sur lequel est basé une partie de l'argumentaire de Singer & Ames (1970), nous pouvons estimer la fréquence de photoproduits pyrimidiques, à partir de trois types d'agrégation génomiques (voir figure B.2, pages 128, 129 et 130).

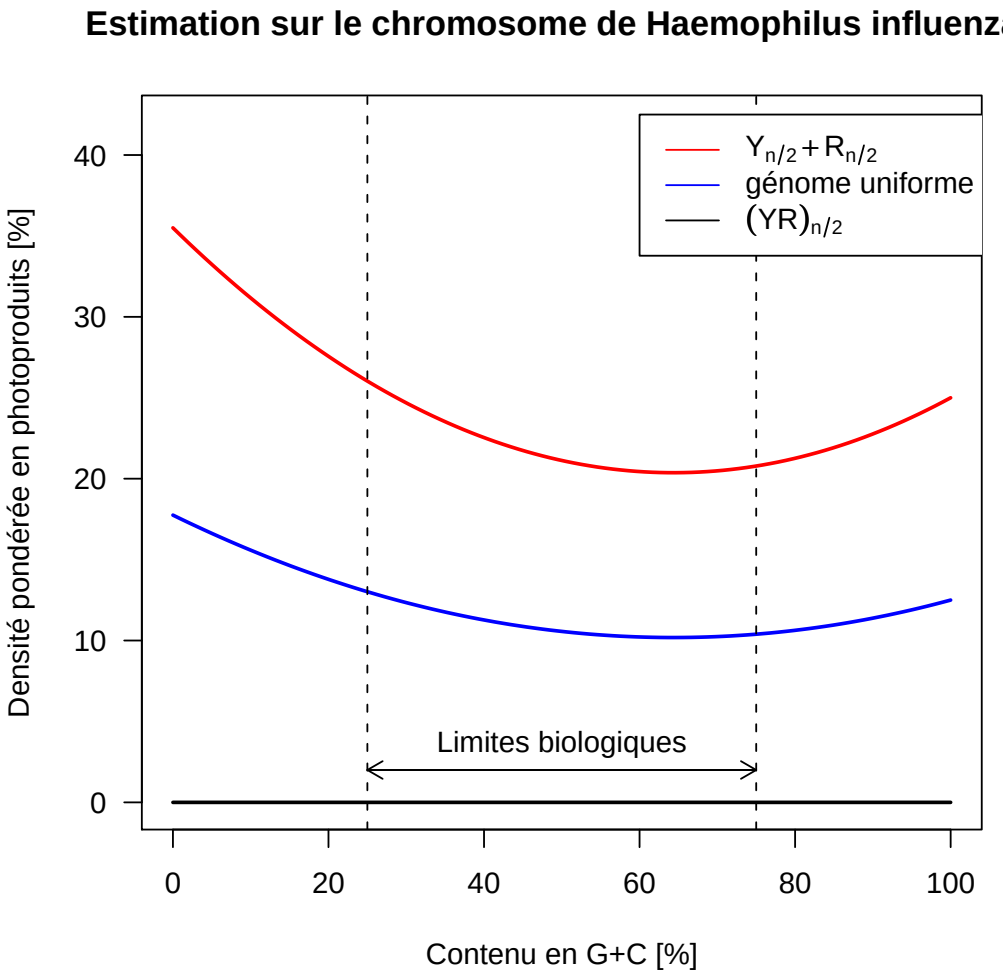


Figure B.2 – (a) *Haemophilus influenzae*. Densité de cibles de la lumière, pondéré par leurs fréquences dans chaque chromosome et estimés pour différents contenus en G+C et pour trois types de génomes : un génome constitué dans sa première moitié uniquement de pyrimidines, et dans sa deuxième moitié uniquement de purines, et ceci pour un même contenu en G+C (noté $Y_{n/2} + R_{n/2}$), un génome possédant une distribution uniforme des quatre bases (noté $uniforme$), et un génome alternant successivement une pyrimidine, une purine, une pyrimidine, ... (noté $YR_{n/2}$).

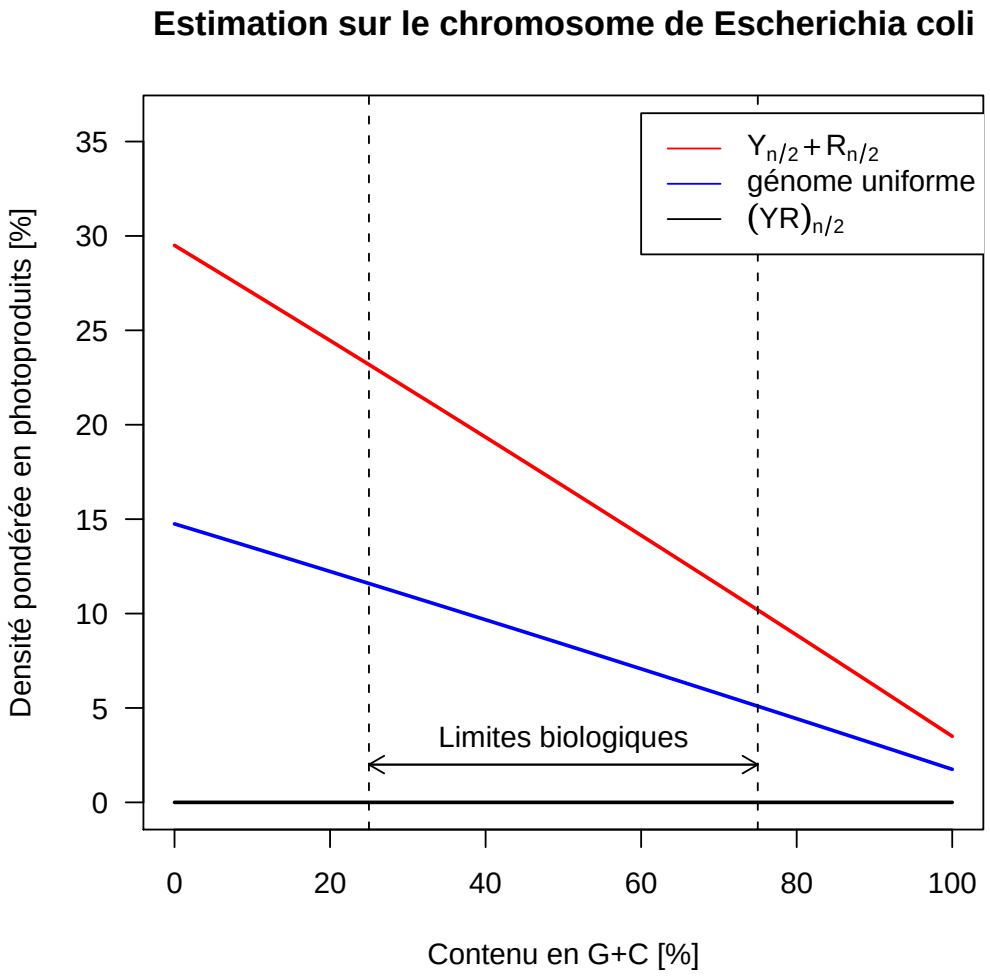


Figure B.2 – (b) *Escherichia coli*. Densité de cibles de la lumière, pondéré par leurs fréquences dans chaque chromosome et estimés pour différents contenus en G+C et pour trois types de génomes : un génome constitué dans sa première moitié uniquement de pyrimidines, et dans sa deuxième moitié uniquement de purines, et ceci pour un même contenu en G+C (noté $Y_{n/2} + R_{n/2}$), un génome possédant une distribution uniforme des quatre bases (noté $uniforme$), et un génome alternant successivement une pyrimidine, une purine, une pyrimidine, ... (noté $YR_{n/2}$).

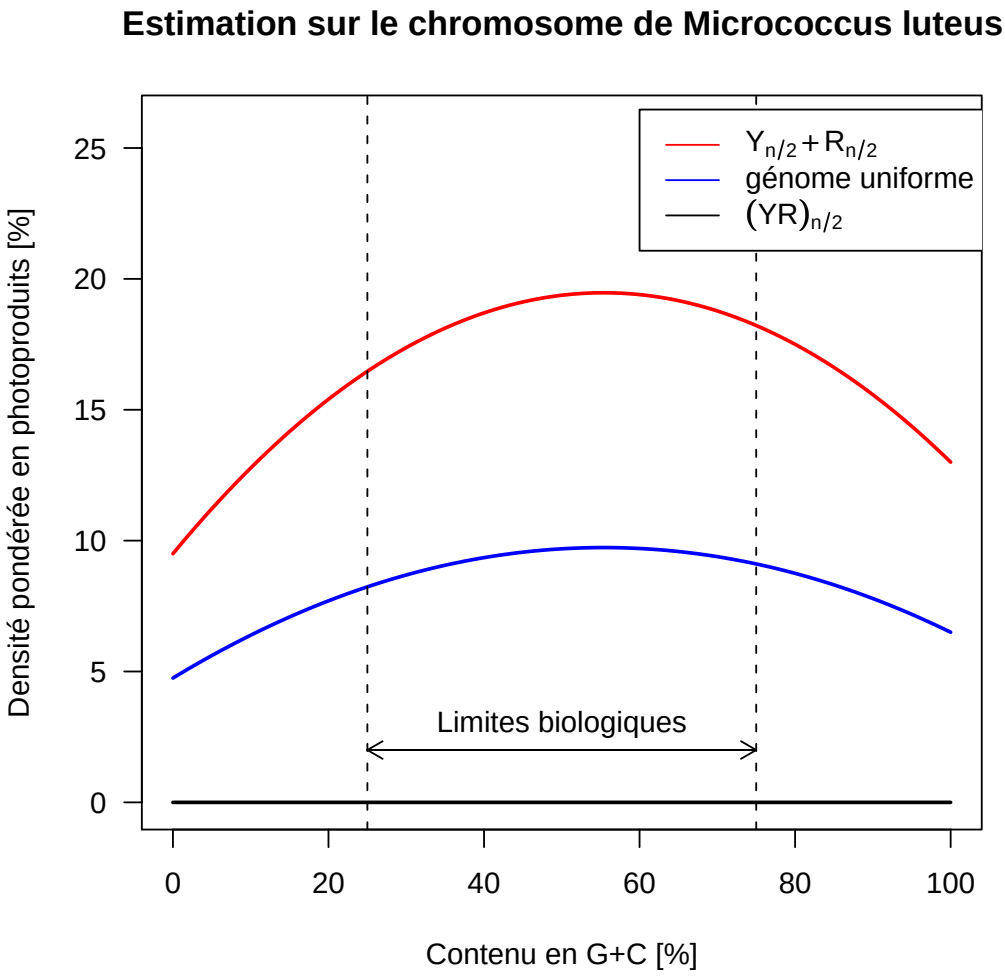


Figure B.2 – (c) *Micrococcus luteus*. Densité de cibles de la lumière, pondéré par leurs fréquences dans chaque chromosome et estimés pour différents contenus en G+C et pour trois types de génomes : un génome constitué dans sa première moitié uniquement de pyrimidines, et dans sa deuxième moitié uniquement de purines, et ceci pour un même contenu en G+C (noté $Y_{n/2} + R_{n/2}$), un génome possédant une distribution uniforme des quatre bases (noté $uniforme$), et un génome alternant successivement une pyrimidine, une purine, une pyrimidine, ... (noté $YR_{n/2}$).

C. Robustesse à l'écart à l'hypothèse de stationnarité

Dans le chapitre 3, page 63, j'ai testé la robustesse de l'écart à l'hypothèse d'indépendance entre sites sur la qualité des méthodes d'inférence phylogénétique en utilisant deux modèles d'évolution : K80 et K80+CpG, sur des séquences ancestrales à contenu variable en G+C. L'utilisation de séquences ancestrales à contenu en G+C différent du contenu en G+C à l'équilibre du modèle entraîne donc une **violation de l'hypothèse de stationnarité**. Toutefois, je vais montrer dans cet annexe, que cette violation n'a quasiment pas de conséquence sur la qualité de l'inférence phylogénétique, et que les résultats présentés à la fin du chapitre 3 sont bien le résultat de la violation de l'hypothèse d'indépendance entre sites.

Pour cela je présente ici les résultats de la simulation de l'évolution de séquences sur 100 arbres sous le modèle de Kimura (1980), à partir de séquences ancestrales à contenu en G+C variable (de 10 à 90 %), et après reconstruction d'arbres phylogénétiques par Phyml. Les résultats de la figure C.1 (pages 132, 133, 134, 135 et 136) sont clairs. Tout d'abord, on note que la reconstruction est plutôt globalement bonne, puisqu'on reconstruit des arbres proches des arbres vrais. La moyenne du score de Robinson-Foulds se situant à 0.03, et la médiane étant légèrement plus faible (autour de 0.027). La deuxième remarque que l'on peut faire concerne la robustesse de cette inférence face à la variation en G+C. En effet, l'écart de G+C à la valeur de 50%, qui correspond à la distribution stationnaire sous le modèle Kimura (1980), n'entraîne pas de modification de la distribution des distances de Robinson-Foulds entre es arbres vrais et les arbres reconstruits. On en conclut donc que l'inférence phylogénétique sous cette méthode semble assez robuste à l'écart à l'hypothèse de stationnarité, et que les résultats présentés dans le chapitre 3 peuvent être interprétés comme une conséquence de la violation de l'hypothèse d'indépendance entre sites.

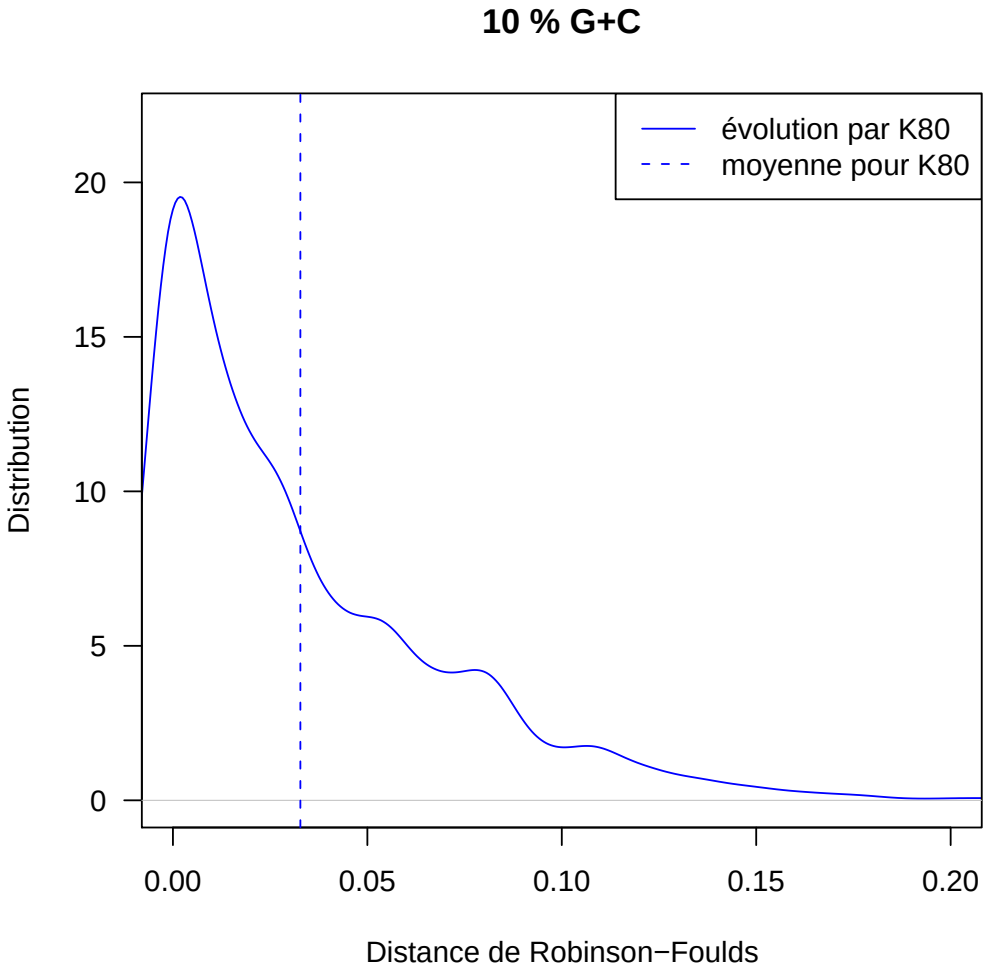


Figure C.1 – (a) Distribution empirique, pour 100 arbres, de la distance de Robinson-Foulds entre l'arbre vrai et l'arbre reconstruit sur un alignement après simulation de l'évolution, sous le modèle K80, d'une séquence ancestrale à contenu en G+C de 10%. La moyenne de la distribution est représentée par une droite en tirets verticale.

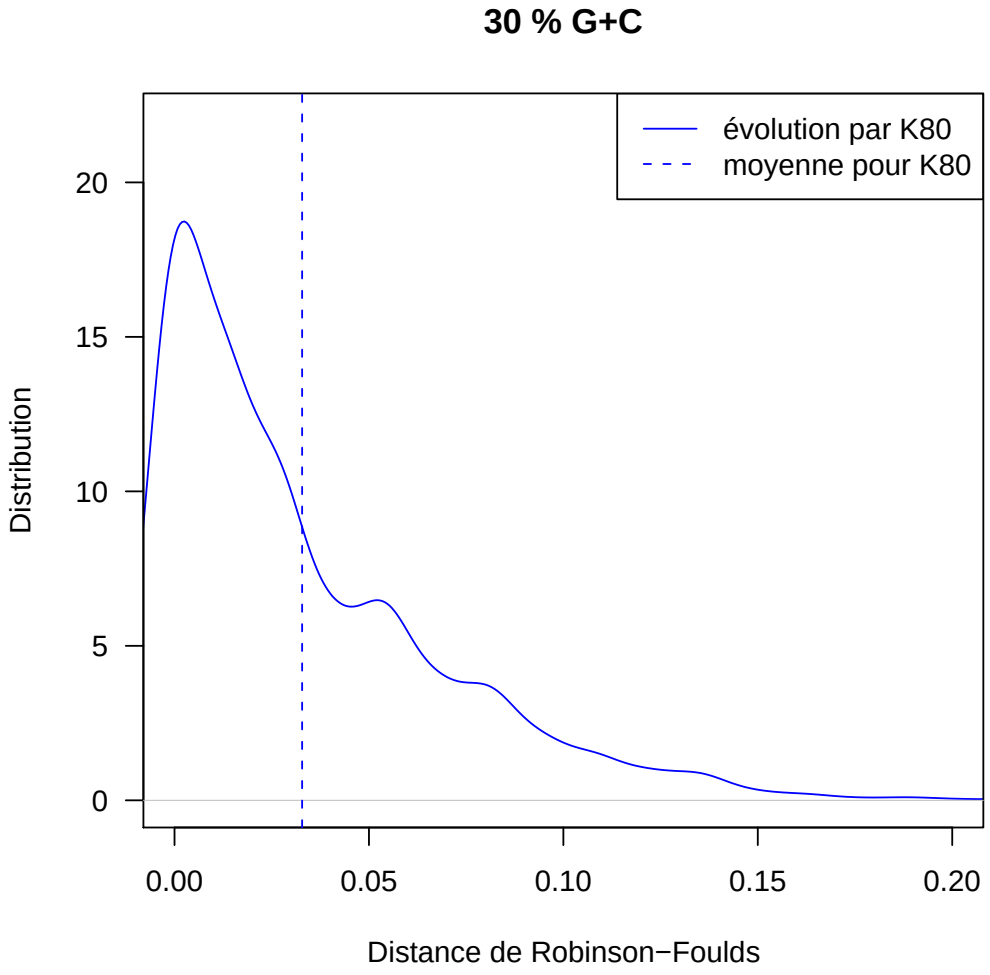


Figure C.1 – (b) Distribution empirique, pour 100 arbres, de la distance de Robinson-Foulds entre l'arbre vrai et l'arbre reconstruit sur un alignement après simulation de l'évolution, sous le modèle K80, d'une séquence ancestrale à contenu en G+C de 30%. La moyenne de la distribution est représentée par une droite en tirets verticale.

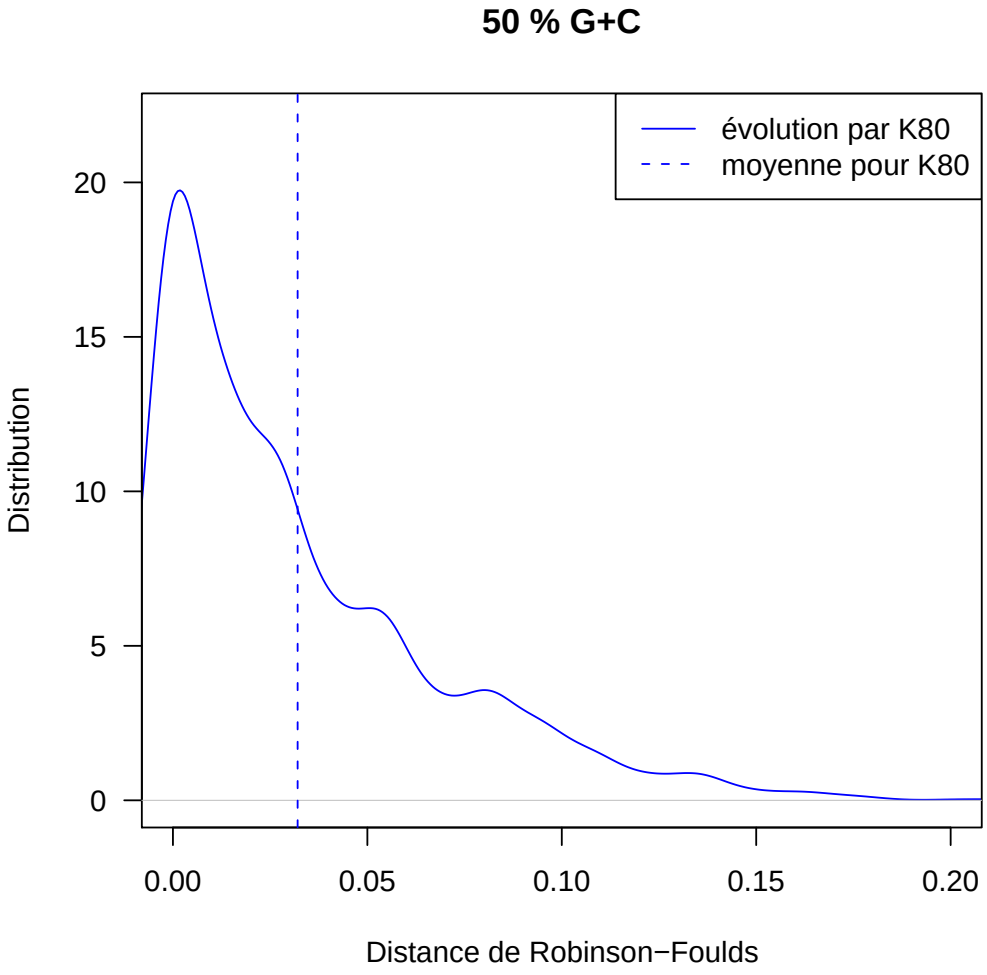


Figure C.1 – (c) Distribution empirique, pour 100 arbres, de la distance de Robinson-Foulds entre l'arbre vrai et l'arbre reconstruit sur un alignement après simulation de l'évolution, sous le modèle K80, d'une séquence ancestrale à contenu en G+C de 50%. La moyenne de la distribution est représentée par une droite en tirets verticale.

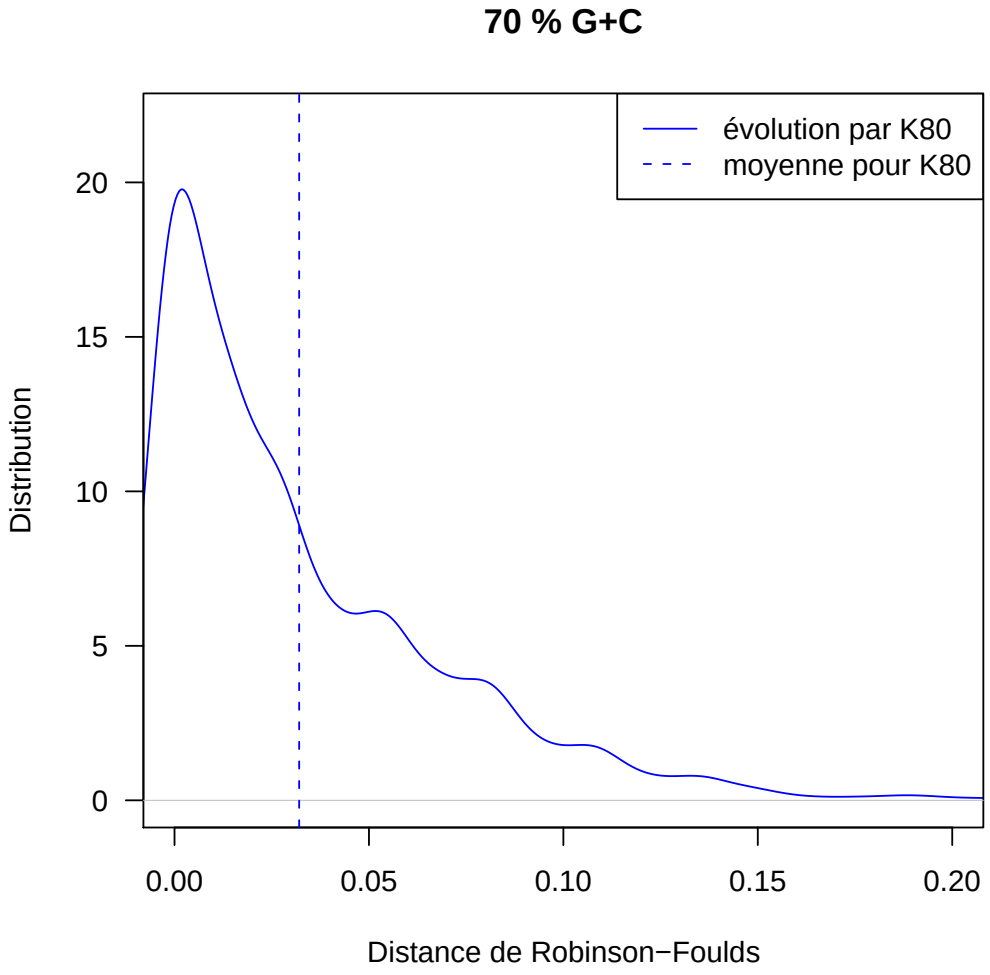


Figure C.1 – (d) Distribution empirique, pour 100 arbres, de la distance de Robinson-Foulds entre l'arbre vrai et l'arbre reconstruit sur un alignement après simulation de l'évolution, sous le modèle K80, d'une séquence ancestrale à contenu en G+C de 70%. La moyenne de la distribution est représentée par une droite en tirets verticale.

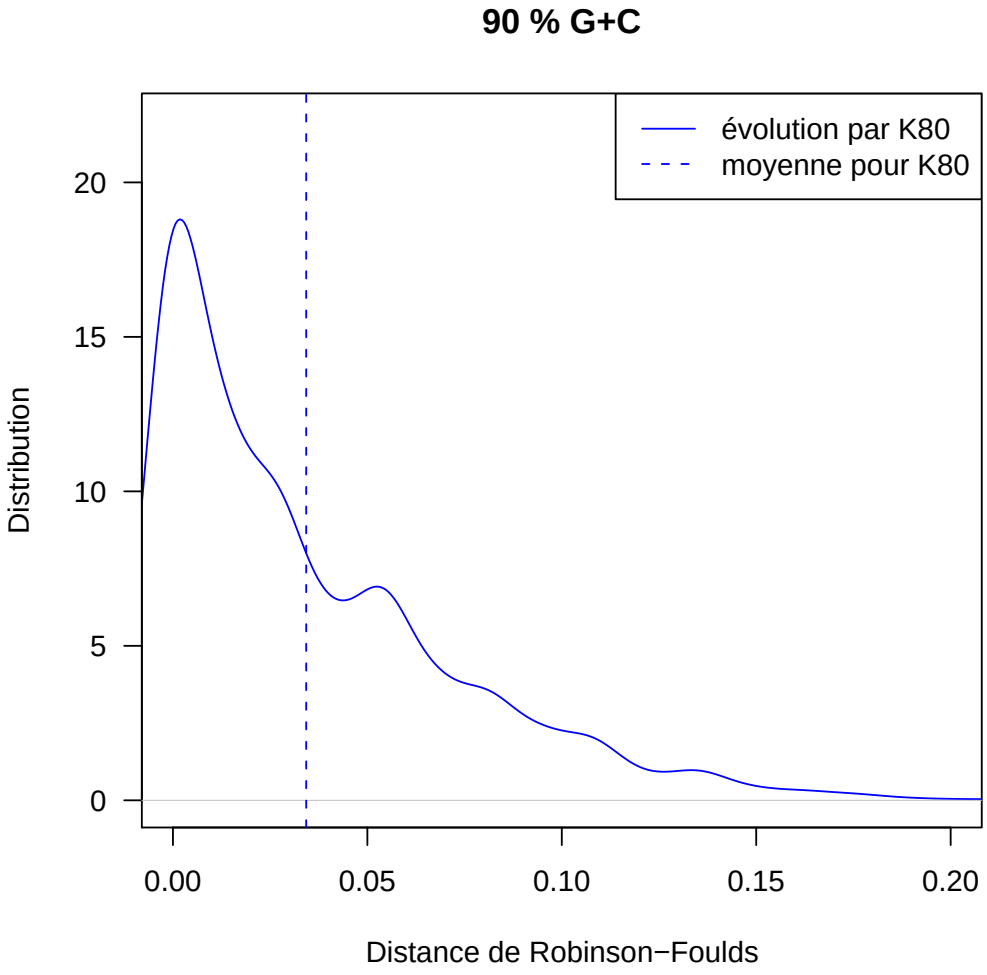


Figure C.1 – (e) Distribution empirique, pour 100 arbres, de la distance de Robinson-Foulds entre l'arbre vrai et l'arbre reconstruit sur un alignement après simulation de l'évolution, sous le modèle K80, d'une séquence ancestrale à contenu en G+C de 90%. La moyenne de la distribution est représentée par une droite en tirets verticale.

D. Article

UV-Targeted Dinucleotides Are Not Depleted in Light-Exposed Prokaryotic Genomes

Leonor Palmeira, Laurent Guéguen, and Jean R. Lobry

Laboratoire de Biométrie et Biologie Évolutive (UMR 5558); Centre National de la Recherche Scientifique (CNRS); Univ. Lyon 1, 43 bd 11. nov, 69622, Villeurbanne, Cedex, France; and HELIX, Unité de recherche, Institut National de Recherche en Informatique et en Automatique (INRIA)

We have investigated the hypothesis that pyrimidine dinucleotides are avoided in light-exposed genomes as the result of selective pressure due to high ultraviolet (UV) exposure. The main damage to DNA produced by UV radiation is known to be the formation of pyrimidine photoproducts: it is estimated that about 10 dimers per minute are formed in an *Escherichia coli* chromosome exposed to the UV light in direct overhead sunlight at sea level. It is also known that on an *E. coli* chromosome exposed to UVb wavelengths (290–320 nm), pyrimidine photoproducts are formed in the following proportions: 59% TpT, 7% CpC, and 34% CpT plus TpC. We have analyzed all available complete prokaryotic genomes and the model organism *Prochlorococcus marinus* and have found that pyrimidine dinucleotides are not systematically avoided. This suggests that prokaryotes must have sufficiently effective protection and repair systems for UV exposure to not affect their dinucleotide composition.

Introduction

Statistical analysis of the global composition of genomes and its link with environmental or metabolic characteristics has been the focus of considerable interest. It has been established that ultraviolet (UV) light—at both UVb (290–320 nm) and UVa wavelengths (320–400 nm)—damages DNA by specific mechanisms. It has been shown that UVb wavelengths are particularly dangerous for DNA and that the damage they most often cause is the formation of cyclobutane pyrimidine dimers by the photoexcitation of adjacent pyrimidines (Setlow 1966). If one of these dimers is formed on a DNA strand, this leads to a local DNA distortion, which blocks both transcription and replication (Setlow 1966; Singer and Ames 1970; Besaratinia et al. 2005). Singer and Ames (1970) have estimated that about 10 dimers per minute are formed in an *Escherichia coli* chromosome by the UV light in direct overhead sunlight at sea level.

The sensitivities of the 4 pyrimidine dinucleotides to UVb wavelengths differ: experiments on *E. coli* DNA show that TpT photoproducts make up to 59% of the target dimers, CpC up to 7%, and that CpT and TpC share the remaining 34% (Setlow 1966). It is noteworthy that most dimers involved are T rich and so a high G + C content will tend to result in genomes with fewer target dimers (see fig. 1).

Singer and Ames (1970) investigated whether bacteria exposed to higher UV radiation have a higher G + C content as the result of environmental adaptation. They found a strong link between the genomic G + C content in bacteria and the amount of UV exposure in their habitat, and they conclude that bacteria exposed to high levels of UV have a higher G + C content than those with less exposure. We wanted to reassess this hypothesis for the following reasons.

First, the results reported were controversial (Bak et al. 1972). Bak et al. (1972) discuss the UV exposure experienced by various bacteria and conclude that some of these species are not as highly exposed as suggested by Singer and Ames (1970). If this is the case, their conclusions would obviously be undermined. They also suggested that the wide variation in the G + C content in bacteria may be mainly attributable to phylogenetic relationships rather than to adaptation to habitat. Since this first controversy, there have been no major follow-up studies, and the question remains open.

Second, the choice of G + C content as an indicator of the impact of UV on pyrimidine dinucleotides is questionable: figure 1 shows that G + C content is a poor indicator of UV-target content in a genome. Indeed, for a given G + C content, the density of target dinucleotides present in the genome can vary considerably, depending on the degree of aggregation of the pyrimidine dinucleotides (see legend of fig. 1). Conversely, a given phototarget density can result from many different G + C contents.

Third, the availability of completely sequenced genomes now makes it possible to investigate the impact of UV exposure on genomes by directly measuring their pyrimidine dinucleotide content. Pyrimidine dinucleotides are the direct targets of UVb wavelengths, and if UV light has a major impact on genomes, we can expect highly exposed microorganisms to display significant depletion of all 4 pyrimidine dinucleotides (CpC, CpT, TpC, and TpT).

Last but not least, recent studies have focused on the relationship between external forces and genomic content (e.g., Naya et al. 2002; Foerstner et al. 2005). In particular, Naya et al. (2002) have shown that aerobic bacteria have a higher G + C content than anaerobic bacteria. In addition, aerobic bacteria are more likely to be exposed to sunlight, which means that it could be difficult to distinguish between the effects of aerobiosis and those of UV radiation.

Materials and Methods

Systematic Study

All 221 bacterial and archaeal genomes available on the European Bioinformatics Institute Genome Reviews database were retrieved on 16 June 2005. Out of the 221 complete genomes extracted, 2 data sets were created: one contained all annotated "CDS" sequences and will subsequently be referred to as "coding sequences," and one contained all sequences other than those annotated as "CDS,"

Key words: G + C content, dinucleotide content, ultraviolet radiation, aerobiosis, *Prochlorococcus marinus*.

E-mail: palmeira@biomserv.univ-lyon1.fr.

Mol. Biol. Evol. 23(11):2214–2219. 2006
doi:10.1093/molbev/msl096
Advance Access publication August 22, 2006

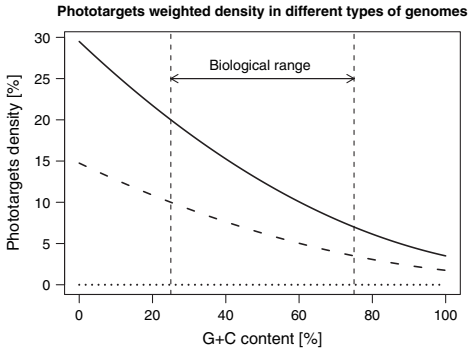


FIG. 1.—Density of phototargets weighted by their frequency in the *Escherichia coli* chromosome and calculated for different G + C contents and for 3 kinds of random genomes. The weights are as follows: $0.59*f_{tt} + 0.34*(f_{tc} + f_{ct}) + 0.07*f_{cc}$ (where f_{xy} is the frequency of dinucleotide xy in the specified genome). Three models of random genomes are analyzed. In the worst case (solid curve), the genome is the concatenation of a sequence of pyrimidines and a sequence of purines: all pyrimidines are involved in a pyrimidine dinucleotide. In the best case (dotted curve), the genome is an unbroken succession of pyrimidine–purine dinucleotides: no pyrimidine is involved in a pyrimidine dinucleotide. In the "random case" (dashed curve), the frequency of a pyrimidine dinucleotide is the result of chance ($f_{xy} = f_x \times f_y$).

"rRNA" (ribosomal RNA coding regions) or "tRNA" (transfer RNA coding regions), and will subsequently be referred to as "intergenic sequences."

We started by a systematic approach and investigated all available complete bacteria and archaeal genomes. We then took a closer look at the genomes of 3 strains of the picocyanobacterium *Prochlorococcus marinus*.

Prochlorococcus marinus as a Model Organism

Each of the 3 strains of *P. marinus* we investigated is adapted to a different depth in the water column (Dufresne et al. 2003; Rocap et al. 2003) and, therefore, exposed to different intensities of UV radiation. This seemed to make it an ideal model organism for investigating this hypothesis.

Dufresne et al. (2003) have shown that the SS120 strain is adapted to living at a depth of 120 m. The MIT 9313 strain is adapted to living at a depth of 135 m, and so both these strains can be considered to be low-light–adapted strains (Rocap et al. 2003). The MED4 strain is adapted to living at a depth of 5 m and can be considered to be a high-light–adapted strain (Dufresne et al. 2003). The residual intensities of 260-nm irradiation (UVb) at various depths in pure water can be estimated from water's absorbance coefficient (Quickenden and Irvin 1980) as follows: 70% of its original intensity at 5-m depth (MED4 strain), 0.0002% at 120-m depth (SS120 strain), and 0.00007% at 135-m depth (MIT 9313 strain).

The accession numbers and references for these strains are as follows: strain CCMP 1375/SS120/SARG (GenBank accession number AE017126) (Dufresne et al. 2003), subsp. pastoris, strain CCMP 1378/MED4 (GenBank accession number BX548174), and strain MIT 9313 (GenBank accession number BX548175) (Rocap et al. 2003).

Statistical Analysis

Our aim was to find out whether pyrimidine dinucleotides are avoided in bacterial genomes. In prokaryote genomes, coding sequences can constitute up to 80% of the entire genome and sometimes even more. These sequences are subjected to strong selective pressure, which makes other effects difficult to detect. Nevertheless, deleting 80% of the data is not a good way to detect small effects. We therefore developed 2 methods for measuring the over- and underrepresentation of dinucleotides: one for coding sequences, the other for intergenic sequences.

The idea was to compute a normalized statistic of the ρ_{xy} statistic (Karlin and Brendel 1992), in order to make it possible to compare the results for sequences belonging to different species and even to different phyla.

$$\rho_{xy} = \frac{f_{xy}}{f_x \times f_y}, \tag{1}$$

where f_{xy}, f_x, and f_y are the frequencies in the studied sequence of dinucleotide xy, nucleotide x, and nucleotide y, respectively.

The normalized statistic is of the following type:

$$z_{score} = \frac{\rho_{xy} - E(\rho_{xy})}{\sqrt{\mathrm{Var}(\rho_{xy})}}, \tag{2}$$

where $E(\rho_{xy})$ and $\mathrm{Var}(\rho_{xy})$ are the expected value and variance of ρ_{xy} according to a given model that describes the sequence. This statistic follows the standard normal distribution. The expected value and the variance can be computed either by simulation or by analytical calculation, if asymptotic results are available.

Naturally, we can propose various models of sequences for calculating the expected value and variance. Each of these models is constructed so as to preserve some of the constraints of the studied sequence for the expected counts. This means that both the expected count and the observed count will share the specified constraints, and the z_{score} statistic will reflect what is over- or underrepresented in the studied sequence once the effects of these constraints have been eliminated. Two of these models will be shown here, which means 2 z_{score} statistics will be presented.

Intergenic Sequence Analysis

The unconstrained base shuffling model describes each sequence as a series of independent draws following the frequencies of the 4 letters as counted on that sequence. In this model, only the base composition of the analyzed sequence is preserved, and asymptotic results are available (Prum et al. 1995):

$$E(\rho_{xy}) = 1 \tag{3}$$

$$\mathrm{Var}(\rho_{xy}) = \frac{(1 - f_x)(1 - f_y)}{n f_x f_y}. \tag{4}$$

The z_{score} statistic computed using this model allows us to answer the following question: is there an anomalously

high or low *XpY* content given the base composition of the studied sequence?

We have used this model on intergenic sequences because we do not know enough about the selective forces involved to be able to develop an appropriate model that would allow us to see what is happening underneath the selective constraints acting on intergenic regions.

Coding Sequence Analysis

Unlike the intergenic regions, a certain number of constraints in coding regions has been identified. In CDS, we know that there is a bias in codon usage (Grantham et al. 1980), and we therefore expect that the dinucleotides present in the preferred codons will be overrepresented in the generic statistic presented in the "Intergenic sequence analysis." We have therefore developed a model that allows us to compute a z_{score}, which erases this codon-usage bias. This statistic enables us to identify over- and underrepresentations that exist in coding sequences, despite the presence of codon-usage bias.

In the codon shuffling model (CS), each sequence is described as a series of independent draws of codons following the frequencies of the codons as counted in that sequence. In the CS model, the codon composition of the analyzed sequence is preserved, which means that the base composition of the sequence analyzed is also preserved.

We can show that computing the z_{score} statistic using this model can be reduced to computing the z_{score} statistic on dinucleotides that overlap 2 codons:

$$z_{\text{score}} = \frac{\rho_{XY_{3-1}} - E(\rho_{XY_{3-1}})}{\sqrt{\text{Var}(\rho_{XY_{3-1}})}} = \frac{XY_{3-1} - E(XY_{3-1})}{\sqrt{\text{Var}(XY_{3-1})}}, \quad (5)$$

where $\rho_{XY_{3-1}}$ is the ρ statistic for dinucleotides overlapping 2 codons, and where XY_{3-1} is the count of dinucleotides that overlap 2 codons. Asymptotic results are available (Gautier et al. 1985):

$$E(XY_{3-1}) = \frac{n_1 n_2 - n_3}{n}, \quad (6)$$

$$\begin{aligned}\text{Var}(XY_{3-1}) = {} & E(XY_{3-1}) - [E(XY_{3-1})]^2 \\ & + \frac{1}{n(n-1)}[(2n_3(n_1 + n_2 - n_1 n_2 - 1) \\ & + n_1 n_2(n_1 - 1)(n_2 - 1))], \end{aligned} \quad (7)$$

where n_1, n_2, and n_3 are the number of codons ending with the letter X, the number of codons starting with the letter Y, and the number of codons starting with letter Y and ending with letter X, respectively. n is the total number of codons in the sequence.

The z_{score} statistic computed using this model allows us to answer the following question: is there an anomalously high or low *XpY* content given the codon-usage bias of the studied CDS?

Reproducibility

All computations were made using R's "seqinR" (Charif and Lobry 2006) package and were conducted using the computation resources available from the IN2P3's Computing Center.

Data and results are available and can be reproduced online at: http://biomserv.univ-lyon1.fr/~palmeira/repro/uv.html.

Results

Systematic Study

No systematic underrepresentation of CpT, TpC, CpT, or TpT dinucleotides was found (see fig. 2). None of the 4 pyrimidine dinucleotides was globally and significantly over- or underrepresented. This clearly does not bear out the initial hypothesis being tested and means that there is no avoidance of these 4 dinucleotides in prokaryotic genomes, despite the fact that they are major targets for photoinduced damage (Setlow 1966).

There is a rather good correlation between the *XpY* content of intergenic sequences and the *XpY* content of coding sequences, which is strong evidence for general DNA mechanisms common to both coding and intergenic sequences. This shows that in highly constrained CDS sequences, our method is able to recover general signals also present in intergenic sequences. This is true not only for the 4 pyrimidine dinucleotides but for all 16 dinucleotides and could be explained by the existence of biased mutational processes acting indifferently on the whole genome and producing genome-wide biases (Chen et al. 2004).

The rather universal overrepresentation of TpT dinucleotides in all genomes is surprising, even though it was not always statistically significant. Unlike eukaryotic mRNA, where poly-A stretches have a known essential function, there is no evidence for major poly-A or poly-T stretches in bacterial DNA. However, ApA and TpT periodical patterns have been reported in both bacteria and eukaryotes (Tomita et al. 1999). This periodicity has been related to DNA coiling and supercoiling and could explain the observed slight overrepresentation.

Very few outliers have been found for CpC dinucleotide, but those that have been found are the 2 chromosomes of the fully sequenced *Burkholderia mallei* and *Burkholderia pseudomallei* genomes. These 2 prokaryotes are both pathogens commonly found in soil and in groundwater, and there is no evidence in the literature to suggest that these 2 strains are exposed to higher levels of UV than the other prokaryotes in the data set. This feature cannot, therefore, be linked to UV exposure and may be particular to this genus.

Prochlorococcus marinus as a Model Organism

No difference was found between the relative abundances of pyrimidine dinucleotides in these 3 strains (see fig. 3). However, these 3 ecotypes have been separated long enough to evolve different G + C contents (30.8% for MED4 at 5-m depth; 36.4% for SS120 at 120-m depth; and 50.8% for MIT 9313 at 135-m depth) (Dufresne et al. 2003; Rocap et al. 2003), and at least 2 of these ecotypes have divergent genomic adaptations (Rocap et al. 2003). These 2 previous studies show that the genomes of the 3 strains have diverged, yet this divergence has had no effect on pyrimidine dinucleotide content. There is, therefore,

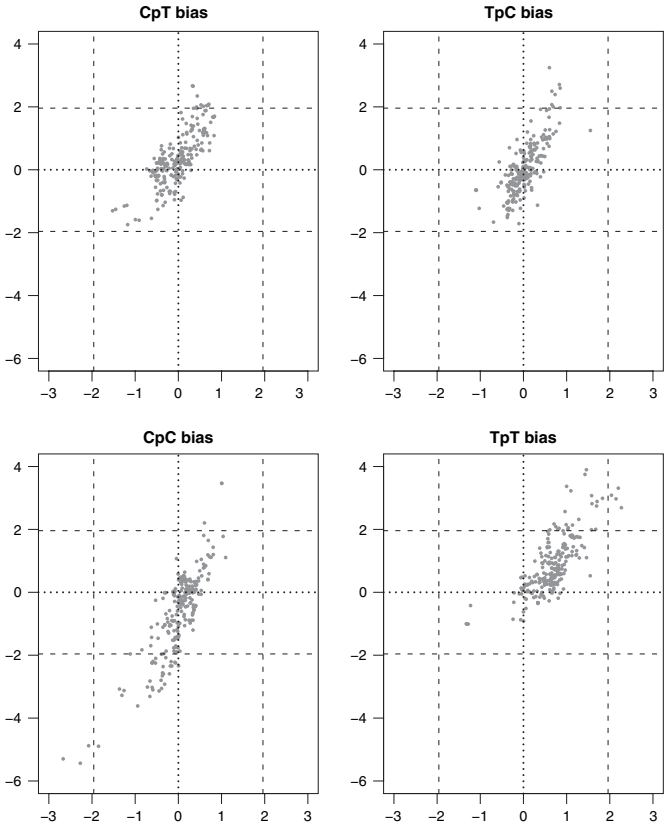


FIG. 2.—Plot of the mean z_{score} statistics for intergenic sequences (x axis) and for coding sequences (y axis), for each of the 4 pyrimidine dinucleotides. On each plot, a dot corresponds to the mean of these 2 statistics in a given prokaryote chromosome. The null x and y axis (dotted lines) and the 5% limits of significance for the standard normal distribution (dashed lines) are plotted as benchmarks. It should be noted (see fig. 3) that the variability within one chromosome is sometimes as great as that between different chromosomes.

no evidence of an impact of UV exposure on dinucleotide content.

One possible exception could be the CpC dinucleotide, which seems to be slightly underrepresented in the high-light–adapted strain, compared with the 2 low-light–adapted strains. For the other 3 pyrimidine dinucleotides, there is no avoidance of the pyrimidine dinucleotide and, therefore, no link between UV exposure and relative pyrimidine dinucleotide abundance. We also note that the variability within one strain can be as great as that between different chromosomes (see fig. 3). This finding is consistent with the lack of any link found between relative dinucleotide abundance and exposure to UV.

Discussion

We have shown that UV exposure has no systematic impact on pyrimidine dinucleotide bias in prokaryotes. This is true not only for all bacteria and archae, but when we looked at strains of *P. marinus*, we once again found no link between UV exposure and pyrimidine dinucleotide abundance. This means that there is no evidence of the avoidance of pyrimidine dinucleotides in microorganisms exposed to UV.

Prokaryotes have developed mechanisms to repair DNA damage. Our findings show that these systems must be efficient enough to make it unnecessary for pyrimidine dinucleotides to be avoided in their genome. This result is in agreement with recent studies on resistance of marine bacteria to UV radiation, see Agogué et al. (2005) and references therein. From an evolutionary perspective, this is probably due to their inheritance of highly efficient repair systems from ancestral organisms living at the time when there was no ozone layer to filter UV light.

The fact that protection and repair systems in bacteria are efficient enough for the genomic content to have

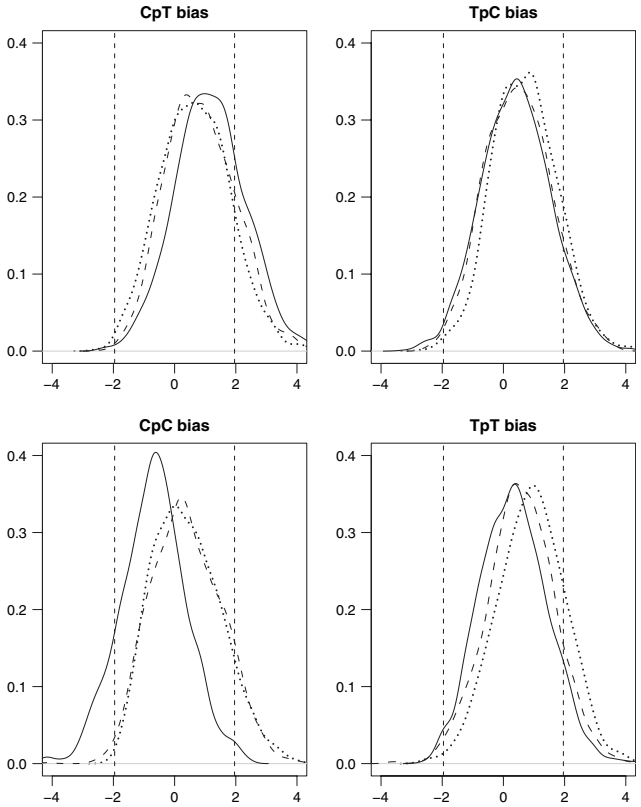


FIG. 3.—Each figure shows the distributions of the z_{score} in all coding sequences corresponding to each of the 3 strains of *Prochlorococcus marinus*. In each figure, the distribution for the MED4 (a high-light–adapted strain) is shown as a solid line; the distribution for the SS120 (a low-light–adapted strain) is shown as a dashed line, and the distribution for the MIT 9313 (a low-light–adapted strain) is shown as a dotted line. The 5% limits of significance for the standard normal distribution (dashed vertical lines) are plotted as benchmarks.

evolved totally independently of UV exposure tends to support the findings of Naya et al. (2002): UV exposure and aerobiosis are not likely to interfere in their analysis.

Acknowledgments

This work was funded jointly by the Action Concentee Incitative "New Interfaces of Mathematics", the Action de Recherche Coopérative "Integrated Biological Networks," and the Agence Nationale de la Recherche "Régularités: Inférence et Statistique" project grants.

Funding to pay the Open Access publication charges for this article was provided by the Action Concentee Incitative, the Action de Recherche Coopérative, and the Agence Nationale de la Recherche.

Literature Cited

Agogué H, Joux F, Obernosterer I, Lebaron P. 2005. Resistance of marine Bacterioneuston to solar radiation. Appl Environ Microbiol 71:5282–9.

Bak AL, Atkins JF, Singer CE, Ames BN. 1972. Evolution of DNA base compositions in microorganisms. Science 175: 1391–3.

Besaratinia A, Synold TW, Hsiu-Hua C, Chang C, Xi B, Riggs AD, Pfeifer GD. 2005. DNA lesions induced by UV A1 and B radiation in human cells: comparative analyses in the overall genome and in the p53 tumor suppressor gene. Proc Natl Acad USA 102:10058–63.

Charif D, Lobry J. 2006. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla H, Porto M, Vendruscolo M, editors. Structural approaches to sequence evolution: molecules, networks, populations. New York: Springer Verlag. Biological and Medical Physics, Biomedical Engineering. Forthcoming.

Chen SL, Lee W, Shapiro L, McAdams HH. 2004. Codon usage between genomes is constrained by genome-wide mutational processes. Proc Natl Acad Sci USA 101:3480–5.

Dufresne A, Salanoubat M, Partensky F, et al. (21 co-authors). 2003. Genome sequence of the cyanobacterium *Prochlorococcus marinus SS120*, a near minimal oxyphototrophic genome. Proc Natl Acad USA 100:10020–5.

Foerstner KU, von Mering C, Hooper SD, Bork P. 2005. Environments shape the nucleotide composition of genomes. EMBO Rep 6:1208–13.
Gautier C, Gouy M, Louail S. 1985. Non-parametric statistics for nucleic acid sequence study. Biochimie 67:449–53.
Grantham R, Gautier C, Gouy M, Mercier R, Pavé A. 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res 8:r49–62.
Karlin S, Brendel V. 1992. Chance and statistical significance in protein and DNA sequence analysis. Science 257: 39–49.
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 55:260–4.
Prum B, Rodolphe F, de Turckheim E. 1995. Finding words with unexpected frequencies in deoxyribonucleic acid. J R Stat Soc 57:205–20.
Quickenden TI, Irvin JA. 1980. The ultraviolet absorption spectrum of liquid water. J Chem Phys 72:4416–28.
Rocap G, Larimer F, Lamerdin J, et al. (24 co-authors). 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature.
Setlow RB. 1966. Cyclobutane-type pyrimidine dimers in polynucleotides. Science 153:379–86.
Singer CE, Ames BN. 1970. Sunlight ultraviolet and bacterial DNA base ratios. Science 170:822–6.
Tomita M, Wada M, Kawashima Y. 1999. ApA dinucleotide periodicity in prokaryote, eukaryote, and organelle genomes. J Mol Evol 49:182–92.

Dan Graur, Associate Editor

Accepted August 14, 2006

TITRE en français

Analyse et modélisation des dépendances entre sites voisins dans l'évolution des séquences d'ADN

RÉSUMÉ en français

Cette thèse a porté, d'une part, sur l'analyse des sur- et sous-représentations en dinucléotides au sein de différents génomes complets, en recherchant les liens éventuels avec des mécanismes connus de dommages causés à l'ADN qui soient liés à des sites avoisinants – particulièrement les voisins directs en 5' et 3'. L'étude de l'effet des UVs sur les génomes de micro-organismes, et sur l'effet de la méthylation sur les génomes de métazoaires en a été un des grands axes. D'autre part, les résultats récents de Bérard *et al.* sur des modèles d'évolution incorporant des dépendances entre bases adjacentes (pyrimidine suivie de purine) ont permis de développer une approche probabiliste d'estimation des substitutions liées au mécanisme de méthylation-désamination spontanée des dinucléotides CG.

MOTS-CLEFS en français

dinucléotides ; effet des CpG ; modèles d'évolution ; modèles de Markov ; reconstruction phylogénétique ; analyse de séquence ; lumière UV

TITRE en anglais

Analyzing and modelling neighboring site dependencies in DNA evolution

RÉSUMÉ en anglais

On the one hand, this study examined dinucleotide over- and under-representations in different complete genomes, in order to determine possible links with DNA damage known mechanims. We focused on direct 5' and 3' neighbors, and analyzed the effect of UV light on the genomes of micro-organisms, and the effect of methylation on the genomes of metazoans. On the other hand, recent results by Bérard *et al.* on models of evolution incorporating neighboring site dependencies (pyrimidine followed by purine), allowed us to develop a probabilistic approach for the estimation of substitution rates due to the methylation-deamination process acting on CG dinucleotides.

DISCIPLINE : Bioinformatique

MOTS-CLEFS en anglais

dinucleotides ; CpG effect ; models of evolution ; Markov models ; phylogenetic recontruction ; sequence analysis ; UV light

INTITULE ET ADRESSE DE L'U.F.R. OU DU LABORATOIRE :

Laboratoire de Biométrie et Biologie Évolutive - UMR 5558 CNRS
Batiment Gregor Mendel - Université Claude Bernard Lyon1
43, bv du 11 novembre 1918 - 69622 Villeurbanne cedex